fluxicon

Disco User's Guide

Author: Dr. Anne Rozinat
Contact: anne@fluxicon.com

Contents

1

Data Extraction Guide

One of the big advantages of process mining is that it starts with the data that is already there, and usually it starts very simple. There is no need to first set up a data collection framework. Instead you can use data that accumulates as a byproduct of the increasing automation and digitization of your business processes. These data are collected right now by the various IT systems you already have in place to support your business.

Sometimes people are worried that they do not have the right data, but in practice this is rarely the case. There are so many ERP and CRM systems, delivery notes, request, complaint, ticket, or order systems, credit checks, etc. etc. – so most organizations have lots of data.

The starting point for process mining is a so-called *event log*. But what exactly is an event log? Where do event logs come from? And how do you know whether your data satisfies the requirements to apply process mining? This is what this introductory chapter is about.

What you will learn:

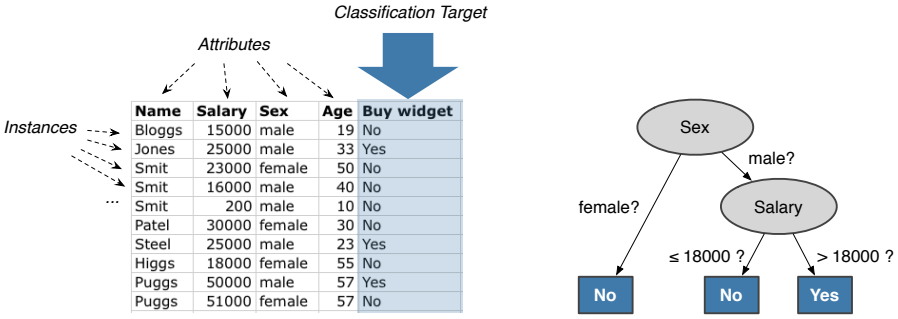
- What kind of data is needed to do a process mining analysis.
- The three key requirements for an event log.
- How to get started with your own data extraction.

1.1 The Mental Model for Process Mining

The core idea of process mining is to analyze data from a *process perspective*. You want to answer questions such as “How does my As-is process currently look like?”, “Are there waste and unnecessary steps that could be eliminated?”, “Where are the bottlenecks?”, and “Are there deviations from the rules and prescribed processes?”. To be able to do that, process mining approaches data with a mental model that maps the data to a process view.

To understand what this means, let us first take a look at another mental model: The mental model for classification techniques in data mining.

Assume that you have a widget factory and you want to understand which kinds of customers are buying your widgets. On the left in Figure 1.1, you see a very simple example of a data set. There are columns for the *attributes* Name, Salary, Sex, Age, and Buy widget. Each row forms one *instance* in the data set. An instance is a learning example that can be used for learning the classification rules.



Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

Fig. 1.1. Data mining example: The classification target class needs to be configured.

Before the classification algorithm can be started, one needs to determine which of the columns is the target class. Because we want to find out who is buying the widgets, we would make the Buy widget column the classification target. A data mining tool would then be able to construct a decision tree like depicted on the right in Figure 1.1.

The result shows that only males with a high salary are buying the widgets. If we would want to derive rules for another attribute, for example, predict how old the customers who buy our widgets would typically be, then the Age column would be the classification target.

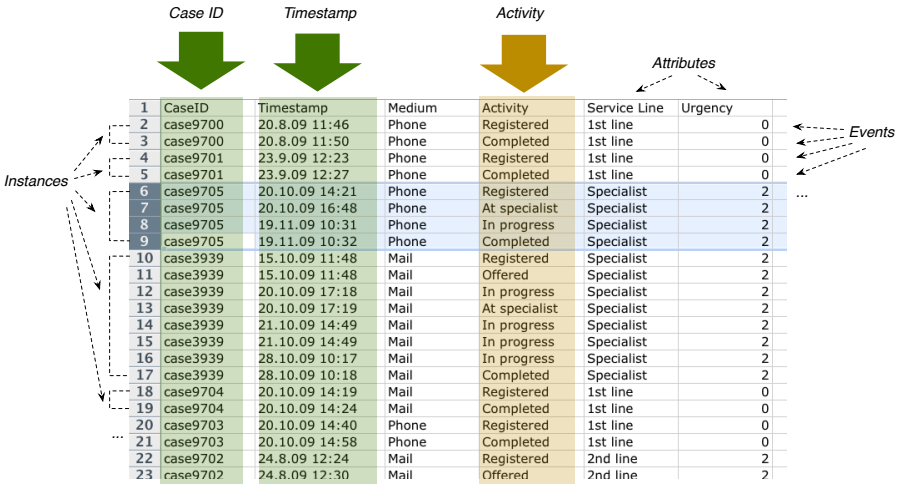
For process mining, we have a slightly different mental model because we look at the data from a process perspective.

In Figure 1.2, you see a simplified example data set from a call center process. In contrast to the data mining example above, an individual row does not represent a complete process *instance*, but just an *event*. Because a data set that is used for process mining consists of events, these data are often referred to as *event log*. In an event log:

- Each event corresponds to an activity that was executed in the process.
- Multiple events are linked together in a process instance or case.
- Logically, each case forms a sequence of events—ordered by their timestamp.

From the data sample in Figure 1.2, you can see why even doing simple process-related analyses, such as measuring the frequency of process flow variants, or the

time between activities, is impossible using standard tools such Excel. Process instances are scattered over multiple rows in a spreadsheet (not necessarily sorted!) and can only be linked by adopting a process-oriented meta model.



	CaseID	Timestamp	Medium	Activity	Service Line	Urgency
1						
2	case9700	20.8.09 11:46	Phone	Registered	1st line	0
3	case9700	20.8.09 11:50	Phone	Completed	1st line	0
4	case9701	23.9.09 12:23	Phone	Registered	1st line	0
5	case9701	23.9.09 12:27	Phone	Completed	1st line	0
6	case9705	20.10.09 14:21	Phone	Registered	Specialist	2
7	case9705	20.10.09 16:48	Phone	At specialist	Specialist	2
8	case9705	19.11.09 10:31	Phone	In progress	Specialist	2
9	case9705	19.11.09 10:32	Phone	Completed	Specialist	2
10	case3939	15.10.09 11:48	Mail	Registered	Specialist	2
11	case3939	15.10.09 11:48	Mail	Offered	Specialist	2
12	case3939	20.10.09 17:18	Mail	In progress	Specialist	2
13	case3939	20.10.09 17:19	Mail	At specialist	Specialist	2
14	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
15	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
16	case3939	28.10.09 10:17	Mail	In progress	Specialist	2
17	case3939	28.10.09 10:18	Mail	Completed	Specialist	2
18	case9704	20.10.09 14:19	Mail	Registered	1st line	0
19	case9704	20.10.09 14:24	Mail	Completed	1st line	0
20	case9703	20.10.09 14:40	Phone	Registered	1st line	0
21	case9703	20.10.09 14:58	Phone	Completed	1st line	0
22	case9702	24.8.09 12:24	Mail	Registered	2nd line	2
23	case9702	24.8.09 12:30	Mail	Offered	2nd line	2

Fig. 1.2. Process mining input data: Case ID, Activity and Timestamp need to be identified.

For example, if you look at the highlighted rows 6-9 in Figure 1.2, you can see one process instance (case9705) that starts with the status Registered on 20 October 2009, moves on to At specialist and In progress, and ends with status Completed on 19 November 2009.

The basis of process mining is to look at historical process data precisely with such a "process lens". It is actually quite simple, and one of the big advantages is that process mining does not depend on specific automation technology or specific systems. It is a source system-agnostic technology, precisely because it is centered around the process-oriented mental model explained above. This way, it can be applied to a wide range of processes, including but not limited to customer service processes, system monitoring, healthcare, IT services, enterprise or financial processes.

1.2 The Minimum Requirements for an Event Log

The data columns determine the analysis possibilities that you have later on and here is where the real process mining requirements come into play. According to the mental model described before, you need to identify at least the following three elements: Case ID, Activity, and Timestamp.

1.2.1 Case ID

A case is a specific instance of your process. What precisely the meaning of a case is in your situation depends on the domain of your process. For example:

- In a purchasing process, the handling of one purchase order is one case.
- In a hospital, this would be the patient going through a diagnosis and treatment process.
- In a call center process, a case would be related to a particular service request number.

For every event, you have to know which case it refers to, so that the process mining tool can compare several executions of the process to one another. So, you need to have one or more columns that together uniquely identify a single execution of your process. They form a case identifier (case ID).

Rule #1: The case ID determines the scope of the process.

Be aware that the case ID influences your process scope. It determines where your process starts and where it ends. In fact, there may be more than one way to set up your case ID. For example, in a sales process you could set up the case ID in two different ways:

1. You can see the processing of a particular lead through the sales funnel as the process you want to analyze. Then the product lead number is your case ID.
2. At the same time, you may want to see the overall sales process for a customer as your process scope—the same customer may have gone through your sales funnel for different products. Then the customer ID is your case ID.

Both alternatives are logical and can make sense, depending on your analysis goals. In your project you can take different views on the process and analyze it from different perspectives. The important part for now is that you have at least one column that can be used to distinguish your process instances and serve as a case ID.

1.2.2 Activity

An activity forms one step in your process. For example, a document authoring process may consist of the steps Create, Update, Submit, Approve, Request rework, Revise, Publish, Discard (performed by different people such as authors and editors). Some of these steps might occur more than once for a single case while not all need to happen every time.

There should be names for different process steps or status changes that were performed in the process. If you have only one entry (one row) for each case, then your data is not detailed enough. Your data needs to be on the transactional level (you should have access to the history of each case) and should not be aggregated to the case level.

Events can sometimes record not only activities you care about, but also less interesting debug information. Look for events which describe the interesting activities for your process. While you can also filter out less relevant events later in the analysis, it is helpful to start off with data that is as clean as possible.

> **Rule #2:** The activity name determines the level of detail for the process steps.

Be aware that the chosen activity influences how fine-granular you are looking at your process. Again, there may be multiple views on what makes up an activity. For example:

1. In the call center example log that comes with the demo logs for Disco (see Figure 1.3) the *Operation* attribute (Inbound Call, Handle Case, etc.) contains the obvious process steps you want to analyze. In this situation, the *Operation* column is your activity name.
   A simplified process model that can be discovered based on just the *Operation* column as process step is displayed in the lower left of Figure 1.3. (Read Section 4.1 for more information on how process maps can be discovered and simplified.)
2. At the same time, you may want to distinguish activities that take place in the 1st level support (indicated by FL, which stands for "Front Line") and in the 2nd level support (BL stands for "Back Line") of the call center. Then the *Operation* attribute plus the *Agent Position* together form your activity name.
   A process map that distinguishes *Operation* process steps based on the *Agent Position* is shown in the lower right of Figure 1.3.

To get started, it is important that you have at least one column that can be used to distinguish your process steps and serve as an activity name.

1.2.3 Timestamp

The third important prerequisite for process mining is to have a timestamp column that indicates when the activity took place. This is not only important for analyzing the timing behavior of the process but also to establish the order of the activities in your event log.

> **Rule #3:** If you don't have a sequentialized log file, you need timestamps to determine the order of the activities in your process.

Sometimes, you have a *start* and *complete* timestamp for each activity in the process (like in the call center example in Figure 1.3). This is good. It allows you to analyze the processing time of an activity (the time someone actively spent on performing that task), also called execution time or activity handling time. Refer to Section 3.1.6 to learn how to include multiple timestamps in Disco.

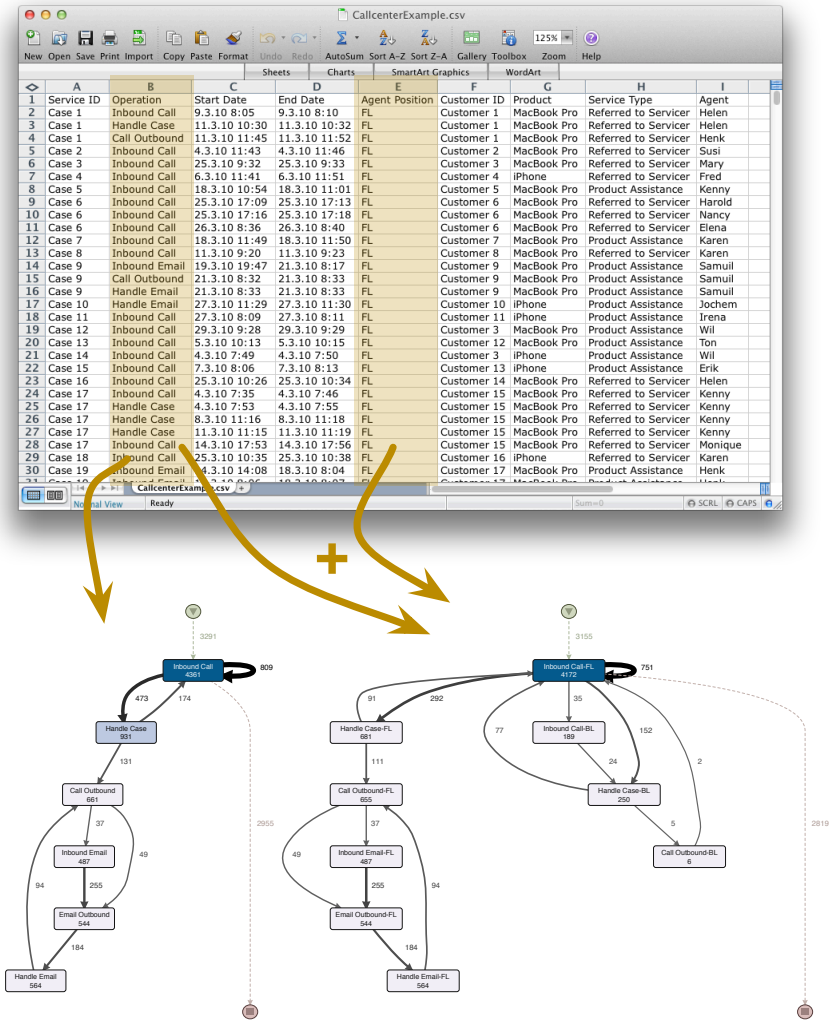


Fig. 1.3. Left: Discovered process model for the call center demo example based on choosing the Operation column as the activity name. Right: More detailed process model discovered based on using both the Operation and the Agent Position column together as the activity name.

If you have just one timestamp then you can still analyze the time between two process steps, although you will not be able to distinguish the inactive waiting time (where nobody actually worked on the process instance) and the active processing or execution times.

1.2.4 Other Columns

Additional columns can be included for the analysis if available. For example, there may be more attributes that describe specific properties that are relevant to answer the questions that you have about your process.

Which attributes are relevant for you depends on your domain. Typical additional attributes are:

- What kind of *product* the service request in a call center was about (or the order in a sales or repair process). Include this attribute if you want to compare the performance for different product categories.
- There may be *process categories* that are already defined. For example, in IT services there are different processes for managing incidents, change orders, and for fulfillment or field service. By including the process type you can separate the data and analyze the corresponding processes in isolation.
- The *channel* through which a lead came in (email or ad campaign, coupon, etc.) is often relevant for sales processes. Similarly, for repair services new requests may come in through the dealer, the call center, or the web portal.
- Processes can vary for different *partners*. For example, you may want to compare the process at different repair shops in service process.
- *Domain-specific characteristics* are influencing processes: In a repair service, there are different requirements for warranty vs. out-of-warranty repairs. In a hospital, the disease of a patient determines the precise diagnosis and treatment process, and so on.
- By which *person* or *department* was the activity handled. This information is needed for organizational handover analysis, which may reveal communication patterns and inefficiencies at the hand-over points between departments.
- If you are analyzing data from a multinational company and want to compare processes in different *countries*, then the country information needs to be pulled out of the source data as well.
- The *value* of an order is relevant for many purchasing processes, because depending on the amount of money that is involved different anti-fraud rules will apply.

These are just examples. Include any attributes you find relevant because they can improve the significance and value of the analysis. However, the Case ID, the Activity name, and the Timstamp information are the only fixed requirements.

1.3 Checklist For Your Own Data Extraction

To help you get started, here is a checklist that you can use as a guide for your own data extraction.

1.3.1 Step 1: Which Process?

At first, you will have to pick a process that you want to analyze. It is best if the process is clearly defined (you know what actions belong to it) and executed frequently. Ideally, you start with a somewhat simple process that is still relevant and could be improved.

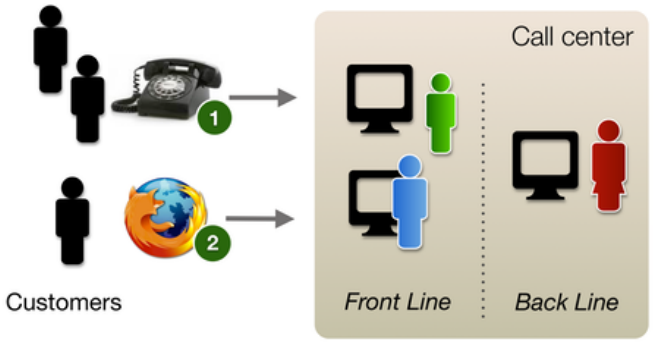


Fig. 1.4. Simplified example of a customer service process at a call center. In the call center, customers can get in touch either via phone or via a web form (which sends an email to the call center).

Also think about the questions that you want to answer: Do you want to know how the process looks like? What the most frequent or slowest paths and activities are? Or do you also want to analyze the organizational perspective, for example, how work is being transferred between different departments? The questions that you have can influence the data that you want to extract.

1.3.2 Step 2: Which IT Systems Are Involved?

Any IT system involved in the execution of the process may contain relevant data! Look out especially for CRM, ERP, workflow and ticketing systems. Their data is often most closely related to the executed process, and contains the most interesting information. However, also custom systems and spreadsheets can be analyzed. Sometimes, even very old legacy systems produce log data that can be used for process mining.

Depending on the type of involved systems you have identified, your data may be stored in a number of places. Databases are typical for large process-supporting systems and custom or legacy systems. Other systems store event logs in flat file formats. You may have to turn the logging functionality of your systems on before data is recorded.

You probably want to sit together with an administrator to help you with the data extraction. While most likely the IT staff will create the data dump for you, you

will have to tell them exactly what kind of data you need, in which format, and so on.

Often, when people ask about the "required format" they mean everything, including content, time frame, etc.—not just the actual file format.

For the file format part, it is recommended to extract a plain Comma Separated Values (CSV) file, where each row corresponds to an activity that happened in the process and each column gives some information about the context of that activity. CSV files can be extracted from almost any IT system or database. Refer to Chapter 3 for further details about the file formats that can be imported in Disco.

For further guidelines on the required columns and the timeframe, read the remaining steps in this checklist.

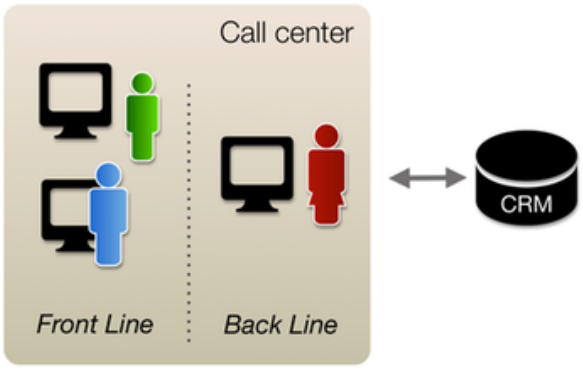


Fig. 1.5. In the call center example, all activities are recorded in a Siebel CRM system.

1.3.3 Step 3: Can The Minimum Requirements Be Fulfilled?

As explained in Section 1.2, the data need to fulfill certain minimum requirements for you to be able to carry out a process mining analysis:

What are your cases? A case is a specific instance of your process. If you look at an ordering process, for example, handling one order is one case. In an administrative process for building permits at a municipality the building permit application number is the case. For every event, you have to know which case it refers to, so that Disco can compare several executions of the process to one another. See also Section 1.2.1.

What are your activities? The activities are the process steps that happen in your process. IT systems may record not only activities you care about, but also less interesting debug information. Make sure you can capture the events that describe the interesting activities (for example, relevant milestones in your process). If less interesting activities are included, you can filter them out later. See also Section 1.2.2.

Do you have timestamps? You need at least one timestamp to be able to bring the events for each case in the right order. If you want to analyze activity durations, you need both a start and a completion timestamp for each activity. See also Section 1.2.3.

If you cannot find a suitable case ID and an activity name in your data, then you will not be able to perform a process mining analysis. So this is very important. If you do not have timestamps, then you can still discover your process if you make sure that the events are correctly ordered (in the order in which the activities have occurred). However, your analysis will be limited to the process flow perspective and you will not be able to analyze the performance of your process.

1.3.4 Step 4: Which Other Attributes Are Available?

To decide which additional attributes you should include beyond the minimum requirements, it is important to know the main goals for your process analysis. For example, do you want to compare processes for different products, or for different channels? Then you need to make sure to include the relevant product category and channel fields from the source data. You find more examples for additional attributes in Section 1.2.4.

Generally, data attributes are often important for the analysis and can be helpful to filter and focus the analysis. By looking at the data attributes that are available, you may also get further ideas for your analysis.

Include anything that might be useful. It is no problem to start out with a data set that contains dozens (e.g., up to 40 or 50) additional columns with process-relevant context information. You can always remove them later on when you import the data in Disco.

1.3.5 Step 5: Which Timeframe Should Your Log Cover?

As a rule of thumb, it is recommended to try to get data for at least 3 months. Depending on the run time of a single process instance it may be better to get data for up to a year. For example, if your process usually needs 5–6 months to complete (think of a legal case in court, or a public building permit process), a 3-month-long sample will not get you even one complete process instance.

So, it really depends on how long a case in your process is typically running. You want to get a representative set of cases and you need to keep some room to catch the usual few long-running instances as well.

If you are still unsure how much data you need to extract, use the following formula based on the expected throughput time for your process:

Formula: $timeframe = expected \times 4 \times 5$

The baseline is the *expected* process completion time for a typical case. The 4 ensures that you have as much data that you could see four cases that were started and completed after each other (of course there will be others in between). The 5 accounts for the occasional long-running cases (20/80 rule) and makes sure you see cases that take up to five times longer in the extracted time window.

For example, if the expected completion time of a typical case in your process is 5 days, then the formula yields $100\ days = 5\ days \times 4 \times 5$, which is approximately 3 months of data. If, however, a typical process is completed in just a few minutes, then extracting a couple of hours of data may be enough.

This formula is just a starting point. The more you know about your process, the better you will be able to judge the amount of data you should extract.

(a) Events extracted based on fixed timeframe

(b) Events extracted based on **cases started in a** fixed timeframe

Fig. 1.6. The difference between extracting data based on events in a specific period and based on events that belong to cases that were started in a specific timeframe.

Another way to make sure you get a good data sample is to choose a timeframe that you want to analyze (say, for example, April this year) and then extract all events for the cases that were *started* that month. This way, you can catch long-running instances even though you are focusing on a shorter timeframe for your analysis.

Figure 1.6 illustrates the difference. Every horizontal bar represents one case over time. The highlighted area stands for the selected timeframe, and the dark blue areas are the events that are covered by the data extraction method.

- In Figure 1.6(a), all events outside of the chosen timeframe are ignored, which leads to incomplete cases in your data set. These incomplete cases can be easily filtered out with Disco and are not a problem as long as you have enough data.
- In Figure 1.6(b), the events for all cases that are *started* within the chosen timeframe are kept, even if they fall outside the selected time period. This leads to a greater number of completed cases and can be useful if the chosen timeframe is short.

The amount of data you should extract also depends on the questions that you want to answer. For example, if you want to understand the typical process, then, at a certain point, adding more data will not give you any new insights. However, if you are looking for exceptions or irregularities that are important from a compliance angle, you probably want to check the data of the whole audit year to catch everything

Process Mining Checklist

Name:

Data Extraction Guide

Questions		Done
1. Which process? _____		☐
2. Which IT systems? _____		☐
3. Minimum requirements fulfilled?	yes / no	☐
Cases: _____ Activities: _____ Timestamps: _____		
4. Other important attributes ?		☐
Attribute #1: _____ Attribute #2: _____ Attribute #3: _____ Attribute #4: _____ Attribute #5: _____ Attribute #6: _____ Attribute #7: _____		
5. Timeframe:		☐
From: _____ To: _____		

that went wrong in the audited period.

You can print out the checklist from the previous page and fill it out to make sure you have identified all the important elements for your data extraction. Use it as a cheat sheet when you talk to your IT administrator or support staff about the kind of data that you need.

Once you have extracted the right data, importing your event log in Disco is really easy. You can just open your file and simply select each column to configure it either as Case ID, Activity, Timestamp, Resource, or Attribute. Then press *Start import* as shown in Figure 1.7. For further details on how to import your data, you can refer to the Import reference in Chapter 3.

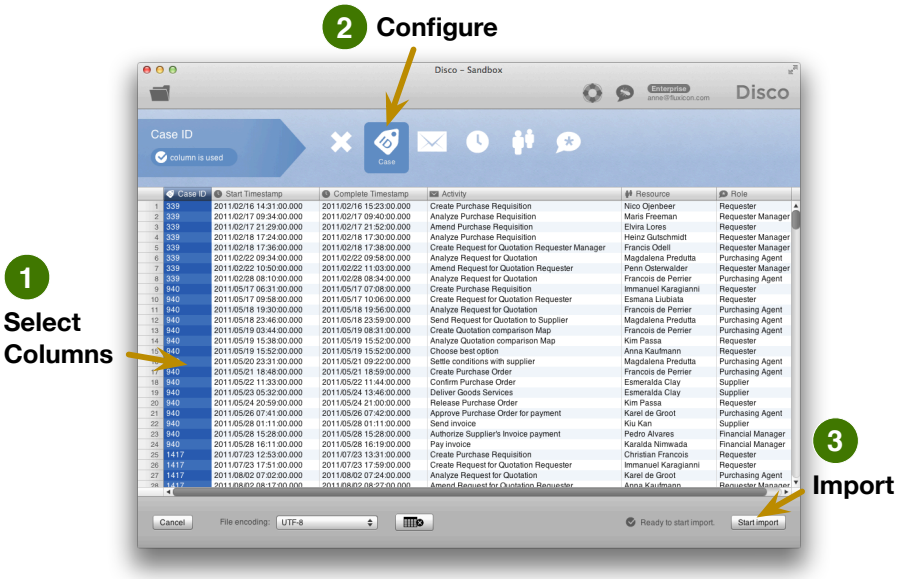


Fig. 1.7. Once you have extracted your data, importing it in Disco is easy: Simply select each column to configure it either as Case ID, Activity, Timestamp, Resource, or Attribute. Then press *Start import*.

Part I

Reference

2

Installation

2.1 Installing Disco on Windows

To install Disco on Windows, please follow these steps:

1. Download Disco from `http://fluxicon.com/disco/`.
2. When you extract the downloaded .zip file, you will find the installer *Disco-Setup.exe* and an example log file[1]. Double-click the *Disco-Setup.exe* file to start the installation (see Figure 2.1).

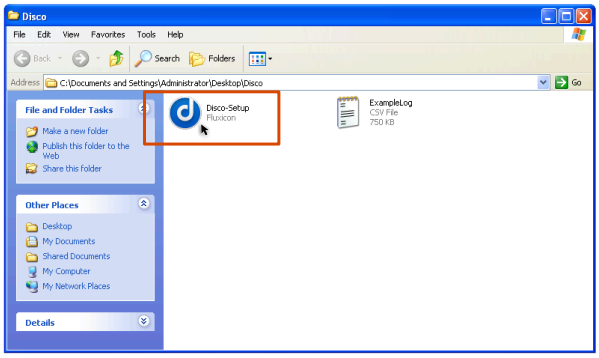
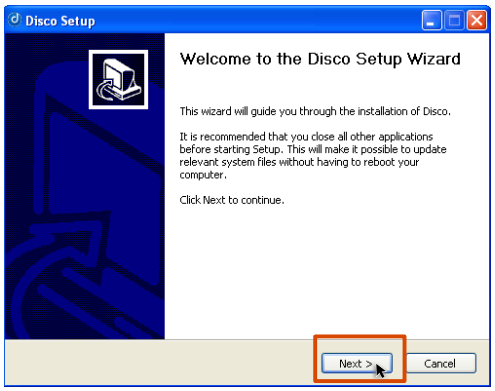


Fig. 2.1. Double-click the *Disco-Setup.exe* file to install Disco.

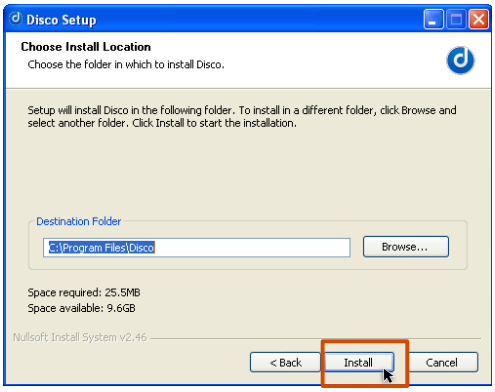
3. In the upcoming dialog click *Run* as shown in Figure 2.2(a).
4. Then click Next as shown in Figure 2.2(b).
5. Finally, click *Install* as shown in Figure 2.2(c) and *Finish*.

[1] You can download this example log and other example files at the bottom of the Disco web page at `http://fluxicon.com/disco/`.

(a) Click *Run* to start the Disco installer.



(b) Click *Next* to continue with the installation.



(c) Click *Install* to start the installation.

Fig. 2.2. Follow the instructions in the installer to complete the installation.

2.2 Installing Disco on Mac OS X

To install Disco on Mac OS X, please follow these steps:

1. Download Disco from `http://fluxicon.com/disco/`.
2. Double-click the downloaded file *Disco.dmg* to mount the disk image. A disk image named "Disco" will appear on your desktop. Double-click to open it (see Figure 2.3).

Fig. 2.3. Double-click the *Disco.dmg* disk image and drag Disco to your Applications folder.

3. The disk image contains the *Disco* application. Drag it to your *Applications* folder. Eject the disk image "Disco" by dragging it onto the trash / eject icon in your dock. Start Disco by double-clicking the *Disco* application in your *Applications* folder.

2.3 Registering Disco (Windows and Mac OS X)

When you start Disco for the first time, you will be asked to accept our license agreement and you need to register your copy of Disco. The setup is easy:

Step 1: Read the software license agreement and tick the checkbox *I have read and understood this license*. Then click the button *I accept these terms* (see Figure 2.4). If you should want to review the software license agreement again at a later point in time, Section 8.3 describes how to find it.

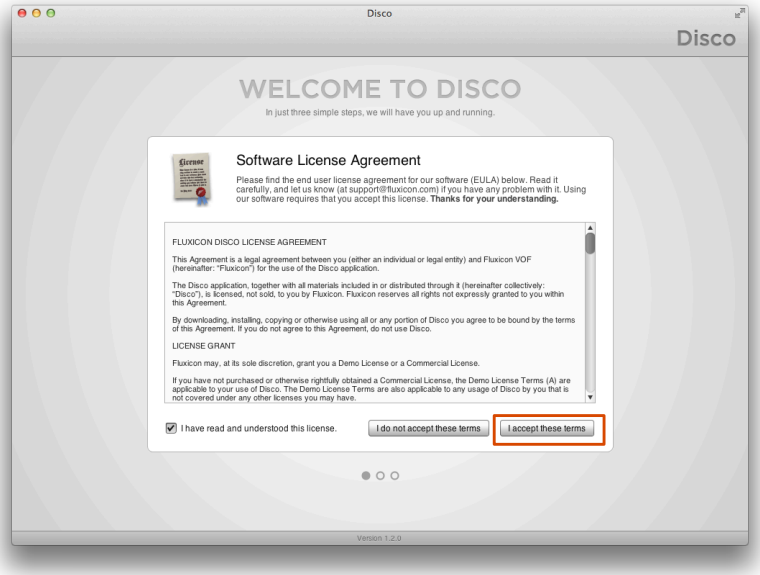


Fig. 2.4. Step 1 out of 3: Accept Disco's software license agreement.

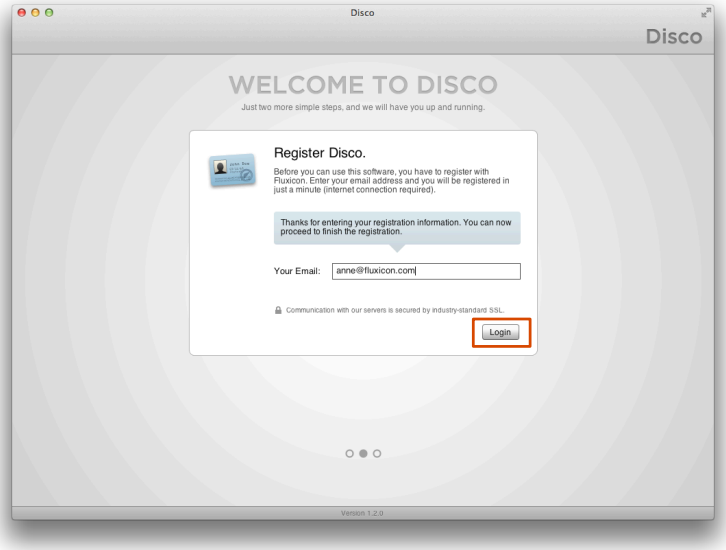
Step 2: To register Disco, fill in your email address and click *Login* as shown in Figure 2.5(a). You will receive an automatic email with the subject *Complete your Fluxicon ID registration*, which contains a personalized registration key. If you can't find the email in your inbox, please check your Spam folder.

Step 3: To activate Disco, provide the registration key that you received via email in the text field and click *Complete registration* as shown in Figure 2.5(b).

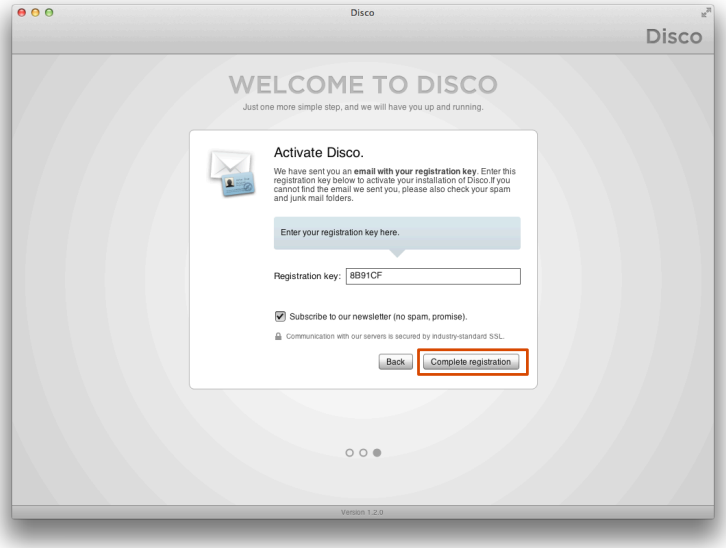
2.4 First Steps after Installation: The Sandbox

To make it easy for you to get started, we built a sandbox project into Disco that can be opened from the empty workspace as indicated in Figure 2.6(a). Simply click on the *Sandbox...* button and the sandbox project will be downloaded from the internet.

The sandbox project is shown in Figure 2.6(b). It introduces you to Disco based on the purchasing example log that is also used throughout this user's guide. Read

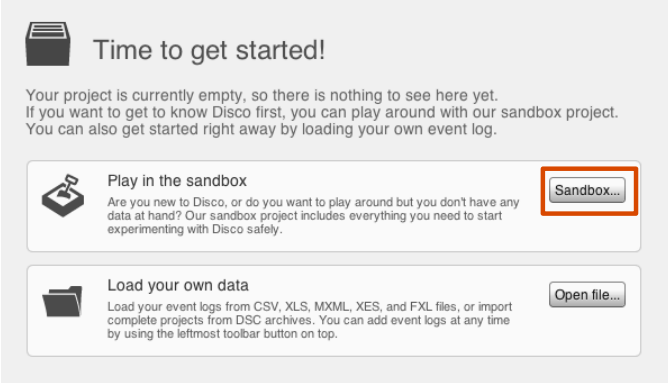


(a) Step 2 out of 3: Request your registration key.

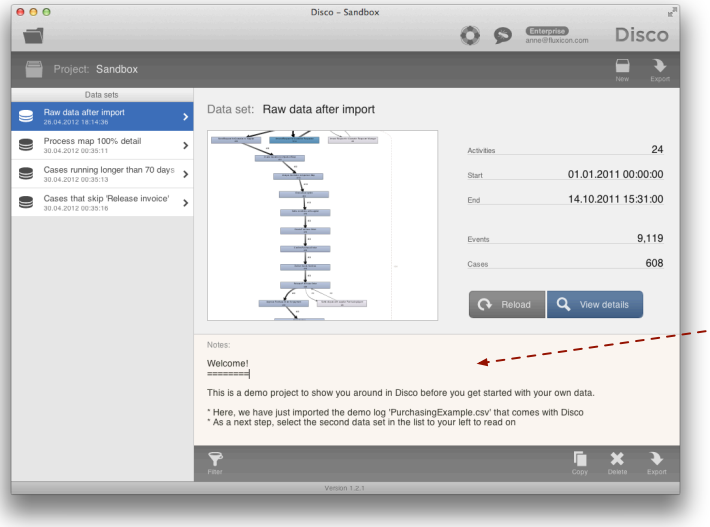


(b) Step 3 out of 3: Complete your registration.

Fig. 2.5. Request and fill in your license key in Disco.



(a) After you have completed the installation, load the *Sandbox* project to play around in Disco.



(b) Read the text in the *Notes* section and follow the instructions.

Fig. 2.6. The *Sandbox* project introduces you to Disco based on a purchasing example log.

the text in the *Notes* section and follow the instructions to get a quick tour of the functionality in Disco.

If you want to get back to the sandbox as a reference after you have imported some of your own data, you just have to clear your current project view by creating a new project (see Section 6.19).

2.5 Automatic Updates – How it Works

Updates are downloaded automatically by Disco. As soon as an update becomes available, a blue bar appears at the top (see Figure 2.7).

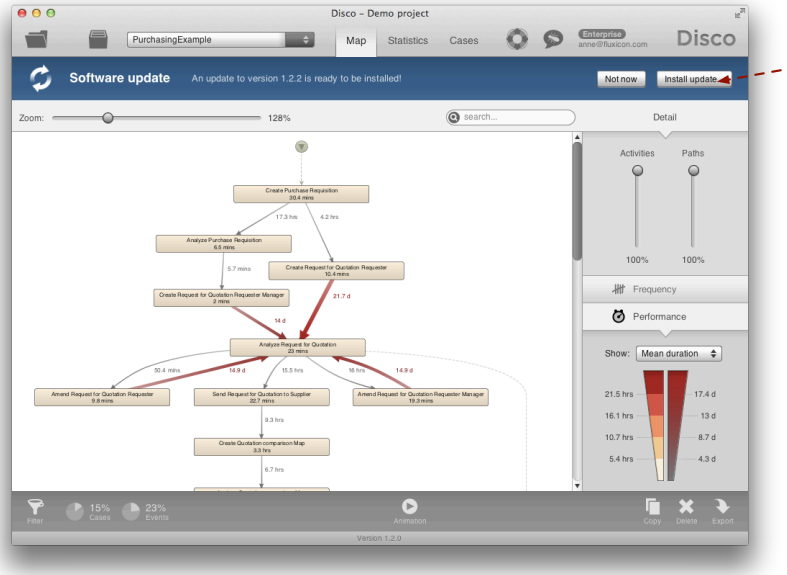


Fig. 2.7. When a new update is available, a blue banner appears at the top of Disco.

You have two choices for how to deal with the update:

- *Not now*. If you click the *Not now* button then the blue bar will disappear and the update will be automatically installed the next time you start Disco. This is useful particularly when you are in the middle of something and don't want to interrupt what you are doing by installing the update.
- *Install update*. If you click the *Install update* button, Disco tells you that you need to quit Disco and re-start it to let it install the latest update. This way, you will be working immediately with the latest version and all the improvements it has brought. To get a sense of what is new, Disco shows you a summary of the changes that have been made (see Figure 2.8). Simply click OK to make them disappear and start working with your updated version of Disco.

If you are not sure which version of Disco you are currently working with, you can always check the version number in the footer at the very bottom of the screen.

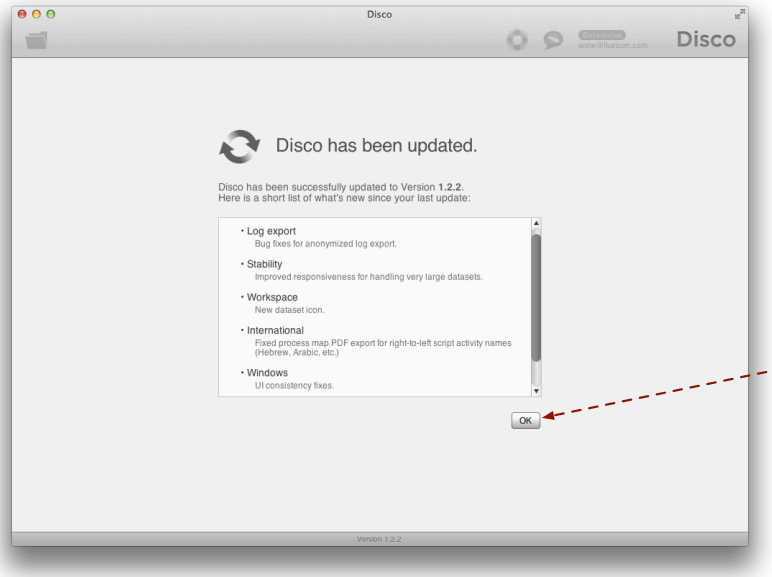


Fig. 2.8. The *Sandbox* project introduces you to Disco based on a purchasing example log. Read the text in the *Notes* section and follow the instructions to get acquainted.

3

Import

In Chapter 1 you have learned about what kind of data is needed to do a process mining analysis.

Disco has been designed to make the data import really easy for you by automatically detecting timestamps, remembering your settings, and by loading your data sets with unprecedented speed. In this chapter you find further details about which kind of files you can load in Disco and how the data import works.

3.1 Importing Data Sets

Importing is symbolized by the folder icon shown in Figure 3.1.

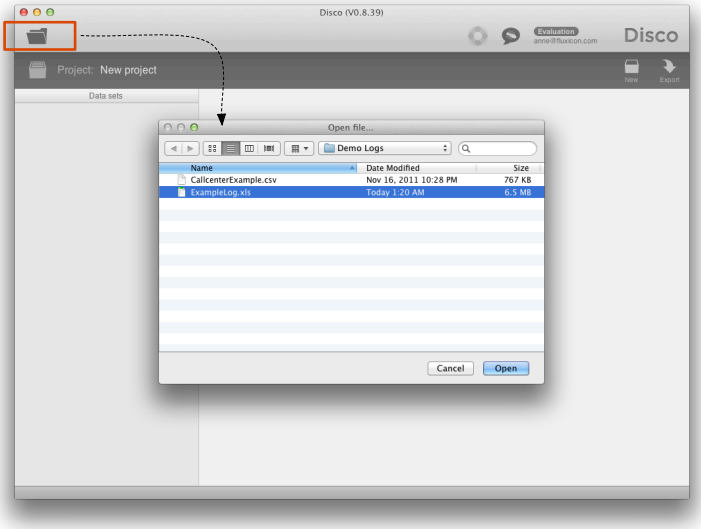
Fig. 3.1. Open symbol in Disco.

Clicking on the folder symbol will open up the file chooser dialog and let you import a data set into your workspace. You find the Open symbol in the top left of your Disco window as shown in Figure 3.2(a).

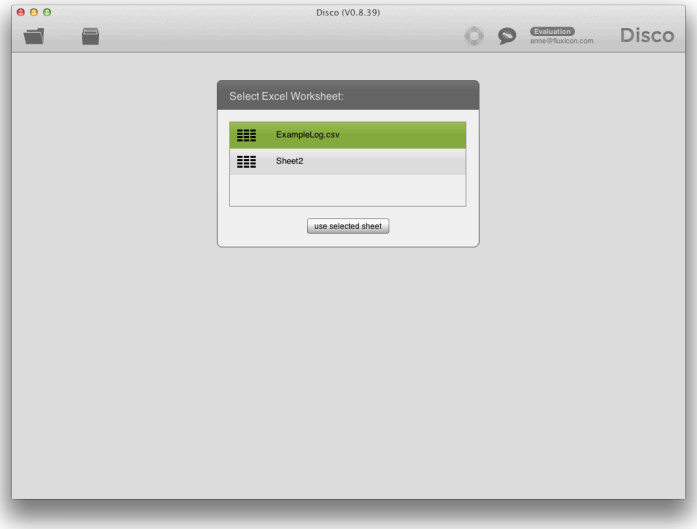
Most of the times you will open and configure files with the following extensions:

- .csv (Comma Separated Values)
- .txt (Text files)
- .xls and .xlsx (Microsoft Excel)

If you open an Excel file that contains multiple non-empty sheets, then Disco lets you choose which one you want to import like shown in Figure 3.2(b). Read



(a) Pressing the open symbol brings up the file chooser dialog. (On Windows you will see the standard Windows file chooser dialog.)



(b) If you have multiple non-empty sheets in your Excel file, then Disco lets you choose which one you want to import.

Fig. 3.2. Opening a file in Disco.

on to learn how the required format for these files looks like, and how exactly the configuration works in Disco.

Furthermore, Disco reads pre-configured files in various standard formats:

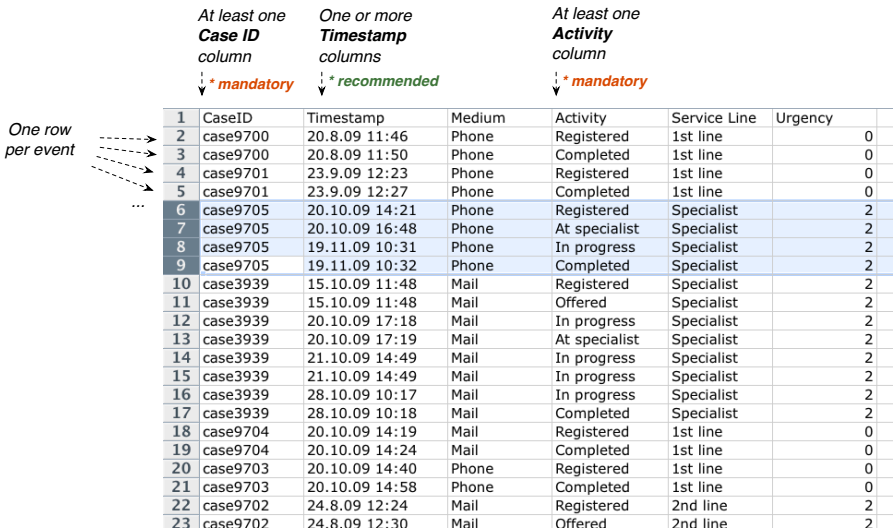
- .mxml and .mxml.gz (see Section 3.2.1 for more info on MXML logs)
- .xes and .xes.gz (see Section 3.2.1 for more info on XES logs)
- .fxl (see Section 3.2.2 for more info on Disco log files)
- .dsc (see Section 3.2.3 for more info on Disco project files)

Jump to Section 3.2 if rather than importing CSV or Excel files what you want is to load files with the extensions above.

3.1.1 Required format for CSV, Excel and TXT Files

You can open files in the format shown in Figure 3.3. Every line or row is expected to contain the information about one event or executed activity in your process. You need at least one case ID column, an activity, and ideally one or more timestamps.

If you don't know what a case ID or an activity is, please read the introduction about event logs in Chapter 1.



	CaseID	Timestamp	Medium	Activity	Service Line	Urgency
1						
2	case9700	20.8.09 11:46	Phone	Registered	1st line	0
3	case9700	20.8.09 11:50	Phone	Completed	1st line	0
4	case9701	23.9.09 12:23	Phone	Registered	1st line	0
5	case9701	23.9.09 12:27	Phone	Completed	1st line	0
6	case9705	20.10.09 14:21	Phone	Registered	Specialist	2
7	case9705	20.10.09 16:48	Phone	At specialist	Specialist	2
8	case9705	19.11.09 10:31	Phone	In progress	Specialist	2
9	case9705	19.11.09 10:32	Phone	Completed	Specialist	2
10	case3939	15.10.09 11:48	Mail	Registered	Specialist	2
11	case3939	15.10.09 11:48	Mail	Offered	Specialist	2
12	case3939	20.10.09 17:18	Mail	In progress	Specialist	2
13	case3939	20.10.09 17:19	Mail	At specialist	Specialist	2
14	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
15	case3939	21.10.09 14:49	Mail	In progress	Specialist	2
16	case3939	28.10.09 10:17	Mail	In progress	Specialist	2
17	case3939	28.10.09 10:18	Mail	Completed	Specialist	2
18	case9704	20.10.09 14:19	Mail	Registered	1st line	0
19	case9704	20.10.09 14:24	Mail	Completed	1st line	0
20	case9703	20.10.09 14:40	Phone	Registered	1st line	0
21	case9703	20.10.09 14:58	Phone	Completed	1st line	0
22	case9702	24.8.09 12:24	Mail	Registered	2nd line	2
23	case9702	24.8.09 12:30	Mail	Offered	2nd line	2

Fig. 3.3. Example file in Excel: You need to have one row for each activity that was performed in the process, plus a case ID, an activity, and a timestamp.

Note that the rows in your file do not need to be sorted. Disco will sort the activities per case based on the timestamps in your file. Only if you use a log that does not have timestamps, the events will be imported in the order in which they appear in the file.

The order and the name of the columns do not matter either. For example, it is not necessary that your case ID column is called "CaseID" as in Figure 3.3. What is important is that you have *any* columns that can be used as a case ID and as activity name. You can tell Disco which column means what in the configuration step (see Section 3.1.2).

The timestamp does not need to have a specific format. Instead, Disco reads the timestamps in the format that you provide (see Section 3.1.3 to learn how timestamp patterns can be configured). If you have multiple timestamp columns that indicate when an activity has been scheduled, started, and completed, then you can make use of that (see Section 3.1.6 for further details).

Finally, you can have as many additional data columns as you like. They will be included as attributes that can be used in your analysis later on.

If your data is not in Excel but in a database or some other information system, then the best format to extract the data is as a Comma Separated Values (CSV) file. Logically, a CSV file is very similar to the Excel table from Figure 3.3: Each line contains one event and the different column cells are separated by a comma (or some other delimiting character). You can open a CSV file in a standard text editor.

Here is an example excerpt of an event log in CSV format:

```
case01,Request Quotes,Tom,2009-06-12 09:45
case01,Authorization,Peter,2009-06-12 13:11
case23,Request Quotes,Mary,2009-06-12 15:29
case01,Compile Result,Jonas,2009-06-13 10:01
case23,Authorization,Peter,2009-06-13 11:25
case01,Check Result,Amanda,2009-06-13 15:09
case23,Compile Result,Andy,2009-06-14 09:15
```

The first line describes an event which:

- Has occurred while executing the case case01,
- Has been triggered by the execution of activity Request Quotes,
- Where this activity had been executed by resource Tom,
- And where the activity had been executed on 12 June 2009, at quarter to ten in the morning.

The delimiting character can be either a comma (","), a semicolon (";"), a tab ("\t"), or a pipe ("|") character. If your delimiting character is contained in the contents of your file, then the content elements need to be grouped by quotes. For example, if you use the comma as delimiting character and your activity names contain commas as well:

```
case01,Request Quotes, Standard,Tom,2009-06-12 09:45
```

then the activity name needs to be surrounded by quotes:

```
case01,"Request Quotes,Standard",Tom,2009-06-12 09:45
```

All this is pretty standard and will be done automatically by most databases or other export functions.

3.1.2 Import Configuration Settings

Once you have opened your Excel, CSV, or text file in Disco, you see an import configuration screen as shown in Figure 3.4.

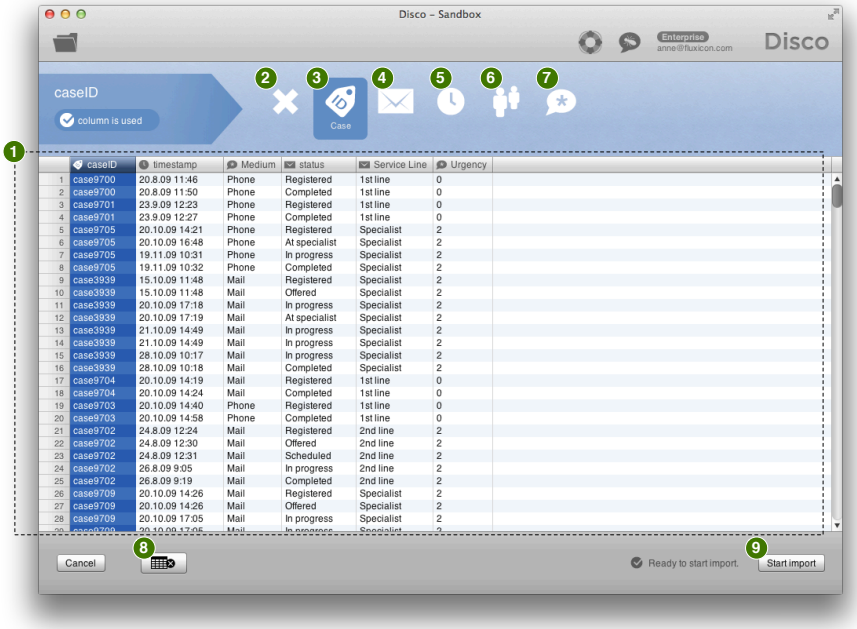


Fig. 3.4. Import Configuration screen in Disco.

On the import screen, you see the same columns as if you had opened the file in Excel (see Figure 3.3 for comparison). Disco starts by guessing what each column might mean (trying to identify case ID, activity, and timestamp), but you can check and correct the configuration before the import proceeds.

❶ *Preview of the loaded file*. A preview of the file you have loaded helps you through the configuration step. Select each column to configure it.

The following configuration options are available :

Remove. ❷ Ignores the selected column (will not be imported at all).

Case. ❸ Set the selected column as the case ID.

Activity. ❹ Configures the selected column as activity name.

Timestamp. ❺ Indicates that the selected column contains the timestamp.

Resource. ❻ Configures column as resource column for organizational analysis.

Other. ❼ Includes the selected column as an additional attribute.

The current configuration for each column is indicated by the little configuration symbol in the header of each column in the previewed data table ❶.

❽ *Exclude/Include all*. With this button you can exclude or include all columns at once. This is particularly helpful if you have many different columns and only want to include or exclude a few of them.

❾ *Start import*. After you have configured your columns, you can start to import the log by pressing this button.

Minimum Requirements

You need to configure at least one case ID column and one activity column before you can start the import of your file.

Disco tells you which configuration step is missing if these minimum requirements are not met: For example, in Figure 3.5(a) you see a configuration where no activity column has been defined yet and, therefore, the *Start import* button is still inactive. In Figure 3.5(b), the status column is used for the activity name and the log is ready for import.

If you are not sure what the case ID or activity should be for your log, you can learn more about event logs in the introductory Section 1.2.

Import Progress

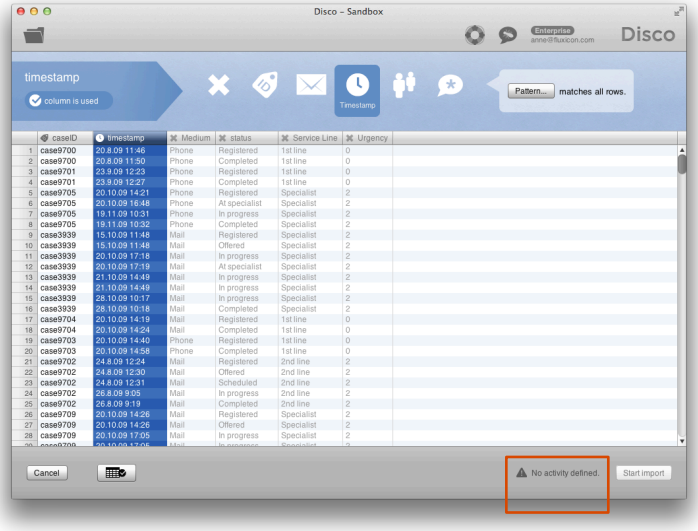
As soon as you have pressed the *Start import* button, Disco starts reading the complete file according to the given configuration. The progress of the import is shown along with an indication of how much data has been already read; see ❶ in Figure 3.6(a).

Disco has been designed for speed and usually loads your data really fast. So, most of the times the import will be finished before you can actually read the progress indicator. However, if it should take longer and you want to abort the import process, then you can do that by pressing the **x** to the right of the progress bar; see ❷ in Figure 3.6(a).

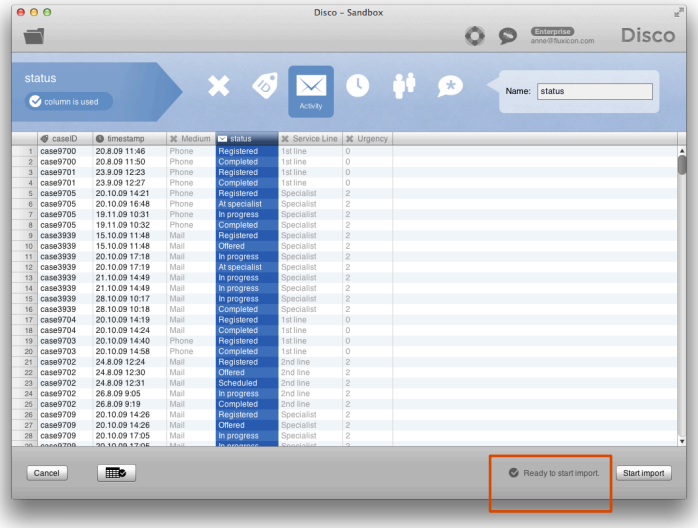
After the import is finished, you are directly taken to the Map view, where you can inspect the process flow and start your analysis; see Figure 3.6(b). Learn more about how to work with process maps in Disco in Chapter 4.1.

Saving the Import Settings

The import configuration is saved automatically and will be restored the next time you load the same or a similar file.

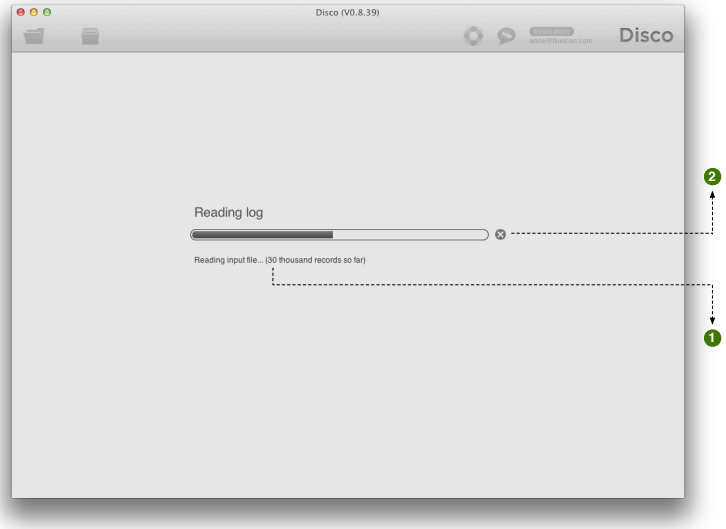


(a) Disco tells you if no activity or case ID column has been configured yet.

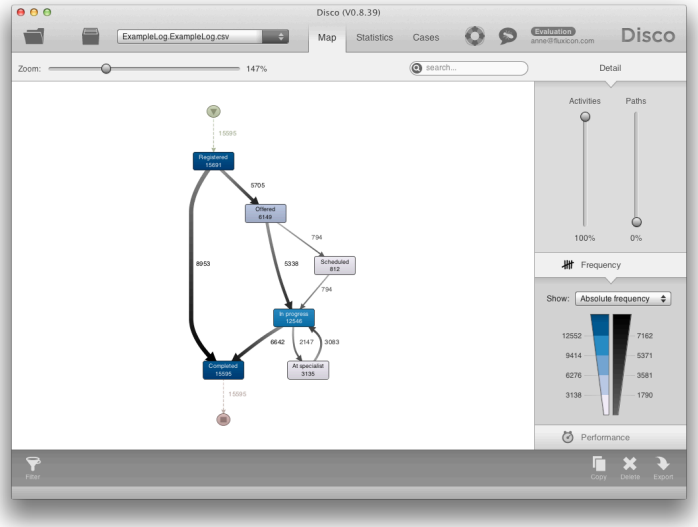


(b) As soon as all required columns are configured, the actual log import can be started.

Fig. 3.5. Disco guides you through the configuration process and tells you which configuration steps are still missing.



(a) After you have configured your columns, the import can be started. You'll see the progress (❶) and can stop the import if needed (❷).



(b) As soon as the import is finished, you are directly taken to the Map view.

Fig. 3.6. After you have completed the configuration, your file is imported and you can immediately inspect the process flow.

3.1.3 Configuring Timestamp Patterns

Timestamps can come in various formats. For example, the date 4 August 2012 might be represented as `04/08/12` or `08/04/12`, as `2012-08-04`, or as `4.8.2012`, and in many other ways. The same holds for the time of the day. Different conventions regarding the order, separating characters, with or without spaces, etc. often make it a pain for data analysts to deal with timestamps.

Disco makes it as easy for you as possibly possible: When it parses the first rows of your data set to make suggestions for how you might want to configure your data columns (see Section 3.1.2), different timestamp patterns are tested against your data to see which one gives the best match. This means that *in more than 90% of the cases your timestamp format is automatically detected* and you do not need to manually configure anything about it at all.

You can verify that everything is in order by selecting your timestamp column in the configuration screen—like in Figure 3.5(a): If Disco says that the pattern matches all rows (see Figure 3.7 below) then everything is OK.

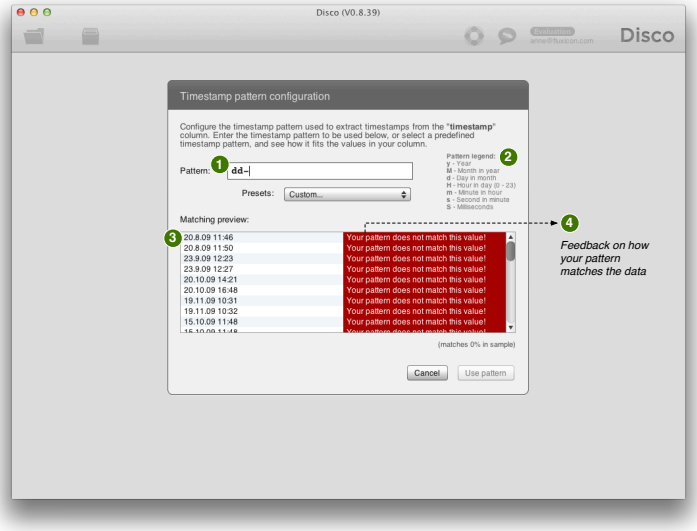
Fig. 3.7. Feedback on how many rows of the selected timestamp column match the current timestamp pattern.

To inspect and change the timestamp pattern press the *Pattern...* button (see ❶ in Figure 3.7). This will bring up the timestamp pattern configuration screen shown in Figure 3.8.

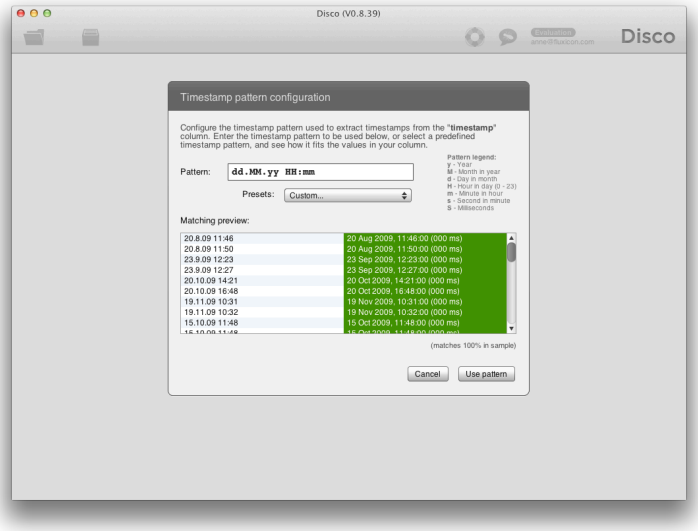
Disco allows you to directly specify the date and time pattern in Java's Simple Data Format[1]. To specify a custom timestamp pattern, you can follow these steps:

1. Look at the first couple of example dates from your file in the preview; see ❸ in Figure 3.8(a). Find out how the pattern starts: Is it the year? The month? For example, in Figure 3.8 it is the day of the month.
2. Check the pattern legend; see ❷ in Figure 3.8(a). Find out which letter represents your date or time entity. For example, the day of the month is represented by the lower case letter "d".
3. Start typing your custom pattern in the pattern entry field; see ❶ in Figure 3.8(a). Type the letter representing your date or time entity as many times as the corresponding entity has digits in the pattern. For example, the date of the month is represented by up to two digits. So, you would start typing `dd` to match the day at the beginning of your timestamp pattern.

[1] Further documentation and examples about the Simple Date Format are available on Oracles's documentation page here: `http://docs.oracle.com/javase/1.4.2/docs/api/java/text/SimpleDateFormat.html`.



(a) If the current pattern (❶) does not match your data (❸) then the interactive preview will give you direct feedback (❹).



(b) Once you have found a matching pattern you can use it.

Fig. 3.8. The timestamp pattern configuration screen allows you to inspect and modify the timestamp pattern to fit your data.

4. Look at the timestamp examples in your preview (❸) again and find out how the pattern separates the different date and time entities. Is it a dot (.) or a dash (-) or just a white space? For example, in Figure 3.8 the day of the month and the month are separated by a dot. So, you would expand your pattern from `dd` to `dd.MM` for matching both day and month.
5. Verify in each step that the pattern that you have provided so far gives the expected result. The preview window gives you interactive feedback while you are typing your pattern; see ❹ in Figure 3.8(a). For example, in Figure 3.8(a) the pattern `dd-` is wrong and does not match the timestamps in the data sample.
6. Continue until all date and time elements are matched completely. For example, the pattern `dd.MM.yy HH:mm` matches all timestamps in Figure 3.8.

Once you have matched all date and time elements for the example timestamps in your preview, you can press *Use pattern* and Disco will remember your custom pattern for the future.

3.1.4 Combining Multiple Case ID, Activity, or Resource Columns

In most situations, there is a clear candidate for the case ID and the activity column. But often there are multiple views that one can take on the data.

In Chapter 1 we have explained the mental model that underlies process mining and how, for example, the activity name determines the level of detail for the process steps. If you refer back to Figure 1.3, then you can see how a more detailed process map for the call center example could be discovered based on using both the Operation and the Agent Position column together as the activity name.

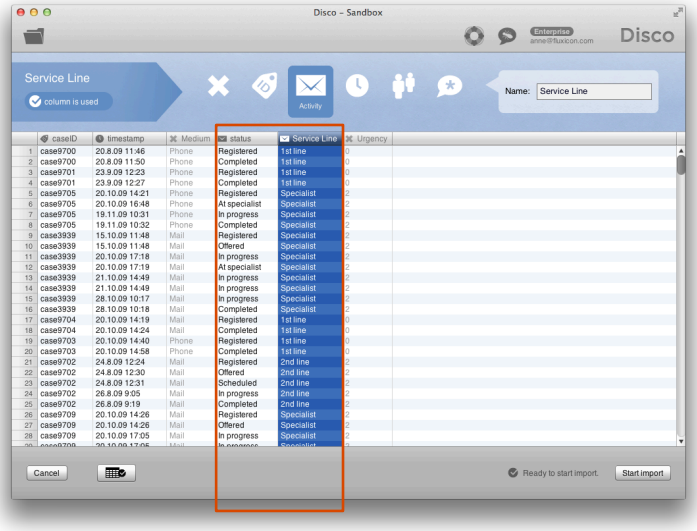
To make it easy for you to explore multiple views, Disco allows you to combine multiple columns for:

- the case ID,
- activity, and
- resource field.

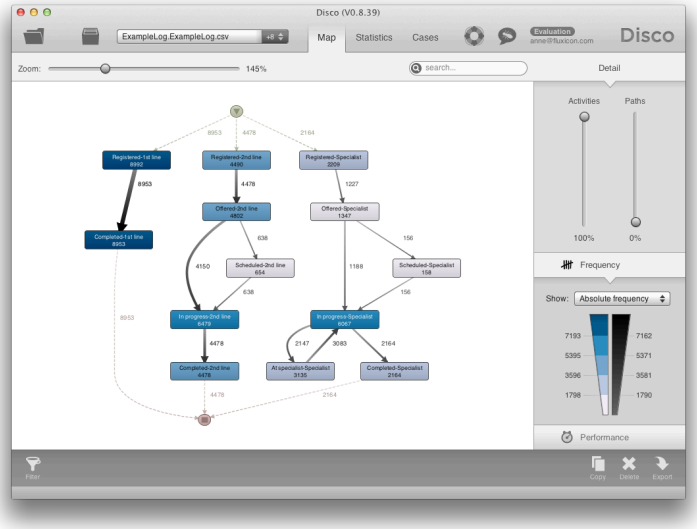
This works by simply assigning not just one but multiple columns as your case ID, activity, or resource. The contents of these columns will be concatenated (joined together).

For example, in Figure 3.4 only the *Status* column was configured as the activity. The *Service Line* column was included as a plain attribute. This leads to the process map shown in Figure 3.6(b), with activity names such as ‘Offered’, ‘Scheduled’, and ‘In progress’. Alternatively, you could configure both *Status* and *Service Line* as activity like shown in Figure 3.9(a). This then gives you a more detailed process map as you can see in Figure 3.9(b), with activity names such as ‘Offered-2nd line’, ‘Offered-Specialist’, and ‘In progress-Specialist’.

In the example in Figure 3.9 the *Service Line* is a static attribute—assigned for each case depending on where it was eventually resolved (so, you don’t see the actual transfer of a case taken in by the 1st line and handed over to the 2nd line). This is due



(a) By including the service line (1st line, 2nd line, or Specialist) into the activity name ...



(b) ... one can distinguish activities that were performed for cases resolved in these different parts of the organization.

Fig. 3.9. Multiple columns can be combined to create a more detailed process view.

to the way the data was recorded (see the call center example in Figure 1.3, where the service line handovers are recorded in a more detailed manner).

As a result, you get to see the different process flows for the 1st line, 2nd line, and Specialist process categories in one view, next to each other. You could do the same for different products, departments, or other interesting categories for your process. A similar result can be achieved by using the Attribute filter (see Section 5.2.5) to explicitly filter different versions of the process based on the *Service Line* attribute (or other categories).

3.1.5 Swapping Cases, Activities, and Resources

Similar to combining multiple activity columns (see Section 3.1.4), different views can be taken on the process by changing the column configuration.

For example, in Figure 3.10(a) the *Resource* column from the purchasing example log is configured as the activity. The result is a process map where you can see how work is transferred between people as shown in Figure 3.10(b).

Similarly, in Figure 3.11 a role-based view on the process is shown. By configuring the *Role* column—see column to the right of the Resource column in Figure 3.10(a)—as the activity, one can see how the process flows between different roles in the organization. The performance visualization in Figure 3.11(b) highlights that the manager is a bottleneck by taking about seven days on average before forwarding the case to the purchasing agent.

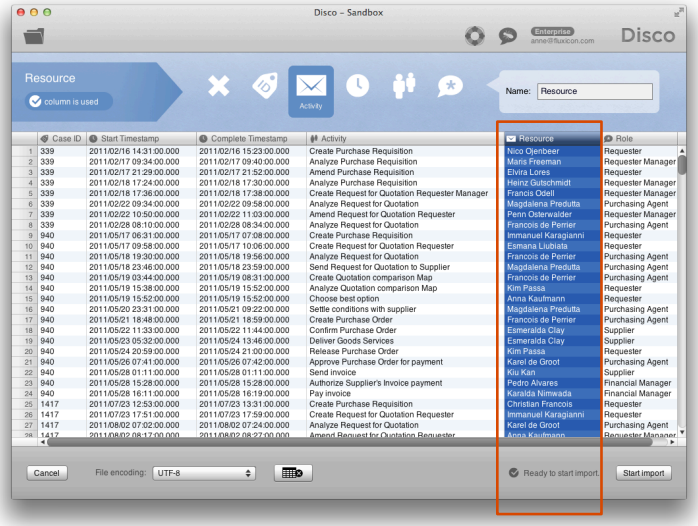
To give a case ID swapping example, imagine a healthcare process where one would normally want to look at the diagnosis and treatment process from a patient perspective. So, the patient ID column would be the natural configuration for the case identifier. However, it can also be interesting to compare how different physicians work to identify and promote best practices. In this case, the physician's name (normally configured as a resource column) would be part of the case ID.

Depending on how you look at your data, multiple views on the process can be taken (see also Chapter 1). By making it easy to interpret columns in different ways as shown here, by combining multiple columns (Section 3.1.4), and by allowing you to go back and adjust the configuration (Section 3.1.7), Disco encourages the exploration of your process in a multi-faceted way.

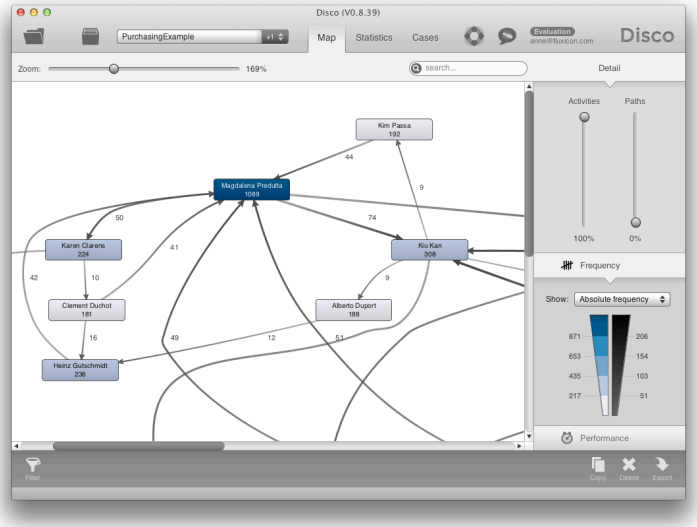
3.1.6 Including Multiple Timestamp Columns

At least *one* timestamp is usually needed to bring the events for each case in the right order, and in many situations one timestamp is all you have (like in the example in Figure 3.4). A single timestamp is enough to discover the process flows, to measure the time between status changes, and the time to run through the whole process from start to finish etc.

However, sometimes you have multiple timestamps for the start and completion of an activity. This is great because it allows you to measure not only the time between different process steps but also the actual execution time for an activity. If you

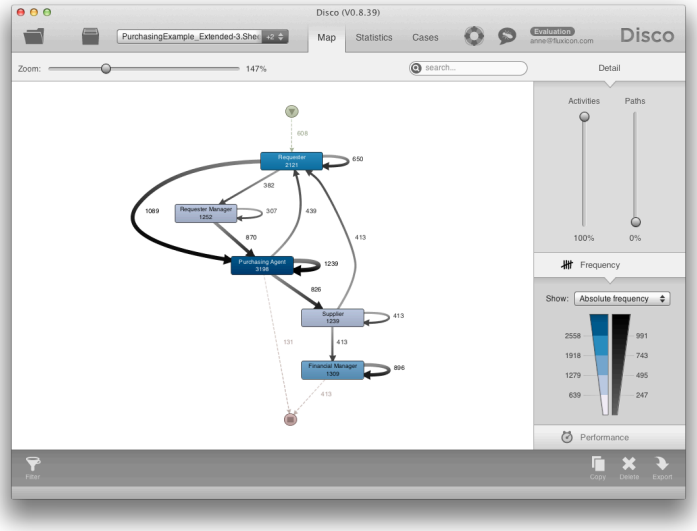


(a) By configuring the resource column as the activity ...

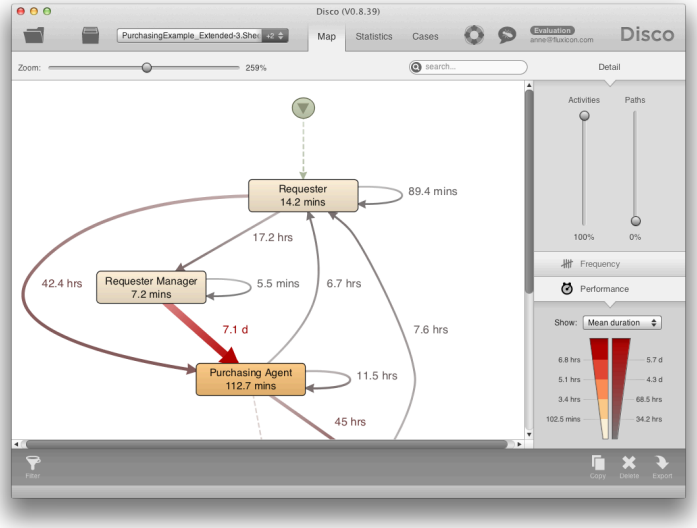


(b) ... you can see how the process flows between people.

Fig. 3.10. By swapping activities and resources in the configuration step, you can see how work is transferred between persons rather than which activities follow each other.



(a) By configuring the role column as the activity, the process flow between functions is visualized.



(b) Due to low availability the manager causes most of the delays in the purchasing process.

Fig. 3.11. A role-based view on the purchasing process.

	Service ID	Operation	Start Date	End Date	Agent Position
1	Case 1	Inbound Call	9.3.10 8:05	9.3.10 8:10	FL
2	Case 1	Handle Case	11.3.10 10:30	11.3.10 10:32	FL
3	Case 1	Call Outbound	11.3.10 11:45	11.3.10 11:52	FL
4	Case 2	Inbound Call	4.3.10 11:43	4.3.10 11:46	FL
5	Case 3	Inbound Call	25.3.10 9:32	25.3.10 9:33	FL
6	Case 4	Inbound Call	6.3.10 11:41	6.3.10 11:51	FL
7	Case 5	Inbound Call	18.3.10 10:54	18.3.10 11:01	FL
8	Case 6	Inbound Call	25.3.10 17:09	25.3.10 17:13	FL

Fig. 3.12. If you have start and completion timestamps, make sure to include both.

have two timestamps per activity, you can simply configure both of them as a timestamp column as shown in Figure 3.12. Disco will interpret the earliest timestamp as the start and the latest timestamp as the completion of the activity for each row.

As a result, the active time (where somebody is actually working on the activity) and passive time (where nothing happens to the case) can be distinguished in the analysis. For example, in Figure 3.13 you see a fragment of the resulting process map, where average durations are attached both to the activities themselves as well as to the (idle) times between them. Refer to Section 4.1.4 to learn more about how the process maps in Disco can be annotated with performance information.

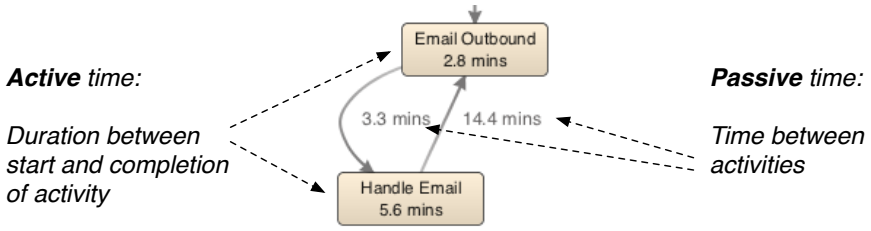


Fig. 3.13. If start and completion timestamps are available, then execution times (active) and waiting times (passive) can be distinguished in the analysis.

In some situations, you may have even more than two timestamps. For example, in Figure 3.14 another call center log fragment is shown. There are three columns that hold relevant timestamp information: `ACT_CREATED`, `ACT_START`, and `ACT_END`:

- The `ACT_END` column always holds the completion timestamp for the agent's activity.
- The `ACT_START` column holds the actual start time for timed activities; for non-timed activities the corresponding field is empty.
- The `ACT_CREATED` column is always filled, but for timed activities the start time does not reflect the actual start time. The time difference between `ACT_START` and `ACT_CREATED` for timed activities is significant for call center analyses as timing is very important.

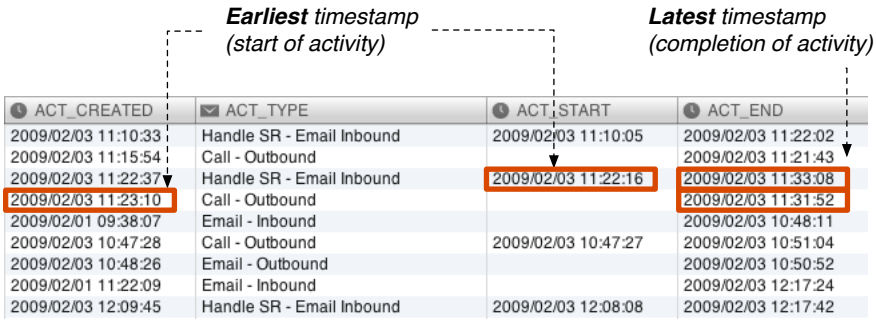


Fig. 3.14. Among multiple timestamp columns per row, the earliest is taken as the start and the latest as the completion of the activity.

With Disco this situation can be resolved very easily. Simply include all three columns as a timestamp column during configuration and the earliest and latest timestamp will be chosen as the start and completion time for each activity.

3.1.7 Adjusting the Import Configuration

As shown in Section 3.1.4 and Section 3.1.5, there are often multiple views that can be taken on your process. To explore these different views, or perhaps to correct a configuration mistake, you sometimes want to change the configuration of your data set after you have completed the import.

In Disco, you can adjust your import configuration in two different ways:

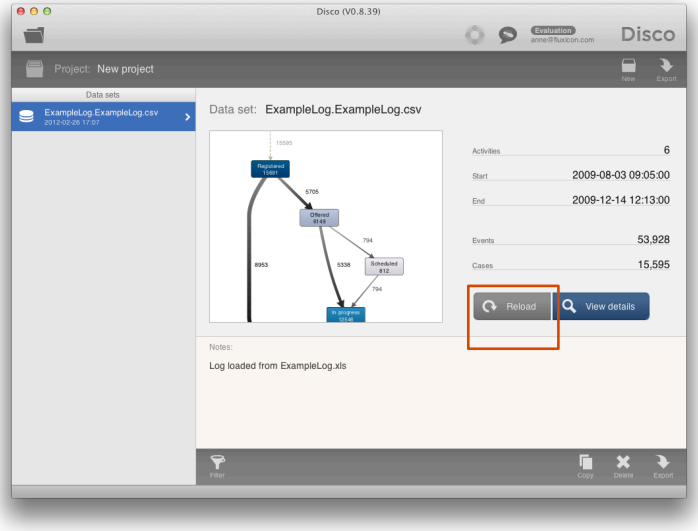
- Press the *Reload* button in the project view and you will be brought back to the configuration screen as shown in Figure 3.15.
- Simply import the same file again and configure it another way. Disco will remember your configuration settings from the last time you imported it. So, a re-import has the same effect as using the *Reload* button.

In the case that you already have done some filtering work (see Chapter 5) that you need to preserve but still want to change the configuration, then re-loading the original file does not help. Instead, you can do the following:

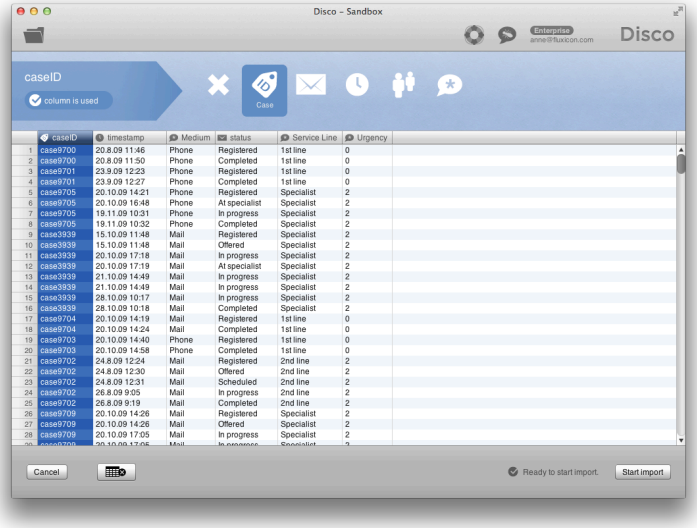
1. Export the currently filtered data set as a CSV file (see Chapter 7).
2. Import the exported CSV file and configure it the way you want it.

3.2 Importing Pre-configured Data Sets

In most situations, the starting point is data in CSV or XLS format as described before. However, you can also import data sets in standard event log formats such as MXML and XES (Section 3.2.1). Furthermore, efficient Disco-compressed log files (Section 3.2.2) and complete Disco projects (Section 3.2.3) can be imported, too.



(a) Pressing *Reload* in the project view ...



(b) ... brings you back to the configuration screen.

Fig. 3.15. The *Reload* button is a quick way to go back to the configuration screen to check how the data looked like, to potentially correct configuration mistakes, or to take alternative views on your process.

3.2.1 MXML and XES

The Mining XML (MXML) format has been around as a standard format for event logs for several years now. The EXtensible Event Stream (XES) format is the successor of MXML and has been approved by the IEEE Task Force on Process Mining in 2010.

Standard formats are useful to facilitate the interoperability between different tools. For example, logs in MXML or XES format can be loaded by other process mining tools such as the academic toolset ProM. Disco places a high value on interoperability and imports and exports all event log standard formats that are in use. Read more about the supported standard event log types in Section 7.2.1.

Even if you do not work with other process mining tools, the MXML or XES import can be still useful for you:

No configuration: Data sets exported in MXML or XES already contain the configuration information (about case IDs, activities, timestamps, etc.). So, if you want to exchange a data set with a co-worker who also uses Disco, you can send her an MXML or XES file to let her skip the configuration step entirely. This way, you do not need to explain which columns should be configured how, but instead she opens the file and directly sees the process map. Or similarly, if you want to keep differently filtered or configured versions of your event log as files for yourself, storing them in MXML or XES standard format saves you the configuration step when you next import these files again.

Faster import: Next to skipping the configuration step during import, reading a pre-configured data set is also more efficient because it is already sorted (in contrast, the rows in a CSV file still need to be sorted by Disco). For larger data sets, the difference can be a 10x speedup (and more) during import.

If you should have MXML or XES files that you would like to re-configure in a different way, you can still export them as a CSV and re-import the CSV again (see also Section 3.1.7).

3.2.2 Disco Log Files

Loading data sets in MXML or XES format is already faster than importing CSV data (see also Section 3.2.1), but nothing is as fast as loading data in the native Disco Log Files (FXL) format. FXL is a proprietary (no standard) format but the best if you need to exchange really large data sets with another Disco user.

For example, for one of our large benchmark logs the import of the CSV file (including sorting) took 3.5 hours, the MXML file loaded in about 20 minutes, and the FXL file was read in just over one minute.

Files in FXL format have also a much smaller file size compared to compressed XML standard logs and CSV.

3.2.3 Disco Projects

Not only individual log files but complete projects can be exported and imported. This way, multiple data sets including all applied filters and notes can be shared with other Disco users or used to backup and re-load previous project work. Refer to Section 6.3 to read more about how to export and import Disco projects as DSC files.

3.3 Troubleshooting

Here are a few typical problems that may occur during the import of your file, and what to do about them.

3.3.1 "Columns are empty when I load my file"

When you import an XLS file (see Section 3.1), it may be the case that the cells of some of the columns in your Excel data sheet are empty because they contain formulas. Formulas or references to other cells cannot be interpreted by Disco.

To import the values (the results of a formula) as your data set into Disco, you can use one of these two alternative solutions:

- *Option 1: Copy and paste values*. You can copy the columns that contain cells with formulas and paste them with the `Paste special - Values` option. Make sure that you only paste the values of your column, because the standard copy / paste functionality in Excel will copy the formulas into your new column!
- *Option 2: Export XLS as CSV file*. A file in CSV format is a plain text file that contains no formulas. If you export your data set from Excel as a CSV, then only the resulting values of all your cell formulas will be exported. The exported CSV file can be imported by Disco without problems.

3.3.2 "My date and time are in separate columns"

If you configure multiple columns as timestamp, Disco will parse each of them to determine the start and completion times for the activities. This means that if your source data has one timestamp distributed over two columns (see *Date* and *Time* columns in Figure 3.16), Disco cannot get the complete timestamp.

As a solution, you can open your data set in Excel and create a third column (see *Date & Time* in Figure 3.16) that combines the two. The `CONCATENATE` function in Excel can be used to do that. After you have created your combined timestamp column, you can save or export the data set and configure the *Date & Time* column as your timestamp column in Disco.

In some cases the `CONCATENATE` formula does not work because the *Date* and *Time* values are interpreted by Excel as a number. To solve this problem, you have two options:

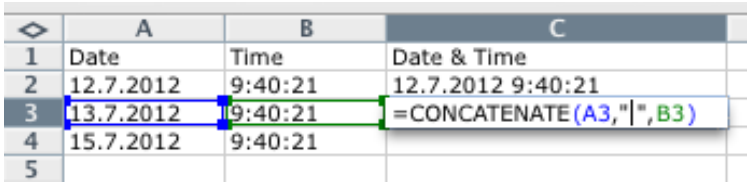


Fig. 3.16. If you have your date and time in multiple columns, you need to combine them into one column before importing your data set in Disco.

- Make sure that the *Date* and *Time* columns are interpreted as text. For example, instead of writing `B3` within the formula, write `TEXT(B3;"HH:MM:SS")`.
- Add the two columns by using the formula `=A3+B3`. Then right click on cell C3 and select *Format Cells*, from the format cell dialog box select the *Number* tab, from the category list select *Custom*, under *Type* write the format *DD/MM/YYYY HH:MM:SS*, and click *OK*.

3.3.3 "I need to merge multiple files"

Currently, only single files can be imported with Disco. If you have parts of your event log distributed over multiple data sets, you need to do some extra pre-processing work before you can start your process mining analysis. Databases or specialized tools can be used to do the merging.

Contact the IT support staff who extracted the data for you, or get in touch with us to help you with this.

3.3.4 "The Start button is greyed out"

As shown in Figure 3.5, you can only start the import of your file after you have configured all the required columns (case ID and activity need to be set).

3.3.5 "Disco has problems reading my file"

If not all lines in your CSV or Excel file can be read, then Disco tells you about it (see Figure 3.17).

To find out what the problem is, open the file and compare the line that is reported with the line before. Can you see a difference? Does the problematic line have more or less columns than all the other lines before? Are there wrong quotes?

For example, the error message in Figure 3.17 was produced when attempting to read the file shown in Figure 3.18. The file is a modified snippet from the example log to test different non-latin scripts such as Korean, Hebrew, and Chinese[2]. When I added the last line to test latin characters in the mix, I forgot to fill in a value for the Resource column in my source file. As a result, you can see that all the other columns are shifted one to the left, and that line No. 8 is missing a cell at the end of the row.

[2] Yes, Disco has full Unicode support and allows to import and analyze logs in any language.

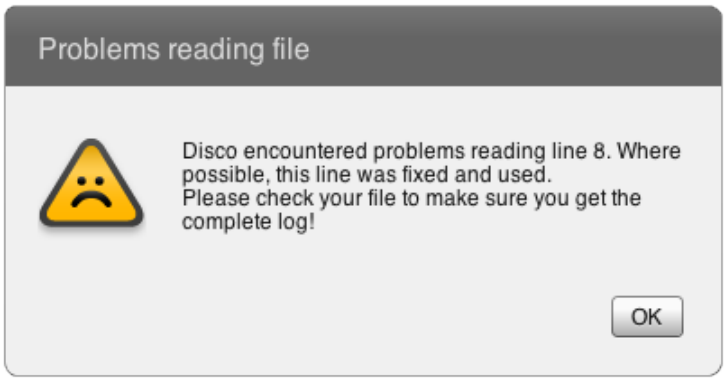


Fig. 3.17. If Disco reports the lines that had problems during the import, make sure to check the source file in the given line number to figure out what the problem is.

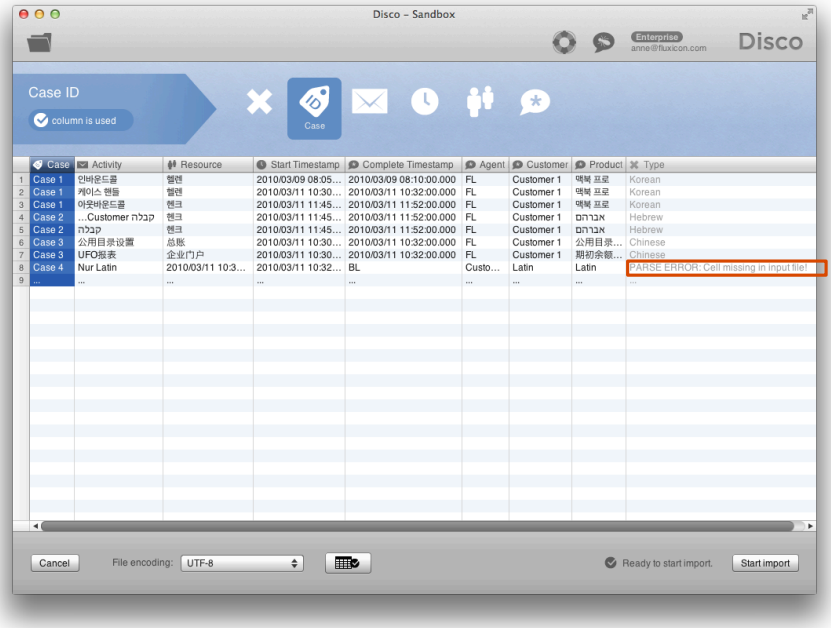


Fig. 3.18. The source file that produced the error message in Figure 3.17: Line 8 has one column less than the other rows. Disco fills up the missing cell at the end of the row but warns you about the mismatch.

3.3.6 "I get weird symbols"

Disco automatically detects the encoding of the text in your data set and can also deal with languages that have special symbols. If you find that characters in your data set are not read properly, make sure that you have saved your file in Unicode (UTF-8) format.

If Disco has not correctly auto-detected your file encoding, you can also try other encodings from the list provided in the lower left of the configuration screen (see Figure 3.19). When you select a new encoding, Disco will reload your CSV file with the chosen encoding and you can see in the preview of the configuration screen whether your characters are now displayed correctly.

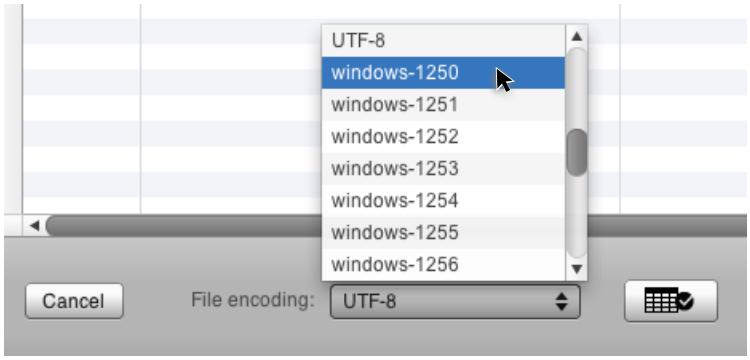


Fig. 3.19. You can also reload your CSV file with an explicit file encoding, in case auto-detection fails.

4

Analyzing Data Sets

Once you have imported a log file (Chapter 3), you are directly taken to your first process map in Disco's analysis view. The analysis view is symbolized by the magnifying glass symbol shown in Figure 4.1.

Fig. 4.1. Analyze symbol in Disco.

From the project view (Chapter 6), you can get back to the analysis by clicking on the magnifying glass symbol (see also Section 6.1.1 for how to navigate from your workspace to the analysis view).

For each data set, there are three alternative analysis views as shown in Figure 4.2: Map, Statistics, and Cases. They show different aspects of the same, underlying data, and you can switch between them by selecting the corresponding tab. If you have applied a filter (Chapter 5), then all three analysis views show you their perspective on the filtered data set.

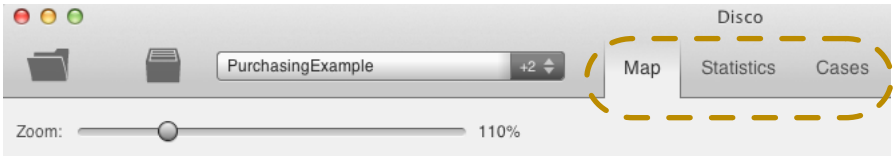


Fig. 4.2. The three analysis views in Disco: Map, Statistics, and Cases.

Once you have imported multiple data sets or created copies to keep bookmarks of your analysis results, you can access them right from within the analysis view by selecting the data set from the quick switch list shown in Figure 4.3. This way, you can rapidly move back and forth to compare different data sets, and to "jump" to bookmarks in your analysis.

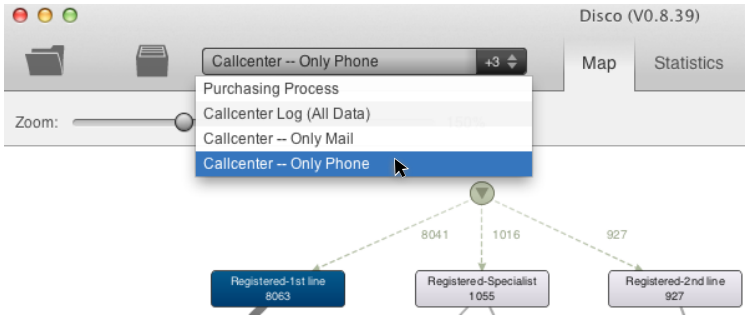


Fig. 4.3. Along with the imported data sets, copies are available through the quick switch list for rapid access from the analysis views.

Most of your time using Disco will be spent in the three analysis views (Map, Statistics, and Cases), because that's where you discover how your process has been running, where you inspect statistics and KPIs, and where you can track down individual cases.

The remainder of this chapter explains in detail how each of the three views work and how they support you in understanding your process at hand.

4.1 Map view

After you have imported your event log, the first thing you usually want to see in a process mining tool is what your process actually looks like. Therefore, Disco brings you right into the Map view (Figure 4.4).

In the Map view, you see a process map that visualizes the actual flow of your process based on the imported log data. The Map view contains the following elements:

❶ *Canvas with process map*. The main area is reserved to visualize your process map. Refer to Section 4.1.1 for further details on how to read the process maps in Disco.

❷ *Zoom slider*. You can zoom in and out in your process map in two different ways:

- The zoom slider (❷) gives you an explicit control to make the process map larger and smaller.
- Alternatively, you can simply use your mouse wheel to zoom in and out.

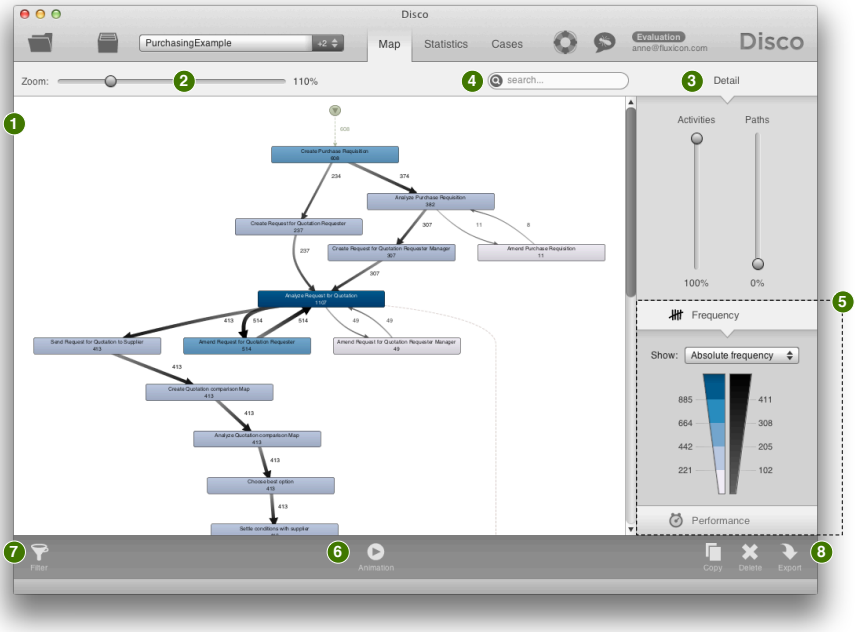


Fig. 4.4. The Map view in Disco.

To move the currently displayed area of your process map around you can either use the vertical and horizontal scroll bars or click and hold your mouse while dragging the process map.

❸ *Map detail controls*. Because real-life processes can become quite complex and confusing when every detail and all exceptional process flows are shown, Disco gives you a quick and easy way to make the process map simpler and only show you the most important flows. Refer to Section 4.1.2 to learn how you can adjust the level of detail of your process maps.

❹ *Search field*. The search field allows you to find a specific activity also in large process maps. Read Section 4.1.3 to see how the search field works.

❺ *Process map visualization options*. In addition to showing you in which order the activities in your process have been performed (the actual process flows), you can also enhance your process maps with several process metrics. Refer to Section 4.1.4 for further details on the process map metrics that are available in Disco.

❻ *Animation*. Animation can be very useful to communicate analysis results to process owners or other people who are no process analysis experts. By showing how the cases in the data set move through the process (at their relative, actual speed), the process will be literally "brought to life". Read Section 4.1.6 to learn more about animation.

❼ *Filtering*. The log filter controls for the current data set can be accessed from each of the analysis views. Filters are really important to drill into specific aspects of your process and to focus your analysis. Read Chapter 5 for detailed information on how filtering works in Disco.

❽ *Copy, Remove, and Export data set*. Data sets can be copied, deleted, and exported right from the current analysis view. Read Section 6.2 for further details on copying and deleting data sets. Process maps can be exported, for example, as a PDF file. The export functionality of Disco is explained in detail in the Export reference in Chapter 7.

4.1.1 How To Read the Process Map

The process map is the most important analysis result in Disco. It shows you how your process has actually been executed. The process flows that you see in the Map view are automatically reconstructed (”discovered”) based on the sequence and timing of the activities in your imported event log data. So, without further knowledge about the process, or any pre-existing process model, you obtain an objective picture of the real process.

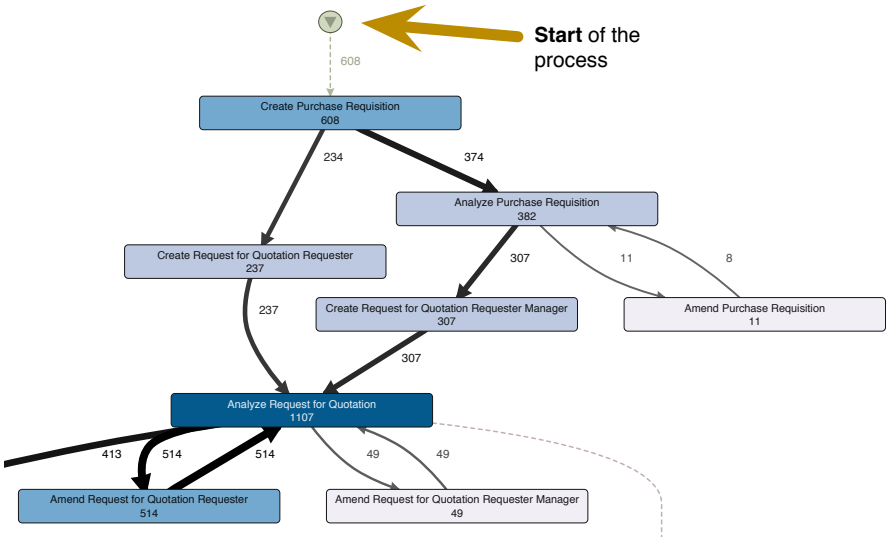


Fig. 4.5. Start of the purchasing process demo example.

The discovered process is visualized in a simple and intuitive way: The start of the process is illustrated by the triangle symbol at the top of the process map (see Figure 4.5). Similarly, the end of the process is illustrated by the stop symbol (see Figure 4.6). Activities are represented by boxes and the process flow between two activities is visualized by an arrow. Dashed arrows point to activities that occurred at

the very beginning or at the very end of the process. By default, the absolute frequencies are displayed in the numbers at the arcs and in the activities (see Section 4.1.4 for how to change the metrics that are displayed in the process map). The thickness of the arrows and the coloring of the activities visually support these numbers.

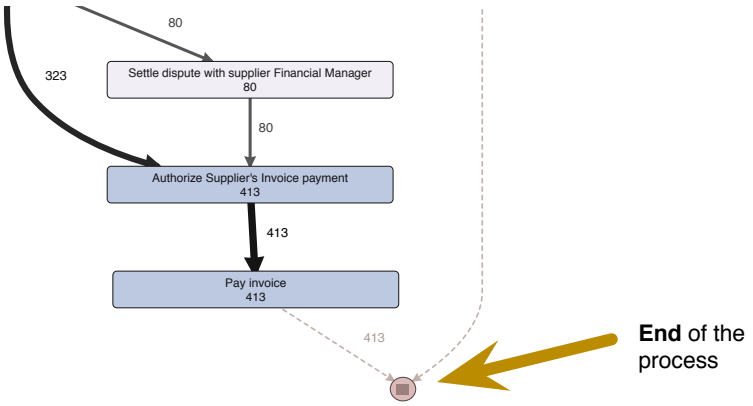


Fig. 4.6. End of the purchasing process demo example.

For example:

- In Figure 4.5 we can see that there are 608 cases (different instances of the purchasing process) in the data set that all start with the activity *Create Purchase Requisition*.
- Afterwards, the process splits into two alternative paths: In 374 cases the activity *Analyze Purchase Requisition* was performed after *Create Purchase Requisition*. The other 234 cases perform activity *Create Request for Quotation Requester* instead. Because the path where 374 cases have "travelled through" indicates the main flow in this part of the process, it is visualized by a thicker arrow.
- In total, activity *Analyze Request for Quotation* is the one that is executed most often (in total 1107 times)—almost twice as much as we have cases in the data set! This comes from the dominant loop with activity *Amend Request for Quotation Requester* (see Figure 4.5). Repeatedly, purchase orders are amended and need to be re-analyzed, which is of course very inefficient and from a process improvement perspective we would need to find out what is going on. Perhaps people don't know what they are allowed to purchase, and we might resolve the problem by updating our purchasing guidelines or providing additional training.
- Figure 4.6 shows another fragment from the end of the purchasing process. We can see that 413 purchase orders are completed and end with *Pay invoice*. Some others are stopped earlier in the process (see other dashed arrow). In fact, in Figure 4.5 we can see that some are stopped after activity *Analyze Request for Quotation* (see dashed arrow at the bottom of the shown fragment).

4.1.2 Adjusting the Level of Detail in Your Process Map

Real-life processes can often become very complex. Therefore, Disco allows you to interactively adjust the level of detail that you want to see. There are two slider controls that you can use to modify the level of detail that is shown in your process map:

Activities The Activities slider influences the number of activities shown in your process map, ranging from only the most frequent activities up to all including the least frequent activities.

Paths The Paths slider determines how many paths are shown in your process map, ranging from only the most dominant process flows among the shown activities up to all (even rare) connections between the activities.

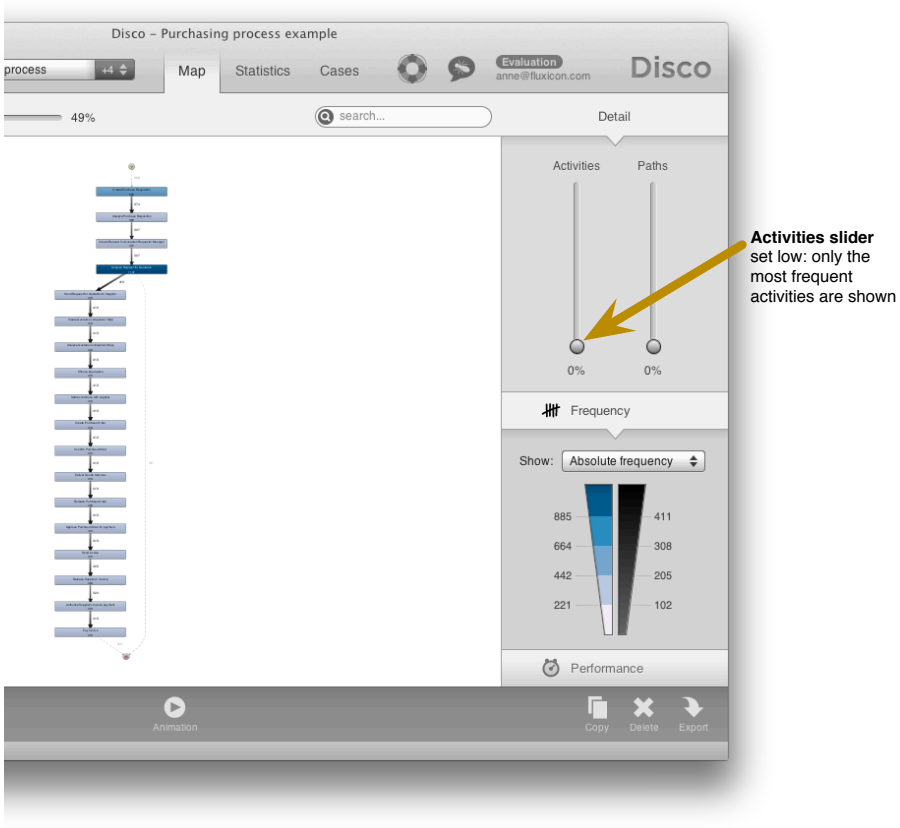


Fig. 4.7. The Activities slider influences the number of activities shown in your process map. It ranges from showing only the activities from your most frequent process flow at the lowest point (0%) up to showing all activities that have ever occurred (100%).

In Figure 4.7 you can see how the purchasing example process from the previous section looks like when the Activities slider is set to the lowest point. Only those activities that occur in the most frequent process variant are shown.

When you move the Activities slider up, increasingly also less-frequent activities are included, up to showing all activities that have ever occurred (even if they just occurred once or twice). For example, in Figure 4.4 and in Figure 4.5 you see a process fragment with 100% of the activities shown: Less frequent activities such as *Amend Purchase Requisition*, which occurred only 11 times, are shown as well.

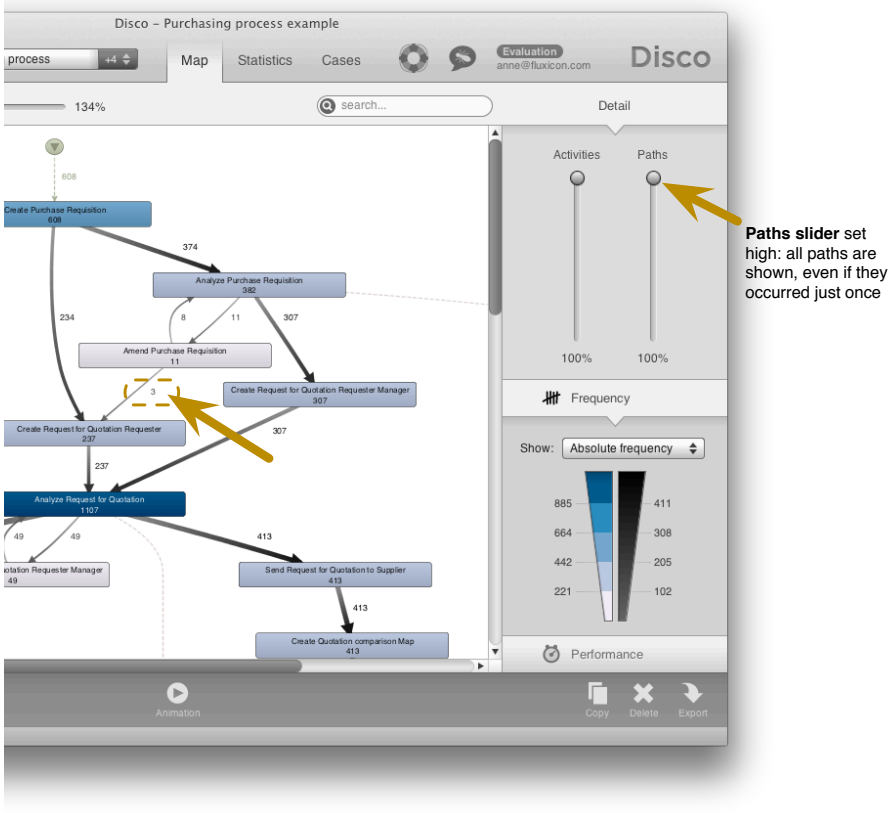


Fig. 4.8. The Paths slider can be used to show only the most dominant paths in your process map (0% slider position) up to all connections between activities that have occurred (100%).

The Paths slider works based on the activities that are currently revealed by the Activities slider (e.g., all or just a subset). If the Paths slider is set to the lowest point, then only the most dominant connections between these activities are shown. This way, Disco makes sure that all your activities are always connected and avoids getting "dangling" process fragments that cannot be put in context with the remaining

activities even if you look at a simplified process map. For example, in Figure 4.4 and in Figure 4.5 you see the purchasing process with all activities (Activities slider pulled up to 100%) and just the most dominant connections between them (Paths slider set to the lowest point).

In Figure 4.5 you can see that not all the details of the process are revealed yet because activity *Amend Purchase Requisition* has been performed 11 times directly after activity *Analyze Purchase Requisition* (see incoming path with number 11), but only 8 times the process has returned to *Analyze Purchase Requisition* afterwards (see outgoing path with number 8). Where did the process go in the other 3 cases? This can be revealed by pulling the Paths slider up to 100% as shown in Figure 4.8. We can see that 3 times the process went directly from *Amend Purchase Requisition* to activity *Create Request for Quotation Requester*.

To determine the right level of detail for your own process, it is recommended to start with the slider position that Disco determines automatically when it creates the first process map for you. Then, start by pulling up the Activities slider until you see all the activities that you want to see. Most of the times, you will be able to reveal 100% of the activities and still get a readable process map. Only then start moving the Paths slider up as much as possible. Stop when the process map becomes too complicated to inspect it properly. This way, you can get reasonably complete and still readable process visualizations for almost any process, no matter how complex it is.

Refer to the Variation filter in Section 5.2.2 to learn about an alternative way to simplify your process map in a more controlled way based on process variants.

4.1.3 Searching Activities in Your Process Map

When you have process maps that contain many different activities, it is sometimes difficult to find a specific activity that you are interested in. In such situations, you can use the search field in the upper right corner of the process map.

Searching works as follows:

- You simply start typing in the search field as shown in ❶ in Figure 4.9. While you type, Disco highlights all activities that match the searched term (see bold red arrows in Figure 4.9) and automatically zooms into the process map to bring all matched activities into focus.
- You can narrow down your search by typing multiple words as shown in Figure 4.10.
- If you want to go back, you can simply delete the word that you typed in the search field by pressing the Delete key.
- If you want to stay in the focused view but get rid of the red arrow markers, press the little **x** button in the right corner of the search field (see in ❷ in Figure 4.9). This will remove the bold red arrows but keep the zoom level and position.

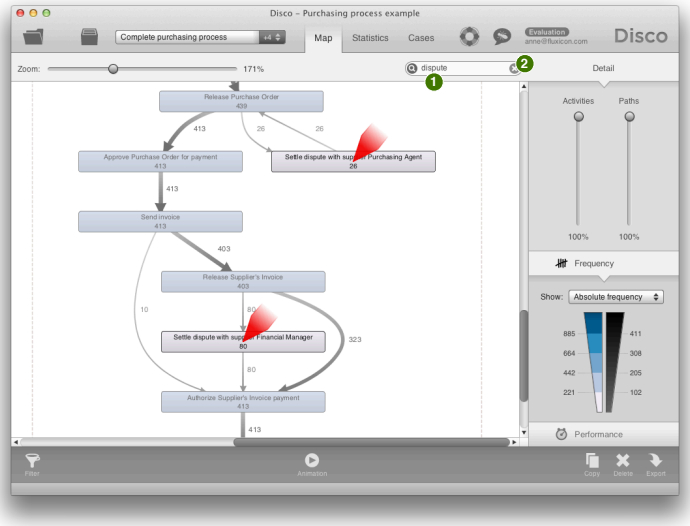


Fig. 4.9. When you start typing in the search field, Disco interactively highlights all activities that contain the current search term for you.

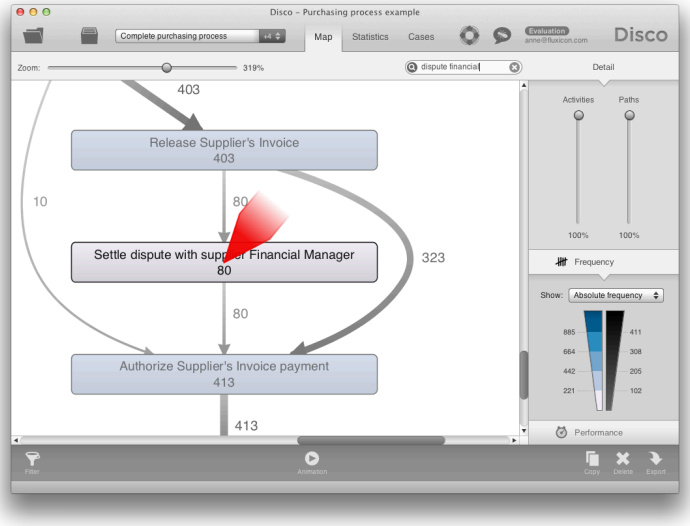


Fig. 4.10. When you type multiple words, you can narrow down your search to activities that contain all of these words.

4.1.4 Displaying Different Metrics in the Process Map

In the default view, the process map is displayed with absolute frequencies. This means that the numbers in the activities and at the paths indicate how many times the activity was performed in total, or how often that particular path has been "travelled", respectively, throughout the whole process. However, you can change the visualization and display also other metrics right in your process map.

The following metrics are available in the Map view of Disco:

Frequency The frequency metrics show you which parts of your process have been executed most often. You can change between them like shown in Figure 4.11.

- *Absolute frequency.* This is the default view based on total frequencies as explained in the beginning of this section (see also Figure 4.11).
- *Case frequency.* If you have repetitions in your process, it can be helpful to ignore them for a minute and just view relative numbers of how many cases passed through which activities and along which path (regardless of whether they came by there just once or multiple times).
- *Max. repetitions.* Displays the maximum number of repetitions within a case.

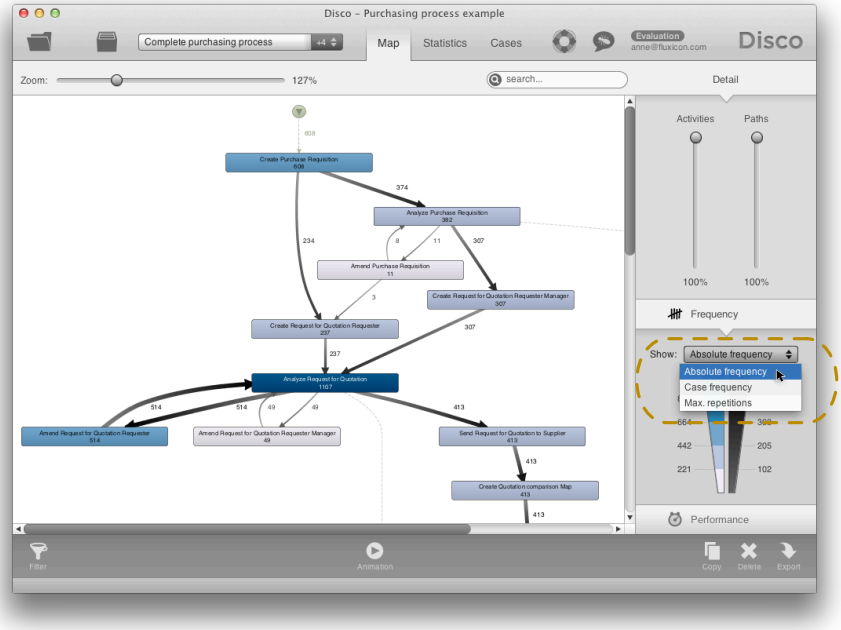


Fig. 4.11. Within the Frequency or Performance view, you can select the metric that you want to display from the drop-down list as shown here.

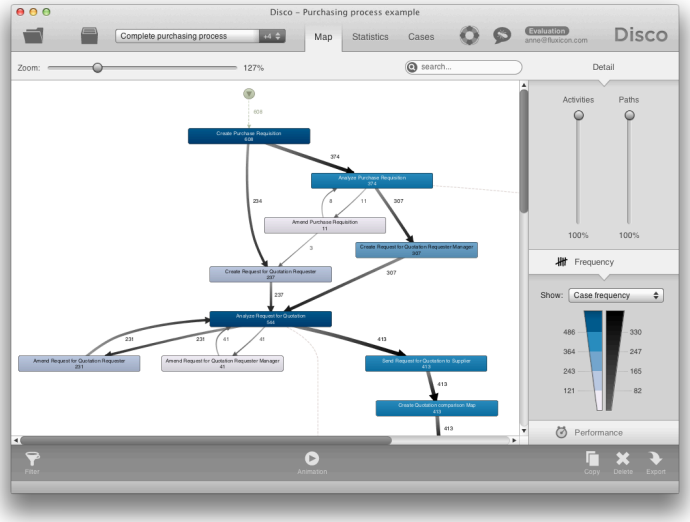


Fig. 4.12. Use the *Case frequency* option to ignore repetitions and just see relative numbers for how many cases passed through which activities and along which paths.

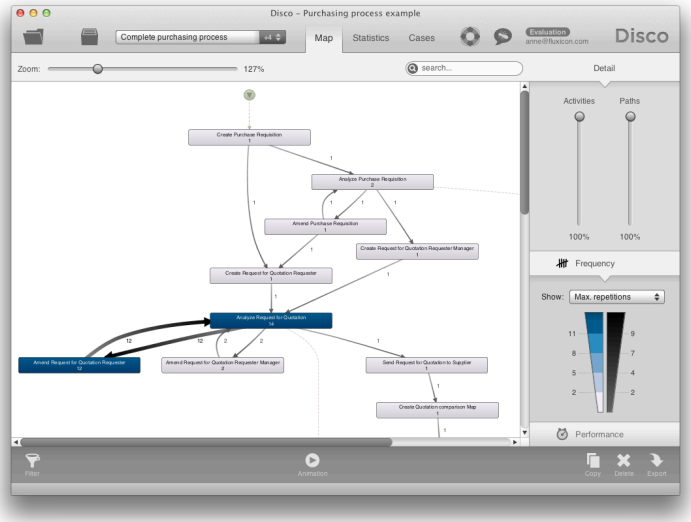


Fig. 4.13. The *Max. repetitions* option lets you see up to how many repetitions occurred in your process within the same case.

An example of the *Case frequency* view is shown in Figure 4.12, where—in contrast to the absolute view in Figure 4.11—it now becomes visible that out of 608 cases 231 went through the loop with activity *Amend Request for Quotation Requester*.

The *Max. repetitions* option does exactly the opposite: For example, in Figure 4.13 we can see that the loop with activity *Amend Request for Quotation Requester* was run through up to 12 times within a single case.

Performance When you have obtained a good understanding about the actual process flow, you often want to know more about the *time* that is spent in the different parts of your process. This is what the Performance metrics are for. You can change to the Performance view by selecting the Performance tab, which brings up the performance metrics and visualization legend (see Figure 4.14).

- *Total duration.* The default metric that is displayed when you get into the performance view is *Total duration*. It shows the accumulated durations (summed up over all cases) for the execution of each activity and for the delays on each path.

  Because the frequencies of the activities and paths are included in this cumulative view, it allows you to quickly spot the high-impact areas in your process: For example, it can have much more impact on your overall throughput time if you can speed up a frequently performed part of your process by just 5 minutes compared to reducing the time spent in a rarely used process path by 1 day.

  An example of the performance view with the *Total duration* option is shown in Figure 4.14. There, we have filtered the data set to only show the process map for those cases that last longer than 70 days (refer to the Performance filter reference in Section 5.2.3 to find out how to do this), and now we want to know where all this time is lost in the process. Clearly, the biggest impact area is the delay between the activity *Amend Request for Quotation Requester* and *Analyze Request for Quotation* (see thick red arrow with the displayed cumulative time of 11 years).

- *Mean duration.* Alternatively, the average time spent within and between activities can be displayed. For example, in Figure 4.15 one can see that while the execution of activity *Amend Request for Quotation Requester* takes on average only 9.8 minutes, it creates a delay of, on average, 14.9 days afterwards.

- *Max. duration.* The largest execution times and delays that were measured.

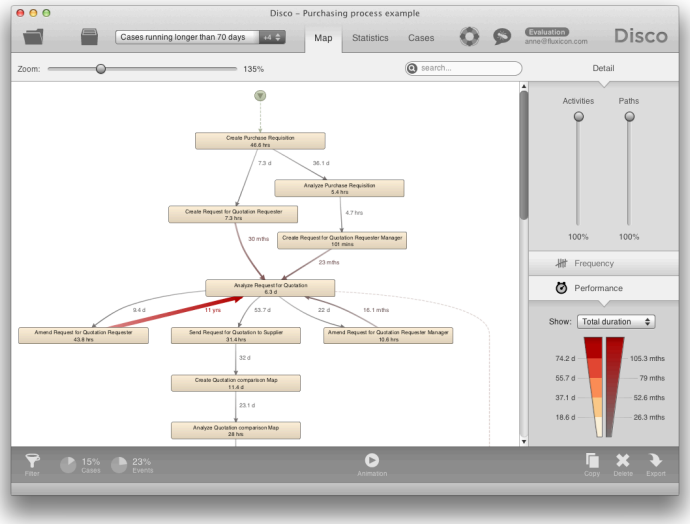


Fig. 4.14. The *Total duration* option lets you see the high impact areas for delays in your process by showing the cumulative times (added up over all cases) for each path and activity.

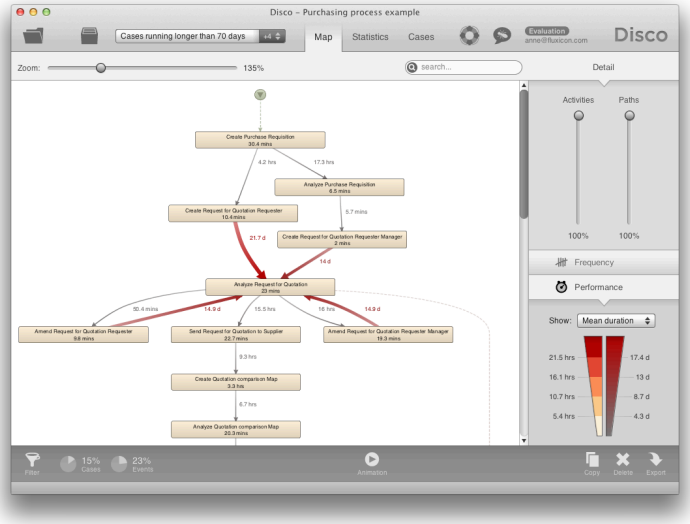


Fig. 4.15. The *Mean duration* option displays the average execution times for each activity and the average idle times on each path.

> **Note:** The execution times for activities can only be displayed if you have both start and a completion timestamps in your data set. If you only have one timestamp in your event log, then you can still analyze the time between activities, but the activity durations will be displayed as *instant*.

Refer to the import reference in Section 3.1.6 for further information on start and completion timestamps, and how they can be included.

Finally, sometimes you just want to see all these metrics at one glance and not switch back and forth. This can be achieved by simply clicking on an activity or path as shown in Figure 4.16 and works regardless of in which metrics visualization view you currently are.

When you want to remove the overview badge again, just click somewhere at the background of your process map (outside or next to an activity or path).

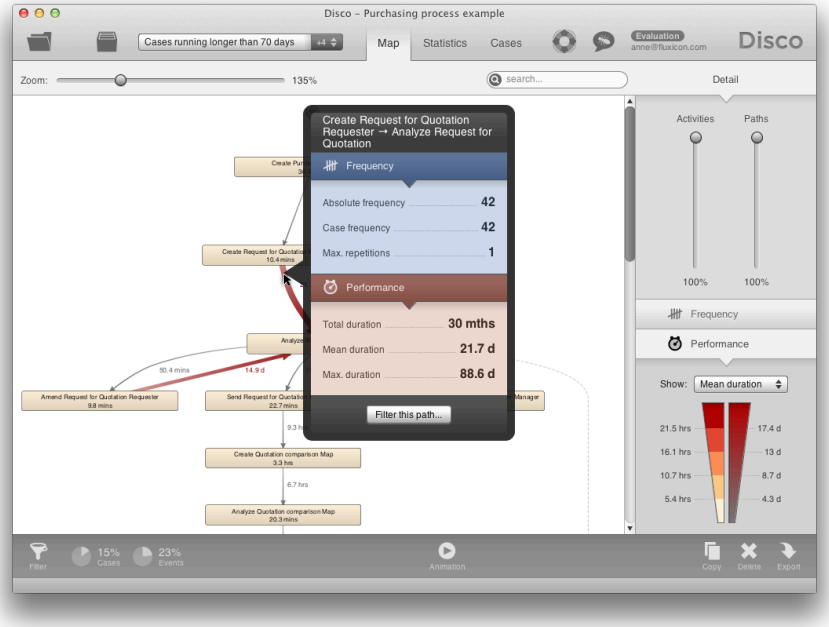


Fig. 4.16. When you click on an activity or a path in your process map, you can see all metrics at one glance in an overview badge.

4.1.5 Filtering Activities and Paths

The overview badge shown in Figure 4.16 also allows you to quickly filter your data set for cases that follow a particular path, or that execute a particular activity.

For example, in Figure 4.17 you can see the end of the purchasing process from before. Clearly, the main process flow runs through the activity sequence *Send invoice* → *Release Supplier's invoice* → *Authorize Supplier's invoice payment* (potentially settling a dispute with the supplier in between). In fact, the activity *Authorize Supplier's invoice payment* is a mandatory process step that was put in place to avoid fraud.

However, one can see that 10 cases move directly from *Send invoice* to *Authorize Supplier's invoice payment*, skipping the required *Release Suppliers invoice* step. This is a compliance issue. As a process analyst we would want to find out which cases have bypassed the prescribed process step to find out what happened there, and to prevent this in the future.

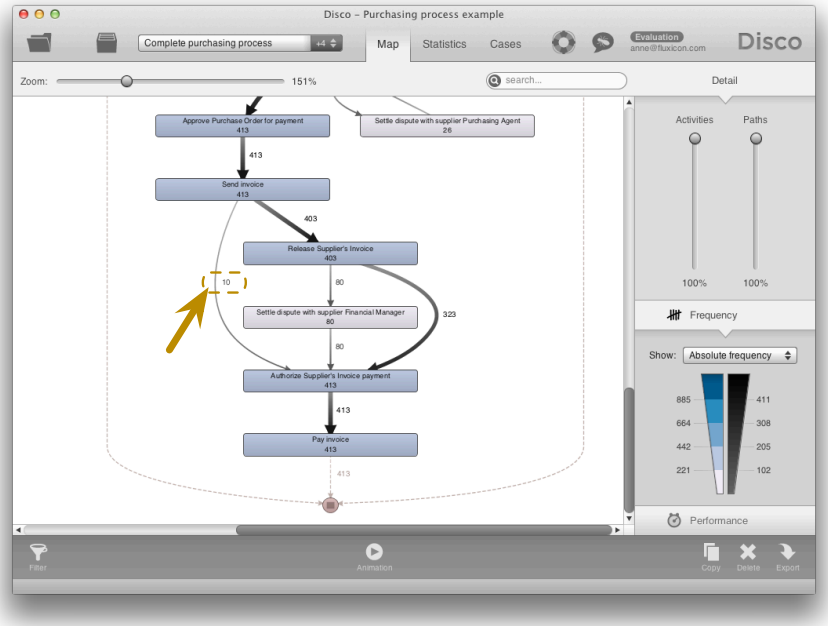


Fig. 4.17. 10 cases bypass a mandatory authorization step at the end of the purchasing process. As a next step, we want to drill down to find which 10 cases these were.

Drilling down into specific cases that follow a specific path is easy: You can simply click on the corresponding path to bring up the overview badge as shown in Figure 4.18.

Fig. 4.18. Step 1: Click on the corresponding path to bring up the overview badge. Click the *Filter...* button to add a filter for this specific path.

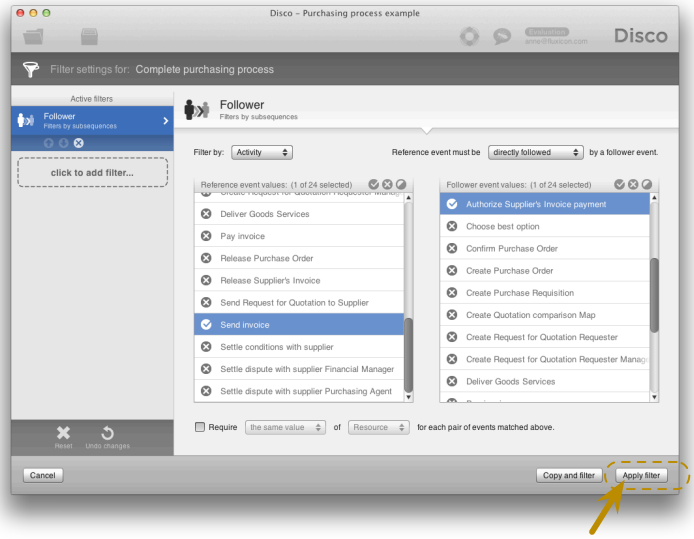


Fig. 4.19. Step 2: Disco automatically adds and pre-configures a Follower filter (see Section 5.2.6). Just press *Apply filter* to activate the filter for you data set.

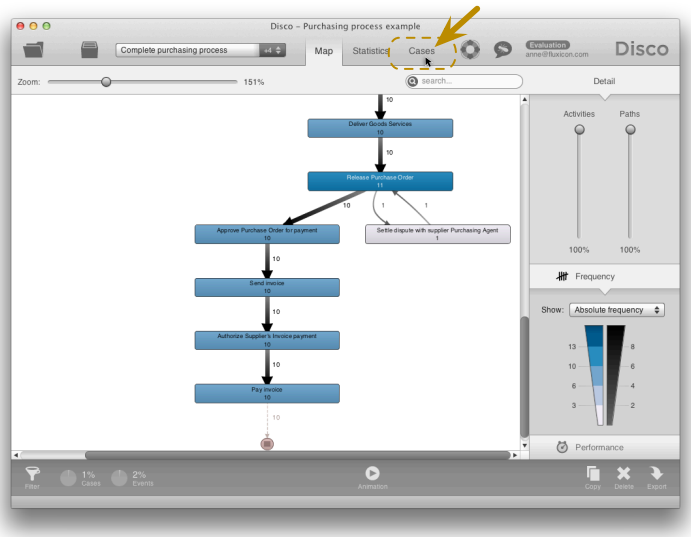


Fig. 4.20. Step 3: Change to the Cases view (Section 4.3) to see detailed information about these 10 cases rather than the process map.

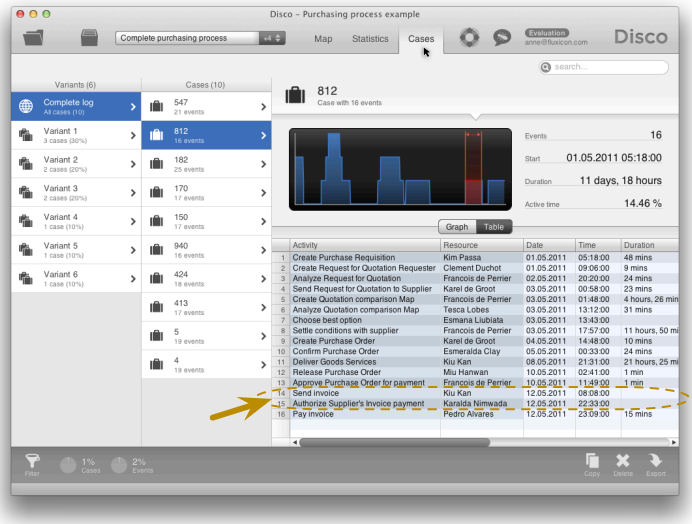


Fig. 4.21. Step 4: The Cases view now shows a list of all cases in the filtered data set. Here, case 812 has been selected and we can inspect the history with all its details.

Once you press *Filter this path...* a pre-configured Follower filter (see filter reference in Section 5.2.6) is automatically added to the filter stack (Figure 4.19). After you have applied the filter you get back to the process map, filtered down just for these 10 cases (Figure 4.20).

However, in this situation the process map is not particularly interesting. Instead, we want to see which cases followed the non-compliant path. So, you can change to the Cases view (Section 4.3) as shown in Figure 4.21 to see all the details about the 10 cases that followed the forbidden path.

In the previous example scenario, a *path* was filtered and a corresponding Follower filter has been added automatically (Section 5.2.6). In the same way, an *activity* can be filtered by clicking on the activity and pressing *Filter this activity...* as shown in Figure 4.22.

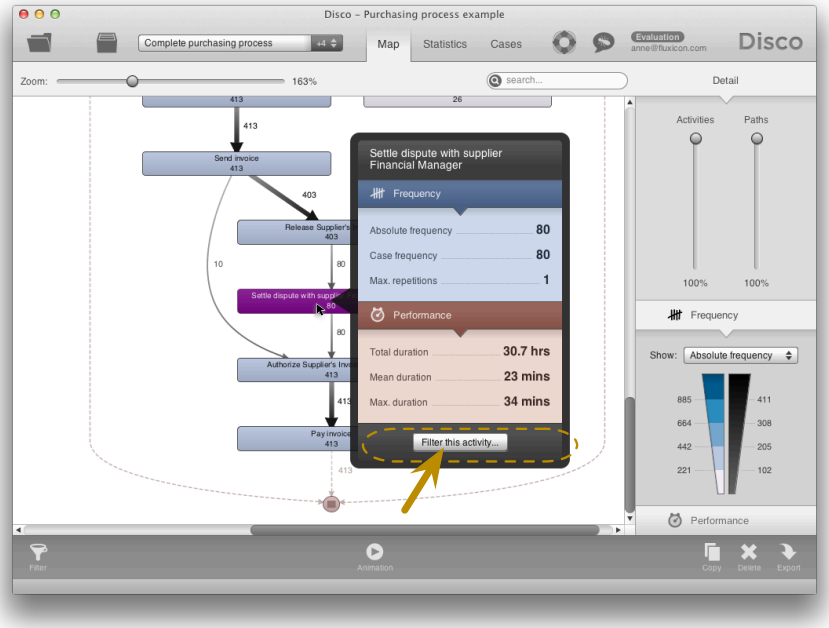


Fig. 4.22. Clicking on any activity brings up the *Filter this activity...* shortcut to add a pre-configured Attribute filter (Section 5.2.5) that only leaves cases that pass through this activity.

This will add a pre-configured Attribute filter that only needs to be applied to narrow down the data set to those cases that perform the activity (here this would be the 80 cases that performed activity *Settle dispute with supplier*). Refer to the filter reference in Section 5.2.5 to learn more about the Attribute filter.

4.1.6 Animation

Animation is a way to visualize the process flow over time right in the discovered process map (a bit like showing a "movie" of your process). Animation should not be confused with simulation. Rather than simulating, the *real events* from the log are *replayed* in the discovered process map as they took place.

Animation can be very useful to communicate analysis results to process owners or other people who are no process analysis experts. By showing how the cases in the data set move through the process (at their relative, actual speed), the process will be literally "brought to life".

For example, imagine that we want to visualize the bottleneck that we discovered in the purchasing demo process, when we analyzed the performance metrics in the process map for cases that take longer than 70 days in Section 4.1.4 (Figure 4.14 and 4.15). To start the animation, simply press the little button with the Play symbol at the bottom of the Map view (see also ❻ in Figure 4.4).

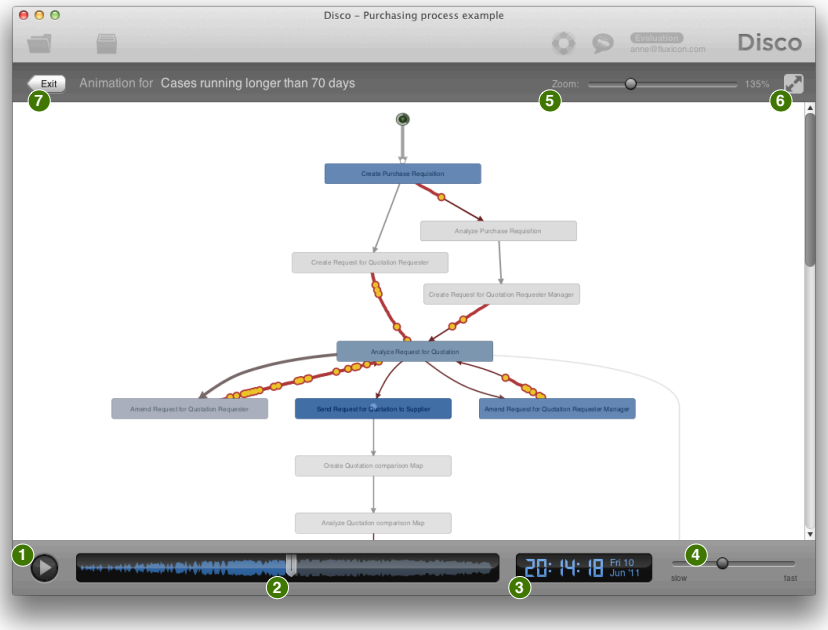


Fig. 4.23. Animation view in Disco. Start the animation by pressing the *Play* button (❶) and observe how your cases flow through the discovered process map over time.

This will bring up the animation view as shown in Figure 4.23:

❶ *Play button.* Start the animation by pressing the Play button in the lower left corner. You will see small "bubbles" (so-called tokens) that start moving through

the process map. Each token represents one case and moves through the process at the relative, actual speed of the corresponding case in your event log.

- Between activities, the cases are visualized in bright yellow color with a red border to make it easy to spot unnecessary delays (like, for example, there is a clearly visible queue of cases in the back loop from activity *Amend Request for Quotation Requester → Analyze Request for Quotation* in Figure 4.23).
- When an activity is performed, the active activity is highlighted in blue. It slowly fades back to grey after the activity has stopped to let you observe activity sequence patterns as they unfold.

  Cases that are currently performing an activity are visualized in a light blue color inside the activity box (for example, in activity *Send Request for Quotation to Supplier* in Figure 4.23 there is currently one case performing this activity). When there are two activities that are performed in parallel for the same case, then this case is visualized by two tokens for that time.

The animation is performed in chronological order based on the timestamps in your data set. It starts with the cases that were active at the beginning and ends with the latest activities in your log.

❷ *Progress indicator*. The progress indicator shows you how much of your overall log timeline has already been replayed by the animation. The replayed part is highlighted in blue and the needle indicates the current replay position. To move around in the timeline simply drag the needle to the left or right.

In addition to the position of the animation in your replay, the timeline visualization also gives you a sense of how much activity occurred in your process over time (the thicker the more activity). This helps you, for example, to jump right into the more "busy" periods of your process with the animation.

❸ *Current replay time*. The current replay position is displayed in this date and time window. For example, in Figure 4.23 the animation shows the state of the purchasing process on Friday 10 June 2011 at 20:14.

❹ *Speed control*. To increase or decrease the speed of the animation, you can move this speed slider to the left (slower) or to the right (faster).

❺ *Zoom control*. You can zoom in and out in the animated process map by using this zoom slider or, alternatively, use your mouse wheel to zoom in and out (works in the same way as in the regular Map view). To move the currently displayed area of your process map around you can either use the vertical and horizontal scroll bars or click and hold your mouse while dragging the process map.

❻ *Full screen mode*. For presentations, it can be useful to focus only on the animation and use as much of your screen as possible to display the process flow. For this, you can press the full screen mode button in the upper right corner. When you want to return from the full screen mode, simply press the same button in the upper right corner again.

❼ *Return to Map view*. To return to the regular Map view of your process, press the Exit button in the upper left corner.

While Disco implements the live animation using efficient and performance-optimized graphics functions, it still remains a resource-intense part of the application. Therefore, it can happen that some larger data sets (for example, with millions of events) are too big to animate completely. If your log is too big to animate in full, Disco will give you the following message (see Figure 4.24).

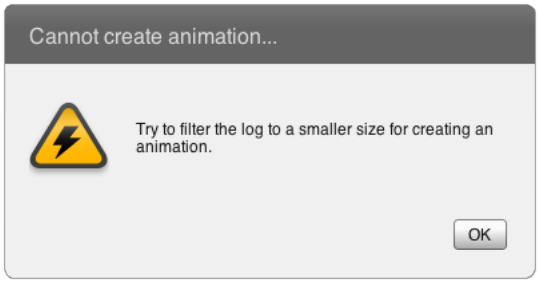


Fig. 4.24. It can happen that your log is too big to create an animation. Disco will tell you this and you can use a filter to created a smaller data set.

In this situation, you can use a filter to reduce your data set and animate only a part of it. For example, try using the Timeframe filter (see Section 5.2.1) to split your data set into two parts, one for the first half and one for the second half, to animate them after each other.

4.2 Statistics view

While the Map view (Section 4.1) gives you an understanding about the actual process flow, the Statistics view provides you with additional overview information and detailed performance metrics about your process. You get to the Statistics view by simply changing to the *Statistics* tab as shown in Figure 4.25.

Depending on the statistics view that you have selected on the left (❹–❼), the following information areas are shown:

❶ *Overview information*. Key overview figures about the selected statistics view. For example, for the global statistics shown in Figure 4.25, you can see how many events and cases are in your data set, and which time frame is covered in the log.

❷ *Performance charts*. A number of pre-generated charts visualize relevant performance metrics for the current statistics view. Charts can be exported as explained in Section 7.3.

❸ *Detailed information*. In the lower part of the screen, detailed statistics are shown in a table format. Every table in Disco can be exported as a CSV file to further process the information with other tools such as Excel or Minitab. Refer to Section 7.3 to learn how to do that.

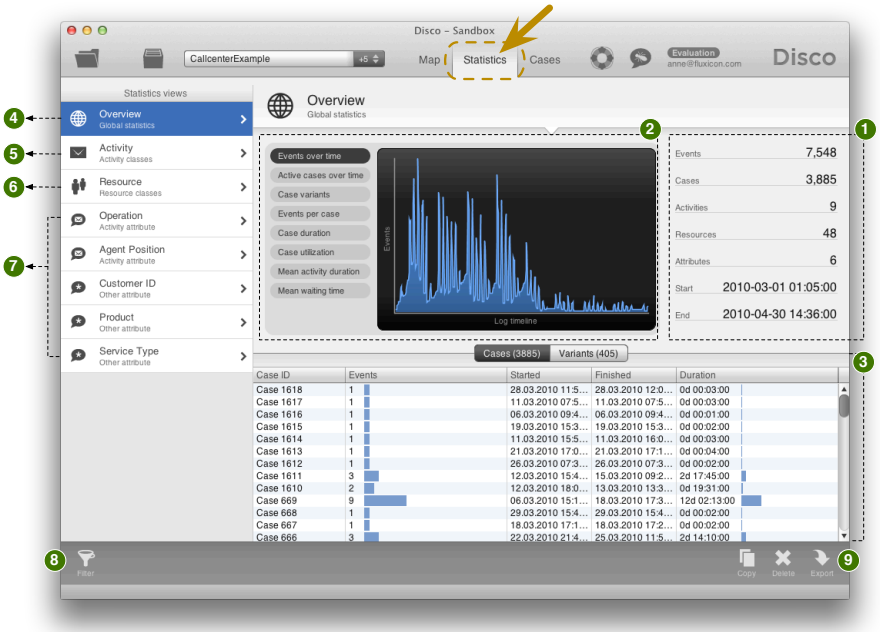


Fig. 4.25. The Statistics view in Disco.

The different statistics that are available are organized into the following views:

❹ *Global statistics*. Overview statistics about the whole log, individual cases, and variants (see Section 4.2.1).

❺ *Activity statistics*. Statistics about the different process steps in your data set (see Section 4.2.2).

❻ *Resource statistics*. Statistics about the people or organizational units in your data set (see Section 4.38).

❼ *Attribute statistics*. Statistics about all further attributes (see Section 4.39).

Finally, like in any of the three analysis views (Map, Statistics, Cases), you have the following functionalities available at the bottom of your screen:

❽ *Filtering*. The log filter controls for the current data set can be accessed from each of the analysis views. Filters are really important drill into specific aspects of your process and to focus your analysis. Read Chapter 5 for detailed information on how filtering works in Disco.

❾ *Copy, Remove, and Export data set*. Data sets can be copied, deleted, and exported right from the current analysis view. Read Section 6.2 for further details on copying and deleting data sets. Process maps can be exported, for example, as a PDF file. The export functionality of Disco is explained in detail in the Export reference in Chapter 7.

4.2.1 Global Statistics

In the lower part of the global statistics screen, you find a list of all cases in your data set (❸ in Figure 4.25).

For each case, you can see the number of events, the earliest timestamp, the latest timestamp, and the duration as shown in Figure 4.26. You can scroll through the list of cases and when you want to inspect the full history for one case (for example, one with a particular long duration), you can search for the case in the Cases view (Section 4.3).

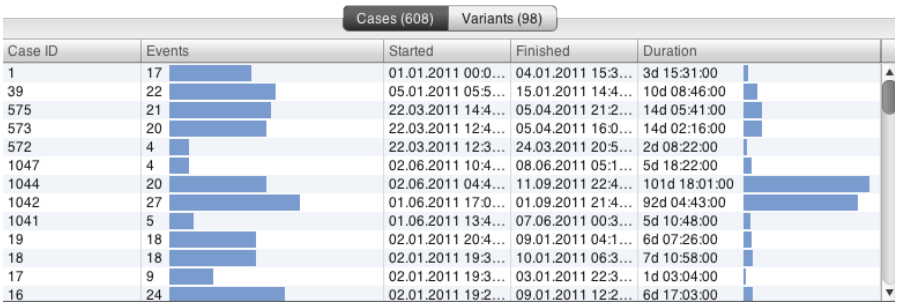


Fig. 4.26. A list of all cases with their number of events, start and end time, and duration.

Alternatively, you can switch to see the variants as shown in Figure 4.27. A variant in Disco is a specific sequence of activities, and there may be multiple cases that follow the same sequence through the process. In the variants overview you see at one glance how many cases follow each variant, how many events make up the sequence, and what the average duration is for all cases that follow this variant.

To see all cases that follow a particular variant, you can change to the Cases view (Section 4.3) and select the variant that you want to see on the left.

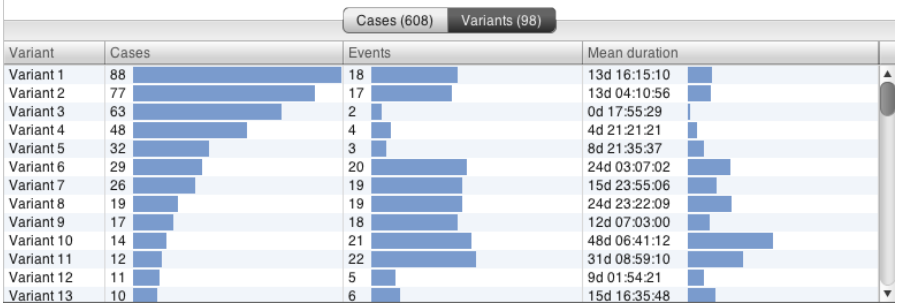


Fig. 4.27. A list of all variants with their frequency, number of events, and average duration.

In the upper part of the screen, a few summary figures and a chart are shown (❶ and ❷ in Figure 4.25). The summary figures are:

- *Events*: Total number of events or activities in the data set.
- *Cases*: Total number of cases or process instances in the data set.
- *Activities*: Total number of different activities (activity types) in the data set.
- *Resources*: Total number of different resources in the data set.
- *Attributes*: Total number of attributes (columns from your original file) that have been imported for analysis.
- *Start* and *End*: The range of time covered by your data set (from earliest to latest timestamp observed).

You can change the chart that is displayed on the left by selecting a metric from the list. Most of these metrics can also be used to filter your data set with the Performance filter, where the same charts are used to guide your configuration. Refer to Section 5.2.3 to learn more about how to use the Performance filter in Disco.

The following metrics are available in the chart view of the global statistics:

Events over time The log timeline on the horizontal axis represents the total timeframe covered by your log (from earliest to latest timestamp observed).

The events over time metric then shows the level of activity in your process by plotting the number of performed activities in the process on the vertical axis. You can hover over the graph with your mouse to inspect the different data points as shown in Figure 4.28.

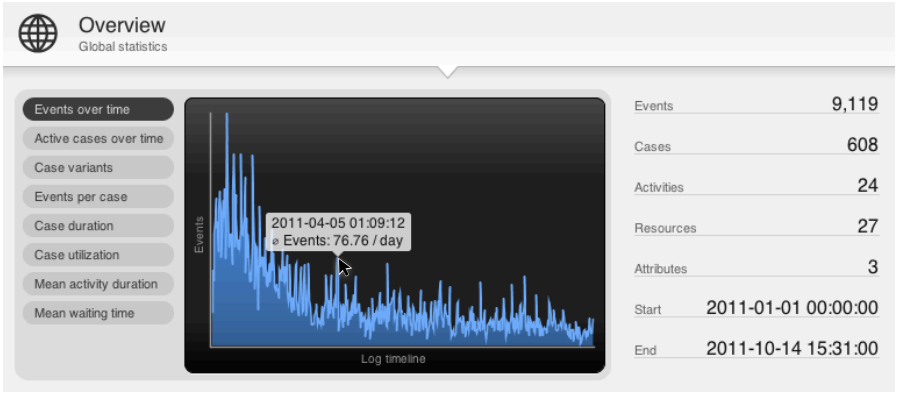


Fig. 4.28. Events over time.

Active cases over time The active cases over time metric shows you the development of the number of cases that are in progress at the same time. When new cases are started, then the value on the vertical axis rises. When cases are completed, then the level of active cases drops.

Note that this chart is also used in the Cases view (Section 4.3) as a visual reference as to over which timeframe the case is active in comparison to the overall log timeline.

Fig. 4.29. Active cases over time.

Case variants In addition to the list of variants shown at the bottom of the global statistics view (Figure 4.27), this chart gives you a visual representation of how many cases follow which variant.

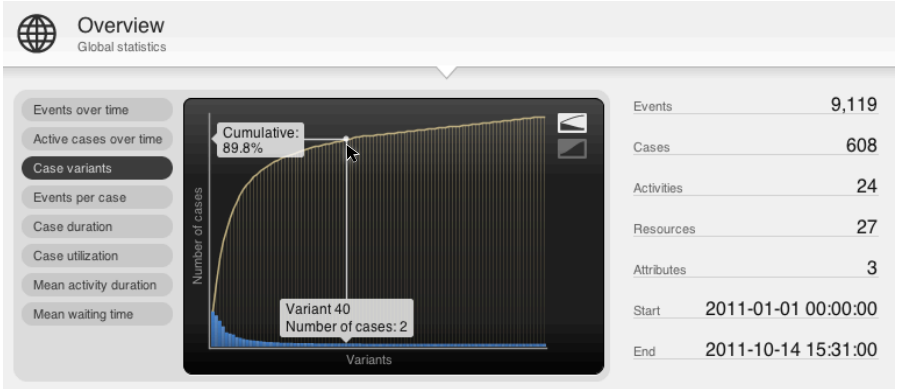
There are two alternative views to inspect the case variant chart:

- In Figure 4.30(a) you can see the pareto view. The variants are lined up from the most frequent on the left of the horizontal axis to the least frequent ones on the right. When you hover over the chart with your mouse, the displayed bubbles show you both the number of cases that follow the activity sequence pattern of the current variant and the total percentage of cases in your data set that are currently covered by all variants from the left up to this point.

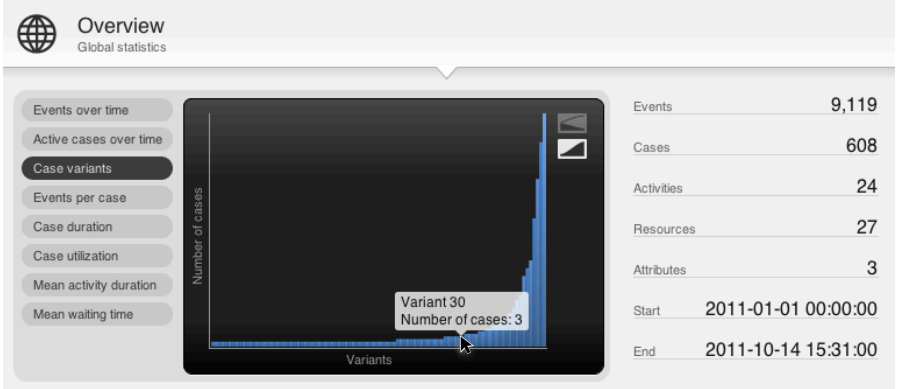
  For example, in Figure 4.30(a) you can see that the most frequent 40 variants cover almost 89.8% of the cases, and that Variant 40 in particular is followed by 2 cases.

- A normal histogram view is shown when you click on the corresponding symbol in the upper right corner of the chart. In the histogram view, the variant frequencies are simply displayed from least frequent to most frequent as shown in Figure 4.30(b), where, for example, the activity sequence of Variant 30 is followed by just 3 cases.

Events per case The events per case metric shows a distribution of how many activities typically occur over the course of the process. The horizontal axis runs from the minimum number of events that has been observed on the left up the maximum number of activities on the right. On the vertical axis the number of events per case is displayed.



(a) Pareto chart showing the cumulative number of cases covered by the variants so far.



(b) Histogram displaying the number of cases for each variant.

Fig. 4.30. Two alternative views for case variants.

For example, in Figure 4.31 you can see that in total 80 cases each performed 17 activities in the process. These 17 activities may have been performed in different sequences or may be different activities in the first place. So, the events per case metric gives you a general sense of how many steps are necessary to complete the process. Again, the Performance filter (Section 5.2.3) can be used to focus your analysis on particular long or short cases.

Case duration The case duration metric shows you the throughput time of the process from the very beginning (start of first activity) to the very end (completion of last activity).

For example, in Figure 4.32 one can see that there are a number of short-running cases and then, towards the right, a smaller number of very long-running cases (the one pointed out by the mouse lasted 74 days and 20 hours). Also here you may want to use the Performance filter (Section 5.2.3) to hone in on particular

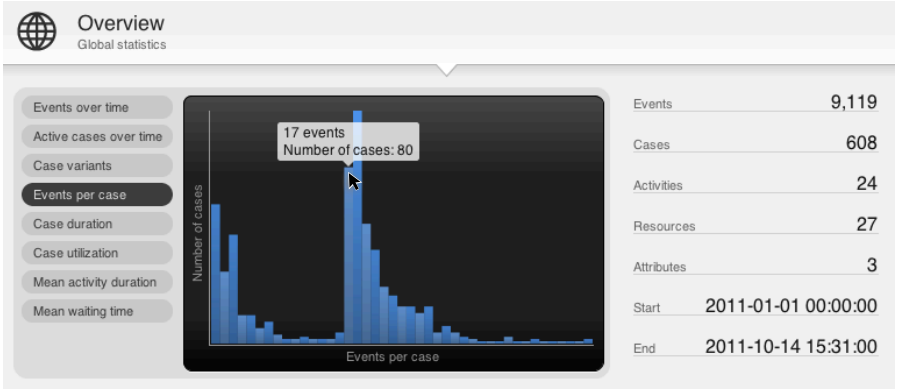


Fig. 4.31. Events per case.

fast or slow cases, or on cases that meet or do not meet a certain service level target for your process.

If you want to see the distribution of time it takes to get just from one specific activity in the process to another one instead of the complete process, you can apply a Trim filter. Refer to Section 5.2.4 to learn how to do that.

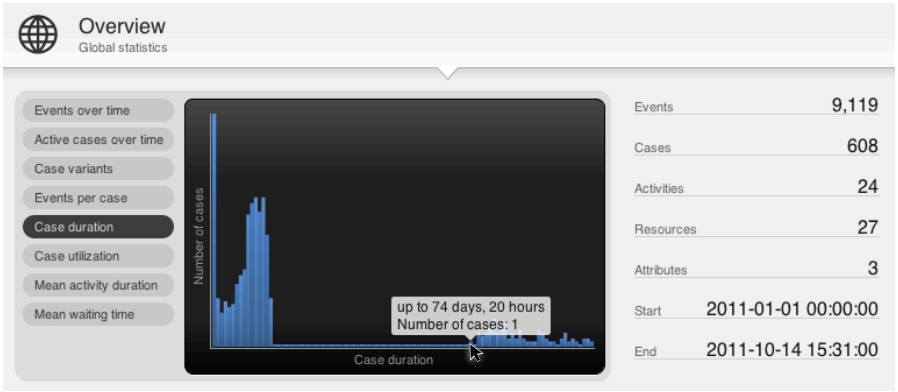


Fig. 4.32. Case duration.

Case utilization (only available if you have start and end timestamps for your activities) The case utilization metric gives you a sense of how much time in your process is spent in activities (active time) relative to the time in between activities (idle or waiting time). If the case utilization is 1.0 (100%), then all the time has been actively spent performing activities (this is trivially true if only one activity was performed).

In many processes, the case utilization is low because there are often much longer waiting times between activities than the time that it takes to actually perform an activity in the process. For example, in the case utilization chart in Figure 4.33 the highest case utilization that was achieved in the process was 41.2%.

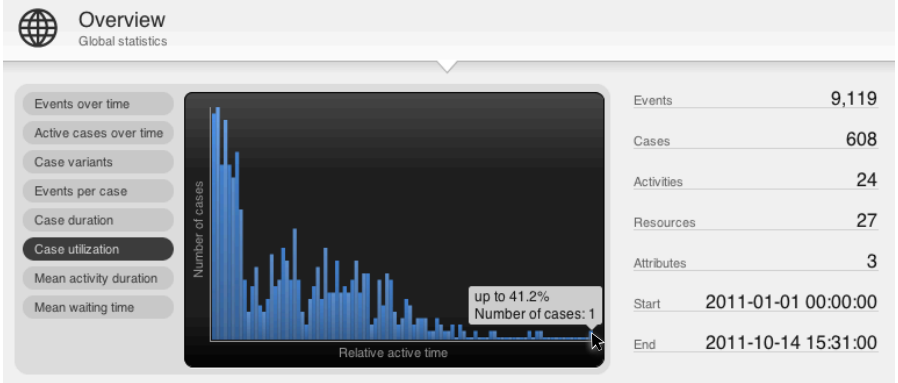


Fig. 4.33. Case utilization.

Mean activity duration (only available if you have start and end timestamps for your activities) The mean activity duration shows how much time was spent—on average–per activity for each case.

For example, in Figure 4.34 the average activity execution time was around 19 minutes for 44 cases in the data set.

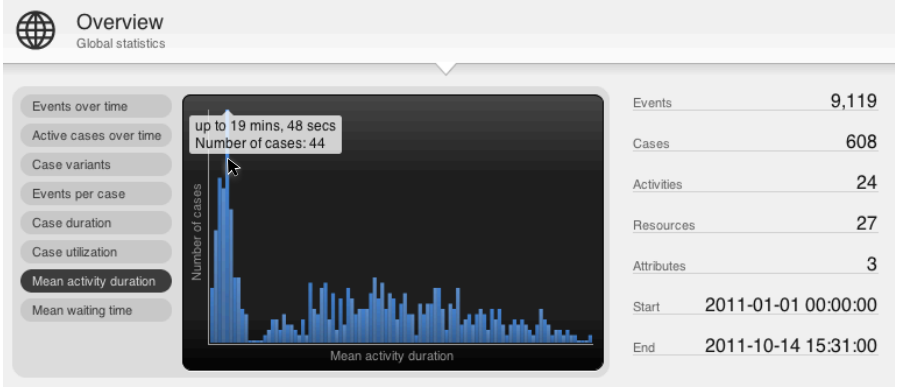


Fig. 4.34. Mean activity duration.

Mean waiting time (only available if you have start and end timestamps for your activities) The mean waiting time indicates the average time that was spent inactively between two activities.

For example, in Figure 4.35 there was one case, where—on average—44 days were spent without the execution of any activity in the process.

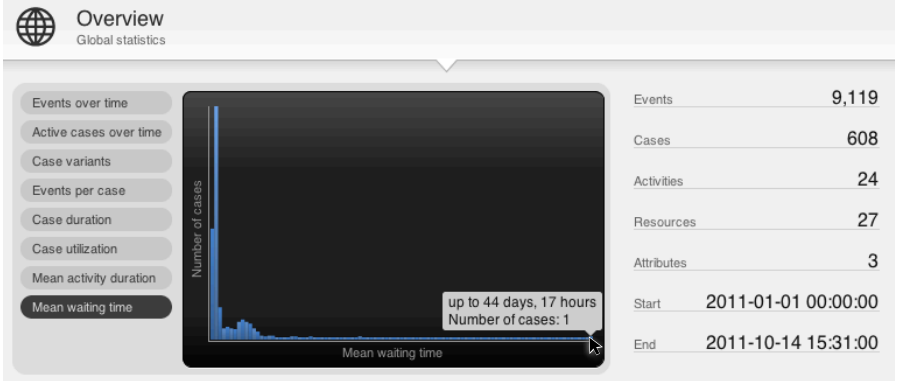


Fig. 4.35. Mean waiting time.

4.2.2 Activity Statistics

The second statistics view is the Activity statistics (❺ in Figure 4.25). It contains performance metrics about the activities in your process.

Depending on your import configuration, the activity may have been composed out of multiple columns. For example, in the call center example in Figure 4.36, the activity name has been composed from the *Operation* and from the *Agent Position* column to distinguish activities that take place in the frontline and back line. Refer to Section 3.1.4 for how to combine multiple activity columns.

The following information is available in the Activity statistics (Figure 4.36):

❶ Like in the Global statistics (Section 4.2.1), there are a number of different metrics available to display in the chart view:

- **Frequency** How often each activity has occurred in the data set. If there are no start and end timestamps for the activity in your data set, then this is the only chart that is displayed. Refer to Section 3.1.6 to learn how to include multiple timestamp columns.
- **Mean duration** (only available if you have start and end timestamps for your activities) The average time between the start and the end of each activity. This is the same metric that is also available as the *Mean duration* performance metric in your process map (see Section 4.1.4), but here you have a sorted chart view of all activities, which sometimes is useful.

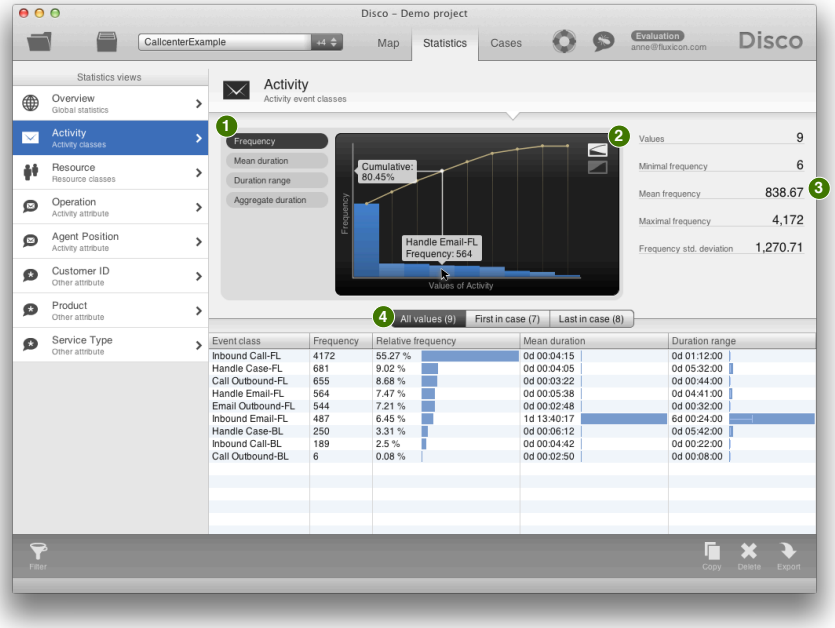


Fig. 4.36. Activity statistics in Disco.

- **Duration range** (only available if you have start and end timestamps for your activities) The time span between the minimum observed execution time and the maximum observed execution time for each activity. By showing how much difference there has been between the fastest and the slowest executions of each your activities, you can see how homogenous or heterogeneous the execution times are.
- **Aggregate duration** (only available if you have start and end timestamps for your activities) The sum of all executions for each activity. This is the same metric that is also available as the *Total duration* performance metric in your process map (see Section 4.1.4). It shows you on which acvitity—in a cumulative view—most of the active time has been spent in your process.

❷ All of the charts can be viewed either as a Pareto chart or as a normal histogram. In Figure 4.36 you see the Pareto chart view for the activity frequencies, where like in some of the Global statistics charts (Section 4.2.1) a yellow line is displayed that shows the cumulative, relative sum (how much out of 100%) for the values below. To change to the histogram view, you can press the corresponding symbol in the upper right corner of the chart (❷).

For example, Figure 4.37 shows the alternative histogram view for the same activity frequencies as Figure 4.36.

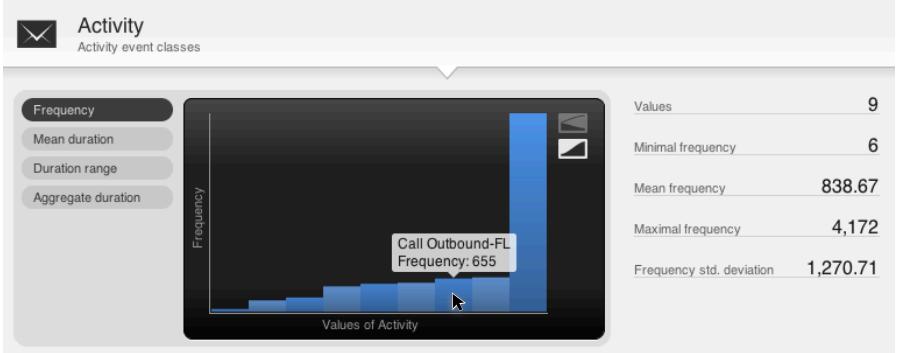


Fig. 4.37. Histogram view of the chart.

❸ In the overview information area, the following statistics are displayed:

- *Values*: The number of different activities.
- *Minimal frequency*: How often the least frequent activity has occurred.
- *Mean frequency*: How often each activity has occurred on average.
- *Maximal frequency*: How often the most frequent activity has occurred.
- *Frequency std. deviation*: The standard deviation for the frequency of activities.

❹ In the detailed table view, you can see a list of all activities sorted by their absolute and relative frequency, along with the mean duration and duration range for each activity (if start and end timestamps for activity are available in your data set).

In addition, you can switch to a view where you only see those activities that have occurred as the first or the last activity in a case. This can be useful when you want to check whether you have incomplete cases in your data set: There are often only a number of valid start and end activities in a process, and all other cases have either been started before the data extraction period began or were not finished yet when the data was extracted.

The Endpoint filter can be used to clean your data of incomplete cases. Refer to Section 5.2.4 to see how the Endpoint filter is used.

4.2.3 Resource Statistics

The Resource statistics (❻ in Figure 4.25) view is built up in exactly the same way as the Activity statistics view, but shows you information about what you have configured as a Resource during import (see Section 3.1.2).

For example, in Figure 4.38 you can see that there are 48 different people working in the purchasing demo process. Refer to Section 4.2.2 for details about the charts, overview statistics and table view.

Fig. 4.38. Resource statistics in Disco.

4.2.4 Other Attribute Statistics

Any column that you have included during import, but which was not assigned as a Case, Activity, Resource, or Timestamp column, will be included as a separate attribute here in the Statistics view (❼ in Figure 4.25).

These additional attributes often provide valuable information about your process. Typical attributes are, for example, product types in a support process, the value of a purchase in a purchasing process, the type of issue in an incident handling process, and so on.

The attribute statistics then give you a breakdown of the frequencies for the different values in each of these attributes. Particularly in combination with filtering, you can use them in two powerful ways:

- *Use attributes to split out different processes:* Sometimes, you actually have different *types* of processes. For example, when a customer service process can be started either via the call center or through a dealer, then the precise process flows that are performed in these two situations often differ, there may be different service level agreements in place, and different parts of your organization may be in charge of them. So, you want to analyze them in isolation.

  This can easily be done if you have an attribute in your data set that indicates the channel. Simply use an Attribute filter (Section 5.2.5) to first create a copy of

your data set for the channel attribute value *call center* and then another copy for the channel attribute value *dealer*. As a result, you will have two separate data sets that can be compared and analyzed independently from each other.

- *Inspect attributes after filtering:* In return, the attribute values can give you feedback about which types of cases have problems after you applied another, for example, performance-oriented filter. Say, for example, you have applied a Performance filter (Section 5.2.3) to focus on cases that do not meet the agreed-upon service level because they took too long to complete. By inspecting the attribute value frequencies after filtering you can see for which channels, products, or other categories this problem is more prevalent than for others.

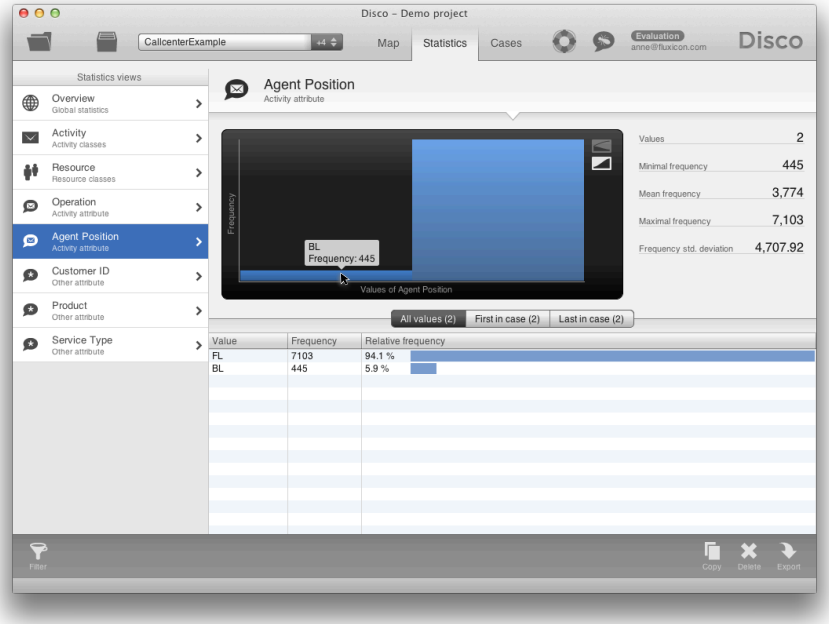


Fig. 4.39. Attribute statistics in Disco.

Figure 4.39 shows a screenshot of an Attribute statistics view for the *Agent Position* attribute from the callcenter example, which has just two values. FL stands for frontline and BL stands for backline. You can see that ca. 94% of the activities occurred in the 1st level support (frontline) and ca. 6% were performed in the 2nd level support (backline).

In fact, in this situation the *Agent Position* attribute has been combined with the *Operation* attribute to form the activity name (Figure 4.36). Although *Agent Position* is used as part of the activity name here, each of the combined columns is still

available through an individual attribute column. To remind you that the attribute is also used as part of the activity, there is a little envelope symbol embedded in the speech bubble symbol. This works in the same way if you choose to combine multiple columns to determine the resources in your process.

Refer to Section 4.2.2 to learn more about when it is useful to combine multiple Activity or Resource columns, and how to do it.

4.3 Cases view

While the Map view (Section 4.1) gives you an understanding about the process flows, and the Statistics view (Section 4.2) provides you with detailed performance metrics about your process, the Cases view actually goes down to the individual case level and shows you the raw data.

You get to the Cases view by simply changing to the *Cases* tab as shown in Figure 4.40.

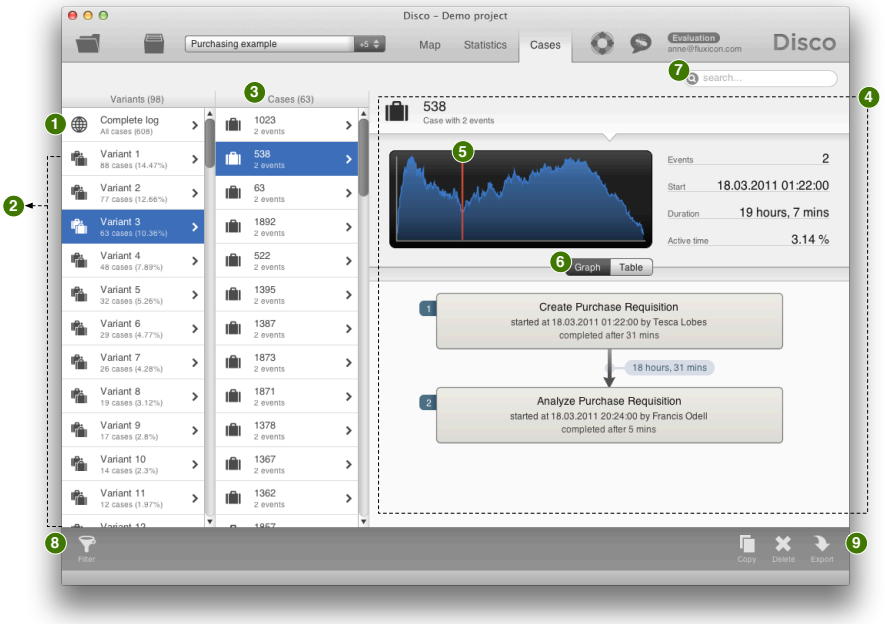


Fig. 4.40. The Cases view in Disco.

To be able to inspect individual cases is important, because you will need to verify your findings and see concrete examples particularly for “strange” behavior that you will most likely discover in your process analysis. Almost always you will find

things that are hard to believe until you have drilled down to an individual example case, noted down the case number, and verified that this is indeed what happened in the operational system.

Furthermore, looking at individual cases with their history and all their attributes can give you additional context (like a comment field) that sometimes explains why something happened. Finally, the ability to drill down to individual cases is important to be able to act on your analysis. For example, if you have found deviations from the described process, or violations of an important business rule, you may want to get a list of these cases and talk to the people involved in them to provide additional training.

The Cases view consists of the following areas:

❶ *Complete log*. Select the Complete log item on the left to see a list of all cases in your data set in the second column to the right (❸ in Figure 4.40).

❷ *Individual variants*. Alternatively, you can select an individual variant, whereas a variant is a specific sequence of activities. Only cases that follow the same activity sequence are then displayed in the list of cases in the second column (❸). The variants are sorted by their frequency.

For example, in Figure 4.40 the third most frequent variant has been selected. There are 63 cases (accounting for 10.36% of the whole log) following the activity sequence *Create Purchase Requisition* → *Analyze Purchase*. One of these cases (case 538) has been selected and is displayed. Refer to Section 4.3.1 to learn more about variants and how to take advantage of them in Disco.

❸ *List of cases*. A list of cases IDs for either the complete log (❶) or the selected variant (❷).

❹ *Individual case*. In the main window you can see further details about the currently selected case from ❸. Refer to Section 4.3.2 for further details about the detailed case view.

❼ *Search*. You can search for specific case IDs or attribute values. Read Section 4.3.3 to see how searching cases works.

❽ *Filtering*. The log filter controls for the current data set can be accessed from each of the analysis views. Filters are really important drill into specific aspects of your process and to focus your analysis. Read Chapter 5 for detailed information on how filtering works in Disco.

❾ *Copy, Remove, and Export data set*. Data sets can be copied, deleted, and exported right from the current analysis view. Read Section 6.2 for further details on copying and deleting data sets. Process maps can be exported, for example, as a PDF file. The export functionality of Disco is explained in detail in the Export reference in Chapter 7.

4.3.1 Inspecting Variants

Variants are an integral part of the process analysis. In Disco, a variant is a specific sequence of activities. You can see it as one path from the beginning to the very

end of the process. In the process map (Section 4.1), an overview of the process flow between activities is shown for all cases together. A variant is then one "run" through this process from the start to the stop symbol, where also loops are unfolded.

Usually, a large portion of cases in your data set is following just a few variants. For example, in the purchasing process shown in Figure 4.40 the top five most frequent variants cover the process flows of ca. 50% of all cases while there are 98 different variants in total.

It is useful to look at the distribution of cases over variants in your data set to:

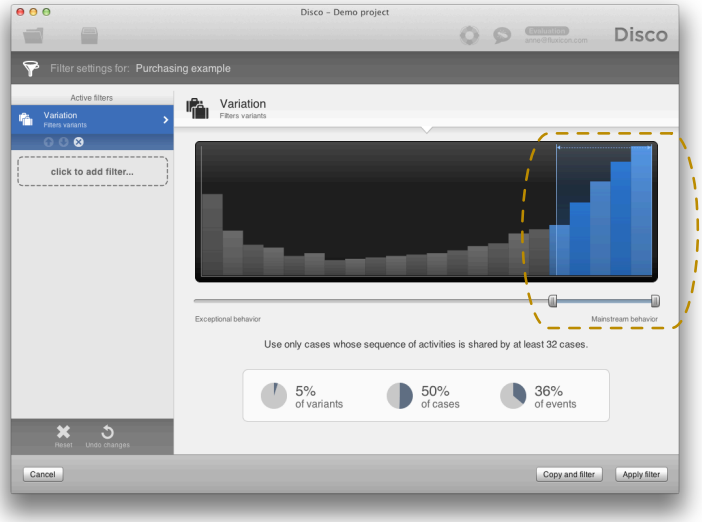
- *Know what the most common activity sequences are.* For example, in Figure 4.40 one can see that the third most frequent variant is the activity sequence *Create Purchase Requisition → Analyze Purchase*. In total 63 cases (10.36% of all cases) follow this particular pattern from beginning to the end of the process.
  From a process improvement perspective this seems strange. Why are so many cases stopped? Should the purchasing guidelines be updated to clarify what people can buy and what not?
- *See how much variation there is.* Besides inspecting the most dominant sequences, it is also useful to know how many variants are there in the first place. For example, popular methodologies like Lean Six Sigma focus on reducing variation to improve quality and process efficiency at the same time. In fact, there is often much more variation in processes than people are aware of. Knowing just how much variation there is (how many variants in total) helps to get a sense of how streamlined your process is.
- *Simplify your process map.* In situations where you want to focus on the mainstream behavior, you can view the process map and statistics for just the most frequent variants. To do this, you can use the Variation filter as shown in Figure 4.41(a). After applying the filter, the process map for just the mainstream process flows are shown in Figure 4.41(b).
  Read Section 5.2.2 to learn how the Variation filter works.

Note that there are certain types of processes, which do not exhibit many common activity patterns. For example, in a hospital the diagnosis and treatment process for almost each patient is unique. These kind of processes are often called "unstructured" and the variants and the variation filter are not of much help in these situations. Instead, you can use the simplification controls of the Map view (Section 4.1) to understand and analyze the process flows.

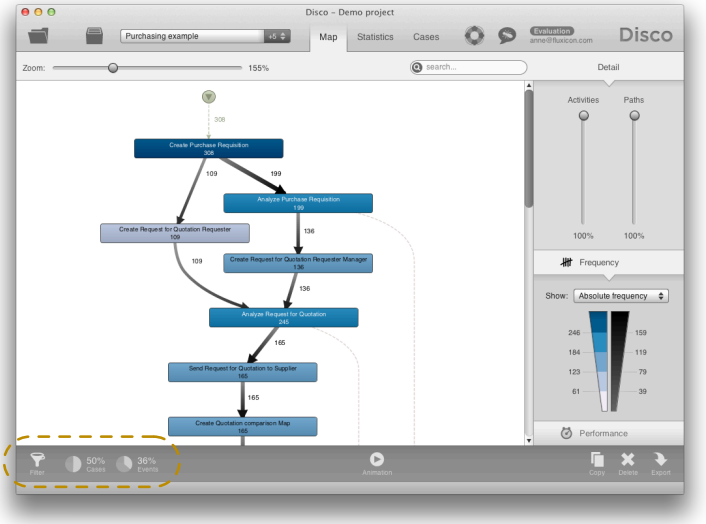
4.3.2 Individual Cases

You can inspect individual cases by selecting one of the case IDs in the list of cases (❸ in Figure 4.40). Information about the currently selected case is shown in the main window (❹ in Figure 4.40).

In this case view, you have some overview statistics about the case in the upper part, and a detailed history view in the lower part of the screen (Figure 4.40):



(a) The Variation filter lets you select a portion of more frequent variants (mainstream behavior) or less frequent variants (exceptional behavior).



(b) The process map of the purchasing process—covering 50% of the cases in the data set—after filtering for just the Top 5 most frequent variants.

Fig. 4.41. Filtering for variants.

❺ *Time frame of case.* This chart shows you where in the log the selected case occurs. The blue chart in the background is the *Active cases over time* chart from the global statistics view (Section 4.2.1). The time frame of the currently displayed case is highlighted in a bright red, so you can easily see where in the context of the overall log timeline (for example, more in the beginning or towards the end) it occurred.

Next to the time frame chart you find the following statistics:

- *Events.* The number of events (i.e., the length of the history) for this particular case. For example, the case 538 displayed in Figure 4.40 has 2 events, while case 1031 from Figure 4.42 has 17 events in total.
- *Start.* The start time for the earliest activity in the history for this case.
- *Duration.* The time between the earliest and the latest timestamp (total duration).
- *Active time* (only available if you have start and end timestamps for your activities). Percentage of time from the total duration of the case that was spent on performing activities (active time).

Fig. 4.42. The table view shows the history of the case in a compact tabular form with all attributes that were included during import.

❻ *Graph and Table view*. You can switch between a Graph and a Table view for inspecting the concrete history for the case. The Graph view is shown in Figure 4.40 and only displays the activity sequence plus time and resource information. Additional data attributes are hidden.

If you want to see the data attribute values, or if you have longer cases and find yourself scrolling around in the Graph view of the case too much, you can switch to the Table view for a more compact representation of the case history including all the attributes that were imported (see Figure 4.42).

4.3.3 Search

The search function of the Cases view allows you to quickly look for cases that contain a specific text fragment either in the case name or in one of the attributes.

For example, imagine that you are analyzing the purchasing process from the demo example and you quickly want to inspect a case where a dispute occurred. Instead of adding a filter to drill down to cases containing a dispute activity (see Section 4.1.5), you can simply search for 'dispute' in the search field of the Cases view (❼ in Figure 4.40) as shown in Figure 4.43.

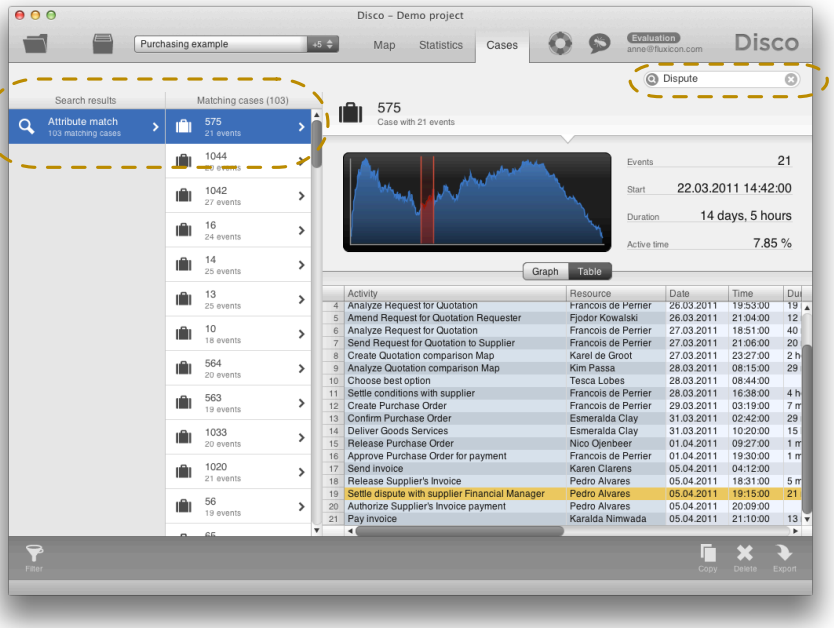


Fig. 4.43. Searching for a text snippet gives you a list of all matching cases and highlights the matching events in the case.

You can not only search for text snippets from the activity names but from all attributes in the data set. The matching events are highlighted in orange to show you where in the case your search has given a result.

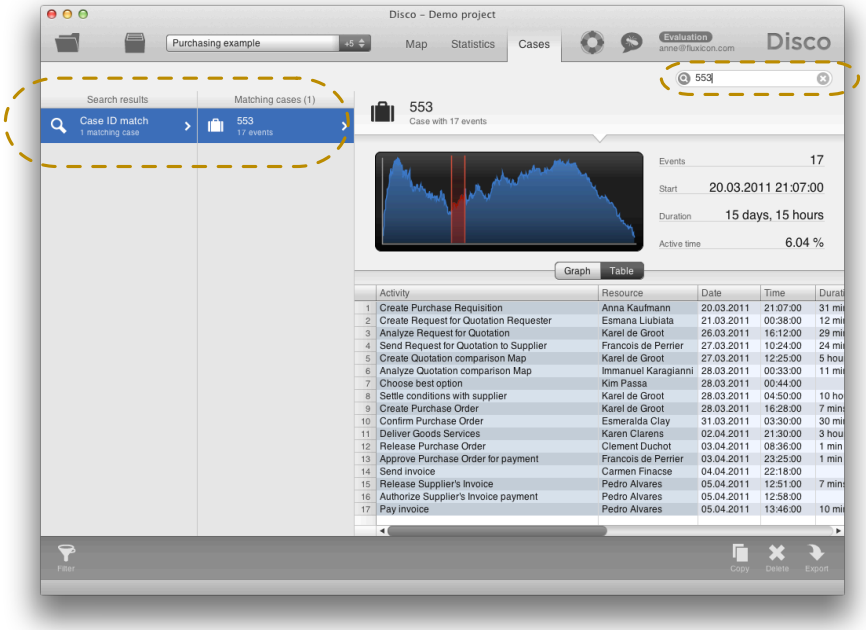


Fig. 4.44. The search for case IDs helps you to find specific cases in your data set.

Finally, you can also search for specific case IDs. Just type the name of the case in the search field and Disco gives you the case history you are looking for. An example case search is shown in Figure 4.44. This can be particularly useful if during your analysis you are comparing the data from the operational system and the analyzed data in Disco to verify your findings.

5

Filtering

In this chapter you find further details about how exactly you can use the log filters in Disco and how they work.

5.1 Working with Filters

Filtering is symbolized by the funnel symbol shown in Figure 5.1.

Fig. 5.1. Filter symbol in Disco.

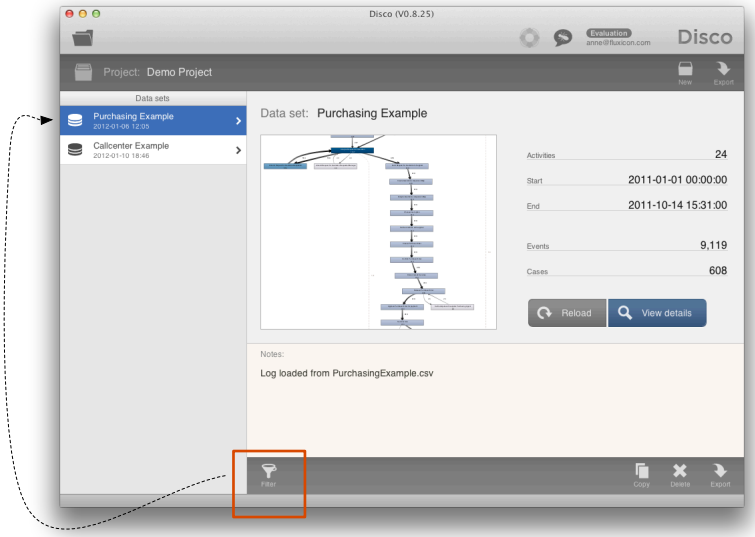
Clicking on the filter symbol will open up the filter settings for the current data set. You can open the filter dialog from multiple places in Disco, making it easy for you to modify or inspect your current filters regardless of where you are in the application.

You find the filter symbol in the following places in Disco:

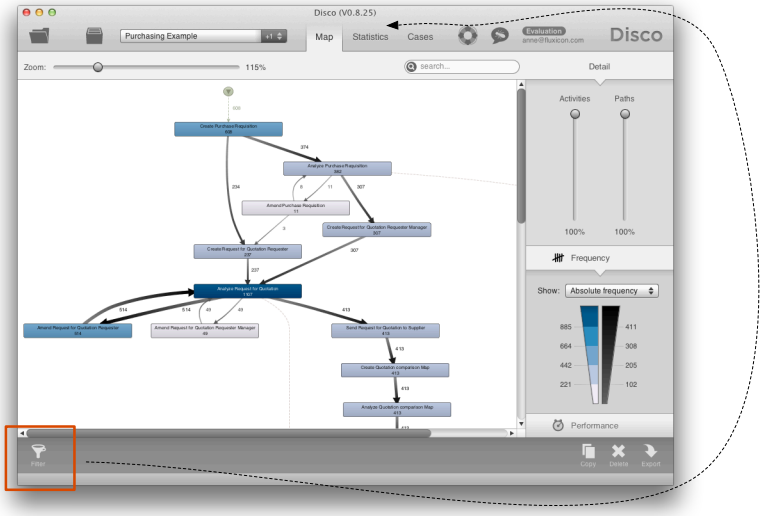
Project view. Pressing the filter symbol on the lower left in the project view – see Figure 5.2(a) – brings up the filter settings for the currently selected data set.

Map view, *Statistics view*, and *Case view*. You can access the filter settings from any of the three data set views by clicking on the filter symbol in the lower left corner – see Figure 5.2(b).

If you click on the filter symbol in any of these places, this will bring up the filter settings dialog. In Figure 5.3 you can see that the empty filter settings contain three main areas, numbered 1–3. These three areas are:



(a) Pressing the filter symbol in the project view brings up the filter settings for the currently selected data set.



(b) The filter settings can be accessed from any of the three data set views: Map view, Statistics view, and Case view.

Fig. 5.2. Opening the filter settings in Disco.

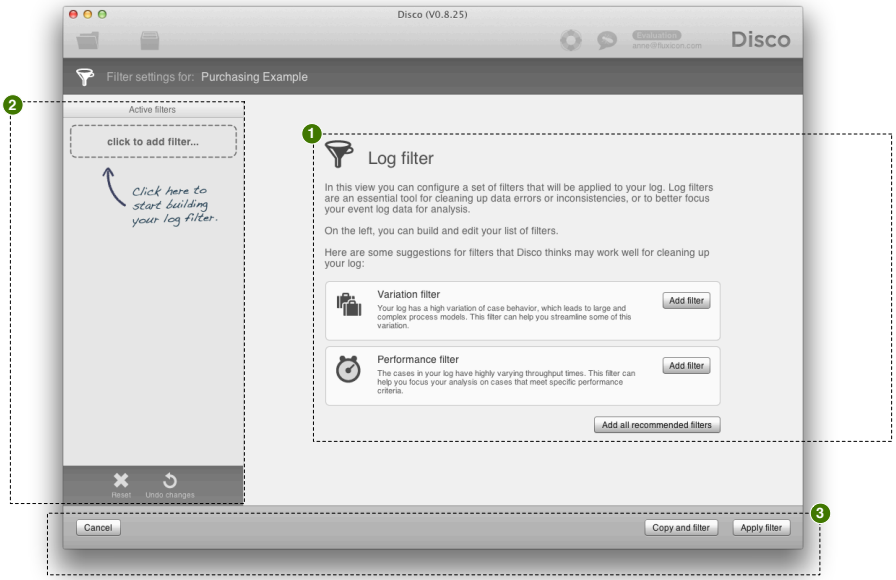


Fig. 5.3. The filter settings dialog in Disco.

❶ *Recommended filters*. You can learn more about Disco's filter recommendations in Section 5.1.1.

❷ *(Still empty) list of active filters*. Learn how to add filters to build your filter list in Section 5.1.2.

❸ *Controls*. After building up or changing your filter list, you can apply the new filter settings. Learn what it means to apply the filter settings in Section 5.1.3.

5.1.1 Filter Recommendations

When you open up the filter settings for a new data set, Disco gives you suggestions for filters that may be useful for your log (see ❶ in Figure 5.3). The recommendations are specific to the data set. So, if you open the filter settings for another data set you most likely will see different recommendations.

For example, in Figure 5.3 we have opened the filter settings for the *Purchasing Example* that comes as a demo event log with Disco. Two filter recommendations are displayed for this particular log:

- Variation filter: Disco detects that there is a high variation of case behavior, which typically leads to large and complex process models. It suggests to use the Variation filter to help you streamline some of the variation. (See Section 5.2.2 for how to use the Variation filter.)

- Performance filter: Disco lets you know that cases in your log have highly varying throughput times. It recommends the Performance filter to focus your analysis on cases that meet specific performance criteria. (See Section 5.2.3 for how to use the Performance filter.)

For another data set you might, for example, get the recommendation to use the Timeframe filter (see Section 5.2.1) because Disco has detected that some of the timestamps lie in the future, which usually hints at a data quality problem.

The recommended filters are meant as a starting point. Read the explanation that comes with the recommendation and decide whether you want to look at the filter. If you want to add one of the recommended filters press the *Add filter* button, or press *Add all recommended filters* to add all at once.

> **Note:** If you add a recommended filter, this does not do anything yet! You still have to configure the filter to take effect. Read through the manual for the individual filters to understand what they do.

There is no guarantee that Disco will recommend all filters that would be useful to you. That's not possible because a lot of the relevant knowledge about a process is only available to a domain expert. Over time, you will build up experience in working with filters. You will know which filter you want to use when.

Read on to learn how to add specific filters and manage your filter stack in the next section.

5.1.2 Adding Filters and Managing the Filter Stack

When you want to add a specific filter to your data set, you can click on the *click to add filter ...* area in the filter list (see ❷ in Figure 5.3). In Figure 5.4 you see how a list of the different types of filters appears in the next step. Move your mouse pointer along this list to choose the type of filter that you would like to add.

After you have chosen a filter type, an instance of this filter will be placed in your filter list. You can either stop there and work with the single filter, or you can add more filters. Figure 5.5 shows an example, where four filters have been added to the data set: A Timeframe filter, an Endpoint filter, and two Attribute filters. As you can see, multiple filters of the same type can be added to the filter stack.

Figure 5.5 also shows you an inventory of the controls (numbered 1–6) that you can use to manage the filter stack in Disco. The following controls are available:

❶ *Settings area of the currently selected filter.* The configuration of the currently selected filter is shown in this area. For example, in Figure 5.5 you can see the current configuration of the first Attribute filter. The settings for the different filter types are explained in detail in Sections 5.2.1–5.2.6.

❷ *Moves up currently selected filter in the list.* If you want to move the currently selected filter up in your list, you can press the little ↑ button. Learn more about the order of your filters in the filter stack in Section 5.1.3.

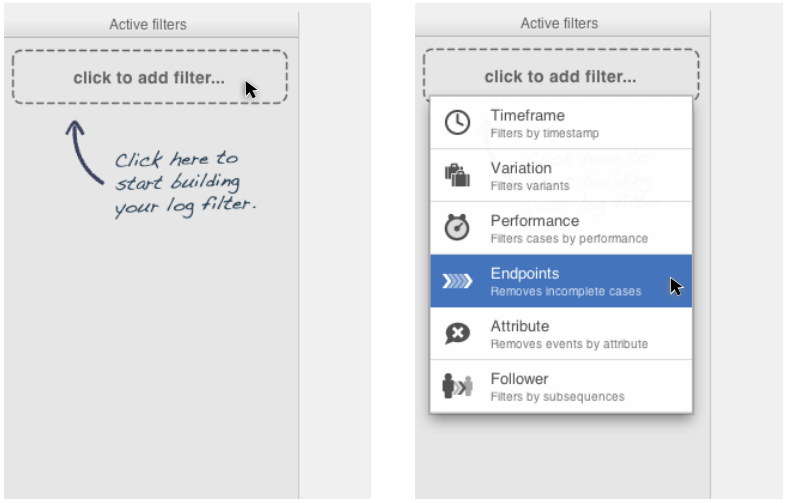


Fig. 5.4. Add new filters by clicking on the *click to add filter ...* area in your filter list. Then, choose the type of filter that you would like to add.

❸ *Moves down currently selected filter in the list.* Press the ↓ button to move the currently selected filter down in your list. Again, read Section 5.1.3 to see how changing the order of your filters in the list affects the outcome.

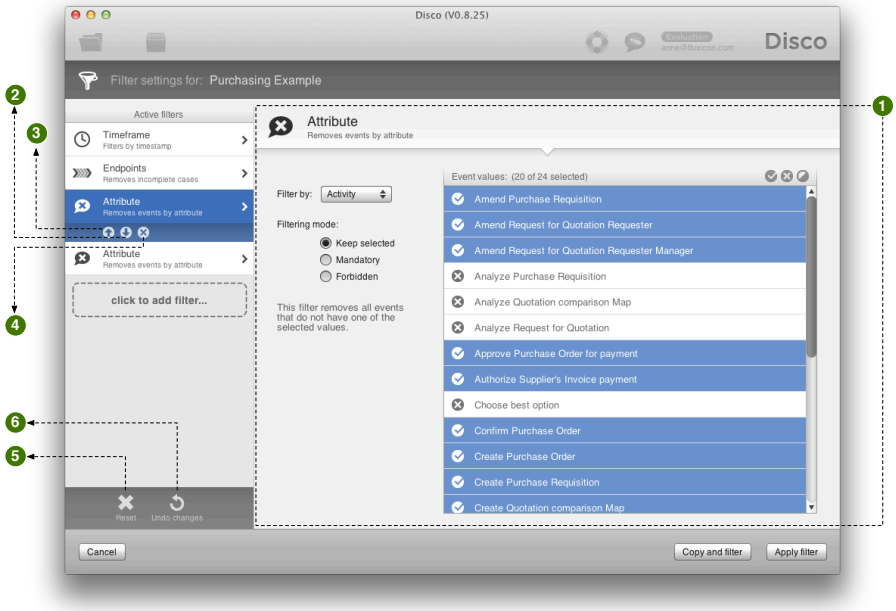


Fig. 5.5. Managing the filter stack in Disco.

❹ *Removes currently selected filter from the list.* If you want to remove the currently selected filter completely from your list of active filters, you can press the small × symbol.

❺ *Clears all filters from the list.* Use this large × symbol at the bottom of your filter list to clear your complete filter stack (remove all filters at once).

❻ *Restores the filter list.* If you changed the settings of a filter but then changed your mind about it, or if you accidentally removed one of the filters from your list, you always have the option to restore the filter list as it was when you entered the filter dialog. This way, you can safely edit and change things, while still having the option to go back if something goes wrong.

Read on to learn how you can apply the current filter settings to your data set, what happens when you apply them, and how you can apply them to a copy of your data set to preserve the original log.

5.1.3 Applying Filters

After you have added and configured your filters, you can do three things in the control area of the filter dialog (see ❸ in Figure 5.3):

Apply filter. If you apply the current filter settings, then each of the filters in your filter stack is applied to the original log, one after the other (see Figure 5.6).
The percentage of cases and events that remain in the filtered data set—compared to the original log—are now indicated next to the filter icon to remind you that you have applied filters. The original data set is still there, and you can go back into the filter dialog and change or remove your filters any time.

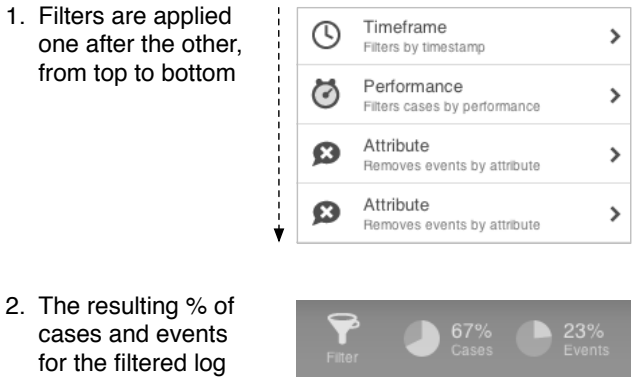


Fig. 5.6. Applying filters.

Copy and filter. If you press *Copy and filter* instead of simply *Apply filter*, then your current filter stack is applied to a newly created copy of your data set and the

original log remains unchanged. This is handy if you realize that you are about to create a subset of your data set that you would like to preserve.

As shown in Figure 5.7 ❶, you can provide a meaningful name (e.g., ”Slowest cases in January”) before you actually create the copy. The created copy of the data set will have an indication of the percentage of cases and events as shown in Figure 5.6 just like the original log would have had, if you had used *Apply filter* instead. The original and the copied data set can be both found in your project view. See the Workspace reference Section 6.3 to learn more about how to manage your data sets in a project.

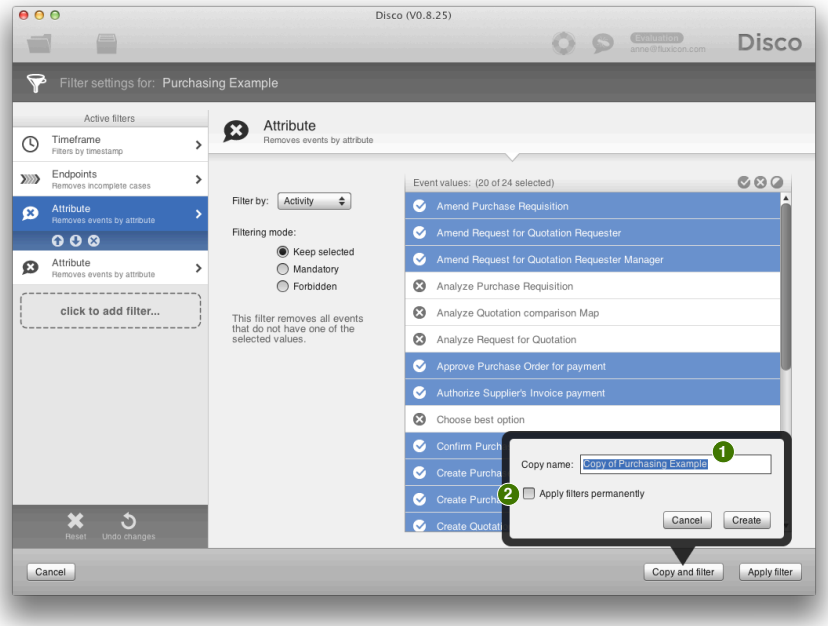


Fig. 5.7. Pressing *Copy and filter* applies the current filter settings to a copy of your data set (and lets the original one unchanged). You can give a new name for the copied data set ❶ and optionally choose to make the current filter stack permanent ❷ (consolidate) in the copied log.

If you select the option *Apply filters permanently* (see ❷ in Figure 5.7), then you will still apply your current filter stack to a newly created copy. However, the filters will be permanently applied and it will not be possible to remove or modify them later on. This is useful, for example, if you have cleaned your log from unimportant events, incorrect timestamps, or incomplete cases, and now want to take the clean log as a new reference point.

As with using *Copy and filter* without the option *Apply filters permanently*, the original data set is still there until you delete it explicitly from your project.

Cancel. Pressing the *Cancel* button simply brings you back to the previous state of the data set. It has the same effect as pressing first *Undo changes* and then *Apply filter*.

After you apply your current filter stack, you are normally just brought back to the project or data set view from where you entered the filter dialog. However, in some situations you might get the sad triangle (see Figure 5.8) telling you that filtering resulted in an empty log. This means that after all filters in the stack were applied, none of the cases remained in the data set anymore.

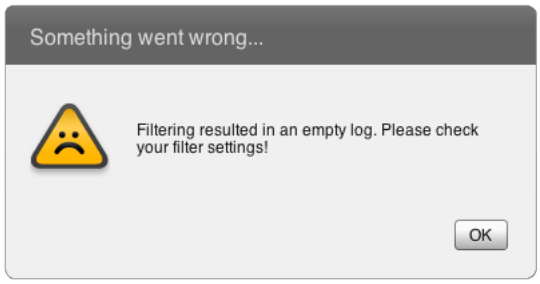


Fig. 5.8. The sad triangle tells you that your filter settings did not give a result.

How can it be that no cases remain in the data set? There are two reasons why you might end up with an empty log:

1. First of all, it could be that a *combination of filters* leads to this situation. Imagine, for example, that you add a Timeframe filter (see Section 5.2.1) that only keeps cases that were started and completed in January 2012. Then you add another Timeframe filter that you configure to keep only cases that were started and completed in May 2012. None of your process instances could have been started and finished both in January *and* in May 2012. Filtering results in an empty log.
2. Secondly, sometimes a *single filter* can give you an empty log result. Imagine, for example, that you have added a Follower filter (see Section 5.2.6) that selects all process instances from your data set in which a client case has been re-opened after it had been closed. If this situation did not occur at least *once*, then your filter results in an empty log. And sometimes, that's exactly what you want. For example, if you are using the Follower filter to check a 4-Eyes rule then an empty log result confirms that no violations of that rule occurred.

If you get an empty log but are sure this isn't what you want, or if you have trouble configuring multiple filters in combination, skip to Section 5.3 to learn how to best resolve such problems.

5.2 Filter Types

There are six different types of filters available in Disco: *Timeframe* filter (Section 5.2.1), *Variation* filter (Section 5.2.2), *Performance* filter (Section 5.2.3), *Endpoints* filter (Section 5.2.4), *Attribute* filter (Section 5.2.5), and *Follower* filter (Section 5.2.6).

Read on to learn more about these filters and how you can use them.

5.2.1 The Timeframe Filter

The timeframe filter enables you to focus your analysis on a certain time period. For example, imagine that you want to analyze four weeks of data to understand the typical process load for one month. Although your data set contains several months of data, you would like to restrict the considered timeframe from Monday 3 February 2011 until Monday 31 February 2011. The timeframe filter allows you to do just that.

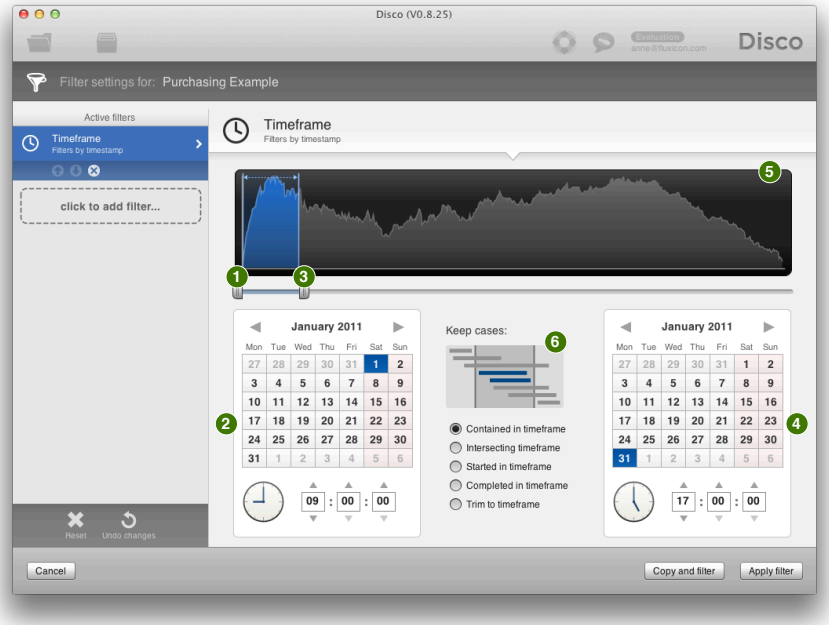


Fig. 5.9. Controls and options of the Timeframe filter.

Figure 5.9 gives you an overview of the options and controls that you have for the Timeframe filter in Disco:

❶, ❷ *Start of selected timeframe*. Choose the beginning of the timeframe that you want to keep in the filtered log by simply dragging the left handle ❶ of the timeframe slider along the time axis. While you are moving the slider, the changing start time will be reflected in the synchronously updating calendar sheet ❷. It also works the other way around: You can explicitly select a specific date (browse through the months by using the ◁ and ▷ buttons at the top) and time in the start time calendar sheet ❷, and this will change the position of the start timeframe slider handle ❶.

For example, in Figure 5.9 the start time has been set to 1 January 2011 at 09:00 o'clock.

❸, ❹ *End of selected timeframe*. Choosing the end of the timeframe works in the same way as choosing the start of the time frame, using the right handle ❸ of the timeframe slider or the end time calendar sheet ❹.

In Figure 5.9 the end time has been set to 31 January 2011 at 17:00 o'clock.

❺ *Visual feedback*. In the chart at the top of the configuration screen, you can see the currently covered timeframe highlighted in blue in comparison to the overall log time line. The reference chart shows the number of active cases over time to give you a sense of how many of the cases are covered by your current selection. The chart that is used is the same one as you can find in the Overview statistics (see Section 4.2.1 for more information on the *Active cases over time* visualization).

❻ *Usage mode*. In the middle of the configuration screen you can determine what you want to do with the selected timeframe: For example, do you want to keep all cases that have been both started and finished between the selected start and end date? Or do you intend to select all cases that have been just started in the configured time period?

To understand the ❻ *Usage modes*, look at Figure 5.10. The blue visualization indicates which cases are kept in the current configuration: Every (grey or blue) horizontal bar represents one case. The horizontal axis represents the timeline, and the position of each horizontal bar shows the earliest and the latest event in relation to the global log timeline. The vertical bars represent your timeframe selection.

The blue bars tell you which cases will remain in your dataset after you apply the timeframe filter, depending on the currently configured usage mode. Here is a description of the ❻ *Usage mode* settings that you can choose from:

- *Contained in timeframe*: The default mode of the Timeframe filter, where only cases that *completely lie within the selected start and end* boundaries are kept.
- *Intersecting timeframe*: Keeps cases that are *overlapping* (intersecting) with the selected timeframe. The reference point is the earliest and the latest timestamp in each case. So, even if nothing actually happened within (no events of the case actually fall into) the selected timeframe, the case will still be kept as long as it started before the end of the selected timeframe (❸, ❹ in Figure 5.9) and ended after the configured start date (❶, ❷ in Figure 5.9).
- *Started in timeframe*: Keeps cases that are *starting* in the selected timeframe.

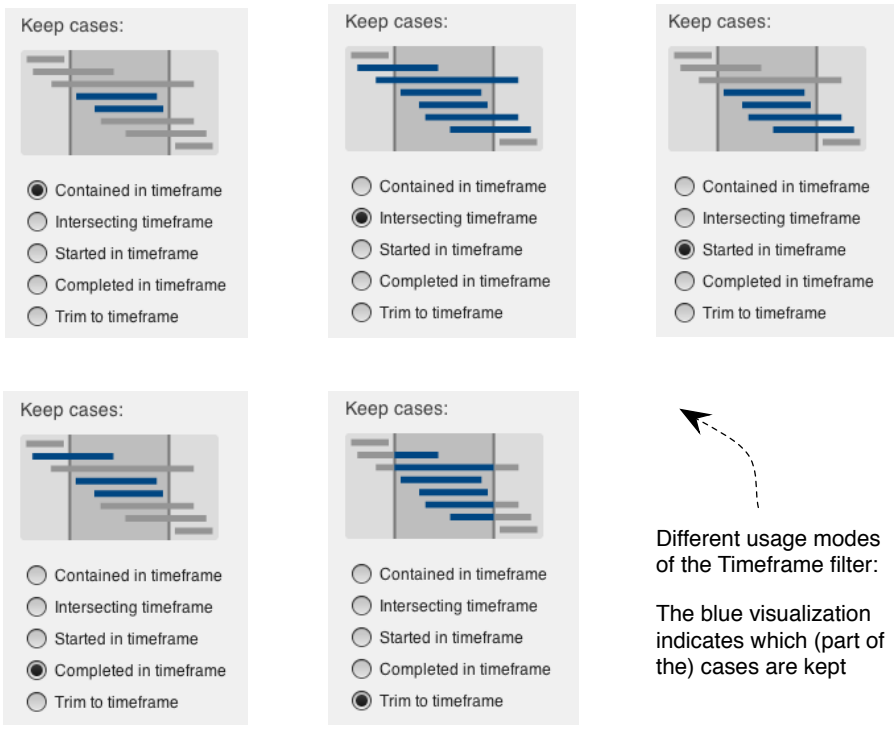


Fig. 5.10. When you change the usage mode, the blue visualizations adapt to help you understand the effect of the mode you are currently using.

- *Completed in timeframe*: Keeps cases that are *completed* in the chosen timeframe.
- *Trim to timeframe*: This one takes all the cases that intersect with the selected timeframe and cuts them off before and after the chosen time period (trims them). So, it removes all events that fall outside of the configured time period.

5.2.2 The Variation Filter

The variation filter is a great way to focus your analysis on either the most common or on the most exceptional behavior in your process.

We'll use the purchase demo example again to explain how it works. In Figure 5.11 you can see the top frequent variants for the purchasing example data set. In Disco, a variant is an activity sequence pattern. This means that all cases that follow the exact same activity sequence belong to the same process variant. If you are not yet familiar with variants, then you can learn more about them in the Cases view reference in Chapter 4.3.1.

In Figure 5.11, one can see that there are in total 98 different variants in the log. The most frequent pattern, Variant 1, is followed by 88 cases (this corresponds to

Fig. 5.11. The most frequent variants for the purchasing demo example.

14.47% of all the cases in the log). The fifth-frequent is Variant 5, which is followed by 32 cases (5.26% of all the cases in the log).

So, now let's say that instead of viewing the process map for all the 98 different variants, you only want to see how the process looks like for the most common behavior. This is exactly, what the Variation filter allows you to do.

Figure 5.12 shows the configuration settings of the Variation filter. It has been configured to include only those variants that are followed by at least 32 cases in the log. In the case of the purchasing demo example, this includes only Variant 1 – Variant 5 (see In Figure 5.11).

❶ *Lower limit*. The minimum number of cases that need to follow the activity sequence pattern of the variant to be passed through to the filtered log.

In Figure 5.12, the minimum number of cases with the same activity sequence (siblings) has been set to 32.

❷ *Upper limit*. The maximum number of cases that are allowed per variant to remain in the filtered log.

In Figure 5.12, all variants with more than 32 cases are included in the selection.

❸ *Coverage feedback*. To get a sense of how many cases in the log are covered by your current selection, you can check the pie charts at the bottom of the configuration screen.

In Figure 5.12, one can see that with the selection of just the most frequent 5% of the variants in fact 50% of all cases (and 3 % of all events) would remain in the filtered log.

Instead of focusing on the *most common behavior*, you can also zero in on the *most exceptional cases*, for example those that occurred only once or twice. In this case

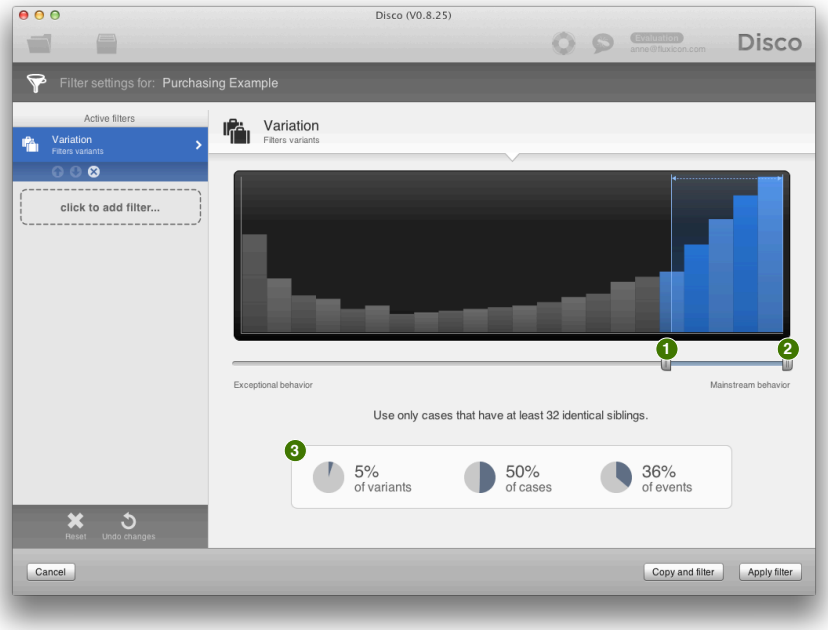


Fig. 5.12. Settings of the Variation filter.

you would move the handle of the variation slider ❶ to the very left and adjust the upper boundary ❷ accordingly.

> **Note:** The variation filter provides a very controlled way of filtering common behavior for structured processes, where some activity sequence patterns are very common and others very rare. Often, 20% of the variants make up 80% of the cases. For unstructured processes, this may be less useful.
>
> Read more about the interactive simplification controls that you can use for any data set to adjust the level of detail in your discovered process map in the Map view reference Chapter 4.1.

5.2.3 The Performance Filter

The performance filter is a case filter that allows you to focus on cases in your data set according to certain performance criteria. For example, if you want to see all cases that exceed a specific throughput time (violating the service level agreement for your process) and find out in which part of the process these cases lose most of the time, then the performance filter helps you to do that.

Figure 5.13 shows the configuration options for the performance filter in Disco. To use the filter, go through the following steps:

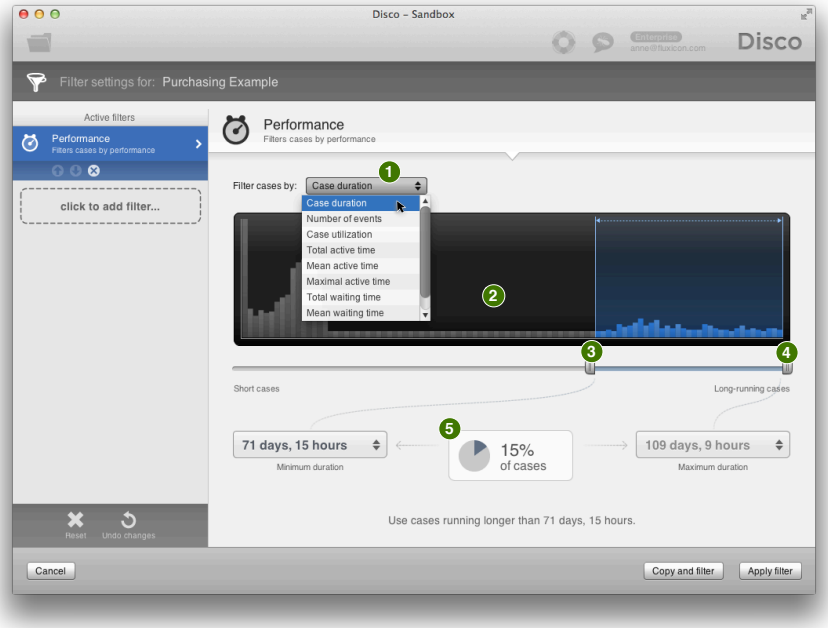


Fig. 5.13. Settings of the Performance filter.

❶ *Performance metric*. First, select the performance metric you are interested in. After you selected the metric, the reference chart ❷ below will update and show the distribution for the metric you have chosen.
The following global metrics are available:

- *Case duration*. Time between the first and the last event in each case. Look up the Statistics view reference in Section 4.2.1 to learn more about what case duration means.
- *Number of events*. The total number of activities that were performed in each case (including repetitions). See also Section 4.2.1.

If you have *start* and *complete* timestamps for your activities in your event log[1], you can also use the Performance filter to look at the active time (time spent in an activity, i.e., between start and end of an activity) and waiting time (inactive time spent between activities, i.e., between the end of an activity and the start of the next activity) in the following way:

- *Case utilization*. Ratio of active vs. waiting time. If the Case utilization is 100%, then there was no waiting time at all. If the Case utilization is 50%,

[1] Refer to Chapter 3 for further details about the import configuration and about how to deal with multiple timestamps.

then half of the total case duration was spent actively on performing activities while the other half was spent idle (i.e., waiting between activities). See also Section 4.2.1.

- *Total active time*. The total active time for the whole case.
- *Mean active time*. The average active time block in the case (active time blocks are separated from each other by idle / waiting times, where no activity was performed at all). See also Section 4.2.1.
- *Maximal active time*. The largest active time block in the case.
- *Total waiting time*. The sum of all idle times for the whole case.
- *Mean waiting time*. The average time for all idle time blocks in a case. See also Section 4.2.1.
- *Maximal waiting time*. The largest chunk of idle time in the case.

❸ *Lower limit*. Determine the minimum value for the performance metric you have chosen by dragging the handle ❸ of the slider to the desired position.
For example, in Figure 5.13 the chosen performance metric is *Case duration* and the label below the selection slider tells you that the current configuration uses all cases that run longer than 71 days and 15 hours.
Instead of using the slider, you can also explicitly set a lower or upper limit by clicking on the minimum or maximum value indicator in the configuration screen as shown in Figure 5.14. This can be useful, for example, if you want to set the boundaries based on specific service level agreements.

Fig. 5.14. You can set an explicit value as the lower or upper limit in the Performance filter.

❹ *Upper limit*. The upper limit determines the maximum value for the performance metric you have chosen.
In the example in Figure 5.13 the upper limit is at its maximum value, and therefore includes *all* cases that run longer than 71 days and 15 hours.

❺ *Coverage feedback*. This pie chart at the bottom gives you an indication of how many cases (compared to the whole log) are included in your current selection. Together with the blue highlighting in the performance chart ❷, this helps you to understand which portion of your data set you are currently focusing on.

5.2.4 The Endpoints Filter

The Endpoints filter allows you to determine what should be the *first* and the *last* event in your process. We'll use an example case (see Figure 5.15) from the purchasing demo event log to show you how this works.

For this particular case, the first event corresponds to the activity *Create Purchase Requisition* and was performed by *Anna Kaufmann*. The last event corresponds to activity *Analyze Request for Quotation* and was performed by *Francois de Perrier*. In total, there are 4 events in this case. You can also see the timestamps for each of the activities, and there could be more attributes related to the case as well. See Chapter 4.3 to learn more about how you can inspect individual cases in Disco.

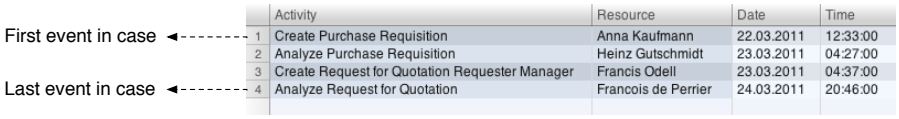


Fig. 5.15. Example case with highlighted start and end events.

The regular end of the purchasing process is *Pay invoice*, so we can see that the case in Figure 5.15 has been stopped earlier—probably because the purchase requisition has not been granted. If we see that many of the submitted purchase orders do not go through, this might hint that people don't know what they are allowed to buy. We might need to offer additional training, or to update the instructions.

To investigate cases based on their start and end event values, you can use the Endpoint filter as shown in Figure 5.16. The following controls are available in the settings screen:

❶ *Event column to be used for filtering*. Most of the time, you want to filter based on the start and end *Activity* in each case. However, you can also use other event columns, such as the *Resource* column (e.g., to keep all cases that were started by Anna Kaufmann) or any other attribute column from your data set.

❷ *Usage mode*. The standard usage mode is *Discard cases*. In this mode, the actual start and end values of the case are used to determine whether the case should be kept or not. In the *Trim cases* mode, you can freely determine the start and end events, which you can read more about later in this section.

❸ *Start event values*. Select the values that you want to allow the cases to start with from the list of available start values.

In Figure 5.16 you can see that all the cases in the purchasing example actually start with the *Create Purchase Requisition* activity. So, there is only one start activity to choose from.

❹ *End event values*. Select the values that you want to allow the cases to end with from the list of available end values.

For example, in Figure 5.16 the end values were configured in such a way that only cases that either end with the *Analyze Purchase Requisition* or with the

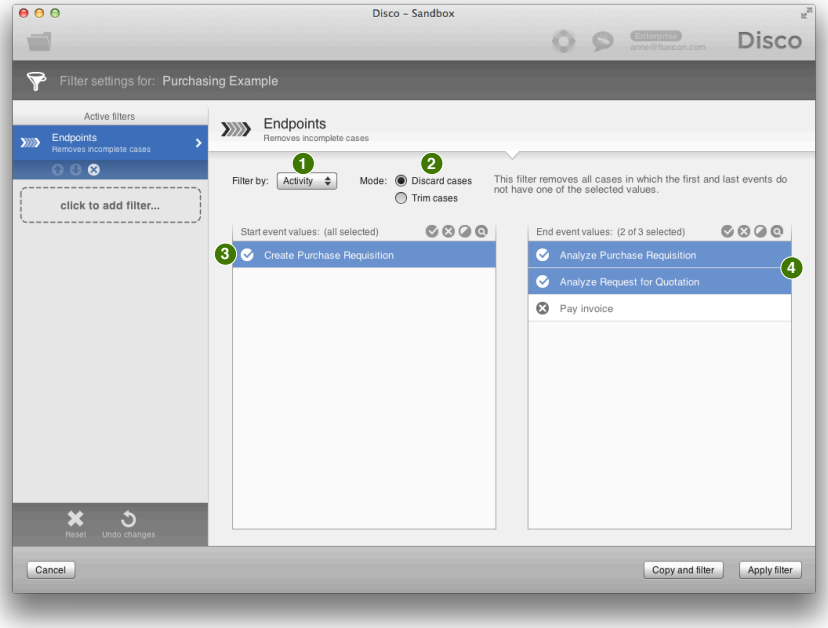


Fig. 5.16. The Endpoints filter in *Discard* mode.

Analyze Request for Quotation activity will be kept. All cases that end with the regular *Pay invoice* activity will be discarded.

Another very common use case for the Endpoints filter in Discard mode is to clean the data set from incomplete cases: In many situations, one gets a data extract of the complete process logging in a particular time frame. So, the event log most likely contains some cases that are incomplete because they were either started before the data extract begins, or they were not yet finished when the data extract stops. So, to clean up your data you should remove those incomplete cases from the log (by selecting only the valid start and end activities in the Endpoints filter).

> **Note:** In *Discard cases* mode, only the values that occur for the first event and for the last event in the case are shown in the list of start and end event values. If you find yourself looking for activities in the middle of the process, you probably want to use the Endpoints filter in *Trim cases* mode (see below).

Trim cases

The *Trim cases* mode can be used to focus your analysis on a part of the process. In this case, the endpoints are used as clipping markers and all events before the first

selected start activity plus all events after the last selected end activity are thrown away during the filtering.

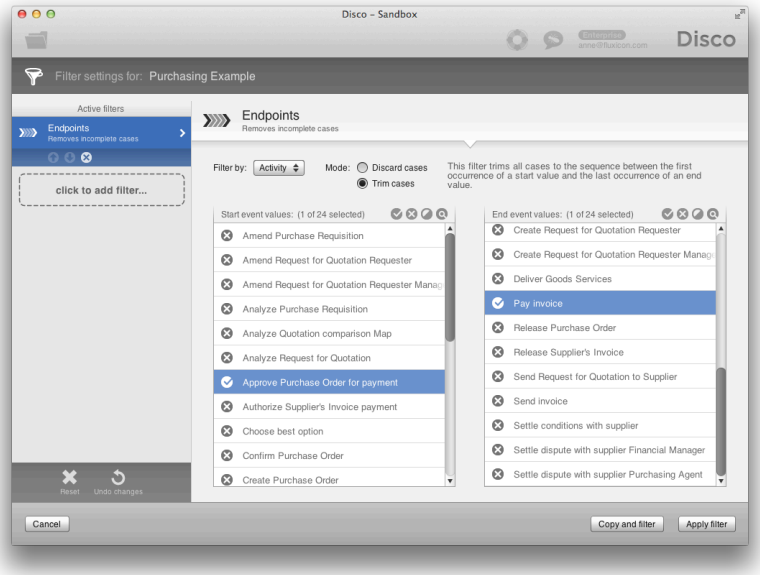


Fig. 5.17. The Endpoints filter in *Trim* mode.

In *Trim cases* mode, you can use any event value as a start or end point to focus your analysis just on this part of the overall process. For example, in Figure 5.17 the *Approve Purchase Order for payment* activity (which happens to be somewhere in the middle of the process) was chosen as a start event and *Pay invoice* was chosen as the end activity.

The effect of the Endpoints filter in *Trim mode* is that in each case all the events before the first start event value (e.g., before the occurrence of activity *Approve Purchase Order for payment*) are removed. Similarly, all events after the last end event value are removed as well (here *Pay invoice* is the last activity in the process anyway).

When you apply the Endpoints filter from Figure 5.17, you can see that the process map is "chopped off" at the indicated endpoints (see Figure 5.18). So, your process, and all statistics like case duration etc., are now just focused on the Approve Purchase Order-to-Pay invoice part.

The trim option can also be used for clean-up purposes if your end activities are not guaranteed to be the last event in the process. For example, sometimes you have a dataset where after a successful completion event there may still be some kind of comment activities, thus making it impossible to use the Discard option for clean-up without removing the comment activities first. Use Trim to directly indicate where your process starts and ends, and it will throw away the rest.

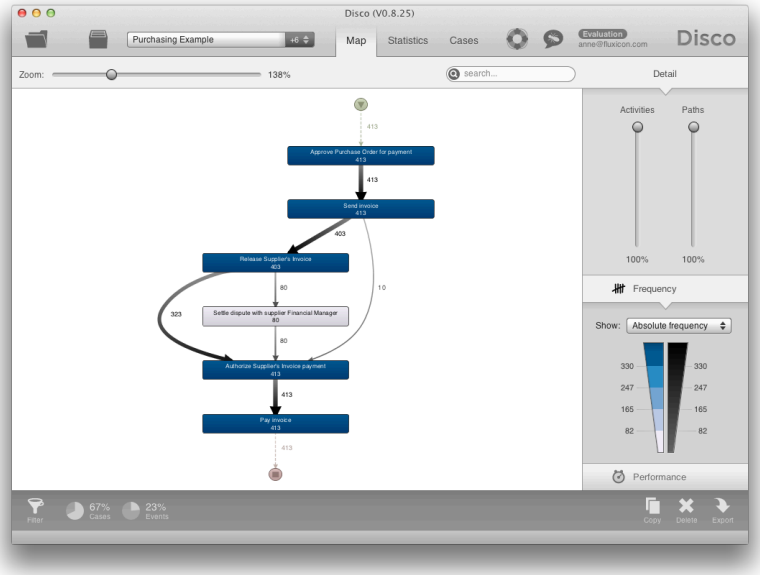


Fig. 5.18. The resulting process map after applying the Endpoints filter in *Trim mode*.

5.2.5 The Attribute Filter

The Attribute filter allows you to filter events or cases based on arbitrary attributes in your data set. Figure 5.19 shows the controls and options that are available in the configuration settings screen.

❶ *Event column to be used*. You can choose to filter based on the *Activity* name, the *Resource* column, or any other attribute column from your data set. Depending on the event column that you choose, all the observed values for that particular column will appear in the list ❸.

❷ *Filtering mode*. The following three filtering modes are available:

- *Keep selected*. Use this filtering mode if you want to *remove events* that do not have one of the selected values (for the Event column ❶ that you have chosen). This can be used for clean-up of unimportant activities, and to focus your analysis.

  For example, in Figure 5.19 the activities *Analyze Quotation comparison Map*, *Choose best option*, and *Create Quotation comparison Map* would be removed from your data set.

- *Mandatory*. Use the mandatory indicator to determine which activities (or resources, countries of origin, product types, etc. - depending on what you have chosen in ❶) *must be present (occur somewhere) to keep a case* in the filtered data set.

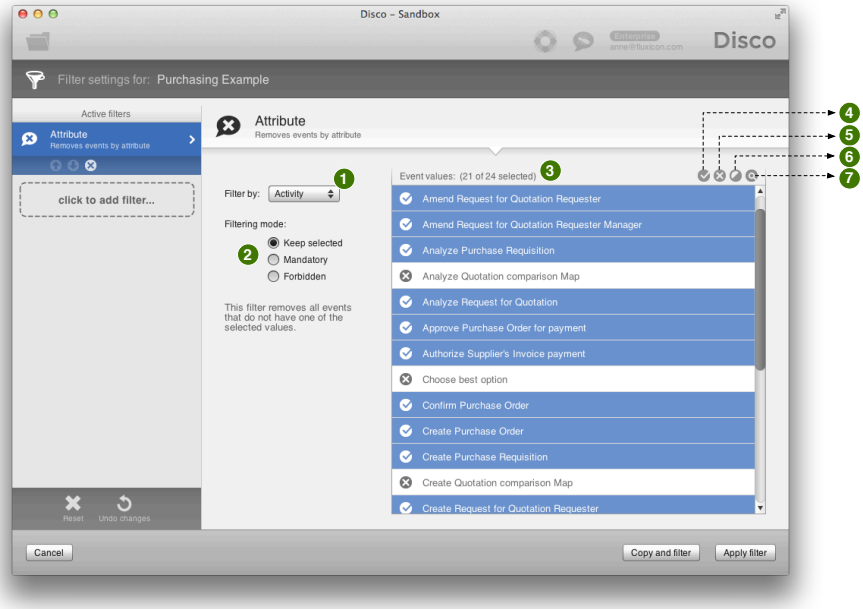


Fig. 5.19. Settings of the Attribute filter.

Often, you want to zero in on just those cases where one particular activity occurred. Therefore, you can use a shortcut to get to this filter directly from the Map view: If you click on the activity in the process map, and then press the *Filter this activity...* button, then you get directly to a pre-configured Mandatory filter. Learn more about this shortcut in the Map view reference Section 4.1.5.

- *Forbidden.* Use the forbidden indicator to determine which activities (or resources or whatever you have chosen in ❶) *must **not** be present to keep a case* in the filtered data set.

  The Forbidden option gives you the inverse result set compared to using the Mandatory option. For example, if you have used the mandatory filtering mode to see the process for all cases where a dispute had to be settled with the supplier, you could use the forbidden filtering mode where no dispute had to be settled with the supplier.

❸ *Event values.* Choose the event values that you want to keep, make mandatory, or make forbidden—depending on the filtering mode ❷ you are using.

❹ *Select all.* Select all values in the list ❸ at once. This can be handy if you work with large lists but only want to deselect a few values from that list.

❺ *Deselect all.* Deselect all values in the list ❸ at once. This can be handy if you only want to select a few values from a large list.

❻ *Invert selection.* Use this button to deselect all currently selected values in list ❸ and vice versa.

❼ *Search.* Reduce the displayed list of values based on specific letter combinations or words that you type in the search field. This is particularly useful if you have hundreds or thousands of different values for an attribute, and finding the value you are looking for manually is cumbersome.

5.2.6 The Follower Filter

While the Attribute filter allows you to select cases based on the occurrence (Mandatory mode) and non-occurrence (Forbidden mode) of activities or other event values, the Follower filter goes one step further: It lets you specify a simple process pattern based on the so-called follower relation.

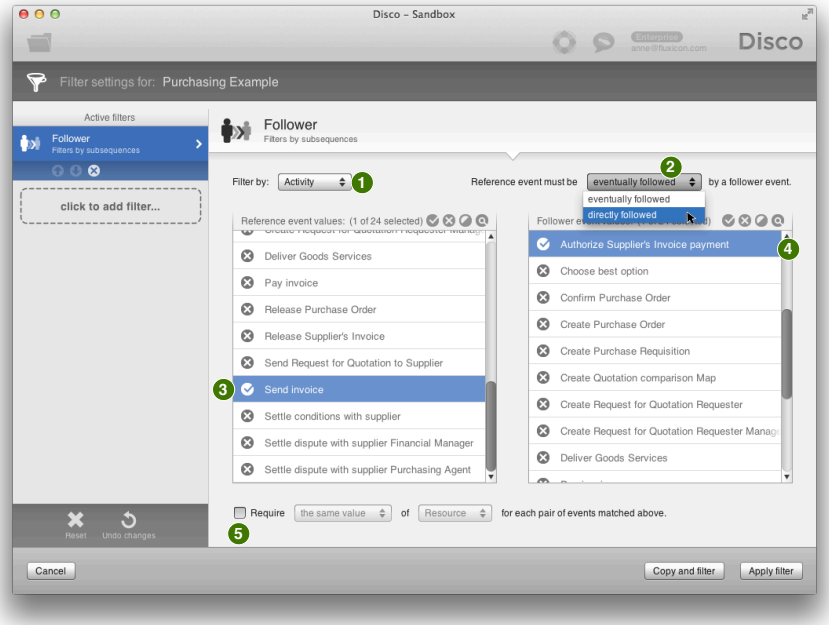


Fig. 5.20. Settings of the Follower filter.

Figure 5.20 shows the main controls that are available in the filter settings:

❶ *Event column to be used.* You can choose to filter based on the *Activity* name, the *Resource* column, or any other attribute column from your data set. Depending on the event column that you choose, all the observed values for that particular column will appear in the list ❸ and ❹.

❷ *Filtering mode*. Two filtering modes are available. Their difference is illustrated in Figure 5.21:

- *Eventually followed*. The eventually follows mode allows you to specify a follower pattern in which a certain activity (or other event value) ❹ must follow the reference event value ❸ *somewhere later* in the same case.

  In Figure 5.21 an example case from the purchasing demo log is shown again, and we assume that *Create Purchase Requisition* has been defined as the reference activity ❸. Then as long as any of the activities *Analyze Purchase Requisition*, *Create Request for Quotation Requester Manager*, or *Analyze Request for Quotation* has been set as a follower activity ❹, this case would remain in the data set after applying the filter in *eventually followed* mode.

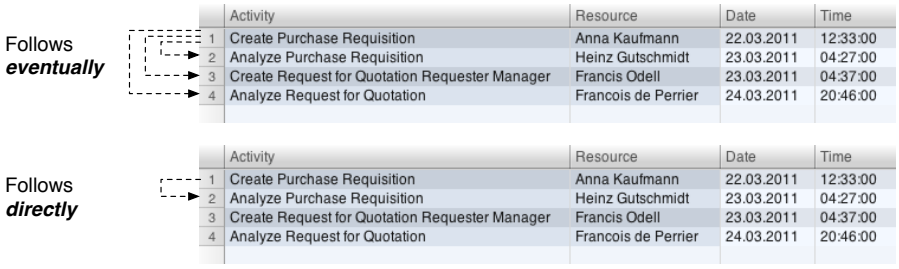


Fig. 5.21. The difference between *directly followed* vs. *eventually followed* lies in whether one only looks at events that occurred directly afterwards or at all following events.

- *Directly followed*. The directly follows mode requires that a certain activity (or other event value) ❹ must follow the reference event value ❸ *directly afterwards* in the same case.

  Again assume that in Figure 5.21 *Create Purchase Requisition* has been defined as the reference activity ❸. Then only if *Analyze Purchase Requisition* has been set as a follower activity ❹, this case would remain in the data set after applying the filter in *directly followed* mode.

Often, you want to zero in on just those cases where one particular path (between activity A and activity B) occurred. Therefore, you can use a shortcut to get to the Follower filter directly from the Map view: If you click on any path in the process map, and then press the *Filter this path...* button, then you get directly to a pre-configured Followers filter. Learn more about this shortcut in the Map view reference Section 4.1.5.

❸ *Reference event values*. The events that provide the reference point for the follows relationship.

❹ *Follower event values*. The events that should (directly or eventually) follow the reference event.

❺ *4-Eyes filter*. If you enable this option, another requirement can be added to the Follower filter based on another dimension. The values for the events in the following relation can be required to be either equal (*the same value*) or different (*different values*) for the chosen event column ❺.

The main use case is to specify a 4-eyes requirement (to ensure so-called segregation of duties) for resources when the Follower specification is based on activities. For example, one could specify that the approval of a travel request cannot be completed by the same person who initially submitted the request.

5.3 Troubleshooting

All filters in the filter stack *refer to the original log when you configure them*, which makes it possible for Disco to efficiently filter also very large data sets. To make your process mining work as explorable, re-usable and efficient as possible, you can quickly change filter settings, combine, move and remove filters, and copy data sets with existing filter stacks as much as you want.

Only once you apply the filter, the actual filtering occurs, in which each filter in the filter stack is applied one after the other. So, the output of the first filter forms the input for the second filter, and so on. See Figure 5.22 for an illustration of how the filter stack works in Disco.

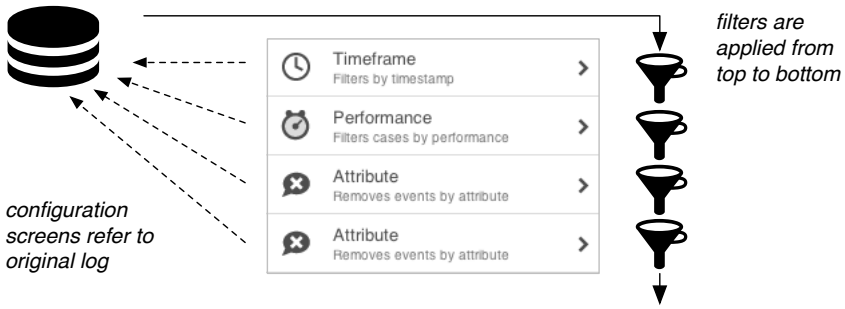


Fig. 5.22. How the filter stack works: Filters can be quickly re-configured and moved around also for large data sets because they refer to the original data set. When the filters are applied, the output of the first filter is the input for the second filter and so on.

While the stacking of multiple filters is powerful and convenient, it can lead to problems in some situations because the combination of filters can add a certain complexity. So, in this troubleshooting section we show you how you can address two problems you might run into.

5.3.1 "I want to see the output of the previous filter"

As shown in Figure 5.22, the configuration screens of all filters in the filter list "see" the *original data set*. Usually, this is no problem because multiple filters are often used to filter orthogonal aspects of the log (timeframe, performance, removing certain activities, etc.).

However, sometimes you want to configure a filter ("see" the configuration settings) *based on the output produced by the filters before*. For example, let's take the Purchasing demo example again and say you added an Attribute filter (see Section 5.2.5) that removes almost all activities except of four activities that you want to focus on. Then, you want to add a Follower filter (see Section 5.2.6) to drill down into specific connections between just these four activities. Normally, the Follower filter configuration screen would show you not only the four activities you are interested in, but all 24 activities from the original log. This makes it difficult for you to find the right activities and you would prefer to see only the four selected ones in the list.

You can resolve this by consolidating your data set in an intermediate step as follows:

1. Add all filters that should be applied before the filter that should "see" the outcome of the previous filters. Then press *Copy and filter* with the *Apply filters permanently* option (see ❶ in Figure 5.23). You can give the new data set a name that reflects this intermediate filtering state, such as "Purchasing activities only" could describe the four activities you wanted focus on in the situation above.
2. Now, you can add the filter that should "see" the outcome of the previous filters to the new data set (see ❷ in Figure 5.23). For example, the Follower filter can now be added to the new "Purchasing activities only" data set. Because the previous filters have been applied permanently, the added filter will now see the output of these filter steps in the configuration screen.

Other situations in which you might want to apply filters permanently are, for example, if you want the percentage indicator in the Performance filter (see Section 5.2.3) to be based on the previous filters, or the variants in the Variation filter (see Section 5.2.2). It is also recommended to apply your filters permanently after cleaning up your log from erroneous timestamps or incomplete cases.

5.3.2 "I am getting an empty log"

Sometimes, an empty log (see Figure 5.8) just tells you that the answer to the question you asked is "no". For example, if you create a filter to find all cases that have been completed this months *and* that took longer than 30 days (to follow up with the customers and give them a present to compensate for their slow service) and the result is empty, then there simply are no such cases in your data set.

However, sometimes you may be getting an empty log and don't know what went wrong. You know that you have made a mistake in the configuration of your filter and need to find it. In this situation, we recommend to track down the problem by looking at your filter stack step-by-step in the following way:

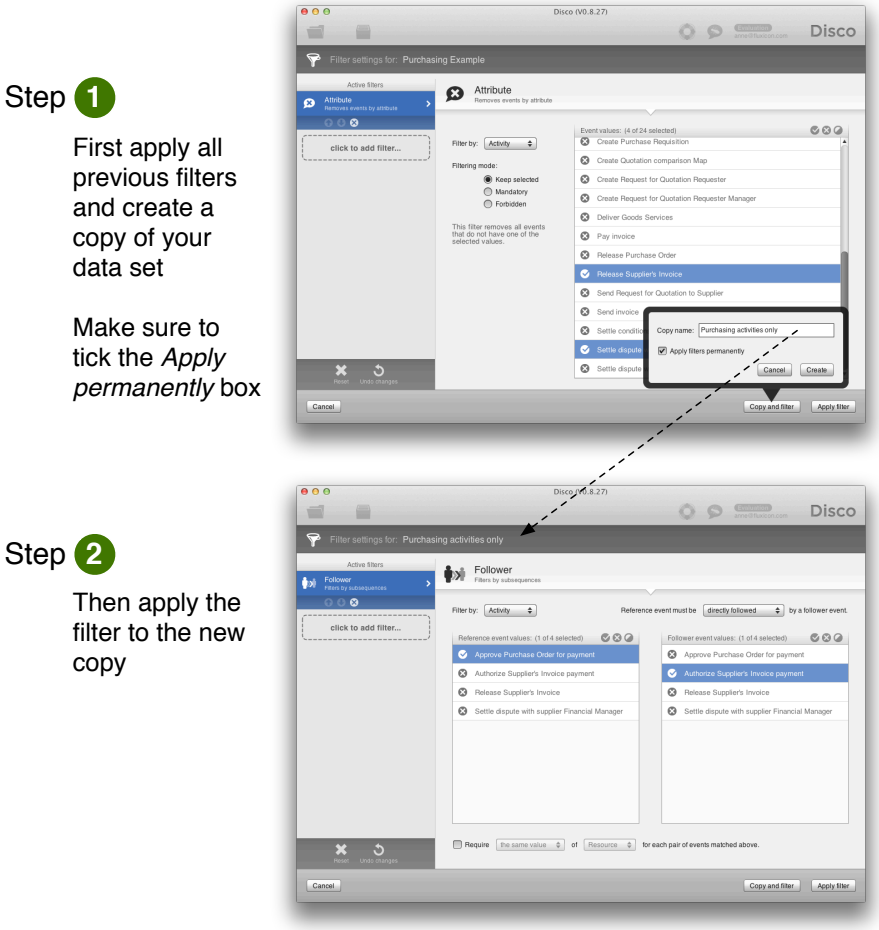


Fig. 5.23. If you want the configuration screen of a filter to "see" the output of the previous filters in your list, then you can achieve this by applying the previous filters permanently before you configure that filter.

1. Make a copy of your current data set for testing the problem (see Section 6.2 if you don't know yet how to copy a data set). You can give the copy a name like "Test", so that you know you can remove it later on.
2. Try each filter in isolation (remove all other filters from the filter stack and just keep the one you are testing). Does it give the expected result? Check the different data set views (Process map, Statistics, Cases view) to inspect the effect this filter has had on your data set. It is likely that you find the problem already in this step.

3. If each of the filters in isolation works just as expected, it's time to look at the combination of your filters: Start by adding just the first and the second filter from your filter list. Does this combination give you the expected result so far? If yes, add and apply the third filter and check the resulting log in the different data set views again. Continue until you find the filter that is not giving you the expected output.
4. If after you found the problematic filter, you still can't see why the filter does not give you the result you expected, make a copy of your test data set by applying all filters *before* this problematic filter using *Copy and filter* and ticking the *Apply filters permanently* option (similar to Figure 5.23). Give the copy a name like "Testing filter X" to remember what you created the copy for.
5. Now add just your problematic filter to the new data set "Testing filter X". Inspect the filter settings. Is anything different in the filter settings after you have applied the previous filters in a permanent manner? Refer to the manual for the corresponding filter in Section 5.2.1–5.2.6 if you are not sure what it does.

6

Managing Data Sets

One of the advantages of Disco is that it supports your project work through the management of multiple data sets in one project view. In a typical process mining project, you will import your log files in different ways, filter them, and make copies to save intermediate results. This results in many different versions and views of your data sets and can easily get out of hand.

The project view in Disco is there to help you keep an overview. It keeps all your work in one place and lets you easily share it with others. In this chapter you find all the details about how projects can be managed in Disco.

6.1 The Project View

A project is symbolized by the filing cabinet icon shown in Figure 6.1.

Fig. 6.1. Project view symbol in Disco.

You can find the project symbol in the upper left corner right next to the open symbol as shown in Figure 6.2. The project view can be reached from anywhere in Disco. Clicking on the project symbol will bring you to the project view (see Figure 6.3). The project view contains the following elements:

❶ *Project name*. You can rename your project as described in Section 6.1.2.

❷ *Data sets*. Each event log that is imported will be placed as a separate data set in your project view. Refer to Chapter 3 for detailed instructions on how to import your data into Disco. When you make copies of your data sets, then these copies

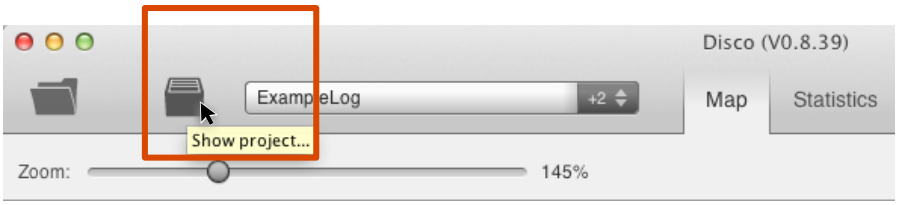


Fig. 6.2. You can return to your project view from anywhere in Disco.

will also appear in your list of data sets here. Learn more about how to make copies in Section 6.2. The main area of the screen (❸–❽) holds information and controls for the currently selected data set.

❸ *Overview and springboard*. Overview information about the currently selected data set is shown here. You see a thumbnail preview of the process map (Chapter 4.1), overview statistics (Chapter 4.2), and cases (Chapter 4.3). This overview information helps you to quickly identify the right data set and also serves as a springboard to jump right into the respective data set analysis view. Section 6.1.1 shows you how this works.

❹ *Reload data*. Pressing the Reload data button brings you back to the configuration screen of your imported file. It provides a quick way to go back and include that attribute you forgot, or to simply look at the source data again. The Import reference chapter explains in more detail how the Reload button works in Section 3.1.7.

❺ *View details*. The View details button brings you to the detailed analysis view (Map, Statistics, or Cases) that you last viewed for the currently selected data set. You can always go to the project view to get an overview, and the View details button brings you back precisely to where you were before. Read also Section 6.1.1 about how to navigate from the project view.

❻ *Notes field*. The notes field is great to remember observations and findings from the analysis as well as for keeping track of ideas and questions that are still open (ToDo items). When you plan to share your analysis results with others (see also Section 6.3), the notes field can hold guiding descriptions for your colleagues about which results can be found in which data set.

❼ *Filtering*. The log filter controls for each data set can be accessed from the detailed analysis views as well as from the project view here. Filters are an important instrument to clean your data and focus your analysis. Read Chapter 5 for detailed information on how filtering works in Disco.

❽ *Copy, Remove, and Export data sets*. Data sets can be copied and deleted in the project view. Read Section 6.2 for further details. The export functionality of Disco is explained in detail in the Export reference in Chapter 7.

❾ *Clear and export project*. If you want to start a new project or export your current work, you can do that from the project view in the upper right corner. Read Section 6.3 for further details on exporting and importing projects.

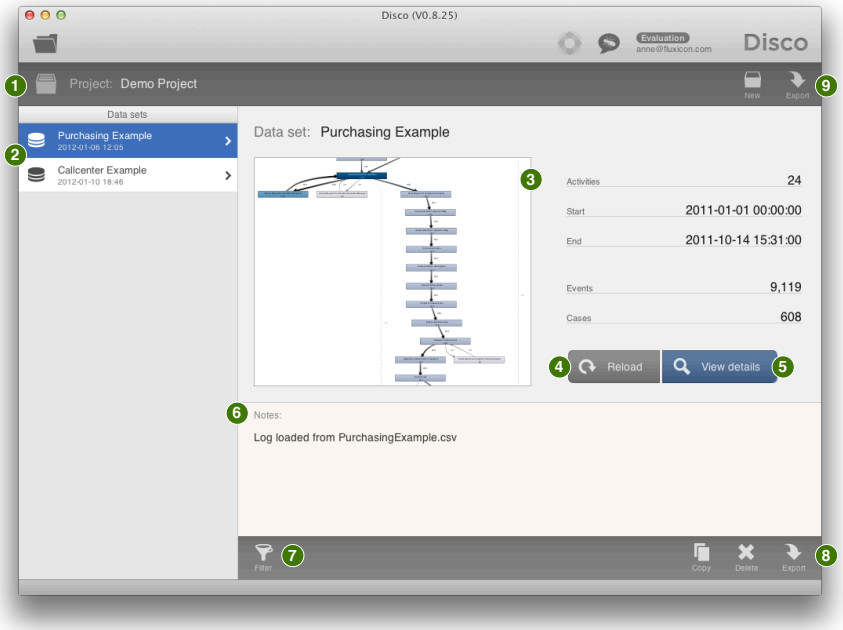


Fig. 6.3. The project view in Disco.

6.1.1 Navigating From the Project View to the Analysis Screens

The project view helps you to keep an overview about the data sets you are working with. Once you want to proceed with the analysis of your process, you can then jump into the detailed analysis views right from your overview information panel (❸ in Figure 6.3). This works as follows:

- If you want to see the process map (Chapter 4.1), simply move your mouse over the thumbnail preview of the process map in the in your project view. A little speech bubble with the text *Click to view map* appears (see Figure 6.4). Once you click in the thumbnail area, you will be directly brought to the Map view for your data set.
- If you want to see the statistics (Chapter 4.2), move your mouse in the overview area until the speech bubble says *Click to view statistics* (see Figure 6.5). One mouse click, and you are in the Statistics view.
- If you want to see the individual cases (Chapter 4.3), move your mouse in the overview area until the speech bubble says *Click to view cases* (see Figure 6.6). Once you click, you will be brought to the Cases view.

If you just want to go back to where you were before in the analysis of your data set, you can press the *View details* button (see ❺ in Figure 6.3). This will bring you to any of the three views (map, statistics, or cases) that you had last explored.

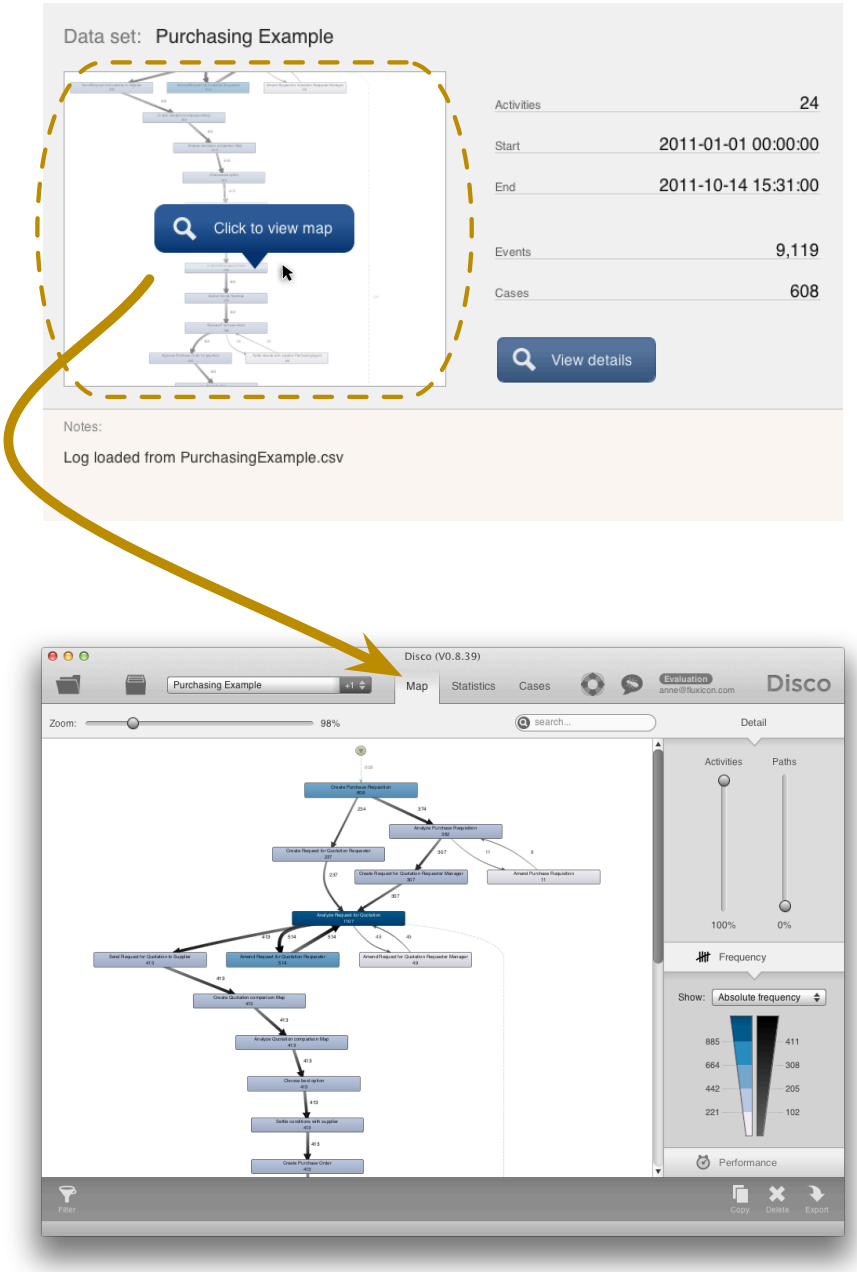


Fig. 6.4. Clicking on the map preview in the project view brings you directly to the process map for your data set.

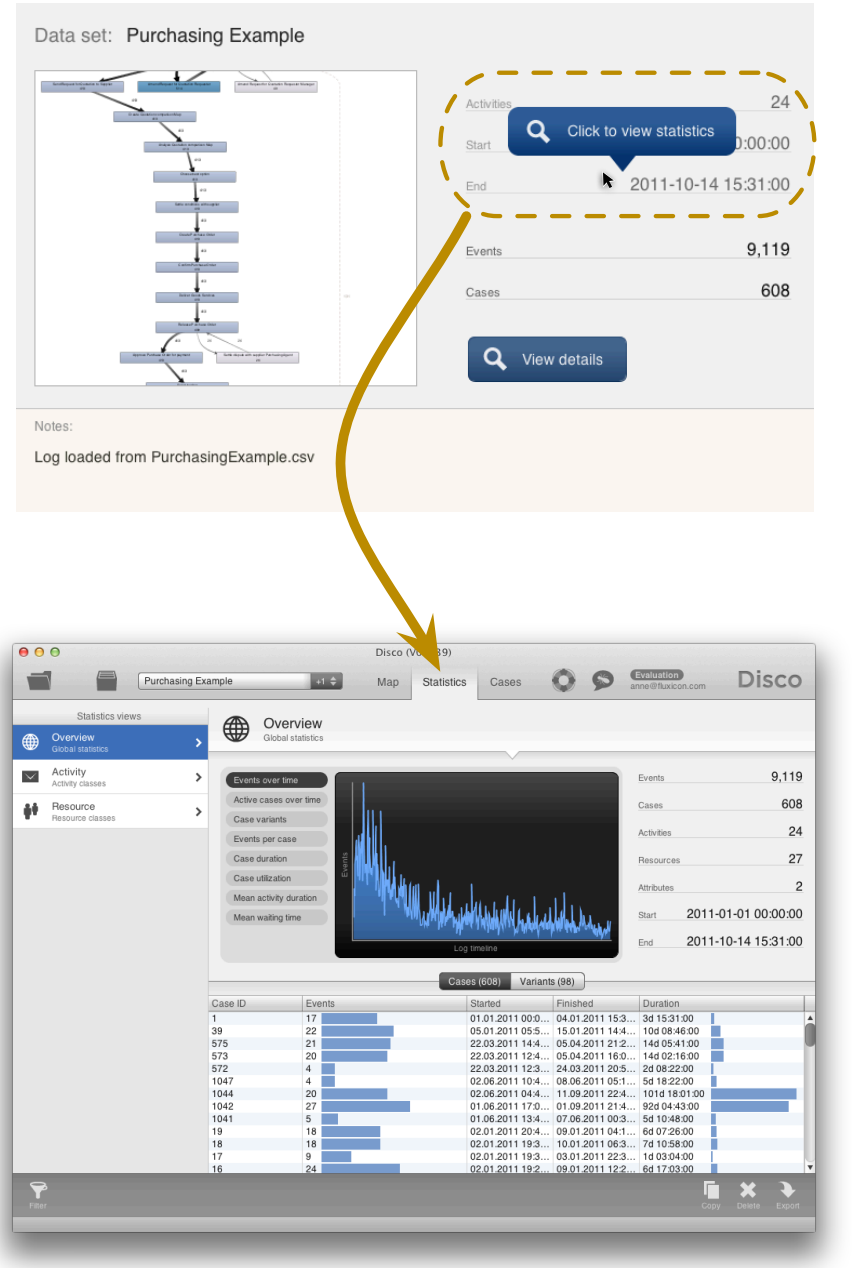


Fig. 6.5. Clicking on the statistics overview in the project view brings you to the detailed statistics for your data set.

Fig. 6.6. Clicking on the case overview in the project view brings you to the cases of your data set.

6.1.2 Renaming Projects and Data Sets

You can rename projects and individual data sets to give them names that better reflect what your project is about.

To rename your project, simply click on the current project name (❶ in Figure 6.3). When you click on the project name, a text field appears and you can type in your new project name (see Figure 6.7).

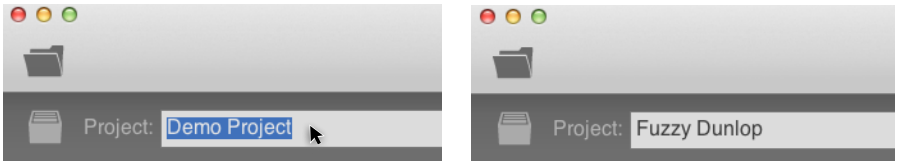


Fig. 6.7. To rename your project simply click on the current project name and type in your new project name.

Renaming data sets works in the same way. First, select the data set for which you want to change the name from the list of data sets in your project (❷ in Figure 6.3). Then, click on the name of the data set in the overview panel as shown in Figure 6.8 to change it.

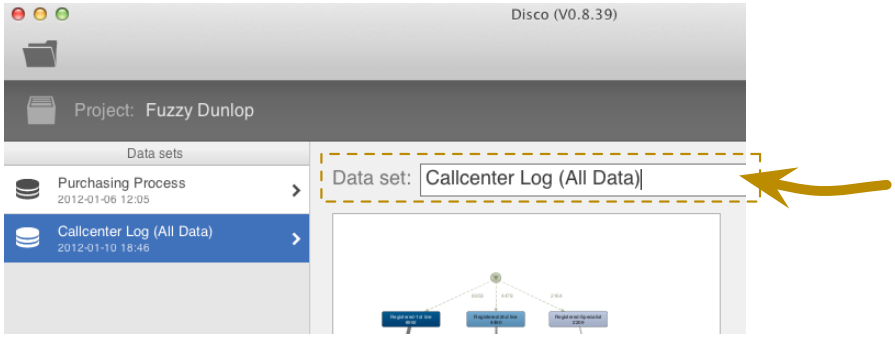


Fig. 6.8. Select the data set you want to rename and click on the current name of the data set to change it.

6.2 Copying and Deleting Data Sets

Copies of data sets are particularly useful to keep "bookmarks" of differently filtered versions of your event log. They provide you with a means to preserve a specific view on your process, store analysis results, and they allow you to easily compare different parts of your process.

You can make copies while you are filtering. This is useful when you are exploring your process through filtering and—along the way—decide that you want to preserve what you did so far and apply the new filter to a copy rather than the current data set. The filter reference in Chapter 5 explains in detail how to do that.

Alternatively, you can make copies right from the project view. This is useful if you already know what you want to do. In both situations, the newly created copy is placed in your project as a new data set. The effect is the same, but the workflow in Disco is different. The following example scenario explains how making copies in the project view works.

6.2.1 Copy Scenario

Let us say that you want to compare how your call center process looks like for service requests that come in through email and those that are initiated on the phone. You know that you have an attribute called *Medium* in your event log that indicates through which channel the request was created. You can use this attribute to create differently filtered copies for 'Mail' and 'Phone' in the following way:

Step 1: You start by taking the complete call center data as a reference point, and you make a copy by pressing the *Copy* button on the lower right of the project view screen. A dialog appears in which you can provide a custom name for the new data set. We give it the name 'Callcenter – Only Mail' because we intend to use this copy as a bookmark for the cases that were initiated by email. After you press *Create*, the copy is placed in the list of data sets. Figure 6.9 visualizes this step.

Although we have named the new copy 'Callcenter – Only Mail', at this point it is still an exact copy of the complete data set. To actually focus the new data set on email requests only, we can add a filter by pressing the filter symbol in the lower left of the data set window (❼ in Figure 6.3).

Step 2: Figure 6.10 illustrates this step. We add an Attribute filter, select the *Medium* attribute, and choose to only keep events with the value 'Mail'. Refer to Chapter 5 for details on the individual log filters that are available in Disco. After the filter is applied, the new copy actually contains only email requests. In Figure 6.10 you can see that the filtered subset contains 35% of the cases compared to the complete call center log.

Step 3: Then, the second copy 'Callcenter – Only Phone' is created. Figure 6.11 shows this step.

During copying, all currently applied filters are copied along with the data set as they are. So, when we use the 'Callcenter – Only Mail' data set as the reference point for our copy, then, again, the newly created data set is already called 'Callcenter – Only Phone' but at this point still contains just email requests.

Step 4: To change the copied filter, you can press the filter symbol and adapt the filter settings as shown in Figure 6.12. After the filter settings have been changed, the new copy contains now only phone requests as desired.

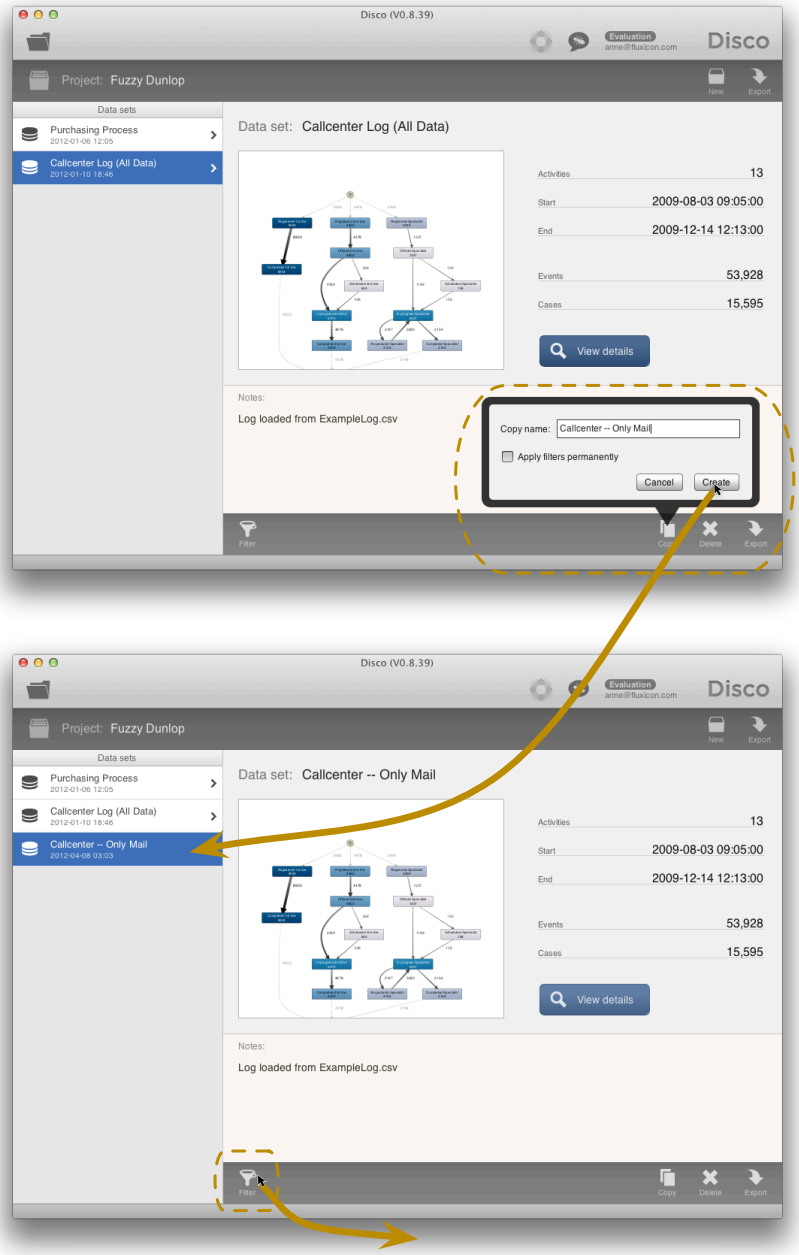


Fig. 6.9. Example Scenario – Step 1: Make a copy of the complete data set.

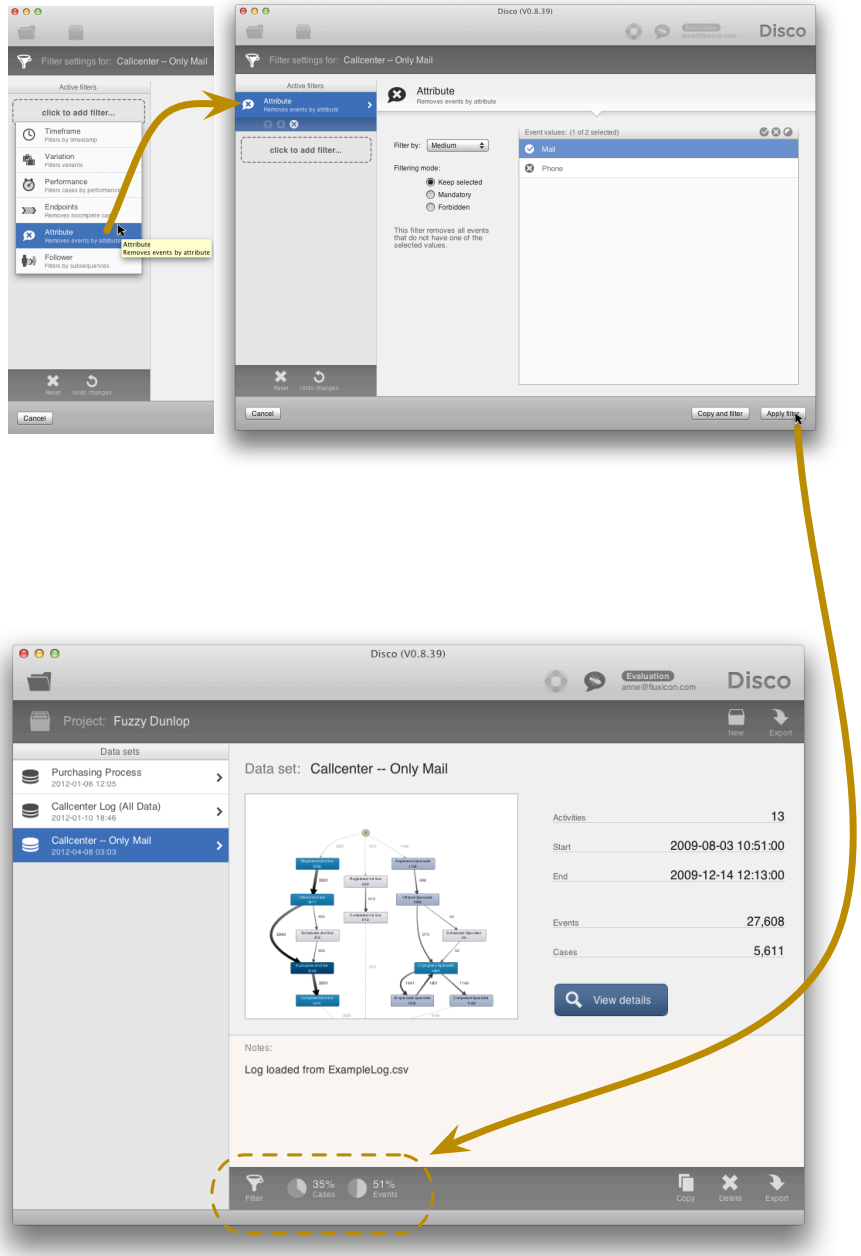


Fig. 6.10. Example Scenario – Step 2: Add a filter to focus on email requests.

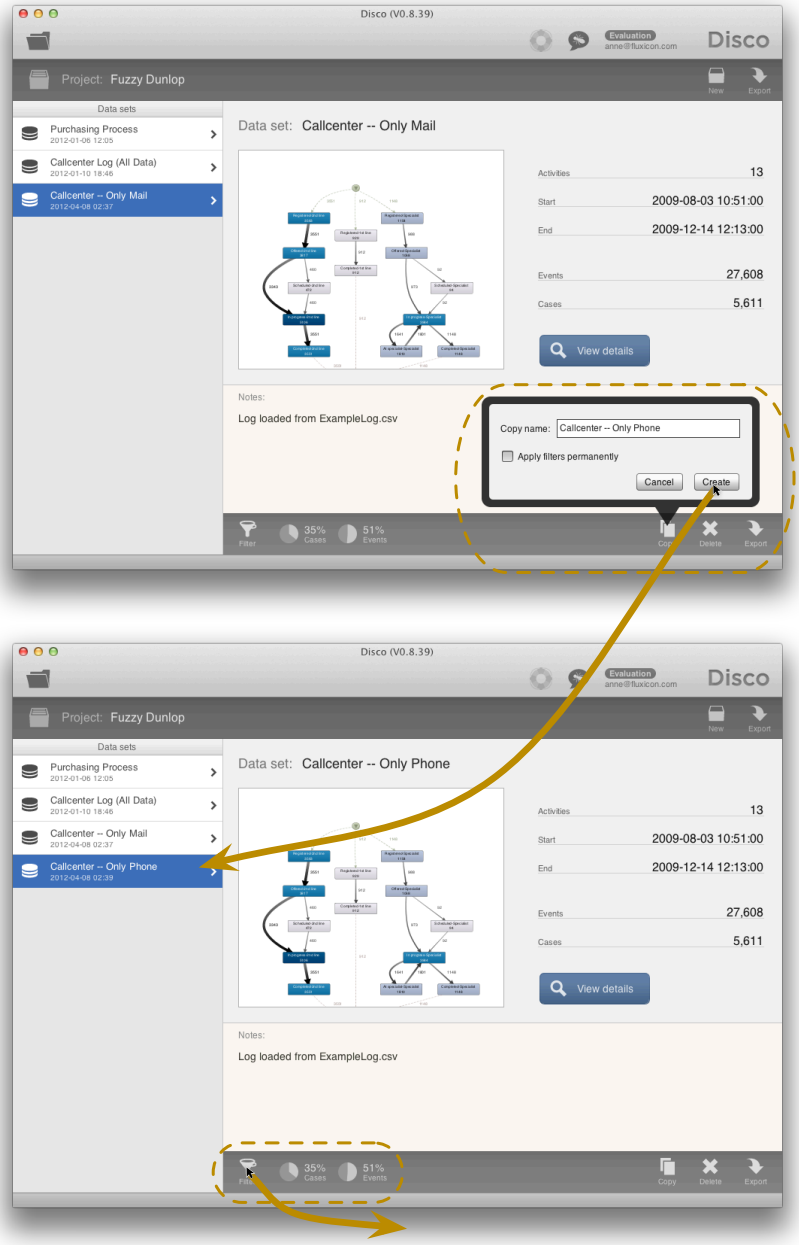


Fig. 6.11. Example Scenario – Step 3: Make another copy for the phone requests.

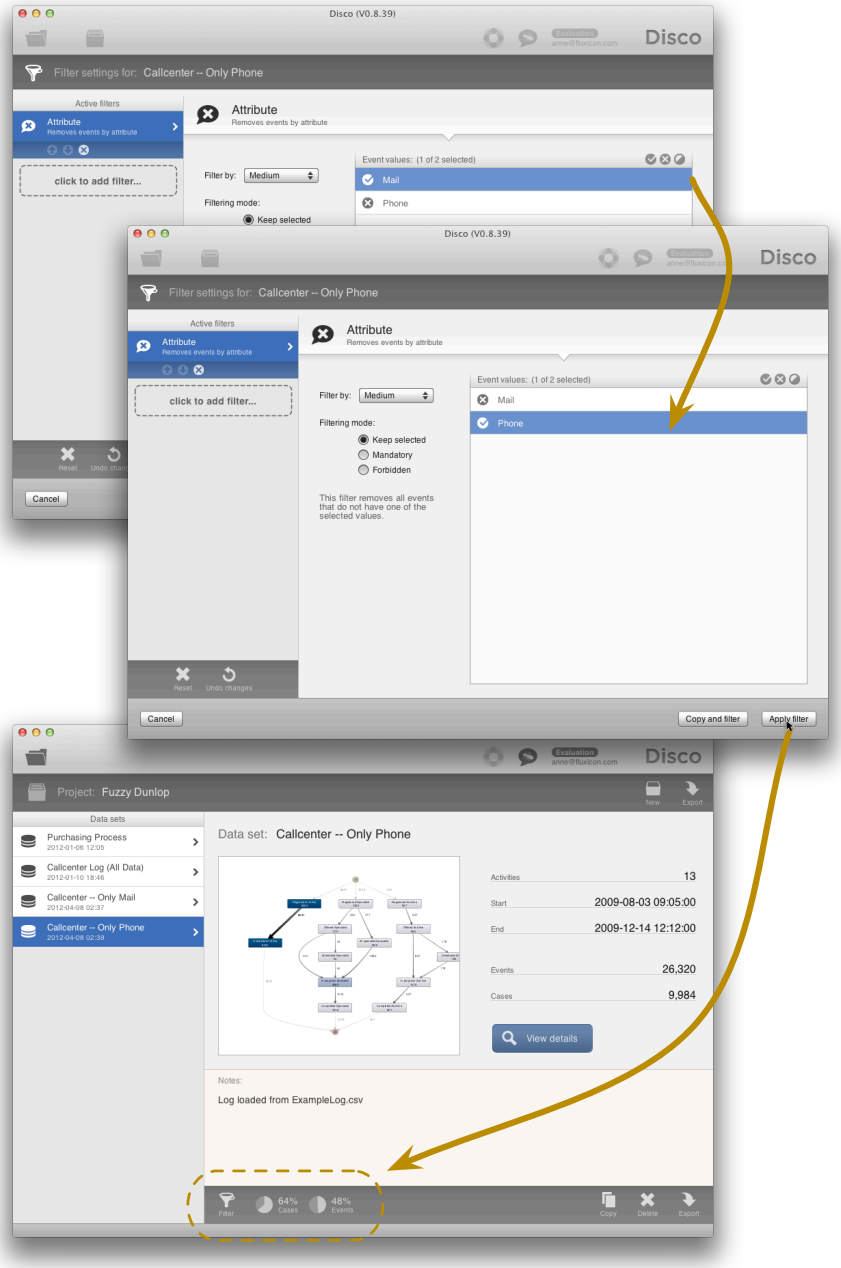


Fig. 6.12. Example Scenario – Step 4: Change the filter to focus on phone requests.

All imported data sets and all created copies can be accessed through the project view as explained in Section 6.1.1. In addition, the data sets are also available through a quick switch list—see Figure 6.13—from within the detailed analysis views (Chapter 4).

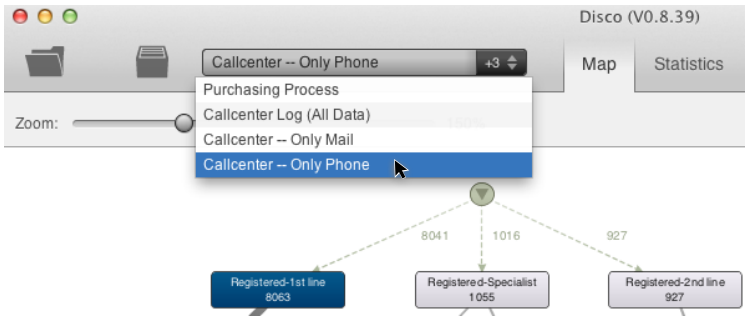


Fig. 6.13. Along with the imported data sets, copies are available through the quick switch list for rapid access from the analysis views.

This way, you can rapidly change and compare different data sets and "jump" to bookmarks in your analysis.

6.2.2 Copying Filters vs. Permanently Applying filters

During normal copying, all currently applied filters are copied along with the data set but can be changed afterwards as shown in Step 4 in the example scenario in Section 6.2.1. Sometimes, however, you want to make a clean copy and permanently apply your filters to a data set. Here is an example scenario where this is relevant.

Imagine that you found out that the purchasing process data you are currently analyzing contains incomplete cases. You want to remove these incomplete cases to not let them disturb your throughput time measurements, which should only be based on the completed instances (not those that are are still running and might have started just recently).

- To remove these incomplete cases, you use the Endpoint filter (Section 5.2.4) and keep only those cases that end with the regular 'Pay invoice' end activity. The result is a filtered data set with 413 completed cases out of the initial 608 cases (67%) as shown in Figure 6.14.

Then you decide that these completed cases should be the new baseline for your further analysis. For example, you want to apply the Performance filter (Section 5.2.3) to verify service level targets and have the filter results reflect the right percentages (like, for example, "15% of the cases do not meet the agreed-upon service level"). The incomplete cases would be in the way.

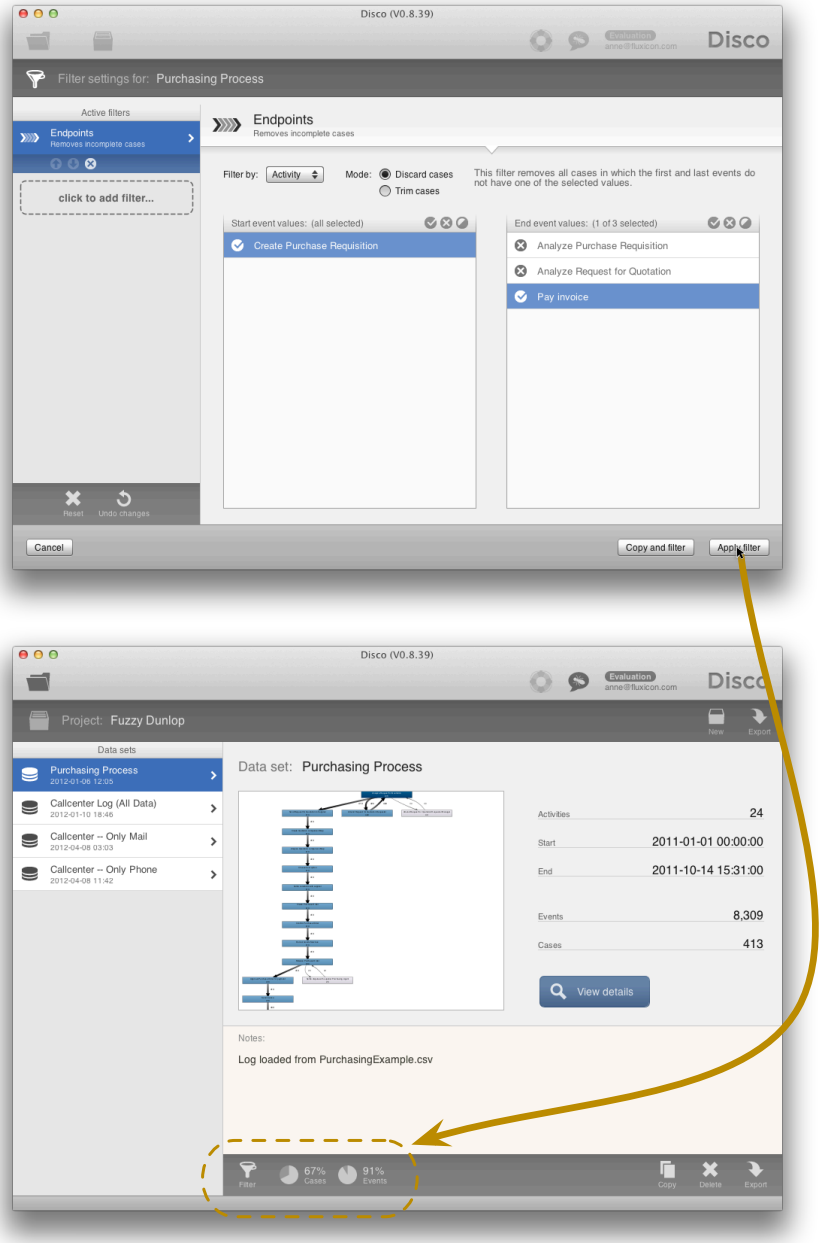


Fig. 6.14. Example Scenario – Step 1: Add filter to remove incomplete cases.

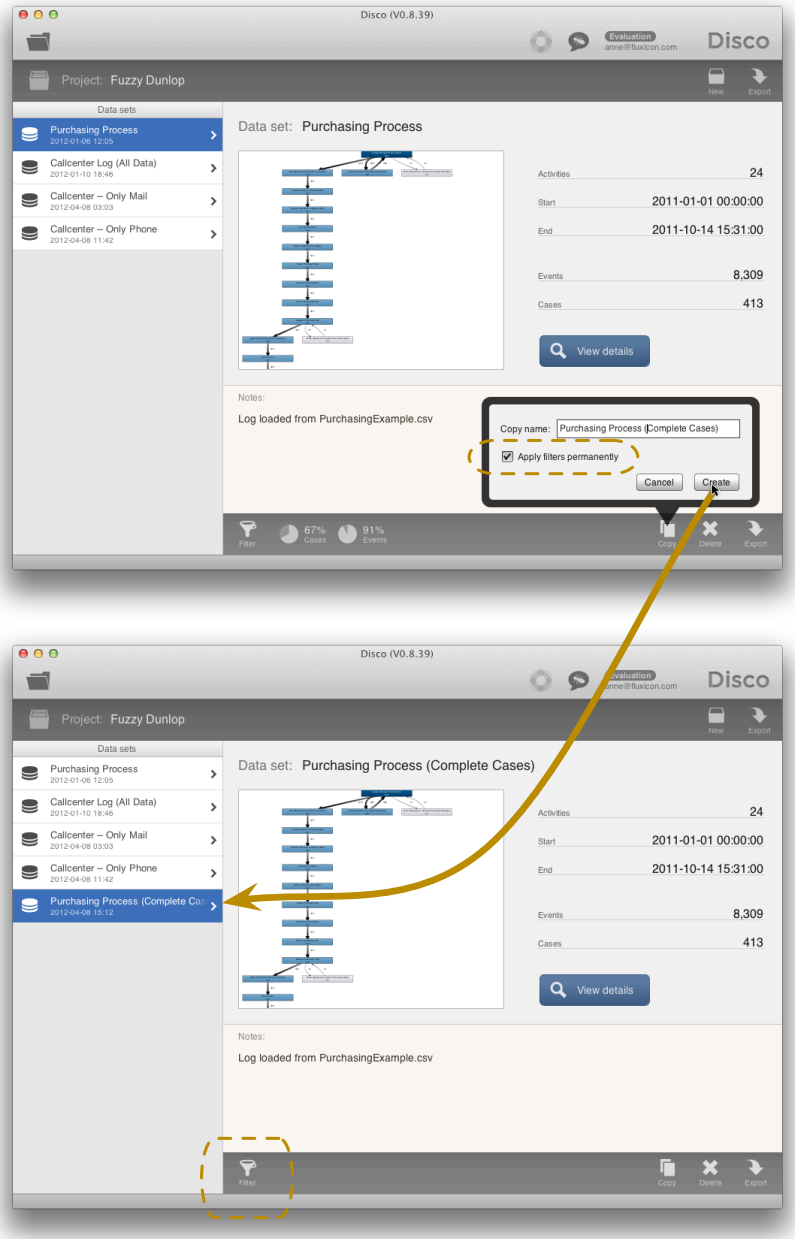


Fig. 6.15. Example Scenario – Step 2: Make copy with *Apply filters permanently* option selected to take the filtered output as a new reference point for the copied data set.

- So, you make a copy of the filtered data set as before but tick the *Apply filters permanently* box before you press the *Create* button as shown in Figure 6.15. The copied data set still contains these 413 completed cases, but the Endpoint filter from before has been applied and "consolidated". It cannot be changed or removed anymore.
- As a result, the filter portion indication in the lower left has disappeared: The previously filtered 67% are now the new 100% for the data set 'Purchasing Process (Complete Cases)'.

In a similar way, filters can also be applied permanently when you create a data set copy for the current configuration of the filter settings. Refer to the filter reference in Section 5.1.3 and 5.3.1 for further details on how and when to create (permanently) filtered copies of your data sets.

6.2.3 Deleting Data Sets

The currently selected data set can be deleted by pressing the *Delete* button in the lower right as shown in Figure 6.16. Be careful when you do this. Deleting data sets cannot be undone.

6.3 Managing and Sharing Projects

You can only have one active project in Disco at the same time. However, you can create multiple projects and work on one of them at a time. If you analyze multiple processes, it is advisable to keep only related data sets together in the same project file to stay organized.

Projects are also useful for sharing your work with other people. By sending them your project file, you will be able to directly share your complete project; including all the data sets, filters, and the notes that you made.

6.3.1 Exporting Projects

You can export your current project from the project view (see also ❾ in Figure 6.3) by clicking on the *Export* button shown in Figure 6.17(a).

A file dialog will appear and you can choose where you want to save your project. Projects are exported as a Disco project file with the `.dsc` file extension. Project files are self-contained. This means that you can just send the `.dsc` file (without sending the original log files) to a colleague and she will be able to import your project in exactly the same state it was when you exported it (see next section).

6.3.2 Importing Projects

Importing projects works in the same way as importing event logs. To load the project, you click the open symbol in the upper left corner as shown in Figure 6.17(b) and locate the project file you want to import.

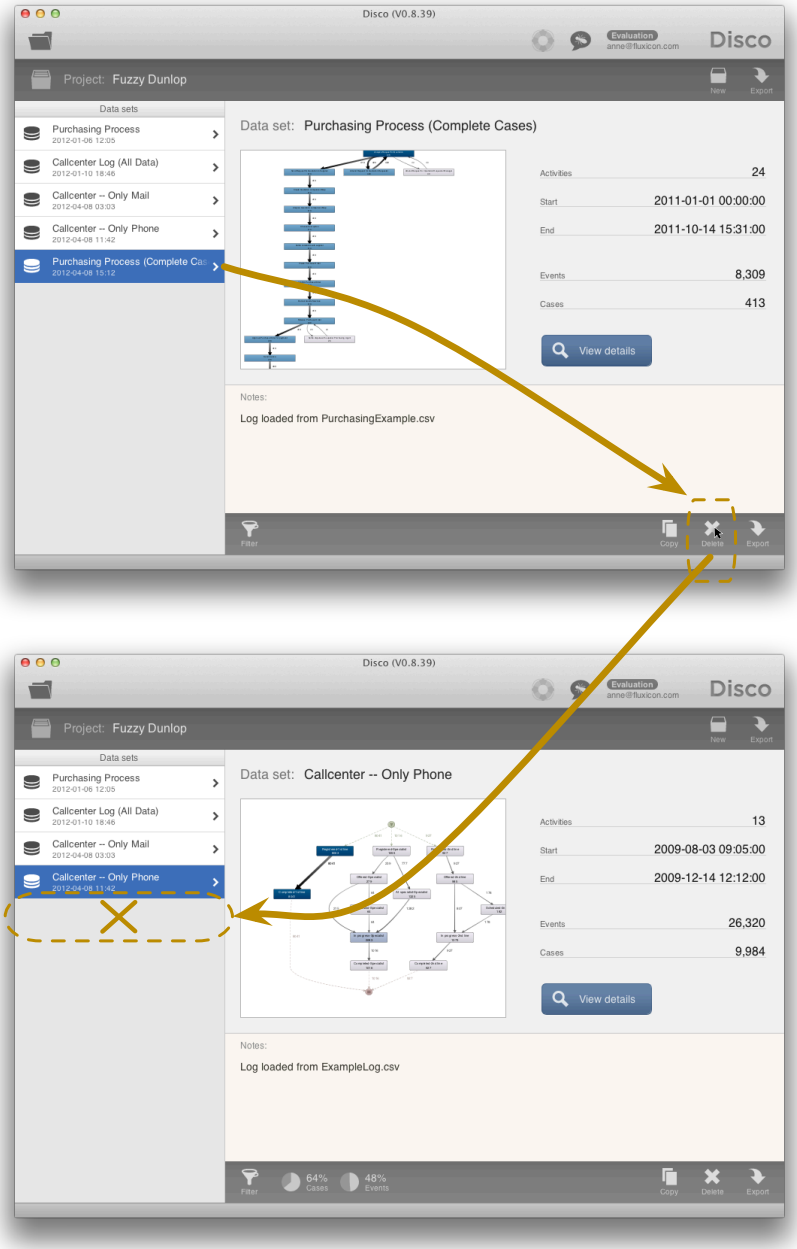
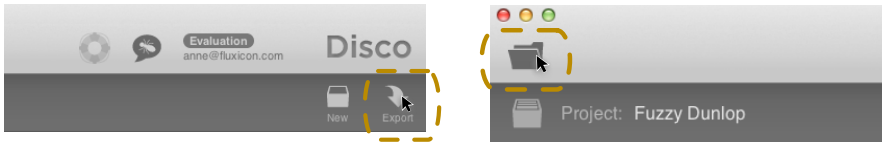


Fig. 6.16. The currently selected data set can be deleted by pressing the *Delete* button in the lower right.



(a) Exporting your current project. (b) Opening an existing project.

Fig. 6.17. Exporting and importing projects in Disco.

When Disco realizes that you want to import a project file (not just another data set), it will ask you whether you want to save your current project first (see Figure 6.18).

If you do not want to lose your current work, make sure to save your current project before you complete the import of the project file. You can do this directly from the dialog shown in Figure 6.18:

- *Save project first*. Press this button to save your current work. Disco will bring up a file dialog that lets you choose the location and name for the exported project. It then first exports your current process to the chosen location and opens the imported project right afterwards.
- *Discard changes*. Press this button if you don't have anything in your current workspace that you feel is worth saving. Disco will directly load the imported project and not save your current workspace.
- *Cancel*. Press this button if you do not want to complete the import of the new project file at all. Disco will bring you back to your current workspace.

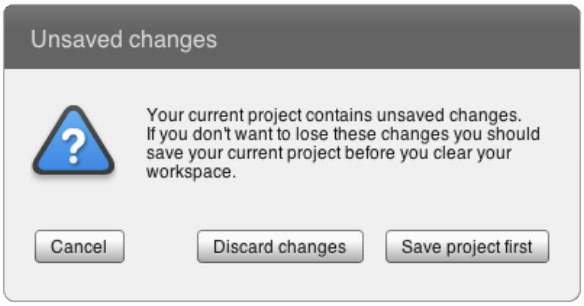


Fig. 6.18. Disco reminds you to save your current workspace before you import or create a new one.

In the Import reference in Chapter 3 you can find further details about the types of files that can be opened with Disco.

6.3.3 Creating New Projects

If you simply want to start a new project, for example, to make a new start for another process that you want to analyze, then you can press the *New* button as shown in Figure 6.19.

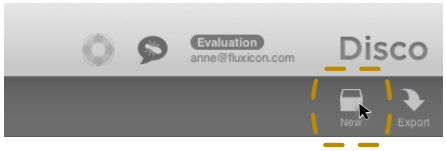


Fig. 6.19. Creating a new project.

Again, you will be asked whether you want to save your current project before clearing your workspace (see Figure 6.18). Press *Save project first* to save your current work before a new, empty project is created.

7

Export

Exports are important to share your analysis results with other people and to further process your data with other analysis tools. In Disco, exports are symbolized by the arrow symbol shown in Figure 7.1.

Fig. 7.1. Export symbol in Disco.

In this chapter, you will learn how to export process maps (Section 7.1), event logs (Section 7.2), charts and tables (Section 7.3), and complete projects (Section 7.4).

7.1 Exporting Process Maps

While it is useful to show the process map to a colleague or client right in Disco because of the interactive filtering and animation possibilities, sometimes you just want to share a picture of the discovered process flows, or copy it into a report or presentation.

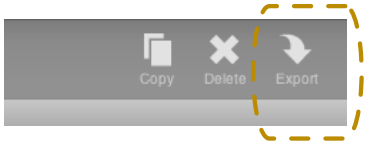


Fig. 7.2. The export symbol in the lower right corner of the screen.

You get to the map export by clicking the export symbol in the lower right corner of either the Project view (Section 6) or any of the three Analysis views (Section 4) as shown in Figure 7.2.

For any data set, the current map view—including the currently displayed metric (see Section 4.1.4) and simplification level (see Section 4.1.2)—can be exported in one of the following formats as shown in Figure 7.3.

- **PDF**. Exporting process maps in PDF format is usually the best choice because PDF is a vector format. This means that you can enlarge the process map (for example, to print it out on a large paper to discuss it in a meeting) without loss of quality.
- **PNG**. PNG is a common pixel-based format that can be used in situations, where PDFs are not supported.
- **JPG**. JPEG is also a common pixel-based format and can be used as an alternative to PNG.

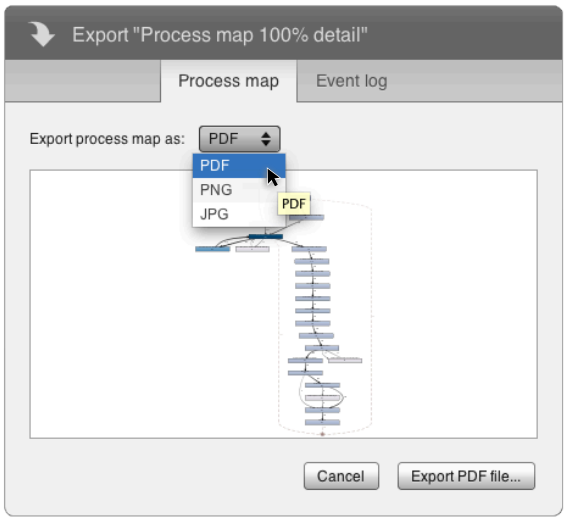


Fig. 7.3. Process maps can be exported in PDF, PNG, and JPEG format.

7.2 Exporting Event Logs

There are several usage scenarios for exporting data sets from Disco:

- Storing an analysis result, such as a list of all cases that you have found to deviate from an important business rule.

- Analyzing the log data further in, for example, other statistics or query tools.
- Exporting a filtered data set to re-import it in Disco with a different view. See Section 3.1.5 for how you can take multiple perspectives on the same data set.
- Saving a specific data set from your analysis for a colleague (or yourself as a backup), so that she can start off where you left rather than re-doing all the filtering steps herself. See also Section 7.4 for how to export complete projects in Disco.

Which export format is suitable for which situation is explained in the Section 7.2.1. Section 7.6 explains the Add endpoints option. Read Section 7.2.3 if you need to anonymize your data to meet privacy regulations in your organization.

7.2.1 Export Types

Like for the export of process maps you get to the event log export by clicking the export symbol in the lower right corner of either the Project view (Section 6) or any of the three Analysis views (Section 4) as shown in Figure 7.2. You then change from the Process map to the Event log export tab as shown in Figure 7.4.

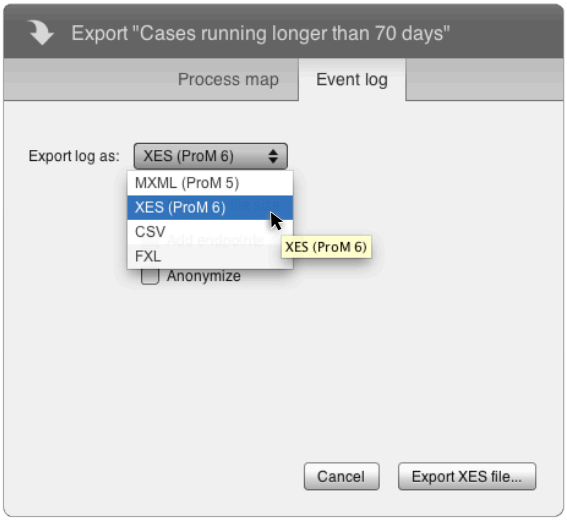


Fig. 7.4. Event logs can be exported in MXML, XES, CSV, and FXL format.

The following export formats for data sets are available in Disco:

- **MXML**. The Mining XML (MXML) format[1] was the first standard format defined for event logs. MXML is supported by the popular academic process min-

[1] The XML Schema definition for the MXML format can be found at the following URL `http://www.processmining.org/WorkflowLog.xsd`.

ing toolsets ProM 5[2] and ProM 6[3], which offer you access to the latest process mining techniques from researchers all over the world.
Even if you do not plan to work with ProM or other process mining tools, exporting your logs in MXML can be useful to save the import configuration step in Disco. Read Section 3.2 for more information on how to take advantage of pre-configured data sets.

- **XES**. The eXtensible Event Stream (XES) format[4] is the successor format of MXML and has been approved by the IEEE Task Force on Process Mining[5].
  Christian from Fluxicon has been the main architect of the standard and also provided the reference implementation and libraries that are currently in use. XES is supported by ProM 6 but not by the older version ProM 5.
- **CSV**. The Comma Separated Values (CSV) format is a good way to exchange data sets with other people, because it can be opened in spreadsheet programs like Microsoft Excel, it can be imported into a database, or loaded into other statistics or query tools for further analysis if needed.
  CSV is just the plain data in columns and rows in the same way as you have probably imported your data in the first place (see also Section 3.1.1). Because CSV—unlike the event log-specific formats MXML and XES—is not XML-based, it is less verbose and easier to read for humans.

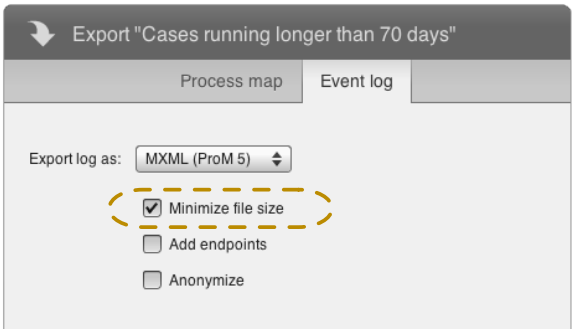


Fig. 7.5. Event logs in MXML, XES, and CSV format can be compressed to reduce the size of the exported file.

MXML, XES, and CSV files can be compressed by ticking the *Minimize file size* option as shown in Figure 7.5. Especially for larger logs it is recommended to use the compression option because the exported files will be much smaller. For MXML logs this will result in files with the ending `.mxml.gz` and for XES in `.xes.gz` files. CSV files will be wrapped in a `.zip` file.

[2] Download ProM 5 from `http://www.promtools.org/prom5/`.
[3] Download ProM 6 from `http://www.promtools.org/prom6/`.
[4] See `http://www.xes-standard.org/` for more information on the XES standard.
[5] IEEE Task Force on Process Mining: `http://www.win.tue.nl/ieeetfpm/`.

Both Disco and ProM can directly read the compressed MXML and XES files, and you can use standard unarchiving tools to open and view the exported files if needed.

- **FXL**. Loading data sets in MXML or XES format is already faster than importing CSV data (see also Section 3.2.1), but nothing is as fast as loading data in the native Disco Log Files (FXL) format. FXL is a proprietary (no standard) format but the best if you need to exchange really large data sets with another Disco user. Files in FXL format have also a smaller file size compared to compressed XML standard logs and CSV.

  Alternatively to saving or exchanging FXL files, you can export and share the whole project workspace altogether. Refer to Section 6.3 to read more about how to export and import Disco projects as DSC files.

7.2.2 Adding Start and End Points

When you select the *Add endpoints* option for any of the event log export types, each case will be exported with a dedicated 'Start' and 'End' event as shown in Figure 7.6. The endpoints are only added if the process has no unique start and end activities.

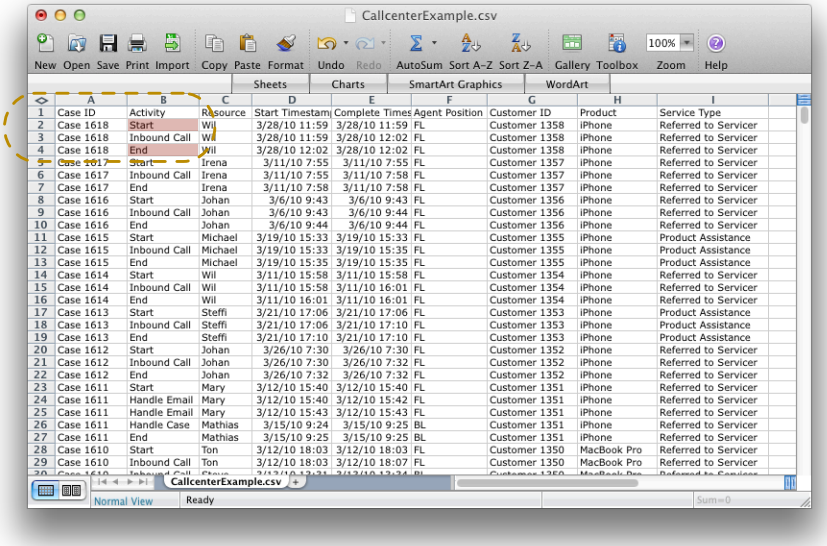


Fig. 7.6. The *Add endpoints* export options adds a 'Start' and 'End' event at the beginning and end of each case.

Having unique end points in your process data can be useful to further process the data in other, not process-aware analytics tools such as Excel. Furthermore, it is

strongly advisable to add start and end events if you want to use mining algorithms such as the Heuristic miner in ProM because they assume that there is an identical start and end event for each case (otherwise the quality of the result may be reduced).

7.2.3 Anonymization

Anonymizing your data set is sometimes necessary to protect the privacy of customers or employees during your process analysis. In Disco, anonymization is possible for Event log export types. When you select the *Anonymize* option, a number of more fine-grained anonymization options appear as shown in Figure 7.7.

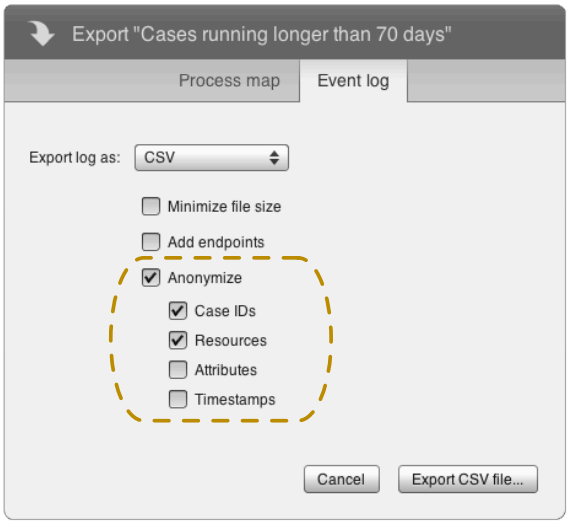


Fig. 7.7. Anonymization options in Disco.

The following detailed anonymization options are available:

- *Case IDs.* The names of the cases are replaced by anonymous *Case 1*, *Case 2*, etc. names. Use this option for de-identification if your case IDs contain sensitive data such as patient names in a healthcare process, or social security numbers in a government process.
- *Ressources.* The names of the people working in the process are replaced by *Value 1*, *Value 2*, and so on. This way, the identity of workers is not revealed in performance-oriented analysis projects.
- *Attributes.* Attribute names and values can reveal sensitive information about the process. Therefore, it is possible to anonymize them as well.
- *Timestamps.* When timestamps are anonymized, then the performance structure of the data set (for example, the execution times of activities) are not changed, but the actual time of when the activities have occurred is obscured.

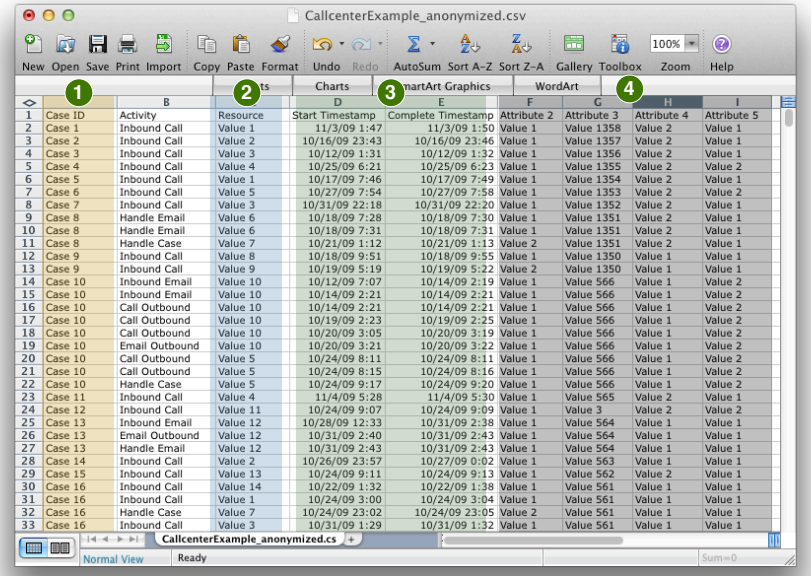


Fig. 7.8. Anonymized call center demo example.

In Figure 7.8, you can see an example of the anonymized call center demo log, where the case IDs (❶), the resource names (❷), the timestamps (❸), and the attributes (❹) have been anonymized.

The activity names are never anonymized, because they are typically needed in a legible form to perform a process analysis. If you do need to change the activity names, you can search and replace them in your original data file before importing it in Disco.

7.3 Exporting Charts and Tables

Next to process maps (Section 7.1) and event logs (Section 7.2), you can also export charts and tables.

Charts can be exported by clicking on them as shown in Figure 7.9. When you click on a table as shown in Figure 7.10, then you get two options:

❶ Press the *Copy* button to copy the content of the current cell to the clipboard. From there you can paste it anywhere, also in other applications. This can be handy if you want to copy individual statistics to a custom report in Excel, or if you want to search for a specific case in the Cases view.

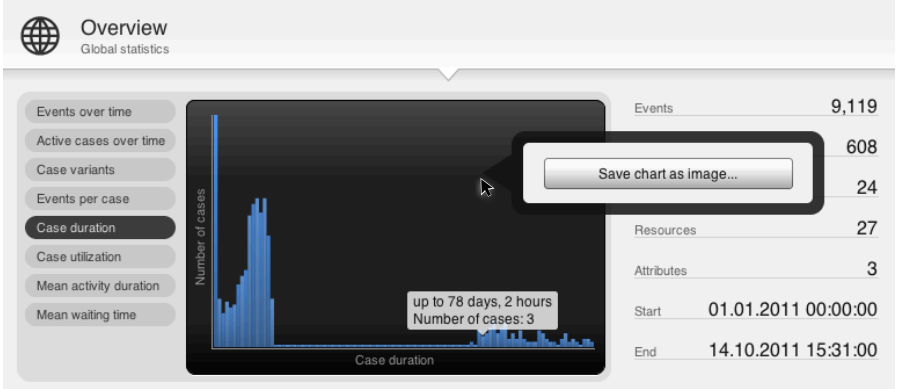


Fig. 7.9. Export a chart by clicking on it and pressing the *Save chart as image...* button.

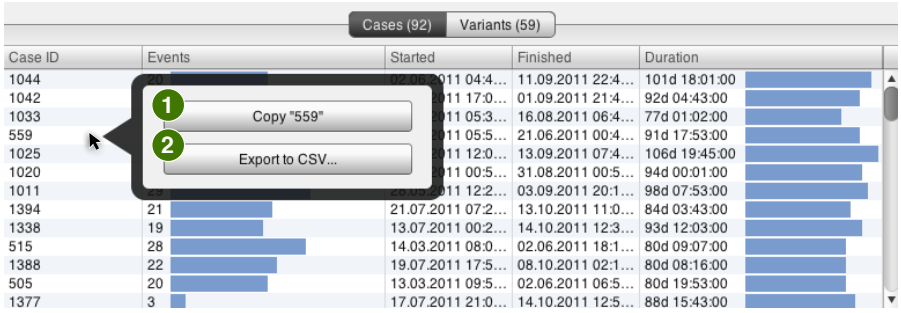


Fig. 7.10. Export any table in Disco by clicking on it and pressing the *Export to CSV...* button.

❷ Press the *Export to CSV...* button to export the complete table in the same way as you currently see it. Any table in Disco can be exported this way. Here are a few use cases for the CSV export of tables:

- You can export the table with the case IDs (as shown in Figure 7.10) for the currently filtered subset of your log where you have found deviations from the prescribed process, and other people will be able to open the file in Excel to view your list.
- Or you might want to copy the tabular frequency overview of an important attribute.
- Another scenario would be to export the detailed history of a particular case as shown in Figure 7.11.

If you need to export more than a few individual cases, you can use the event log export (see Section 7.2) to export your current data set in CSV format.

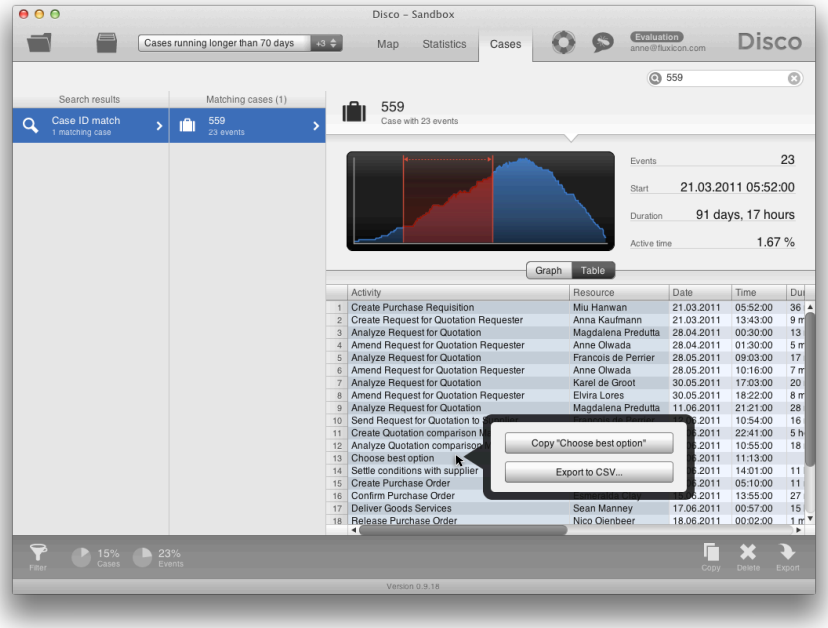


Fig. 7.11. Also the history table of individual cases can be exported as a CSV file.

7.4 Exporting Projects

Finally, you can export your complete Disco project, which bundles all data sets including all copies, filters, notes and current views in one DSC file.

You can use project files for yourself to save your work and keep all related files and notes together. You can also export a project file to share your work with colleague (see also Section 6.3).

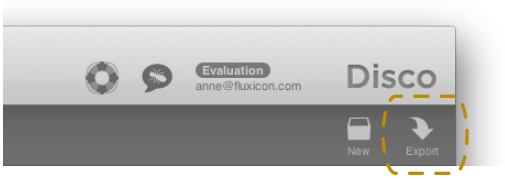


Fig. 7.12. You find the project export in the upper right corner of the project view.

8

The Toolbar

In the upper right corner of Disco, you find a toolbar that is accessible from any view in Disco (see Figure 8.1).

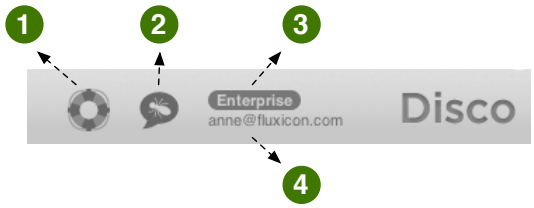


Fig. 8.1. Toolbar in the upper right corner of Disco.

The toolbar gives you quick access to the following elements:

❶ *Help*. Clicking on the lifebelt symbol brings up the helpful sticky notes again that were shown the first time you started Disco (Section 8.1).

❷ *Feedback*. The feedback button allows you to send in-app feedback right from within Disco (Section 8.2).

❸ *License*. The badge above your email address shows your current Disco license (Section 8.3).

❹ *Email address*. The email address with which you have registered Disco serves as your Fluxicon ID (Section 8.4).

8.1 Helpful Sticky Notes

When you started Disco the first time, you probably noted the yellow sticky notes that introduced you to the key elements in the application in every view (see Figure 8.2). They serve as a quick introduction and orientation help while you are exploring the functionality of Disco.

Fig. 8.2. You can bring the sticky notes back by clicking on the lifebelt symbol in the toolbar.

If you are lost at a later point in time, you can bring back these sticky notes anytime by clicking on the lifebelt symbol in the toolbar (❶ in Figure 8.1). You can them disappear again clicking on the lifebelt symbol again or by clicking anywhere else in the Disco application screen.

8.2 Send Feedback

The feedback button in the toolbar (❷ in Figure 8.1) brings up a pop-over as shown in Figure 8.3, which allows you quickly drop some notes that will be sent right to the inbox of Disco's developers.

We closely listen to what our users have to say about Disco and encourage you to send any feedback that comes to mind. For example, you can let us know about:

- *Bugs*. If Disco is not working as expected based on your data set, please let us know about it, so that we can fix it.
- *Improvements*. For example, do you find yourself repeating a certain task in multiple steps all the time and you feel there would be a better way? We would love to hear from you and many suggestions have already found their way into Disco.
- *Something that is not clear to you*. Our goal is to make Disco as simple to use as possible to allow you to focus on your work and your analysis. If something is not clear, tell us and we will explain. This also helps us improve the usability in the long term.

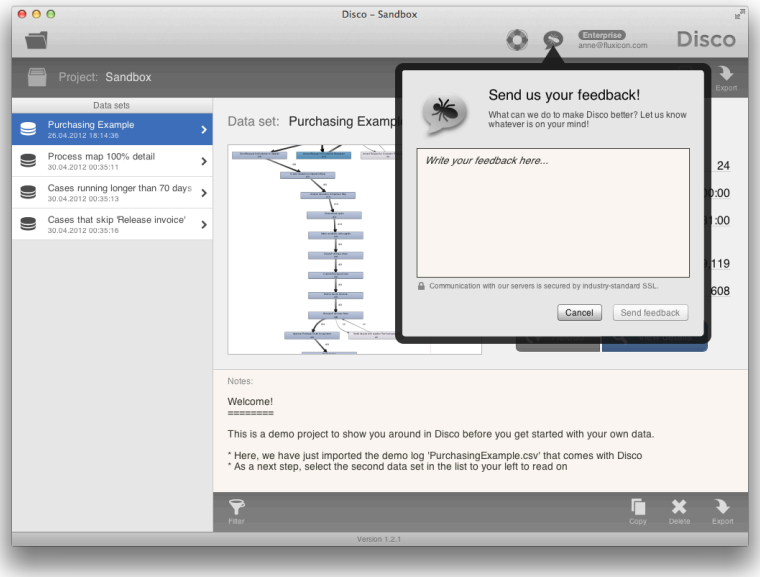


Fig. 8.3. The feedback button sends your comments or improvement suggestions to us right from Disco.

- *Additional functionality that you would find useful*. We have many ideas about where we want to take Disco in the future, and getting your input will help us prioritize new functionality in the right way.
- *Things you like about Disco*. Knowing what people value about Disco will help us to ensure that it stays that way.

Just be sure that there is nothing too small or insignificant that we would not want to hear about it!

8.3 Your Disco License

When you install Disco for the first time, you automatically get a demo license. So the badge above your email address in the upper right corner will say *Demo*.

The *Demo* license limits the import to 100 events per file, cannot be commercially used, and puts a note on exported process maps that says that the map was created with the demo version of Disco. As soon as you want to start using Disco productively, you can purchase a commercial license and the new license type will displayed.

When you click on the badge with your license type (❸ in Figure 8.1), then you get an overview of the licensing terms and the detailed license agreement that you

accepted during the installation of Disco as shown in Figure 8.4 for the *Enterprise* license.

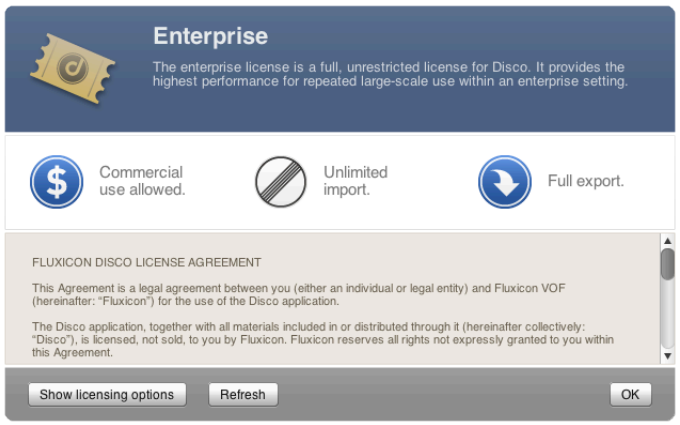


Fig. 8.4. When you click on the license badge in the toolbar, you get an overview of the licensing terms and the license agreement that you accepted during the installation of Disco.

In the license view, you have the following three options:

- *Show licensing options*. Clicking this button brings you to an overview of the different licensing options that are currently available. We have several options for licensing Disco and can help you pick the right package for your own situation. Simply contact us at `http://fluxicon.com/disco/buy.php`.
- *Refresh*. Licenses are automatically downloaded from our server. If you license has been updated while you are running Disco, you can use the *Refresh* button to get the latest license.
- *OK*. Clicking *OK* closes the license view.

8.4 Your Email Address

The email address that you used to register Disco is displayed in the toolbar (❹ in Figure 8.1). When you click on your email address, the screen shown in Figure 8.5 will be shown.

If you want to change your email address, you can click the *Switch email address...* button and re-register Disco with your new email address.

Fig. 8.5. When you click on your email address in the toolbar, you can change your Disco registration to another email address.