

INFOWATCH

InfoWatch Data Discovery. User Guide

31/07/2024

© JSC “InfoWatch”

Tel/Fax +7 (495) 229-00-22

http://www.infowatch.com

Table of Contents

1 Who should read this guide 6

2 Set of documents 7

2.1 Technical support 7

3 Overview 8

4 Investigation Centre Interface 9

5 The Main Page 12

5.1 Dashboards and Widgets 12

5.1.1 Dashboards 12

5.1.2 Widgets 13

The Violations Widget 13

The Empoloyee under Supervision widget 13

6 Single Filter 15

6.1 Filters 15

6.2 Filters in the File Sharing section 15

6.3 Filters in the Events and Analytics → Violation Statistics section 17

6.4 Managing Single Filter 24

6.4.1 Adding and removing filters 24

6.4.2 Setting filter values 24

A condition with multiple filters 24

"Pattern search" value type 25

Including and excluding filter values 25

Resetting filter values 27

6.4.3 Saving and loading filter sets 27

7 Events Section 28

7.1 Adding Events to the Investigation 29

7.2 Viewing and Exporting Event Data 30

7.2.1 Viewing used contacts 30

7.2.2 Viewing attachment data 31

7.2.3 Exporting event data 31

7.3 Filtering events by persons and contacts 33

8 Analytics Section 35
8.1 Statistics Widgets 35
8.2 Exporting Reports with Statistics Data 37
9 File Sharing Section 39
9.1 "Tasks" Tab 39
9.1.1 Actions available in the section 39
9.1.2 Scanning task 39
Resource Scanning 40
Traffic Monitor Event 42
9.1.3 Creating, Configuring and Running Tasks 44
Creating a task 44
Starting a scanning task 44
Task settings 45
Adding and Configuring Hosts 60
Adding Scanning Paths 73
Examples of Task Settings, Host Settings and Adding Scanning Paths 75
9.1.4 Editing Tasks and Hosts 83
9.1.5 Copying Tasks 83
9.1.6 Deleting Tasks, Hosts and Scanning Paths 83
9.1.7 Task Execution Statistics 85
Statistics on scanning tasks 85
9.2 "File Information" Tab 91
9.2.1 File Information Table 92
How entries are added to the detailed file information table within a task 92
Structure of the detailed file information table 92
Identification of persons and groups 95
Actions with files 96
9.2.2 Table Display Configuration 97
9.2.3 Search for Duplicates 97
Search by template 97
Resetting search results 98
9.2.4 Exporting Reports with File Information 98
9.2.5 Example of Managing the Table 100
Configuring the filters 100
Results of filtering 102
9.3 "Duplicate Files" Tab 103
9.4 "Similar Files" Tab 105
10 Persons Section 108

10.1 Viewing Person's Dossier 108
10.2 Personal Information Tab 110
10.3 File Sharing Tab 110
10.4 Creating, Editing and Deleting Dossier 111
10.5 Adding Contacts to Person 112
10.6 Changing Person's Manager 112
10.7 Merging Dossiers 113
11 Investigation Section 114
11.1 Creating and Archiving Investigations 115
11.2 Adding Materials to Investigation 115
11.3 Collaborative Investigations 116
11.4 Editing, Exporting, Printing and Deleting Investigation 116
12 Reports Section 117
12.1 Reports Archive 117
12.2 Tasks for Regular Sending 117
13 End-user license agreement 119
14 Glossary 123

This guide describes how to work with InfoWatch Data Discovery (also referred to as the System or Data Discovery). Data Discovery is software that is designed to search for confidential information on shared network resources, workstations, servers and in document storages.

1 Who should read this guide

This guide is designed for users who are going to work with the System (set up configuration, analyse information resources, etc.). This guide is intended for users familiar with the basics of working in a Linux operating system environment and PostgreSQL DBMS.

2 Set of documents

The documentation includes:

- "InfoWatch Data Discovery. Installation Guide". The document describes how to install, update, and uninstall the System.
- "InfoWatch Data Discovery. Administration Guide". The document contains information on how to configure the System.
- "InfoWatch Data Discovery. User Guide". The document describes how to work with scanning tasks.

Related documentation for InfoWatch Traffic Monitor (also referred to as Traffic Monitor) includes:

- "InfoWatch Traffic Monitor. Installation Guide". The guide describes how to install, configure, update, and uninstall InfoWatch Traffic Monitor.
- "InfoWatch Traffic Monitor. Administration Guide". The guide describes how to administrate InfoWatch Traffic Monitor (database, back end).
- "InfoWatch Traffic Monitor. User Guide". The guide describes how to work with InfoWatch Traffic Monitor (set up configuration, export/import data, prepare policies for processing objects).
- "InfoWatch Traffic Monitor. Configuration Files Guide". Provides explanations for frequently used configuration files.

2.1 Technical support

If you have any problems or questions related to the operation of the System, please contact our technical support service:

- If you purchased the product from InfoWatch partner, please contact the technical support service of the partner
- If you purchased the product from InfoWatch directly, please contact InfoWatch technical support at support@infowatch.com.

Technical Support Service is available to assist you Mon-Fri, 7 AM – 9 PM (Moscow time, GMT+3), except weekends and public holidays in the Russian Federation. You can also visit technical support section on our website: http://infowatch.com/support.

3 Overview

Investigation Centre is a web console that is designed to work with InfoWatch products. You can install one or more products in the cluster. Installing all the products allows you to get the most complete information about violations, persons, risks, etc.

Data Discovery is designed to search for confidential information on shared network resources, workstations, servers and in document storages . Data Discovery works as follows:

- Connects to various resources,
- Scans files on the resources and collects file metadata,
- Sends data to InfoWatch Traffic Monitor and to InfoWatch Data Analysis Service for analysis if a processing mode is selected in the scanning task,
- Displays metadata and results of grouping the scanned files.

With Data Discovery, you can:

- Connect to shared network resources and to distributed file systems (DFS) using the SMB protocol:
    - Connect to SMB shares on MS Windows hosts
    - Connect to SMB shares on Linux hosts
- Connect to Linux-based workstations and servers via SSH, including secure authorization with SSH keys
- Connect to SharePoint Server 2016 document libraries via HTTPS
- Connect to file shares via FTP/FTPS protocols
- Connect to Nextcloud data storage via WebDAV protocol
- Scan files on resources:
    - detect new and modified files
    - collect file metadata
    - create file copies to be analyzed in Traffic Monitor
    - extract text contents from file copies for analysis using InfoWatch Data Analysis Service
- Send data to Traffic Monitor server for analysis
- View statistics on the execution of scanning tasks
- View information about files
- Search for duplicates among scanned files automatically or by a loaded template
- Identify documents with the same terms (similar ones) and group them, using InfoWatch Data Analysis Service
- View extracted text contents
- Export reports with information about files, duplicates or similar ones

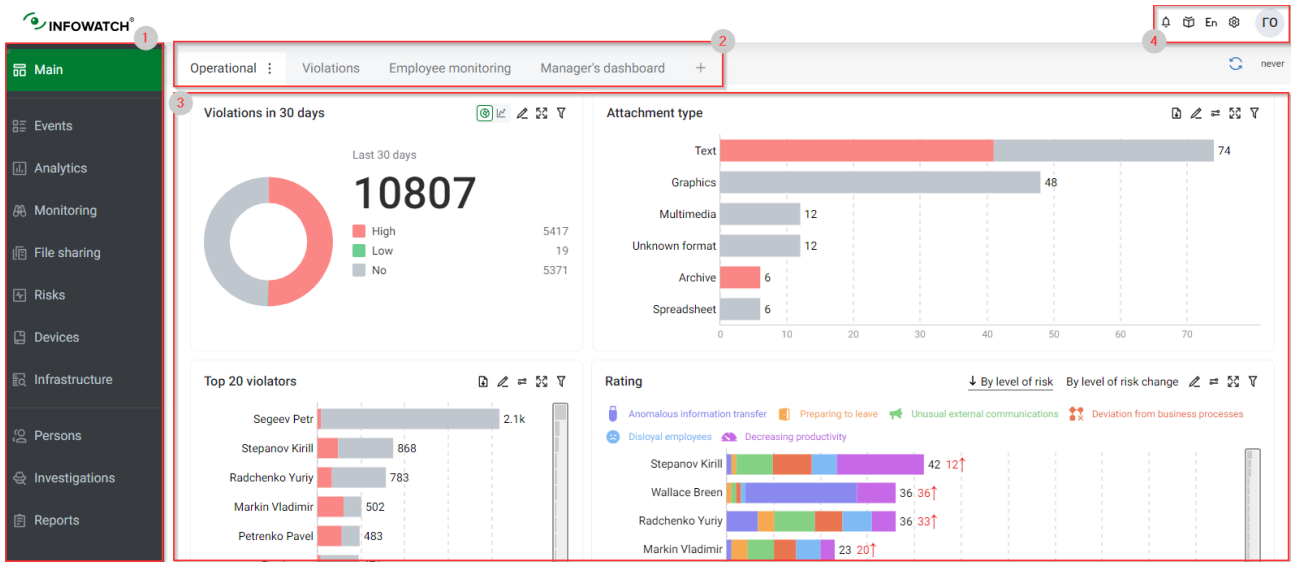
4 Investigation Centre Interface

You can work with Data Discovery in InfoWatch Investigation Centre.

The Investigation Centre is a single graphical environment where you can view and analyze data of installed InfoWatch products:

- Activity Monitor
- Data Access Tracker
- Data Discovery
- Prediction
- Vision

The Investigation Centre interface has several areas. In the picture below, you can see the **Main** page.



The interface elements:

Number in the scheme	Element	Purpose of the element
1	Navigation bar	Working sections of the Investigation Centre
2	Dashboard panel	The list of preset dashboards
3	Working area	Here you can view widgets of InfoWatch products and work with them
4	User menu	The user menu, where you can: • edit your account details, • change your password, • enable hotkeys, • find contacts of the InfoWatch company and the technical support service

Number in the scheme	Element	Purpose of the element
		• log out of the System
	Settings	The Settings section (for details, see "*InfoWatch Data Discovery. Administration Guide*")
	Notifications	A notification folder where messages from the System to the user go (for details, see "*InfoWatch Data Discovery. Administration Guide*", the "*Notifications*" article)
	Change language	Choosing the interface language: *Russian* or *English*
	Guidebook	The Investigation Centre guidebook where you can find: • *Scenarios for using the System* to use the product more efficiently; • The *Administration Guide*, which describes the product settings; • The *User Guide,* which describes working with the product; • Information about new features of the product (*What's new)*

To work in the Investigation Centre, you can use the following sections:

Section	Description
Main	The main section that contains preset dashboards with widget sets. The dashboards allow you to quickly obtain and display information about violations and violators on each dashboard according to filters. You can also create custom dashboards with the needed set of widgets
Events	The list of events according to set filters
Analytics	Statistical information about violations/violators according to set filters. The information is displayed on: • Widgets with graphs, charts and tables, • The connection graph with event visualization. The *Violation statistics* tab is available to all products. On the *Connection Graph* tab, you can find all functions of InfoWatch Vision (see "*InfoWatch Vision. User Guide*"). On the *Activity statistics* tab, you can find functions of InfoWatch Activity Monitor (see "*InfoWatch Activity Monitor. User Guide*")

Section	Description
Monitoring	Monitoring employee activity as well as online monitoring of employee workstations. A section of InfoWatch Activity Monitor (see "*InfoWatch Activity Monitor. User Guide*")
File Sharing	Creating and running network resource scanning tasks, viewing and analyzing detailed information about files including file owners and searching for file duplicates and similar files (for details, see "File Sharing Section")
Risks	Calculation of employee ratings to predict information security risks. A section of InfoWatch Prediction (see "*InfoWatch Prediction. User Guide*")
Infrastructure	Tracking of changes to directory service objects and monitoring the status of accounts. The section of InfoWatch Data Access Tracker (see "*InfoWatch Data Access Tracker. User Guide*")
Persons	A catalogue of persons who work in the company and a catalogue of external contacts
Investigations	Carrying out investigations using accumulated data on violations
Reports	A table with reports where you can export report files and a list of tasks for regular sending

5 The Main Page

The Main Page of the Investigation Centre contains dashboards. These are tools that allow you to get up-to-date information about violations. Dashboards consist of a widget set of one or more InfoWatch products. You can create custom dashboards and add any available widgets to them or select preset dashboard with the most used widgets. The set of available widgets depends on installed InfoWatch products.

Preset dashboards:

- **Operative** - represents key indicators to monitor violations:
  - Graphs with the dynamics of violations for the last 3 days or the last 24 hours, charts with distribution by levels of violations,
  - Top violators among employees who send data,
  - Top most violated security policies,
- **Employee monitoring** - represents statistics of violations involving employees:
  - Top violators among employees who send data,
  - Employee rating (if InfoWatch Prediction is installed),
  - Employee under supervision,
- **Files** - represents the most common files in violations and audit of actions with them:
  - Top triggered protected objects,
  - Top files used in operations,
  - Top file types used in operations,
  - Statistics of actions with files (if InfoWatch Activity Monitor is installed)
- **Violations** - shows anomalies in employee actions, behavior changes, comparative statistics on violations:
  - Graphs with the dynamics of violations for a selected date or period of time, diagrams with distribution by violation levels,
  - Top most violated security policies,
  - Top violators among employees who send data,
- **Manager's dashboard** - presents statistics on employees' actions to the head of the department:
  - Graphs with the dynamics of violations for a selected date or period of time, diagrams with distribution by violation levels,
  - Top violators among employees who send data,
  - Top event types with violations,
  - User decision on events.

5.1 Dashboards and Widgets

5.1.1 Dashboards

Dashboards with widgets are avaialbe on the **Main** page. On the working area, the Security Officer can place preset dashboards, create custom dashboards and then add widgets to them. You can use widgets of any installed InfoWatch product.

For this:

1. On the dashboard panel, click .
2. Choose one of preset dashboards or **New dashboard**. If necessary, enter a new name and click **Save**. The new dashboard will be added to the panel.
3. Click  and select **+ Add widget**.

4. In the window on the left, specify, which widgets and from which installed InfoWatch product you want to add. Or you can choose **All**. To quickly find the desired widget by name, use the search bar.
5. Next to the needed widgets, click **+ Add** and specify their number. You can add up to 10 widgets of the same type to the dashboard.
6. Select other widgets you want to add
7. Click **Apply**. Selected widgets will be added to the working area.

Single Filter conditions specified in other sections of the Investigation Centre do not apply to dashboard widgets on the **Main** page. You can configre any widget regardless of others. To configure a widget:

1. In the corner of the required widget, click .
2. Set required filter values. If necessary, add additional filter attributes (see "Single Filter").
3. Click **Save**.

Data on widgets is updated according to the set interval. To set the update interval, select a value in the right upper corner. You can choose a value from 1 minute to 24 hours. If you do not want to update the data, click **Not update**.

You can also update data manually. For this, click . Please note that only widgets located in the visible area of the dashboard are updated. To update all data on all widgets, scroll through all widgets on the screen and click .

5.1.2 Widgets

You can place widgets of the **Analytics** and **Risks** sections on dashboards.

The set of widgets that can be added to dashboards depends on pruducts that are installed in the Investigation Centre cluster.

The following widgets are available only on the **Main** page: **Violations** and **Employee under supervision**.

The Violations Widget

This widget processes data received from Traffic Monitor and can display:

-  - the total number of violations with distribution by their levels
-  - violations in dynamics over a period of time

In both cases, you can customize the display. For this, use the legend where you can enable or disable graphs of violation levels:

The Empoloyee under Supervision widget

The widget collects information about an employee. Based on this information, the Security Officer can then start an investigation and start monitoring the employee more closely.

To set up the widget:

1. Click  and specify attributes for filters that can apply to the widget.
2. In the **Person** filter, specify a person from the list of employees. If necessary, add other filters and specify their values.
3. Click **Save**. The widget will display:

- Personal information about the employee and security policies, which include the employee. The widget can contain up to 8 policies with the most events with a high violation level.
- The latest event related to the person, with the data, action and policy
- Other persons from events related to the employee,

4. Click  to change the widget display:
    - Employee photo size,
    - Diagrams of policies with high violation level.

6 Single Filter

A Single Filter is a universal tool that allows you to manage your searches and data display in InfoWatch Investigation Centre.

Different sections of the Investigation Centre contain different filter sets.

For Data Discovery, the Single Filter is available in the following sections: **File sharing**, **Analytics** (the **Violation statistics** tab), **Events** and **Persons**.

Whether the filter is available or not also depends on the User Role settings.

Using the Single Filter, you can:

- control the display of entries about scanned files:
  - select specific entries to display,
  - exclude irrelevant entries from the file information table,
- save the filter settings to get the required information quickly.

Important!

When working with the Single Filter, please keep in mind that:

- All settings of the Single Filter apply only within one browser tab,
- Clicking the **Back**/**Forward** browser tabs returns the filter to the state it was in at that step,
- When you log out of the System, all filter attributes are reset to their original state.

6.1 Filters

The Single Filter is a panel that contain a group of filters.

6.2 Filters in the **File Sharing** section

Note:

The Data Discovery filters are unique and work only in the **File Sharing** section, on the **File Information**, **Duplicate Files**, **Similar Files** tabs, and in the dossier: the **Available files** and **File owner** (the **File sharing** tab) subsections. Their values are not transferred when switching to other section and affect only Data Discovery data filtering.

A user who has access to the **File Sharing** section can use Data Discovery filters without any additional settings.

The filter set correspond to table fields with detailed information about files:

Filter name	Description	Setting values
Discovered	The time period in which the file was discovered	Select the time period: • Last 24 hours • Current day – from 0:00 to the current time

Filter name	Description	Setting values
Created	The time period in which the file was created	• Last 3 days • Last 7 days • Last 30 days • This month • Last month • Custom period: • Period (specify the start and end dates of the period) • Custom date (specify one or more dates) The default value: **Any date**
Modified	The time period in which the file was modified	
Task	Scanning task name	To select a value, do either of the following: • Select a value from the list. To search through the list, enter the value in the search bar. The display of the list items does not depend on conditions set in the other filters. • Enter a value in the search bar and select the value of "pattern search" type". The default value: **All**
Host	Host name	
Host IP address	Host IP address	
Location	Full path to the directory	
File name	File name	
Owner	The person or group who owns the file	The list displays names of persons and groups identified by SID. You can filter entries both by groups and by persons at the same time. Persons and groups are selected separately; to do this, use the **Person**/**Group** switch in the filter. All selected values are combined using a logical OR. To choose a value, do either of the following: • Select a person from the list. To search through the list, enter the person name or SID in the search bar • Select a group from the Active Directory tree. To search the tree, enter the group name in the search bar For details on identifying persons and groups, see "File Information Table". The default value: **All** The **Search by rights assigned to the person explicitly** option defines should the filter consider inclusion of a person in groups as well as nesting

Filter name	Description	Setting values
Access granted to	Persons/groups that have access to the file	of groups. If a group, but not a particular person, has access to the file and you have enabled the option, then search by the person included in the group returns no results. Similarly for nested groups: with the enabled option, search by the child group does not return files as a result if the parent group have ownership/access to the files. If you disable the option, in the search results, you see both files to which the person or group has direct ownership/access and files to which the parent group has ownership/access. **Note:** In the filter, you can see all the groups added when integrating with Active Directory. If the group does not have SID, you cannot select it in the filter
Size	File size limits	In the **from** field, enter the minimum file size. In the **to** field, enter the maximum file size. Select the unit of measure. The default value: **All** **Note:** If only the minimum file size is specified, files whose size is greater than or equal to the minimum size will be displayed. If only the maximum file size is specified, files whose size does not exceed the maximum size will be displayed.
Connection type	Host connection type	Select a value from the list. The default value: **All**

Warning!

The **Owner** and **Access granted to** filters may not work correctly after the System update. Edit or create again all the filter sets (**Requests**) that include these filters.

6.3 Filters in the **Events** and **Analytics** → **Violation Statistics** section

The table below lists filters and widgets that are available in the **Events** and **Analytics** → **Violation statistics** section.

Note:

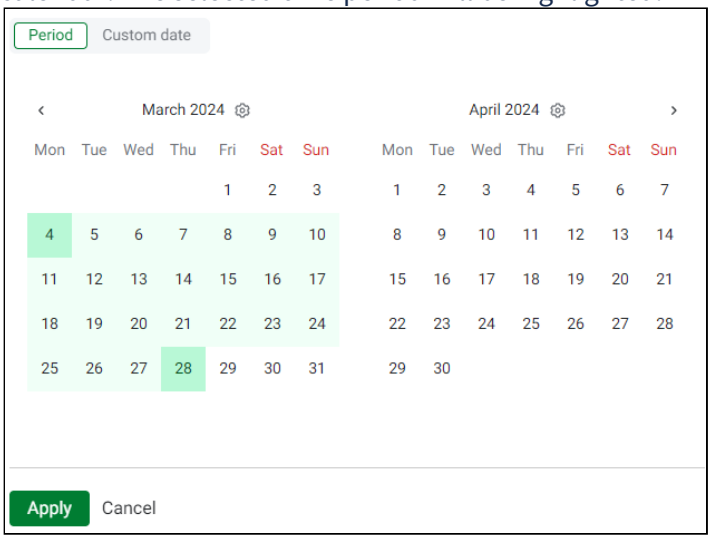
The following widgets listed in the table below are available by default:

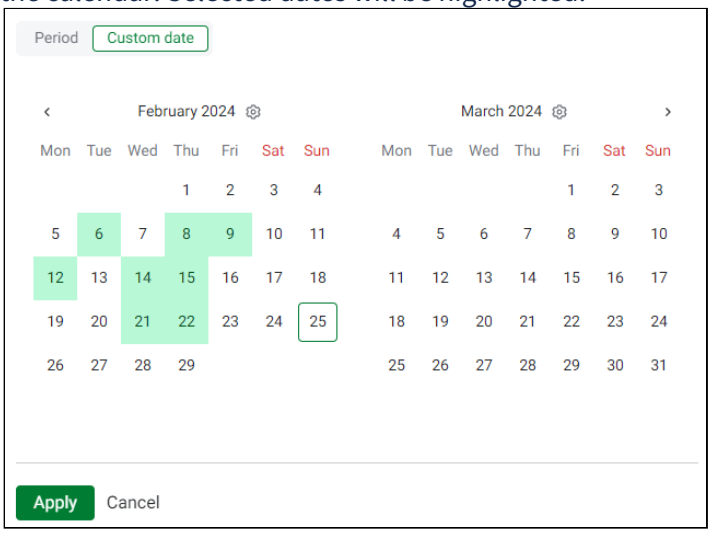
- Date
- Senders
- Event types
- Violation level
- Recipients

Other widgets listed in the table are available only if the cluster includes Vision.

Important!

To make data filtering available, make sure that the user role has access to the **Events** and **Analytics > Violation statistics** section. For more information about configuring roles, see the *Roles* article of the "*InfoWatch Data Discovery. Administration Guide*".

Filter name	Description and possible values
Date	The **Date** widget corresponds to this attribute. Specify the period of time when the data you are interested in is available. • *Last 24 hours;* • *Current day – from 0:00 to the current time;* • *Last 3 days;* • *Last 7 days;* • *Last 30 days;* • *This month;* • *Last month;* • *Custom period* (you may specify the beginning and the end of a period as well as a list of custom dates): **To specify the required time period:** 1. Use your mouse to select the start and end dates in the calendar. The selected time period will be highlighted:  2. Click **Apply**. **To specify one or more custom dates:**

Filter name	Description and possible values
	1. Use your mouse to select one or many required dates in the calendar. Selected dates will be highlighted:  2. Click **Apply**.
Hours	The **Time** widget corresponds to this attribute. Specify the time when the event happened.
Violation level	The **Violation level** widget corresponds to this attribute. Select a violation level (complies with similar in Traffic Monitor): • *High* • *Medium* • *Low* • *No*
Event type	The **Event type** widget corresponds to this attribute. Create event type sampling. For this, select the event type in Traffic Monitor: • *Work with applications:* • *Clipboard* • *Keyboard input* • *Mail:* • *Mail on client* • *Webmail* • *Printer and MFD:* • *Printing* • *File storage:* • *Data Discovery* • *Web activity:* • *Web message* • *Messenger:* • *Facebook* • *ICQ*

Filter name	Description and possible values
	• *MS Lync* • *Mail.Ru Agent* • *MS Teams* • *Skype* • *Telegram* • *Viber* • *WhatsApp* • *VK* • *XMPP* • *Yahoo!* • *File exchange:* • *Cloud storage* • *Removable device* • *FTP* • *Network resource* • *Terminal session* • *File operations:* • *Read file* • *Write file* • *Others:* • *Photography* • *SMS*
Protected objects	The **Protected object** widget corresponds to this attribute. Select from the list of protected objects created in Traffic Monitor.
Policy	The **Policies** widget corresponds to this attribute. Select from the list of policies created in Traffic Monitor.
Verdict	The **Verdict** widget corresponds to this attribute. Select verdicts you are interested in from Traffic Monitor list: • *Allowed* • *Blocked* • *Quarantined*
User decision	The **Statistics** widget where **User decision** value is set for column corresponds to this attribute. Select required user decisions on TM events from the list: • *Violation* • *Requires additional processing* • *No decision* • *No violation*
Violations	Specify the search type to search events with or without violations

Filter name	Description and possible values
File name	The **Top files** widget corresponds to this attribute. Type a file name into the search bar or select files from uploaded attachments in Traffic Monitor events
File type	The **File type** widget corresponds to this attribute. Select categories and/or types of attachments in Traffic Monitor events
Branch	The **Source** widget corresponds to this attribute. Select branches and/or integrations with them
Integration	Select one or more integrations with the source systems
Contact type	The filter enables displaying contacts by different types of contacts used in the System. Select one or more values: • *YFired - filter by events where one of participants is a former employee. If personal contacts of the employee were not automatically added during integration with other systems, they should be added manually (for details, see* "Adding Contacts to Person"), • *External* - filter by events where at least one participant is not from an internal domain, • *Internal* - filter by events where at least one participant is from an internal domain, • *New* - filter by events where a given identified or unidentified contact received from Traffic Monitor appears for the first time on the exact date or during the time period specified in the ***Date*** filter, • *Personal* - filter by events containing messages sent to personal addresses, including emails sent by one employee to a personal address of another employee, • *Self send* - filter by events where the sender and the recipient is the same person. When selecting *Personal* and *Self send* at a time, then the System will display the events with emails sent by the employee to their own personal address
Investigations	The filter allows you to display events that are involved in one or more investigations. Select one of the following values: • *Beyond the investigation* – a filter by events that are not involved in investigations; • *In the investigation* – a filter by events that are involved in investigations. When selecting this value, specify one or more investigation names in the search bar. If you leave the field empty, the search will be carried out on all investigations

Filter name	Description and possible values
Perimeter, In perimeter, Out of perimeter	Select a perimeter from the list. Set a modifier, if needed: • *Perimeter* - search through all the events that have specified perimeters as their attributes • *In perimeter* - search through the events that have specified perimeters in the **In perimeter** attribute • *Out of perimeter* - search through the events that have specified perimeters in the **Out of perimeter** attribute Selecting *Unknown perimeter* will find the events for which entering or leaving the perimeter was not identified in Traffic Monitor.
Number of recipients	The filter allows you to display events with a specified number of recipients. By setting the *From* and *To* values, you can exclude bulk mailing (e.g., the number of recipient is up to 7) or display personal communications only (the number of recipients is up to 1)
Attachments	The filter allows you to display events with or without attachments as well as specify total size of attachments. For example: • Specify the number of attachments including attachments within archives • Specify the attachment size larger than 3 KB to ignore company logos attached to emails • Specify the attachment size larger than 5 KB to find events with large attachments
Subject of email	The filter allows you to carry out a full-text search by email subject in events. You need to enter at least three characters and choose the search method: • Including morphology – the System will search through all entered words including words with the same root; words can be in any order • By exact match If the System finds more than 1000 events, modify the query by specifying more parameters or choosing different values. The filter is available only in the **Statistics** section
Contacts*, Contact of sender*, Contact of receiver*	The filter allows you to display events with specified email addresses or device identifiers. In the search bar, enter first characters of employee name, their contact data or device. In addition, you can configure the filter using modifiers: • **Contains value** - events involving at least one of addresses or device IDs will be displayed;

Filter name	Description and possible values
	• **Exact match** - events in each of which there are all selected addresses or device IDs will be displayed. Does not work with pattern search. Use the **Person** filter to search employees by names
Person*, Sender*, Receiver*	The filter allows you to display events with specified persons as participants. You can specify either the employee's name or their contacts in the search bar. In addition, you can configure the filter using modifiers: • **Contains value** - events involving at least one of selected persons will be displayed; • **Exact match** - events in each of which there are all selected persons will be displayed
Group*, Group of receivers*, Group of senders*	The **Analytics by groups/persons** widget corresponds to this attribute. It creates a selection of AD and Traffic Monitor groups. Specify groups within the company hierarchical structure that are synchronized from Traffic Monitor and Active Directory
Title* Title of senders*, Title of receivers*	Specify all the company's job titles synchronized from Traffic Monitor or AD you are interested in
Department*, Department of senders*, Department of receivers*	Specify all the company's departments synchronized from Traffic Monitor or AD you are interested in
Status*, Status of senders*, Status of recipient*	Specify all the employee statuses synchronized from Traffic Monitor or AD you are interested in
* – For person-related attributes, you can set the following modifiers: • *Any* • *Sender* • *Recipient* If you set the *Sender/Recipient*, modifier, then selected persons, person groups or contacts will be considered as senders or recipients in events. If you set *Any*, all participants of the event will be selected: all persons, person groups and contacts.	

For more information on configuring filter panels, see "Managing Single Filter".

6.4 Managing Single Filter

6.4.1 Adding and removing filters

You can use one or more filters in the filters panel.

To add a filter to the filter panel:

1. Click .
2. In the window that opens, select the required filter from the list.
3. Click outside the filter list window.

The filter will be added to the filter panel.

To remove a filter from a filter panel:

1. Click .
2. Deselect the filter in the list.
3. Click outside the filter list window.

The filter will be removed from the filter panel.

To move a filter to a different location on the panel, hold down  next to the selected filter and drag it to another place.

To hide the Single Filter panel, click 

6.4.2 Setting filter values

At least one value must be selected in the filter. If no value is selected, the default value is used.

If you change the filter value, the table in the **File Storage** > **File Information** section is automatically updated. The table displays only those entries that match the filter conditions.

> **Warning!**
>
> When you log out from the System, filter values are reset. We recommend saving filters before logging out (see "Saving and loading the filters").

A condition with multiple filters

If you set values for multiple filters, then the conditions of these filters are combined using logical "AND".

Example:

Set values for multiple filters:

- in the **File name** filter, specify the value: **report.doc**
- in the **Task** filter, specify the value: **Scanning network shares**
- in the **Size** filter, specify **3** in the **to** field and select the **MB** measurement unit

As a result, the file information table will display only entries that simultaneously meet all the above conditions.

The table displays entries about files named **report.doc** from the **Scanning Network Resources** task, if the file size does not exceed 3 MB.

"Pattern search" value type

The "pattern search" value type is available for the following filters:

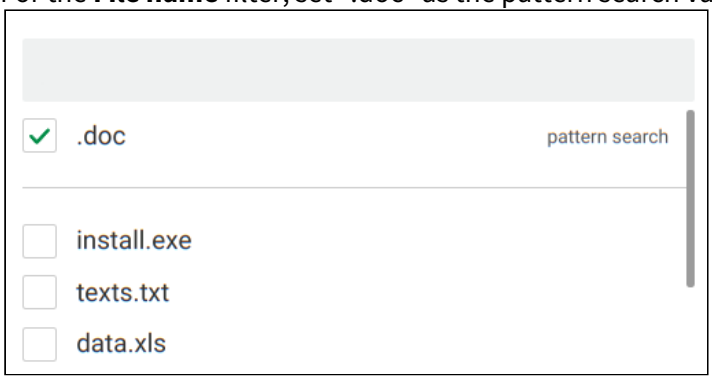
- Task
- Host
- Host IP address
- Location
- File name
- Owner
- Access granted to

To use a value of the "pattern search" type, enter the value in the search bar and select the "pattern search" value type.

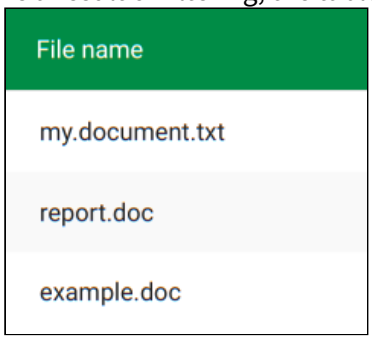
When filtering, the System checks whether the search string value is included in the full value of the file entry field. Values of the "pattern search" type are not case-sensitive.

Example:

1. For the **File name** filter, set ".doc" as the pattern search value:



2. As a result of filtering, the table displays only those files whose name contains the ".doc" string:



Including and excluding filter values

In the **File Storage** > **File Information** section, the following filter values may have multiple state:

- Task
- Host
- Host IP address

- Location
- File name
- Owner
- Access granted to
- Connection type

The following states are available for the value of the above filters:

- ☐ – the value is not selected (the default state),
- ☑ – the value is included,
- ☒ – the value is excluded.

To change the state of the value, click the state box.

You can include and/or exclude multiple values. In this case:

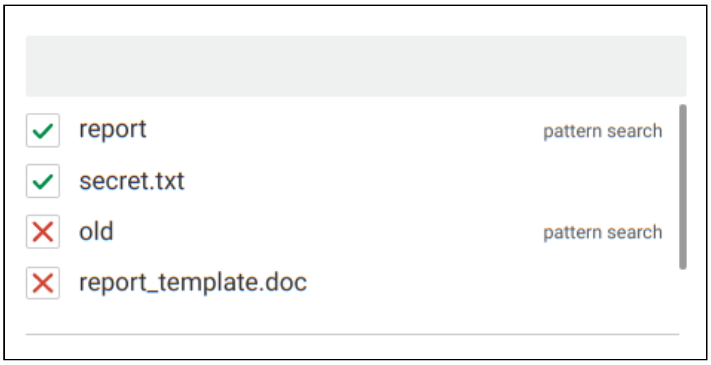
- Multiple included values are combined using logical "OR",
- Multiple excluded values are combined using logical "OR",
- A condition of multiple included values and a condition of multiple excluded values are combined using logical "AND".

Example:

The table contains 5 entries with the following values in the **File name** field:

- report_2021.docx
- old_report.doc
- new_report.pdf
- secret.txt
- report_template.doc

Include and exclude several values in the **File name** filter:



The table will display the file entries whose **File Name** field value meets both conditions at the same time:

- **Condition on the included values**:
  The condition is met if: the file name contains the **report** substring OR the file name is **secret.txt**
- **Condition on the excluded values**:
  The condition is NOT met if: the file name contains the **old** substring OR the file name is **report_template.doc**

As a result, the table displays entries with the following values in the **File name** field:

- report_2021.docx
- new_report.pdf
- secret.txt

Resetting filter values

To reset a filter values, click  in the filter panel. The filter value will be set to default.

To reset values for all filters, click .

6.4.3 Saving and loading filter sets

To use the specified filters values again in the future, you can save them. For this:

1. Open the section where you can use the Single Filter: **Events**, **Analytics** or **Persons**.
2. Set the required filter values and click .
3. In the window that opens, specify the new filter name.
4. Select the filter type:
   - *Personal* – the filter will be available only to the user who created it,
   - *General* – the filter will be available to all users of the System.
5. Click **Save**.

To use a previously created filter:

1. In the Single Filter area, click **Filters**.
2. Click the required filter to activate it. The Main Officer also can:
   -  - edit personal and common filters,
   -  - delete personal and common filters.

Other Security Officers can edit and delete personal filters only.

7 Events Section

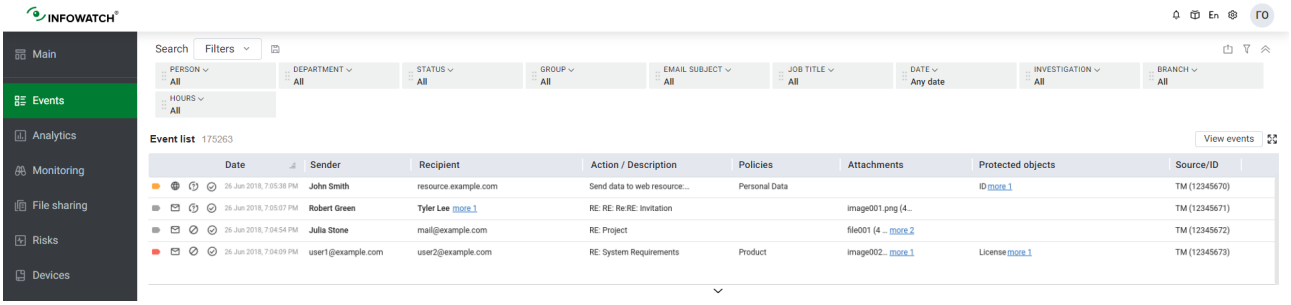
> **Note:**
>
> Data Discovery does not use events from Traffic Monitor.

In this section, you can view data of events uploaded to the Investigation Centre after synchronization with Traffic Monitor:

- Violation level
- Event type
- Security officer's decision on the event
- The System's verdict on the event
- *Date* – event creation date in Traffic Monitor
- *Sender* – a person who committed the violation
- *Recipient* – persons who received data from the sender

> **Note:**
>
> The icon next to the *Sender* or *Recipient* indicates that this is a new contact who appeared in the System on the date/time period specified in the ***Date*** filter.

- *Attachments* – whether the event has files attached, their name and size
- *Action/Description* – sent data description and action taken
- *Policies* – policies that triggered on this violation type
- *Protected objects* – protected objects that were sent
- *Source/ID in source* – event ID in Traffic Monitor



To get extended information on an event, select the required event and click . Together with the main data of the event, the System will display other data such as: messages in the chat, message text, attached images.

In event details there are highlighted words which are related to protected objects detected.

To get the required event sampling:

1. Configure filters in the Single filter panel (for details, see "Single filter").
2. In the list, select events using:
   - the *Ctrl* key – to select required events one by one;
   - the *Shift* key – to select a range of events (from – to).
3. Click **View events** to open the **Events** section of Traffic Monitor for detailed analysis. The section will display only 300 events of all events. If events belong to several sources, then events for each source will open in different browser tabs. To deselect the events, click **Reset selected events**.

Tip:

If you often switch between the Investigation Centre and Traffic Monitor, we recommend configuring pass-through authentication, which allows you to avoid re-entering your username and password. For details on how to configure it, see "*InfoWatch Traffic Monitor. Administration Guide*", section "*Configuring Single Sign-On in InfoWatch products*".

Note:

Table columns with events are displayed according to the role assigned to the user. So it is necessary to grant the user required access rights:

- to view the *Date*, *Event type*, *Sender*, *Recipient* columns, select the **Events** and **Violation statistics** checkmarks on the **Functions** tab of the role settings,
- to view *Description*, *Attachments*, *Source*, *Event ID* columns, enable **Event details**,
- to view *Violation level*, *User decision*, *Verdict* columns, enable **Violation details**,
- to view *Policies*, *Protected objects*, *Perimeters* columns, enable **Policies**, **Protected objects** and **Perimeters**.

User actions:

- Viewing and Exporting Event Data
- Viewing Person's Dossier
- Filtering events by persons and contacts
- Adding Events to the Investigation

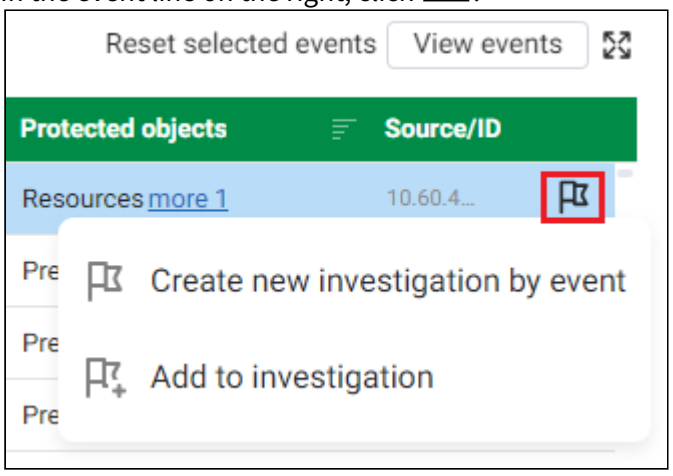
7.1 Adding Events to the Investigation

You can add any events from the list to an investigation or creare a new investigation based on the events.

To add an event to an investigation:

1. Click on the required event in the list.

2. In the event line on the right, click .



3. Choose:
   - **Add to investigation** – start typing the name of the investigation to which you want to add the event. Then choose the investigation from the drop-down list and click **Add**,
     OR
   - **Create new investigation** – a new investigation will be created automatically.

To add one or more events to an investigation:

1. Select required events using *Ctrl* or *Shift* keys.
2. On the right side of the screen above the event table, click:
   - **Add to investigation** – start typing the name of the investigation to which you want to add the events. Then choose the investigation from the drop-down list and click **Add**,
     OR
   - **Create new investigation** – a new investigation will be created automatically.

Events that are involved in investigations are marked with . Such events can be added to new or already existing investigations.

To switch to an investigation from the event:

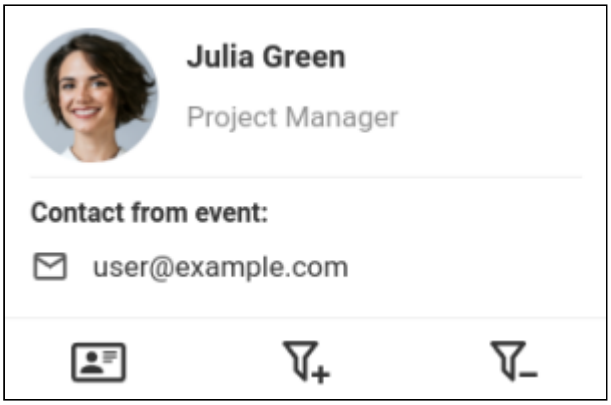
1. Click .
2. Select an event from the list. It will open on a new browser tab.

To view events that are included in investigations, use the **Investigations** filter (for details, see "Single Filter").

7.2 Viewing and Exporting Event Data

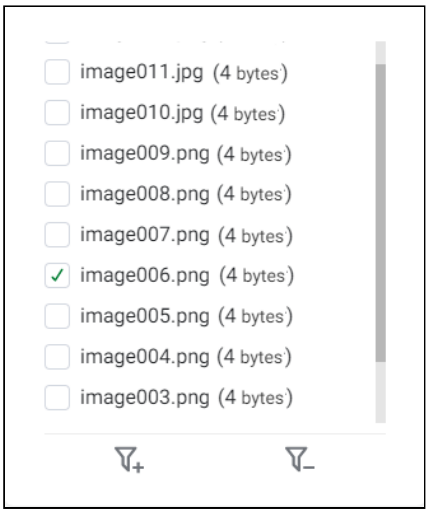
7.2.1 Viewing used contacts

The person card in the **Events** section contains contacts of employees who were involved in the selected event:



7.2.2 Viewing attachment data

To view the attachment details for the required event, click **more**. You will see a list with all attachments of the event:



You can add one or more attachments to the filter. The list will include events with this attachment only.

> **Note:**
>
> For correct operation, you should allow pop-up windows from specified sources to open in the browser.

7.2.3 Exporting event data

Export event data if you need to report an incident, conduct an investigation or to make a compilation of events. You can export event data table to Excel or save it to the dossier. To export data:

1. In the **Events** section, set values of the filter attributes. The event list will be updated.
2. At the top right of the screen, click .
3. Select **Unload report**.
4. Specify the report content and click **Save**.
5. Wait for the report to be generated.

6. Go to the **Reports archive** tab of the **Reports** section and download the report XLSX file by clicking the link.
7. Open the XLSX file. The **Additional information** sheet contains the date and time the report was created, applied filters and their values, as well as the link to the section where the report was created. The **Event list** sheet contains the event table. You can edit the table manually if necessary.

To set up regular report sending:

1. Configure the filter (see "Single Filter") and sorting of entries.

   > **Warning!**
   >
   > If the table contains more than 10,000 entries after filtering, only the first 10,000 entries are added to the report.

   The data in the configured view will be exported to the report.
2. In the upper right corner of the screen, click .
3. Select **Set up regular sending**.
4. Specify the report content.
5. Specify the sending parameters:

Parameter	Description
Report creation period	Specify the period for which to create a report: • **Last 24 hours** • **Current day** • **Last 3 days** • **Last 7 days** • **Last 30 days**
Recipients	Specify the report recipients
Sending frequency	Specify, with what frequency should Officer receive the report: • **Every day** • **Every week** You can select days of the week when Data Discovery will send the reports. • **Every month** Report will be sent on the 1st day of the month following the reporting one
Time	Specify the time of sending report

6. Click **Save**. A task for regular sending will be created. You can view the task list in the **Reports** section, on the **Tasks for regular sending** tab. For details, see "Tasks for Regular Sending".

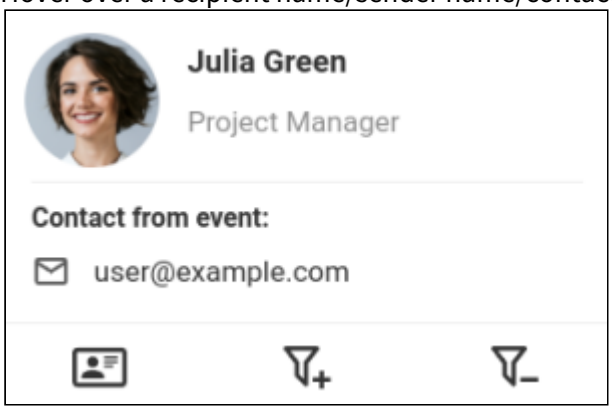
Important!

For regular sending reports, you should configure the connection with the mail server in Investigation Centre (for details, see "*InfoWatch Data Discovery. Administration Guide*", the "Notification Services" article).

7.3 Filtering events by persons and contacts

To filter events by required recipients or senders:

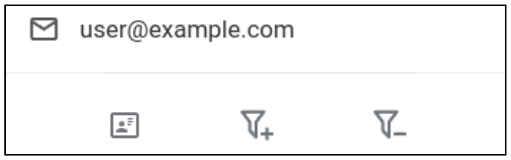
1. Hover over a recipient name/sender name/contact.



2. Click either of the following:
   -  – the recipient/sender will be added as an included attribute of the **Person** filter, and then the list of events will be updated;
   -  – the recipient/sender will be added as an excluded attribute of the **Person** filter, and then the list of events will be updated.

To filter events by undefined contacts:

1. Hover over an undefined contact in the event:



2. Click either of the following:
   -  – the contact will be added as an included attribute of the **Contacts** filter, and then the list of events will be updated;
   -  – the contact will be added as an excluded attribute of the **Contacts** filter, and then the list of events will be updated.

To assign an unidentified contact to a person:

1. Hover over an undefined contact in the event.
2. Click .
3. If you need to create a new dossier with this contact:

a. Click **Create dossier**.
b. Enter all required data as described in "Creating, Editing and Deleting Dossier".

4. If you need to add the contact to an existing dossier:
   a. Click **Add to dossier**.
   b. Select the dossier of the required employee and click **Add**.

8 Analytics Section

The **Analytics** section provides the Security Officer with visual statistics in the form of tables, charts and graphs. These are built according to specified conditions of the Single Filter and based on all events received from Traffic Monitor. The section contains the following tabs:

- **Incident statistics** – statistics widgets (see "Statistics Widgets");
- **Connection graph** – the Vision Connection Graph (see *"InfoWatch Vision. User Guide"*);
- **Activity statistics** – the Activity Monitor widgets (see *"InfoWatch Activity Monitor. User Guide"*).

8.1 Statistics Widgets

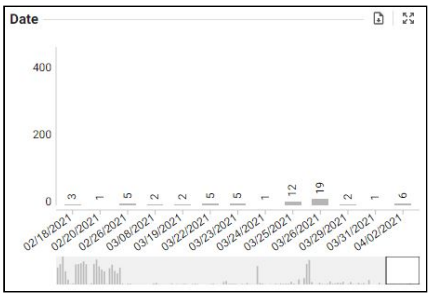
Widgets on the **Incident statistics** tab display the data from Traffic Monitor stored in the System.

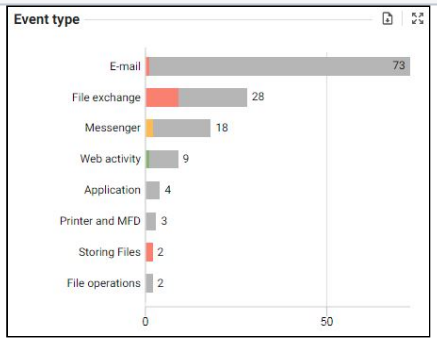
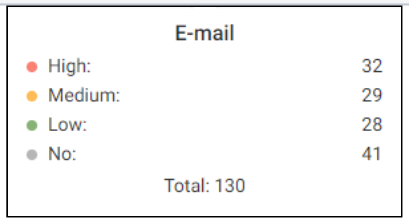
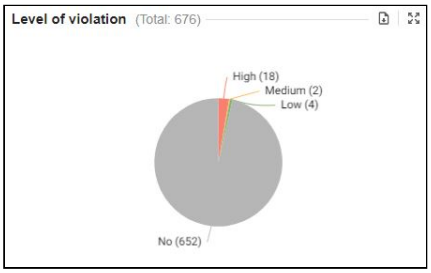
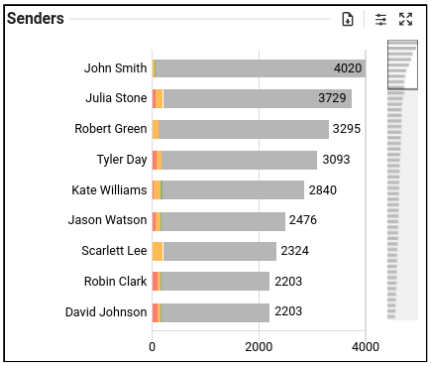
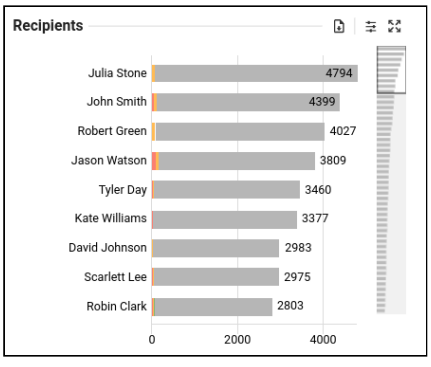
Note:

In the table below, you can find statistics widgets that are always available:

- Date
- Event type
- Level of violation
- Senders
- Recipients

Other statistics widgets are available only if Vision is installed in the cluster.

Widget name	Description
Date 	Displays the events distribution by specified dates according to violation levels. You can specify one or more dates and filter events for the selected day
Event type	Displays the distribution of Traffic Monitor events with different violation levels by types. **To get extended information:** When hovering over a source name, detailed information on the number of events with different violation levels is displayed:

Widget name	Description
	
Level of violation 	Displays the distribution of events by violation level
Senders 	Displays the list of persons with the highest number of violations. The first 50 top violators sorted by events with highest violation levels are displayed. **To customize the widget display:** 1. In the corner of the widget, click ⇄. 2. In the window that opens, specify: a. the number of senders to be displayed: from *50* to *200*; b. how to sort them: *by violation level* or *by the total number of events* 3. Click **Save**.
Recipients 	Displays a list of recipients, identified or unidentified, who are involved in events. The first 50 most popular contacts are displayed. **To add a contact from the event to the filter:** 1. Select a contact in the event with a mouse pointer. 2. Click ⊽+. The selected contact will be added to the Single filter, and the data on the events will be rebuilt accordingly. To exclude a contact from filter, click ⊽−. **To add an unknown contact to a dossier:** 1. Hover over an unknown contact in the diagram.

Widget name	Description
	2. Click . 3. Enter all required data as described in "Creating, Editing and Deleting Dossier".

8.2 Exporting Reports with Statistics Data

You can export data on widgets as a report as well as you can set up regular report sending.

To create and export a report:

1. In the **Events** section, specify all required values of the filter attributes. The events list will be refreshed.
2. In the right upper corner of the screen, click .
3. Select **Unload report**.
4. In the **Report unloading** window, you can choose which data to include to the **Detailed** report, which contains data from widgets in the form of tables.
5. If you want all widgets to be included in the report, set  next to the **Pages to unload** option.
   If you need only specific widgets to be included in the report, set  next to their names.
6. Wait for the report to be generated.
7. Go to the **Reports archive** tab of the **Reports** section and download the report XLSX file by clicking the link.
8. Open the XLSX file. On the **Additional information** sheet, you can find the date and time when the report was generated, applied filters and the link to open the section where the report was created. The **Event list** sheet contains the event table. You can edit the table manually if necessary.

To set up regular report sending:

1. Configure the filter (see "Single Filter") and sorting of entries.

   > **Warning!**
   >
   > If the table contains more than 10,000 entries after filtering, only the first 10,000 entries are added to the report.

   The data in the configured view will be exported to the report.
2. In the upper right corner of the screen, click .
3. Select **Set up regular sending**.
4. Specify the report content.
5. Specify the sending parameters:

Parameter	Description
Report creation period	Specify the period for which to create a report: • **Last 24 hours** • **Current day**

Parameter	Description
	• **Last 3 days** • **Last 7 days** • **Last 30 days**
Recipients	Specify the report recipients
Sending frequency	Specify, with what frequency should Officer receive the report: • **Every day** • **Every week** You can select days of the week when Data Discovery will send the reports. • **Every month** Report will be sent on the 1st day of the month following the reporting one
Time	Specify the time of sending report

6. Click **Save**. A task for regular sending will be created. You can view the task list in the **Reports** section, on the **Tasks for regular sending** tab. For details, see "Tasks for Regular Sending".

Important!

For regular sending reports, you should configure the connection with the mail server in Investigation Centre (for details, see "*InfoWatch Data Discovery. Administration Guide*", the "Notification Services" article).

9 File Sharing Section

The **File Sharing** section contains features of InfoWatch Data Discovery. Here, the user can create and run scanning tasks and then view the results: information about files, duplicates and similar ones, names of file owners.

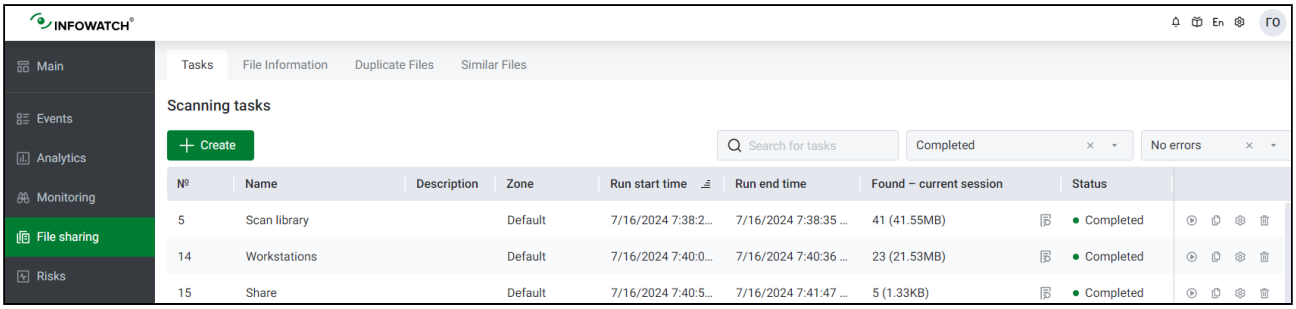
The section has the following tabs:

- **Tasks** – creating and running scanning tasks, viewing task execution statistics
- **File Information** – tables with file data that was received as a result of scanning resources
- **Duplicate Files** – results of automatic search for duplicate files, by hash or by size
- **Similar Files** – results of automatic search for documents with the same terms (similar files) using Data Analysis Service

9.1 "Tasks" Tab

The tab contains a list of scanning tasks and detailed information on each task.

For information on scanning tasks , see "Scanning task".



9.1.1 Actions available in the section

In this section, the following main actions are available:

- Creating, configuring and running tasks (see "Creating, Configuring and Running Tasks"):
  - Adding and configuring hosts (see "Adding and Configuring Hosts")
  - Adding scanning paths (see "Adding Scanning Paths")
- Editing tasks and hosts (see "Editing Tasks and Hosts")
- Copying tasks (see "Copying Tasks")
- Deleting tasks, hosts and scanning paths (see "Deleting Tasks, Hosts and Scanning Paths")
- Viewing statistics and task details (see "Task Execution Statistics")

9.1.2 Scanning task

A *scanning task* is an operation that checks if resources (scanning paths) contain confidential data. Two modes of task execution are available (see "Creating, Configuring and Running Tasks", the **Task execution** parameter):

- Continuous execution. The System scans resources in cycles until it is stopped manually by the user or automatically according to the schedule.
- One-time execution. The task stops automatically if one pass through each resource has been performed.

When the task is started again, the System does not scan all resources anew. Processing resources continues from where it ended during the previous scanning.

A host is a workstation/server, to which the System connects to scan files. The host contains a list of resources called *Scanning paths*. When creating a scanning task, you should add hosts to the task and for each host you should add scanning paths (for details on creating and configuring tasks, see "Creating, Configuring and Running Tasks").

A *pass* is a complete scanning of a single resource. A pass may consist of several sessions.

A *session* is a connection to a resource while the System passes through it.

For more information about resource processing, see "Resource Scanning". For information about events that the System sends to Traffic Monitor, see "Traffic Monitor Event".

Resource Scanning

Types of resources

After you create a task, you can scan:

- Shared network resources and distributed file systems (DFS) via SMB protocol:
  - SMB shares on MS Windows hosts
  - SMB shares on Linux hosts
- File systems of servers and workstations running Linux via SSH protocol
- SharePoint Server 2016 document libraries via HTTPS
- File shares via FTP protocol
- Nextcloud storages via WebDAV protocol

> **Note:**
>
> Data Discovery does not validate the server certificate when connecting to SharePoint.
>
> Scanning Nextcloud version 25 and later is supported.
>
> When connecting via WebDAV protocol, TLS/SSL certificates are not checked.

Account for scanning

The account Data Discovery uses for scanning resources must have the following permissions:

- The permission to read a network share (for SMB connection)
- The permission to read directories
- The permission to list the contents of directories
- The permission to read files

Parameter values for the account are specified in the task settings and in the host settings (for details, see "Creating, Configuring and Running Tasks").

Processing files within a task

During the first pass through the resource, Data Discovery performs the following actions:

- Collects file metadata and adds it to the file information table (see "File Information" Tab).
- If in the **Processing mode** block, you have selected at least one of the modes - creates a file copy in the task storage.

- If the **Send files to Traffic Monitor for analysis** mode is selected - sends events to Traffic Monitor for analysis. Each event consists of a file copy and information about the file.
- If the **Identify similar files** mode is selected - extracts text contents and sends them to Data Analysis Service for analysis.

> **Note:**
>
> If a file size exceeds 4 GB, Data Discovery collects only metadata of the file.

When Data Discovery connects to the resource again, it processes only new and changed files. To detect changed files, Data Discovery checks the file metadata for changes:

- Path to the file
- File name
- File size
- Date and time the file was modified
- File owner
- Users/groups that have access to the file

> **Warning!**
>
> Checking the following metadata for changes is supported only when scanning SMB shares on hosts running Windows:
>
> - File owner
> - Users/groups that have access to the file

If the file is new or changed, Data Discovery:

- Adds file metadata to the file information table (see "File Information" Tab).
- If in the **Processing mode** block, you have selected at least one of the modes - creates a file copy in the task storage.
- If the **Send files to Traffic Monitor for analysis** mode is selected - sends events to Traffic Monitor for analysis. Each event consists of a file copy and information about the file.
- If the **Identify similar files** mode is selected - extracts text contents and sends them to Data Analysis Service for analysis.

> **Note:**
>
> If in advanced settings of the task, you have enabled the **Verify unchanged files** parameter, then every pass through the resource, Data Discovery downloads a file and checks its contents for changes (for details on the parameter, see "Creating, Configuring and Running Tasks"). If the content of the file has changed, the file is sent as part of the event to Traffic Monitor.
>
> If the file size exceeds 4 GB, Data Discovery collects only metadata of the file.

Scanning errors

Errors may occur during scanning, for example:

- Data Discovery cannot connect to a resource;
- used account does not have permissions to process a directory/file.

If Data Discovery cannot connect to a resource, it moves on to process another resource. Within the task, Data Discovery will periodically try to connect to this resource.

If Data Discovery cannot process a directory/file, Data Discovery will skip it. Data Discovery will try to process the directory/file again the next time it connects to the resource.

Traffic Monitor Event

Data Discovery sends events with information about files to Traffic Monitor. File copies are sent as attachments to events.

> **Note:**
>
> The event type and interceptor specified in Traffic Monitor events depend on the TM version:
>
> - For Traffic Monitor version 7.3 and lower, the event type is *Crawler*, and the interceptor is *Crawler;*
> - For Traffic Monitor versions above 7.3, the event type is *Data Discovery* and the interceptor is *Data Discovery*.

Traffic Monitor applies a special anti-garbage filter to Data Discovery events. You can disable this filter and select the desired mode. For details, see "*InfoWatch Data Discovery. Administration Guide*", the "*Settings for automatic deletion of Data Discovery events in Traffic Monitor*".

Attributes of events that are created by Data Discovery and then sent to Traffic Monitor:

Event attribute	Description
Senders	Data Discovery sends the SID of the file owner, for example: `S-1-5-21-1922741149-3613679056-2300624754-1000` The System retrieves the value from the file system of the resource. **Note:** Traffic Monitor can identify a person by the received SID
Recipients	Data Discovery sends a list of SIDs of users and groups that have access to files. The System retrieves the value from the file system of the resource. **Note:** Traffic Monitor can identify a person or a group by the received SID
Computer	Host IP address or domain name of the host where the file was found. **Note:** Traffic Monitor can identify a workstation by received information about the host

Event attribute	Description
Task name	Name of Data Discovery scanning task within which the file was downloaded from the resource
Task start date	Start date and time of the session within which the file was downloaded from the resource
File creation date	Date and time the file was created on the host. The System retrieves the value from the file system of the resource. **Note:** Determining the file creation date is not supported if: • SMB shares are scanned on hosts running Linux • hosts running Linux are scanned via SSH protocol In these cases, the *File Modification Date* attribute value is used as the value of the *File Creation Date* attribute
File modification date	Date and time when the file was modified on the host. The System retrieves the value from the file system of the resource
Path to file or receiver address	Full path to the file on the host **Note**: Full path includes the scanning path of the resource. For files from SharePoint hosts, the full path also includes the site path.
Interception date	Date and time when the file was downloaded by the System
Interception server	Data Discovery server that processed the resource

Warning!

Obtaining the SID of the owner of a file and the SID of users/groups that have access to a file is supported for:

- SMB shares on hosts running Windows operating systems;
- SharePoint Server 2016 document libraries.

The *Senders* and *Recipients* attributes can be filled incorrectly if:

- an SMB share is scanned on a Linux host;
- an SMB share being scanned has been created on an SMB server that does not support file attributes to determine the owner of a file and users/groups that have access to a file;
- an SMB being scanned has been created on a file system that does not support file attributes to determine the owner of a file and users/groups that have access to a file.

Getting the SID of the file owner and SID of users/groups that have access to a file is not supported when scanning hosts:

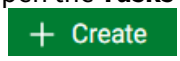
- Running Linux via SSH protocol
- Via FTP protocol
- Via WebDAV protocol

In this case, the *Senders* and *Recipients* fields have "iw data discovery" value.

9.1.3 Creating, Configuring and Running Tasks

Creating a task

To create a new scanning task:

1. In the **File Sharing** section, open the **Tasks** section to see the scanning task list.
2. In the left upper corner, click 

.
3. In the window that opens, configure the task and click **Save** (see "Task settings").
4. Add hosts to the task (see "Adding and Configuring Hosts").
5. For hosts, add scanning paths (see "Adding Scanning Paths").

Warning!

You cannot start a task if there are no scanning paths in the task.

See also "Examples of Task Settings, Host Settings and Adding Scanning Paths".

Starting a scanning task

A new task can have one of the following statuses:

- **New**, if the task was not configured to run automatically according to a schedule;
- **Awaiting scheduled launch**, if the task was configured to run automatically according to a schedule.

To start a task, in the scanning task list, click  in the **Actions** menu. The task status will change to **Running**, and the start button will change to . Click  to stop the task.

Warning!

You can start a scanning task only if there is an active license in the System (see "*InfoWatch Data Discovery. Administration Guide*", "*Licensing*").

You can find more details about started tasks by viewing statistics (see "Task Execution Statistics").

Task settings

> **Note:**
>
> Required fields are marked with * icon.

The following settings are available:

1. The **General** tab:
    a. Main task settings:

Parameter	Description
Name	Task name
Description	Short description of the task
Task execution	Task execution mode. **Continuous**: • The System scans resources in cycles until it is stopped manually by the user or automatically according to the schedule. • When the task is started again, the System does not scan all resources anew. Processing resources continues from where it ended during the previous scanning. **One-time**: • The System performs one pass through each resource, and then the task is considered completed. The task automatically stops and gets the **Completed** status. • If the task that has the **Completed** status is started, the System: 1. processes new and changed files on each resource 2. stops the task and sets the **Completed** status • A task gets the **Completed** status if the System has passed through each available resource. A task with the **Completed** status may contain resources with the **Error** status that the System was unable to process. • If the **Processing mode** is enabled, the task will get the **Completed** status: • when all the data has been sent to Traffic Monitor - for the **Send files to Traffic Monitor for analysis** mode;

Parameter	Description
	• when the notification that there is data for processing has been sent to Data Analysis Service - for the **Identify similar files** mode. • The task can be started and stopped according to a schedule. The task stops if the System has completed one pass through each resource or if it is time to automatically stop according to the schedule. **Warning**: In both cases the task gets the **Awaiting scheduled launch** status.
Processing mode	Mode of processing files. Within a scanning task Data Doscovery collects file metadata and adds it to the file information table. You can view all the detailed information about files on the "File Information" Tab. As well, Data Discovery creates file copies for analysis. For this, you should select at least one of the processing modes: • **Send files to Traffic Monitor for analysis** - to send file copies for analysis, in the **Settings** → **Products settings** → **Zones and servers** section, you should specify Traffic Monitor server available for connection. • **Identify similar files** - for extracting text contents and searching similar files, Data Discovery uses InfoWatch Data Analysis Service (hereinafter - DAS). About installing and configuring DAS, see "*InfoWatch Data Analysis Service. Installation, Configuration and Administration Guide*". To make the mode available: • DAS version 1.2.1 or later must be installed in one cluster with Data Discovery; • protoclusterization functionality must be enabled in DAS; • DAS must be configured for processing of the node, on which Data Discovery is installed. If you do not select any processing mode, Data Discovery collects only file metadata without creating file copies. **Note**: For a file to be processed for identifying the similar ones, the extracted **text content** must be at least 200 bytes. The content with size less than 200 bytes will not be sent for analysis to DAS but will be available for viewing in Investigation Centre web interface, in the **File sharing** section.

b. The **Authorization to access resources** section:

Parameter	Description
By login and password	

SMB group	Domain or group The parameter is used for hosts with SMB connection
Login	Username **Warning**: For connection to SharePoint, specify the value in the format: domain\user
Password	Password
By SSH key (only for SSH connection type)	
Login	Name of the user that has obtained the public key
Passphrase	Password for the encrypted private key
Private key	Private key. You can specify it manually or upload from the file using the  button. See the example of using SSH keys in the "Using SSH keys for access to resources" article

Note:

If authorization parameter values are not specified, then by default anonymous access will be used in the task to connect to SMB hosts.

To speed up scanning of Nextcloud storage, use the application password to connect. For details, see the note for the paragraph Third-party WebDAV clients of Nextcloud official documentation.

2. The **Files filter** tab – filters by file size and file name:
   a. The **Size** filter allows you to set file size limits. If the file size matches the condition set in the filter, the file will be scanned:
      - In the **From** field, enter the minimum file size. You cannot specify a value greater than 4 GB.
      - In the **To** field, enter the maximum file size. You cannot specify a value greater than 4 GB.
      - To the right of each of the specified filter fields, the current unit of measurement is displayed. To change the unit of measurement for a field, click on the unit of measurement and select the desired value from the drop-down list.

        Note:

If only the minimum file size is specified, then files whose size is greater than or equal to the minimum size will be scanned.
If only the maximum file size is specified, then files whose size does not exceed the maximum size will be scanned.

Note:

If a file size exceeds 4 GB, the System does not download the file and does not send an event to Traffic Monitor. The file entry is added to the file information table (see "File Information Table").

b. **List of added masks** contains file name masks:
   - The System scans a file if the file name matches a mask.

Warning!

If you do not specify any file masks, the System will scan files with any names, including system files. Sending a large number of files to the Traffic Monitor server can significantly increase the load on the Traffic Monitor.

   - Actions with masks:
     - **To add a mask**:
       i. Click on the **Add file mask** field. In the list that opens, preinstalled file categories and masks for formats are displayed.
       ii. Perform one of the following actions:
         - Select masks/categories in the list. To search, start typing a mask or a category name in the field.
         - Enter a mask in the field and select that mask in the list.
       iii. Click outside the list.
     - **To delete a mask/category**, click  next to the mask/category.
     - **To export the mask list to a file**, click .
     - You can import the mask list from a file. A .txt file is required, each mask must be specified on a separate line.
       To import the mask list:
       i. Click .
       ii. In the window that opens, select a file.

Note:

After the import, the task will contain only masks specified in the file. All other masks will be deleted.

   - Format of the file name masks:
     - To make a name mask, use the following characters:
       - "*****" means any sequence of characters, including no characters;
       - "**?**" means any single character.

Mask examples

***.doc** is for files whose names end with '.doc' (example.doc, report.doc)

report is for files whose names contain 'report' (new_report.txt, reports.pdf)

doc????.pdf is for files with name pattern: 'doc' + 4 characters + '.pdf' (doc2012.pdf, docs_22.pdf)

file.txt is for files with the 'file.txt' name

- File masks are not case-sensitive.

3. The **Folders filter** tab:
   - To exclude a directory and all its subdirectories from scanning, enter the full path to this directory in the **Exclude** field and click . The path to the directory will be added to the **List of excluded folders**.

     Note:

     By default, the **List of excluded folders** contains system directories for Windows administrative shares.

   - To include a directory and all its subdirectories in scanning, enter the full path to this directory in the **Include** field and click . The path to the directory will be added to the **List of included folders**. If the **List of included folders** does not contain any directory, then all directories will be scanned, except those excluded from scanning.
   - To remove a directory from the list of included/excluded directories, click  to the right of the directory in the list.

Warning!

Excluding directories from scanning has a higher priority than including directories in scanning. For example, if you exclude the `share/documents/folder` directory and include the `share/documents/folder/example` directory, then the `share/documents/folder/example` directory will not be processed.

- **Directory path format:**
  - A full path to the directory is a sequence of directory names separated by "/", which indicates the location of the directory in the file system.
  - For **SMB** resources, the full path to the directory must include the resource name. The full path must not begin with "/".
  - For **SSH** resources, the full path to the directory must include the scanning path. The full path to the directory must begin with "/".
  - For **SharePoint** resources, the full path to the directory must include the document library name. The full path to the directory must not begin with "/".
  - For **FTP** resources, the full path to the directory must include the scanning path. The full path to the directory must begin with "/".
  - For **Nextcloud** storages via **WebDAV** protocol, the full path to the directory must include the scanning path. The full path to the directory must begin with "/".
  - The directory filters are not case-sensitive.

Sample paths to directories

For SMB host:

- If you need to scan the `/folder/example` directory and all its subdirectories on a shared network resource `share`, add the `share/folder/example` path to the **List of included folders**.
- If you need the `/system/files` directory and all its subdirectories on a shared network resource `resource` to be excluded from scanning, add the `resource/system/files` path to the **List of excluded folders**.

For SSH host:

- If the scanning path on the SSH host is set as `/home/folder` and you need to scan the `/home/folder/example` directory and all its subdirectories, add the `/home/folder/example` path to the **List of included folders**.
- If the scanning path on the SSH host is set as `/home/documents` and you need to exclude the `/home/documents/old_files` directory and all its subdirectories from scanning, add the `/home/documents/old_files` path to the **List of excluded folders**.

For FTP and WebDAV protocols, the format for specifying directory paths is similar to SSH.

For SharePoint host:

- If you need to scan the `/folder/example` directory and all its subdirectories in the `library` document library, add the `library/folder/example` path to the **List of included folders**.
- If you need the `/outdated/reports` directory and all its subdirectories in the `documents` library to be excluded from scanning, add the `documents/outdated/reports` path to the **List of excluded folders**.

- **Masks for directory names:**
  When setting a full path to the directory, you can specify the directory name in mask format.
  To compose a name mask, use the following characters:
  - "*****" – means any sequence of characters, including no characters;
  - "**?**" – means any single character.

Warning!

The " ***** " and " **?** " characters can be used only within the name of one directory.
For example, the `share/folder/subfolder/example` path does not match the condition of the `share/*/example` mask, however, it matches the condition of the `share/*/*/example` mask.

Examples of paths using masks

Host type	Path example	Description	Directories that match the condition
SMB	`*/Example`	The **Example** directory and all its subdirectories if **Example** is in the root directory of the resource	`C$/Example` `share/example`
	`*/*/*/Example`	The **Example** directory and all its subdirectories if **Example** is located three levels below the root directory of the resource	`myshare/path/to/Example`
	`share/Doc*`	The directory and all its subdirectories if the directory is in the root directory of the **share** resource and its name begins with "**Doc**"	`share/Doc` `share/documents`
	`?$/Users/*/Documents`	The **Documents** directory and all its subdirectories for each user on a hidden resource	`C$/Users/Ivanov/Documents` `D$/Users/Petrov/Documents`
	`share/documents/*/*`	All subdirectories for the **documents** directory on the **share** resource, but not the **documents** directory itself. This type of paths is only supported for the include filter (**List of included folders**)	`share/documents/reports` `share/documents/templates`
Share Point	`*/Example`	The **Example** directory and all its subdirectories if **Example** is in the root directory of the document library	`my_library/example`
	`*/*/*/Example`	The **Example** directory and all its subdirectories if **Example** is located three levels below the root directory of the document library	`documents/2021/folder/Example`
	`Library/Doc*`	The directory and all its subdirectories if the directory is in the root	`Library/Documents`

Host type	Path example	Description	Directories that match the condition
		directory of the **Library** document library and its name begins with "**Doc**"	
	`storage/documents/*/*`	All subdirectories for the **documents** directory in the **storage** library, but not the **documents** directory itself. This type of paths is only supported for the include filter (**List of included folders**)	`storage/documents/reports` `storage/documents/templates`
SSH, FTP, WebDAV	`/home*`	The directory and all its subdirectories if the directory is in the root directory of the resource and its name begins with "**home**"	`/home` `/HomeFolder`
	`/home/*/*/Example`	The **Example** directory and all its subdirectories if **Example** is located three levels below the home directory	`/home/path/to/Example`
	`/home/*/Documents`	The **Documents** directory and all its subdirectories for each user on a workstation running Linux (for SSH host)	`/home/ivanov/Documents`
	`/home/*/*`	All subdirectories for the **home** directory, but not the **home** directory itself. This type of paths is only supported for the include filter (**List of included folders**)	`/home/username` `/home/folder`

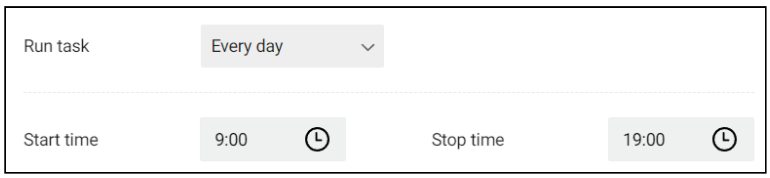
4. The **Schedule** tab – settings for automatic start and stop of the task:

> **Warning!**
>
> The task will not start automatically if there is no active license in the System (see "InfoWatch Data Discovery. Administration Guide", "Licensing").

- If you do not need to start and stop the task automatically, then in the **Run task** field, select **Manually**.

- If you want the task to start and stop automatically every day at the same time, then in the **Run task** field, select **Every day**. In the **Start time** and **Stop time** fields, specify the time H:mm or HH:mm format.
Daily Schedule Examples

1) The task starts automatically every day at 9:00 and stops at 19:00:

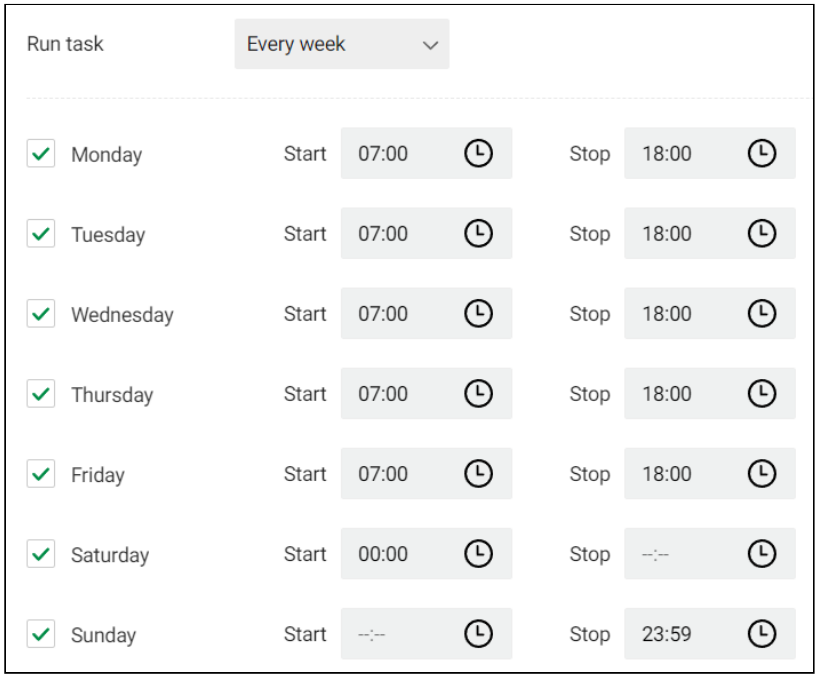


2) The task starts automatically every day at 18:00 and stops at 07:00 the next day:

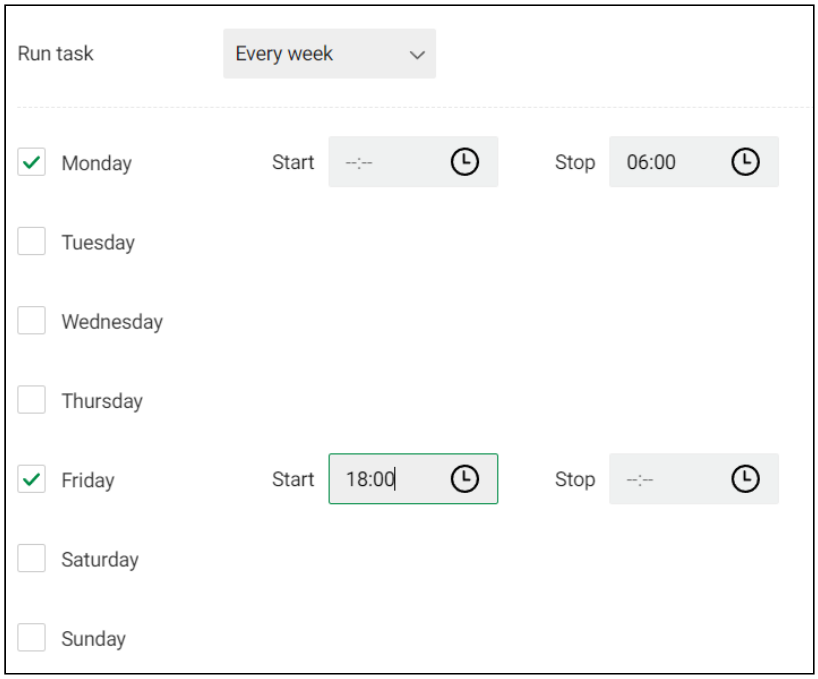


- If you want the task to start and stop automatically on different days of the week, then in the **Run task** field, select **Every week**. Check the box to the left of the day of the week on which you want to start or stop the task. In the **Start** and/or **Stop** field to the right of the selected day, enter the time H:mm or HH:mm format.
Weekly Schedule Examples

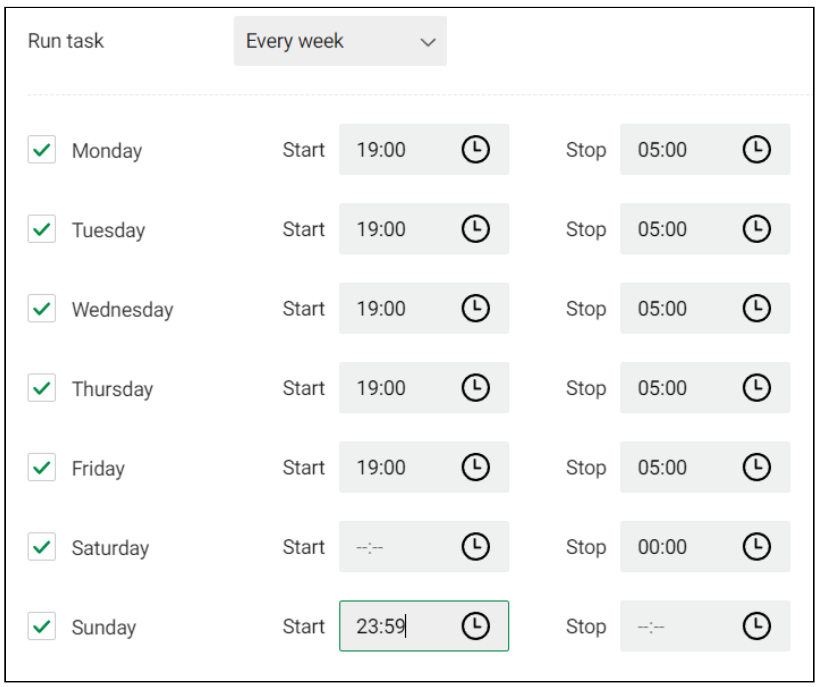
1) The task starts automatically every weekday at 07:00 and stops at 18:00. The task also starts at 00:00 on Saturday and stops at 23:59 on Sunday:



2) The task starts automatically on Friday at 18:00 and stops on Monday at 06:00:



3) The task is executed every weekday between 00:00 – 05:00 and 19:00 – 23:59:



Note:

When you save a schedule, the specified time is added to the System in the time zone of the workstation from which you connect to the Data Discovery web interface.

When viewing a schedule, the time is displayed in the time zone of the workstation from which you connect to the Data Discovery web interface.

Note:

You can also start and stop scheduled tasks manually. The automatic start and stop of the task will be on schedule.

5. The **Advanced settings** tab – infrastructure-specific task settings:

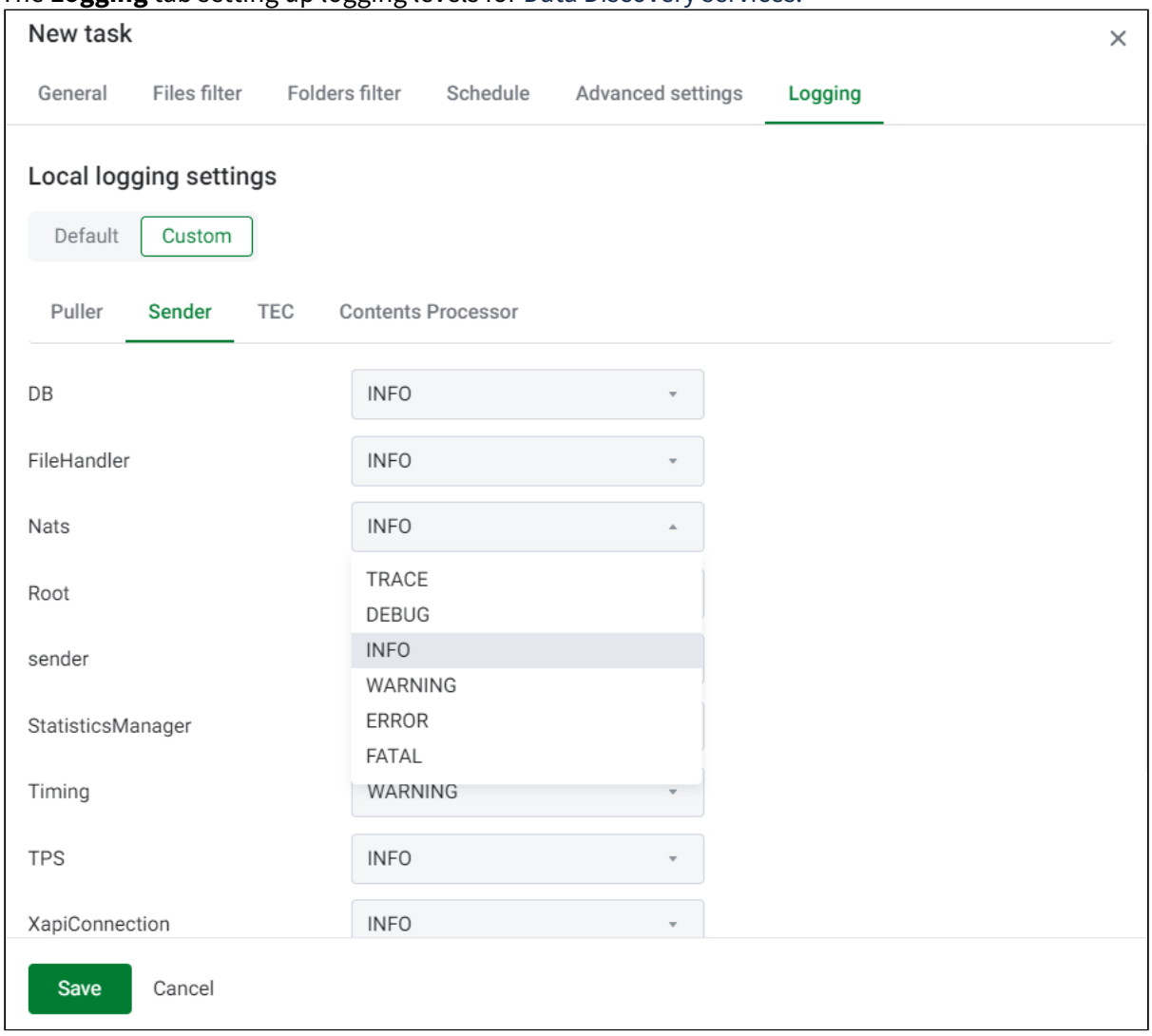
Parameter	Description
Pullers count	The number of scanners (instances of Puller service). Data Discovery uses Scanners for scanning task resources and downloading files from them. If the value **0** is set, the System: • does not scan files and does not download new files to storages; • sends files from storages to Traffic Monitor. **Note**: To scan one host, only one scanner is used within a task. We recommend not specifying more scanners than the number of hosts in the task. If a host has a large number of scanning paths, we recommend distributing the scanning paths to different tasks. For this: a. Create several tasks. b. Add a host to each task. c. In each task, specify several scanning paths for a host. Each scanning path should be specified in one task only. If the host is specified in several tasks, then it will be processed by several scanners. It may significantly increase the load on the host during scanning.
Senders count	The number of transmitters (instances of Sender service). Data Discovery uses Transmitters for sending data to Traffic Monitor. We recommend that you specify a value that is greater than or equal to the number of Data Discovery servers in the task zone. The parameter is available if the **Send files to Traffic Monitor for analysis** processing mode is enabled.
Extractors count	Count of the extractors. Each extractor is a pair of services: Ripper and Contents Processor. Data Discovery uses Extractors for extracting text contents from files, saving them to the file storage and notificating Data Analysis Service that there is data for processing. The parameter is available if the **Identify similar files** processing mode is enabled.

Parameter	Description
Processing interval (in seconds)	The pause after a resource pass is completed. The System rescans the resource only if the specified interval ends and the queue reaches the resource again. The parameter allows you to reduce the load on the resources
Processing depth (in days)	Size of the scanning window for the resource. If the file modification date is outside of the scanning window, the file will not be processed during the current session. The parameter is used to enable scanning resources evenly. The lower the value is, the more often the scanner will switch between resources. The parameter is available if at least one of processing modes is enabled. During the first scanning session, when the System passes through the resource, the scanning period is considered a period from *Newest* to *Newest − Depth*, where: • *Newest* is the latest modification date for files on the resource; • *Depth* is the number of days specified in the parameter. In each subsequent session, the scanning window is a period from *Last* to *Last − Depth*, where: • *Last* is the earliest modification date for files processed during the previous session; • *Depth* is the number of days specified in the parameter. New and modified files are processed in **each** session. **Example of using the parameter** There are four files on the resource: • File A that was modified on 28.05.2021; • File B that was modified on 03.05.2021; • File C that was modified on 15.04.2021; • File D that was modified on 19.03.2021. If the parameter value is 30 days, then: a. In the first session, the scanning window is a period from 28.05.2021 to 28.04.2021. The System processes files A and B. Then the scanner switches to another resource. b. In the second session, the scanning window is a period from 02.05.2021 to 02.04.2021. The System processes file C. Then the scanner switches to another resource.

Parameter	Description
	c. In the third session, the scanning window is a period from 14.04.2021 to 15.03.2021. The System processes file D and finishes passing though the resource. **Note:** If you specify a value of 0, Data Discovery will scan all files regardless of the date of change
Processing duration (in minutes)	Maximum session duration. The session includes getting a list of files that are stored on a resource and processing the files. When the time specified in this parameter expires, processing the resource stops and the scanner starts processing another resource. The parameter is used to enable scanning resources evenly. The lower the value is, the more often the scanner will switch between resources. **Note**: When the scanner returns to the resource whose processing was stopped as the maximum time expired, the processing continues where it was stopped in the previous session. If you specify a value of 0, Data Discovery will scan the resource until competion, without time limit. **Warning**: Each time the scanner connects to the resource, it gets a list of files that are stored on the resource. While getting a list of files, the resource has the **Getting info** status. Processing the resource starts after getting a list of files. If there are a lost of files on the resource, then getting a list of files may take some time. In some cases, getting a list of files may take all the time allotted for one session of processing the resource. In this case, files on the resource are not processed. We recommend increasing the value of the **Processing duration** parameter when scanning resources with a large number of files.
Configuring DNS	You can choose: • **Use system settings** - to use default values; • **Specify DNS addresses manually** - for example, if Data Discovery server is not in the same domain as the hosts you want to scan, specify DNS servers, using which host addresses will be defined. To add an additional field for entering the DNS server address, click  . You can specify up to three DNS server addresses.
SMB Authentication Type	For an SMB connection, you can select the authentication protocol:

Parameter	Description
	• **NTLM** - the default value; • **Kerberos.** For the **Kerberos** protocol you must specify: • **Realm name** - name of the Kerberos Realm in which the hosts reside, The realm name is case sensitive. • **KDC address** - address of the key distribution center (KDC, Key Distribution Center) for the specified realm.
Zone	Task zone. • The task services run on the servers that are added to the zone, • The System downloads files to the storages of the zone servers, • The task uses the zone settings for data export to Traffic Monitor. You can change the task zone only if the task has the **New** status. To change the zone in the other cases: i. In the task settings on the **Schedule** tab, select **Manually** in the **Run task** field and save the settings. ii. Copy the task. iii. Change the zone of the task copy. For more information on zones and servers, see "*InfoWatch Data Discovery. Administration Guide*", "*Zones and servers*".
Verify unchanged files	By default, the scanner checks only metadata for files that were downloaded during previous sessions. If the file metadata is not changed, then the file is not processed again. If the parameter is enabled ( ), then each time the scanner passes through the resource, it downloads files and checks their contents for changes. This allows you to detect files whose content has changed but the metadata has not changed. If the content of the file has changed, the file is sent as part of the event to Traffic Monitor. The parameter is available if at least one of processing modes is enabled. **Warning:** If the parameter is enabled ( ), then the volume of traffic and the time required for reprocessing the resource increases.

6. The **Logging** tab setting up logging levels for Data Discovery services.



The **Default** and **Custom** settings are available.
If you select **Custom** settings, you can set the logging level for each service component of the following services:

- **Puller**;
- **Sender**;
- **TEC** (Task Execution Controller):
- **Contents Processor**.

> **Note:**
>
> Data Discovery saves log files to the `/var/log/infowatch/disc`.

Logging level values:

- `trace` -the most detailed logging, including all messages with `debug`, `info` and higher levels,
- `debug` - output debugging information, usually for diagnosing errors and communicating with development/support,

- `info` - information messages, can be disabled on stable systems,
- `warning` - warnings;
- `error` - errors.

Adding and Configuring Hosts

> **Warning!**
>
> Selecting hosts from LDAP directories and importing host list are only available for hosts with SMB connection. When you add a host, the **Connection type** default value is **SMB**. Changing the value of this parameter is not available when editing added hosts.
>
> The **Inherit from task** parameter is enabled by default for all hosts.

To add a hosts to a task:

1. Click on the task to open the host list.
2. Do either of the following:
   - Click  to add a host and to specify its address manually.
     In the window that opens, configure the host and click **Save** (see "Host settings").
   - Click  to add hosts from a LDAP directory.
     The window that opens displays data from LDAP directories that was received by the System as a result of synchronization with Active Directory. If you delete an integration, the data of this integration will not be displayed in the window. For more information about adding data to the System, see "*InfoWatch Data Discovery. Administration Guide*", "Integrations".

     > **Note:**
     >
     > If the System is installed together with other Platform products, the window may also display data that was received as a result of synchronization with Traffic Monitor.

     The data in the window is presented in the form of a tree. The following icons are used for LDAP directory items:
     -  for a container,
     -  for a group,
     -  for a computer.

     To expand a tree item, click  to the left of the item. To search for items by name, enter the text in the search bar.

     > **Warning!**

The search is performed only by the data displayed in the window. Before you search, it is recommended that you expand the items in the tree and make sure that the items are loaded and displayed in the window.

To add computers from a directory to the task:

a. In the tree, select computers and/or items that include computers.
b. Click **Add**.

When a computer is added to a task as a host, the fully qualified domain name of the computer is automatically specified in the **Address** parameter of the host.

- Click  to **add hosts from a JSON file**.
  In the window that opens, select a JSON file with the description of hosts.
  For hosts to be added correctly:
  - maximum allowed size of a JSON file is **1 MB**;
  - the file contents must match the format (for details, see "JSON File Format for Importing Hosts").

Note:

You can export hosts to JSON file in order to import them to another Data Discovery server.

To export hosts, click . The browser will save the JSON file with a description of hosts to your disk.

Features of exporting hosts:

- Filters do not affect the export of hosts.
- If you select hosts in the list, the System will export only selected hosts.
  If you haven't selected any host, the System will export all hosts in the task.
- Exported descriptions of hosts **do not include passwords** of user accounts under which Data Discovery connects to resources for scanning.
  To scan hosts that have been imported from a JSON file, you need to specify passwords. You can do this in one of the following ways:
  - Before importing, add the `password` or `ssh_key` field to a description file for each host and specify the field value (for details, see "JSON File Format for Importing Hosts").
  - After importing, you can add passwords in the **Authorization parameters** section. You can find this section in the description of hosts in Data Discovery web interface.

Added hosts are displayed in the host list.

At the top of the host list screen, you can see controls to manage the task. Click  to return to the task list.

Host settings:

Note:

Required fields are marked with *.

1. The **General** tab:
   a. Main settings of the host:

Parameter	Description
Address	Domain name or IP address. The System gets IP addresses for all the domain names: • at task start; • regularly during the task: 30 seconds after all the addresses were received. **Note**: For SharePoint connection, you need to specify the address where the SharePoint web application is available. In some cases, the SharePoint web application can be accessed only if the address is specified in the domain name format.
Port	Connection port If the connection port is not specified, then the default value is used: • for SMB connection – 445 • for SSH connection – 22 • for FTP connection – 21 • for WebDAV connection – 8080 This field is required for SharePoint connection.
Connection type	Possible values: • SMB – for scanning shared network resources • SSH – for scanning servers and workstations running Linux • SharePoint – for scanning SharePoint Server 2016 document libraries via HTTPS • FTP – for scanning file shares via FTP protocol • WebDAV (Nextcloud) – for scanning Nextcloud data storages using WebDAV protocol Changing this parameter is not available when editing a host.
Site path	The path within SharePoint to the site. The site must contain the document libraries you want to scan. The parameter is available if SharePoint connection is selected for the host. **Example**: The **documents** library has the address: `sp.example.com/my_path/site/documents`. In this case the site path is `my_path/site`
Protocol	The parameter is available for FTP and WebDAV (Nextcloud) connection types. Possible values:

Parameter	Description
	• **For FTP:** • FTP • FTPS (Explicit FTP over TLS) • **For WebDAV (Nextcloud):** • HTTPS • HTTP
Connection type	To connect to Nextcloud via WebDAV protocol, paste a part of the link from the repository into the *Connection path* field (for details, see "Getting a Link to Connect to Nextcloud") The parameter is available if the connection type for the host is WebDAV (Nextcloud)
Description	Short description of the host

Note:

When connecting via the FTPS (Explicit FTP over TLS) protocol, certificates of authenticity are not checked.

When connecting via the WebDAV protocol, TLS/SSL certificates are not checked.

b. The **Authorization parameters** section:

Parameter	Description
Inherit from task	If the parameter is enabled ( ), then the authorization parameters specified in the task settings are used to connect to the host. To specify other values, disable this parameter. When adding a host, the parameter is enabled by default.
By login and password	
SMB group	Domain or group The parameter is available if SMB connection is selected for the host
Login	User name **Warning**: For connection to SharePoint, specify the value in the format: domain\user
Password	Password
By SSH key (only for SSH connection type)	

Login	Name of the user that has obtained the public key
Passphrase	Password for the encrypted private key
Private key	Private key. You can specify it manually or upload from the file using the  button. See the example of using SSH keys in the "Using SSH keys for access to resources" article

> **Note:**
>
> In the following cases, anonymous access is used to connect to an SMB host:
>
> - If the **Inherit from task** parameter is disabled and authorization parameters for the host are not specified
> - If the **Inherit from task** parameter is enabled and authorization parameters are not specified in the task settings

2. The **Advanced settings** tab:

Parameter	Description
Timeout (msec)	Host connection timeout. For SharePoint hosts, it is not recommended to specify a value less than **60000** (60 seconds). If errors like "Connection timed out / Connection timeout" occur during resource scanning, it is recommended to increase the value of the parameter. For details about resource statuses, see "Task Execution Statistics".
Retry attempts	Maximum number of connection attempts
Debug level	The parameter is responsible for presenting debug information. **For SMB connection:** For troubleshooting, the value **2** is recommended. For details, see smbclient documentation (the `debuglevel` parameter). **For SSH connection:** The following values are recommended for troubleshooting: • **8** – authentication information; • **64** – SFTP debug information; • **128** – information about errors; • **512** – information about working with sockets. To get debug information for several values at the same time, enter the sum of the values. For example, to get SFTP debug information (parameter value: **64**)

Parameter	Description
	and information about errors (parameter value: **128**), in the **Debug level** field, specify **192** (64+128). For details, see libssh2 documentation. **For SharePoint connection:** The following values are recommended for troubleshooting: • **1** – socket information • **2** – information about HTTPS requests and responses • **8** – HTTP authentication information • **16** – HTTP authentication information (plaintext) • **128** – information about HTTP-response body • **256** – information about SSL/TLS To get debug information for several values at the same time, enter the sum of the values. For example, to get authentication information (parameter values: **8** and **16**) and information about SSL/TLS (parameter value: **256**), in the **Debug level** field, specify **280** (8+16+256). **For FTP or WebDAV connection**: We recommend using the following values for troubleshooting: **0** – general information about query execution and errors when executing queries **1** – additional data received from the host **2** – additional data sent to the host **3** – data from executed queries. If encryption was used, data will be presented in undeciphered form **4** – data for queries that were sent to the host **5** – SSL/TLS data sent to the host **6** – SSL/TLS received from the host For FTP and WebDAV, debug information can be received for only one value at a time.
Pause before retry (sec)	Time interval between connection attempts
Chunk size (bytes)	Fragment size of the transmitted data

Using SSH keys for access to resources

Data Discovery can connect to the scanning resources using secure way of connection with SSH keys.

Important!

Consider that key generation specifics depend on used OS settings.

Below is an example of generation of keys in Red Hat Enterprise Linux 7.9 OS. In this OS, the RSA encryption algorythm is used by default.

To use authorization with SSH keys when scanning, complete the steps:

1. Generate a pair of SSH keys by running the command:

```
ssh-keygen
```

> You can generate them directly on the Data Discovery server or on the other computer.
>
> If necessary, you can use another tool or specify your preferred encryption algorythm.

2. You can specify the name of the file with keys or just press **Enter** to create files with the default names.
3. Next, you will be asked to specify the key phrase (passphrase) for using an encrypted key.
   If you do not want to use the encrypted key, press **Enter** without entering the key phrase.
4. If you used the default filenames, two files with keys will be created:
   - `/<user_name>/.ssh/id_rsa` – authorization private key;
   - `/<user_name>/.ssh/id_rsa.pub` – public key.
5. Next, you should transfer the public key (`id_rsa.pub`) **to all hosts**, on which Data Discovery will authorize using SSH keys.
   For this, use the `ssh-copy-id` tool by running the following command:

```
ssh-copy-id <host_user_name>@<host_address>
```

Example of a command of public key transferring:

```
ssh-copy-id ERipley@10.70.12.34
```

> **Note:**
>
> After transferring public key to the host, it is highly recommended to check the possibility of authorization with this key.

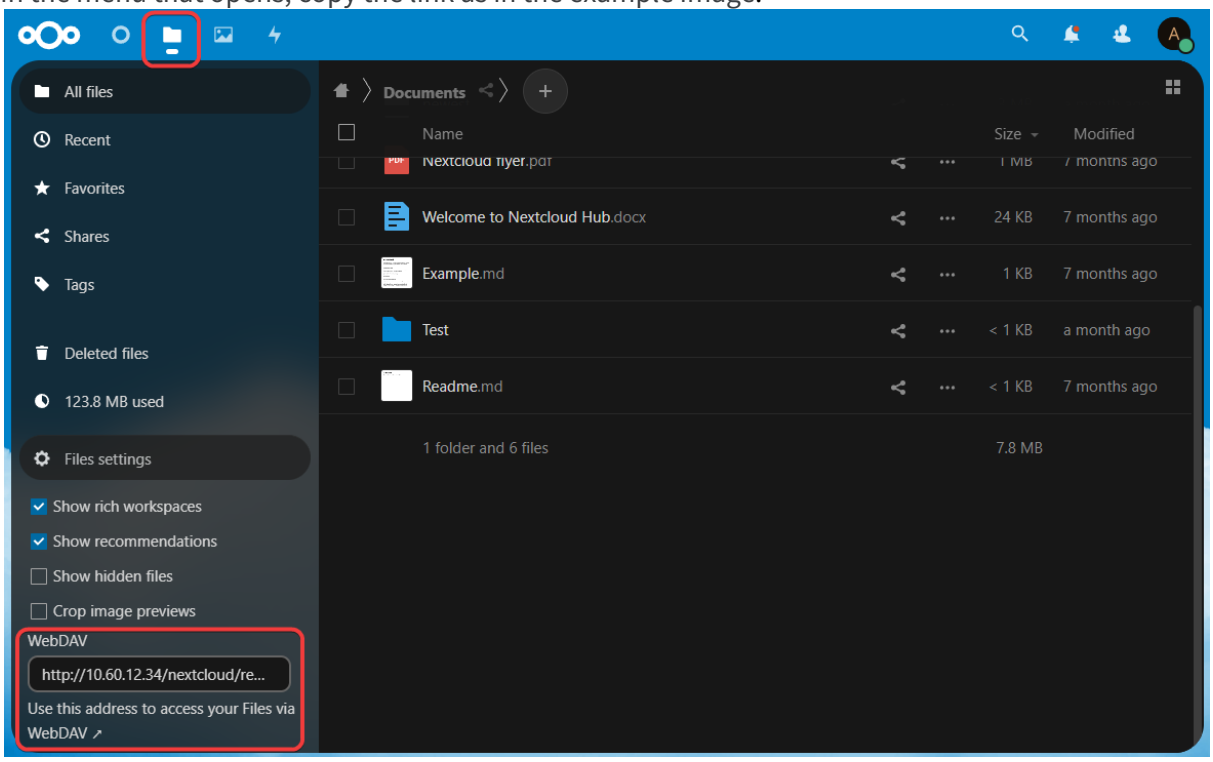
6. In the window of creating a scanning task or when adding a host, specify the following parameters:
   a. **Login** – name of the host user, for whom you have copied the public key .
   b. **Passphrase** – enter the key phrase, if you have specified it when generated keys.
   c. **Private key** – copy the `id_rsa` file content or upload it using the  button.

Getting a Link to Connect to Nextcloud

For Data Discovery to connect to Nextcloud storage via WebDAV protocol:

1. Open the Nextcloud storage web interface.
2. Log in using the account under which Data Discovery will connect.
3. On the top panel, select the **Files** tab.
4. In the bottom left corner, click **File settings**.

5. In the menu that opens, copy the link as in the example image:



6. From the link you have copied, paste the following data into the host adding window:
   a. In the **Address** field, specify the storage IP address.
      In our example, it is `10.60.12.34`
   b. In the **Protocol** field, specify the transfer protocol.
      In our example, it is `http`
   c. In the **Connection type** field, add the rest of the link.

Example link to connect to NextCloud
`/nextcloud/remote.php/dav/files/admin/`

The link leads to the root directory of the storage available to this user. For this host, you will need to specify scan paths relative to the root directory.

JSON File Format for Importing Hosts

In Data Discovery, you can add hosts from a JSON file to the task. With this feature, it will take you less time to configure the System so that it could start performing tasks.

> **Note:**
>
> Maximum allowed size of imported JSON file is 1 MB.

For importing, you need to use a file where the IP address or domain name of the host is specified.
If some parameters are not specified, Data Discovery will use default values.

In the table below, you can see the JSON file contents:

Code (example of file contents)	Format/possible values	Description
`{`		{
`"ver": "1.0",`		JSON schema version
`"hosts": [`		General section that includes descriptions of all hosts
`{`		
`"host": "10.60.44.194",`	String, up to 255 characters	IP address or domain name of the host
`"type": "WEBDAV",`	• `SMB` • `SSH` • `SHAREPOINT` • `FTP` • `WEBDAV`	Connection type. The default value is `SMB`
`"protocol": "HTTP",`	• `HTTPS` • `HTTP` • `FTP` • `FTP`	The parameter is available for FTP and WebDAV connection types. Possible values: • **For FTP:** • FTPS • FTP • **For WebDAV:** • HTTPS • HTTP
`"site": "nextcloud/remote.php/dav/files/admin/",`	String, up to 255 characters	This parameter is used for all connection types: • SHAREPOINT – path to the site, • WEBDAV – connection path. **Important!** To connect to Nextcloud storage via the WebDAV protocol, you need to insert a link from the repository, but **without an IP address** (for details, see "Getting a Link to Connect to Nextcloud")

Code (example of file contents)	Format/possible values	Description
`"resources": [ "archive/cases", "demo", "lv-426" ],`		Scanning paths
`"port": 2121,`	Integer from 0 to 65535	Host connection port
`"note": "Data archive",`	String, up to 255 characters	Host description
`"group_name": {},`	String, up to 255 characters	User group for SMB connection type
`"user_name": "admin",`	String, up to 255 characters	Login for connection
`"auth_type": "SSH_KEY",`	• `PASSWORD` • `SSH_KEY`	Authorization type. Possible values: • PASSWORD - login and password authorization; • SSH_KEY - SSH key.
`"ssh_key": "-----BEGIN RSA PRIVATE KEY-----\nProc-Type: 4,ENCRYPTED\n DEK-Info: AES-128-CBC,9EDACF3CCDBC0F30ECEA3D4CCFCF8F34\n MIIEogIBAAKCAQEApAT92vPvEUIuHMImewQHI WyJbwLNE9r2lW5Zyz7sQpg+p08w\n Ppg/ 5DBVCyOFKBjJVRb3CRVK7l8qdo5cxVFQqhjx2 DvGcdtn+JKGh9+oSVS6r9rh\n GSF4e32z2Ozc5zQdv5FQ+WrUNaElNI7aAG6GC /R6Imich7Igy0Jdw/iuiP1KxG4F\n jFVmvcZwKzMGVuyAKtM1YKP+sJN8TCk99dFM/ hGKGh4GSx+9ELyVGMyJBhps7eHh\n Eaz+2364RYYyCUzwCv5eXs8dSszDtlmJTjt86 w4fNfAF919P1WokkwU/u5FrY7mU\n 1nhG0wZFi+SzcZmCcGpTOr+Y464XYKIN8WRaI QIDAQABAoIBAGNZhq7mjQvGK4pN\n 6+MduHf9d23fo3sAadZv5kIfVhqsf172r4Fq1 WCSCiFaHaEBCorZAz/kPuQOjvkL\n GJSIwbSWsX9KDdUPknXVw1lwF0573S0CBtEjH ginRpYCO3hscAyGqCa8F9h4SMIC\n OXiDefyp1IUdGptzQTv9cjBvZMEiu7A02Tp5R`	String	Authorization private key. **Important!** • You must specify the key as one string. For this, after copying, replace all line breaks in it with " `\n` ". • If authorization private key is specified in the file, then authorization type will be `SSH_KEY`, even if the `auth_type` parameter has the `PASSWORD` value. This parameter is not exported. When importing, you need to specify it manually

Code (example of file contents)	Format/possible values	Description
U7HTkavsLbP4sB9wngJuhGhsBMd\n oMqrbSDKzIhgWuFFLeyGBEqUxfUDCKfWIP6y3 VDb5ZBgRP59qxrFx0TAeaXHoiK8\n dqywSIrGhGyE87Otb6pOQXXwxT8MdT5F3vmVR MUc2eegz694zm2mMaJDiSOLDowA\n djETpsECgYEAz73B3N3ibYWgU+tm5mgbdEO5P coEeWKKOeyrueniw+kP+FMdkqEa\n iGHyWbJHkNHgvt7fg79zpAhalYkJr4t+AsZ4F vaa6GzJ8mvbuUccYn3A1tc+eNlF\n LKAzcLntGJkC6Vbvt/ 0JwmkTNNDprDpBidtF6qp7XbtQRjOklzoR81k CgYEAyh8h\n KFgR63NiUQOFKluAa1mIcVFzlYIrlPpDvbxxH PPpPgZmsLLPSsqVu67/ZBuIYJvY\n 2RG8h98sX+U5ji2lE47fw04tkxOrihLuRX4dr tjMIVqndrxrpHAjKFa6ttrwCd1O\n DlqvN13zWv4EqSbEVOoOgZ7uHEynI9LmTFf7r AkCgYBKqyX3QTmVrCYIg+AkiN71\n VzJJ50bcanLMhyjiVhLP4AGBfD5w6b8jrjDx1 IasowNpF6q3LCVgnDAS94Knuje8\n v6GCf3kGebPMksATRGALdq/ NiOhHf7DBbYOPzQRPhw/ o1vqq4ySS4DmrDjwvT3En\n H81tn4FgG2erE/ zg7lptYQKBgCnXfea7LKKAVWcOAoA9Rv1yIHP M98+IjfseP+3E\n V+o1eCB19MY5jf4eqBadaJNRu868FI0DRfMeu TH04xBS1KyS1DG7OmYLGWTQ/HB4\n Zi6vBLsW6DoFUYdOC4krQyXddvmjzafdEA63L Uy170MRGm21smEVBJfpKmSZIUHG\n qPv5AoGAeDLKkqUg4rlUM9khzaa1i04g8g8mm h2zTfQMiImr/G1K8UCXGLsIIg/M\n ipvF+Bsuxb5X6YxeqGZCvfaCQXgmfjzOzrISW gh/X262zzLTNYzdGpyuIsB3b2M+\n Xc0cBaFm8hNc6SKc/ hXLzoae33ITtUhTcAhZGTjxujpuRPzkVUg=\n -----END RSA PRIVATE KEY-----",		
"ssh_key_pass": "city17",	String, up to 255 characters	Key phrase. You should specify the key phrase, if you use encrypted authorization private key. **Important!** This parameter is not exported. When importing, you need to specify it manually

Code (example of file contents)	Format/possible values	Description
`"password": "12pass34",`	String, up to 255 characters	Connection password. **Important!** This parameter is not exported. When importing, you need to specify the password manually
`"inherit_from_task": false,`	• `true` • `false`	Whether to inherit the connection login and password from the task. If a file being imported contains the login (`user_name`), then the `inherit_from_task` value will be ignored and authorization parameters will not be inherited from the task even if the parameter value is `true`. The default value is `true`
`"timeout_msec": 5000,`	Non-negative integer	Timeout for connection or host operations, in milliseconds. The default value is `5000`
`"retry_count": 5,`	Non-negative integer	Maximum number of connection attempts. The default value is 5
`"pause_sec": 3,`	Non-negative integer	Time interval between connection attempts, in seconds. The default value is 3
`"io_chunk_size": 32768,`	Non-negative integer	Size of a fragment of transmitted data, in bytes. The default value is `32768`
`"client_debug_level": 0`	Non-negative integer	Debug level. The default value is `0`
`}`		
`]`		

Code (example of file contents)	Format/possible values	Description
`}`		

Example of JSON file for importing multiple hosts

```
{
  "ver": "1.0",
  "hosts": [
    {
      "host": "10.60.3.195",
      "type": "SSH",
      "protocol": "HTTPS",
      "resources": [
        "share"
      ],
      "user_name": "root",
      "auth_type": "PASSWORD",
      "password": "SSHpass54321",
      "inherit_from_task": false,
      "timeout_msec": 5000,
      "retry_count": 5,
      "pause_sec": 3,
      "io_chunk_size": 32768,
      "client_debug_level": 0
    },
    {
      "host": "10.60.44.194",
      "type": "WEBDAV",
      "protocol": "HTTP",
      "resources": [
        "Photos"
      ],
      "site": "nextcloud/remote.php/dav/files/admin/",
      "user_name": "NCadmin",
      "auth_type": "PASSWORD",
      "password": "NCpass42",
      "inherit_from_task": false,
      "timeout_msec": 5000,
      "retry_count": 5,
      "pause_sec": 3,
      "io_chunk_size": 32768,
      "client_debug_level": 0
    },
    {
      "host": "10.60.44.233",
      "type": "SMB",
      "protocol": "HTTPS",
      "resources": [
        "demo",
        "shares",
        "smoke"
      ],
      "inherit_from_task": true,
      "timeout_msec": 5000,
```

```
    "retry_count": 5,
    "pause_sec": 3,
    "io_chunk_size": 32768,
    "client_debug_level": 0
  },
  {
    "host": "10.70.229.170",
    "type": "FTP",
    "protocol": "HTTPS",
    "resources": [
      "/dd"
    ],
    "port": 2121,
    "user_name": "admin",
    "auth_type": "PASSWORD",
    "password": "SDpass",
    "inherit_from_task": false,
    "timeout_msec": 5000,
    "retry_count": 5,
    "pause_sec": 3,
    "io_chunk_size": 32768,
    "client_debug_level": 0
  }
]
```


Adding Scanning Paths

To add a scanning path:

1. Click on the host to go to the scanning path list.
2. In the field in the upper left corner, specify the scanning path:
   - For a host with SMB connection, specify the network share name, for example: `share`

     > **Note:**
     >
     > For hosts with an SMB connection, it is not possible to specify a subdirectory of a network resource (for example: share/folder) as a scanning path.
     >
     > To scan a subdirectory of an SMB share, configure filters for directories in the **Folders filter** tab in the task settings.

   - For a host with SSH, FTP or WebDAV connection, specify the full path to the directory, for example: `/home/username/Documents`

     > **Note:**
     >
     > For the WebDAV connection type, specify scanning paths relative to the root directory. You can find the root directory in the **Connection path** field of the host description.

- For a host with SharePoint connection, you can specify a value in the following formats:
  - name of the document library, for example: `Shared Documents`
  - directory path within the library, for example: `Shared Documents/folder/example`

3. Click .

As a result, added scanning path is displayed in the list.

At the top of the scanning path list screen, you can see controls to manage the task and the host. Click  to return to the host list.

Automatic detection of scanning paths

Warning!

Detecting scanning paths is only supported for hosts with an SMB connection. Scanning paths are detected only for the hosts that have no scanning paths specified.

To automatically detect scanning paths:

1. Go to the host list.
2. Select hosts. To select all hosts on the page, click  next to the **Host** column name.

Note:

If you do not select any hosts, the System will search for scanning paths for all hosts in the task.

3. Click  (the action is available if the task is not running).
4. Click **Launch**.

Note:

You can also launch the scanning path detection for a host from a page with the path list:

1. Click on the host to go to the scanning path list.
2. Click  (the action is available only for SMB-hosts that have no paths specified).

The System detects and adds available scanning paths for hosts. If you launch detection for a large number of hosts, then scanning paths detection may take some time. In this case, the task will have the **Detecting scanning paths** status. You cannot start a task if scanning paths detection is in progress.

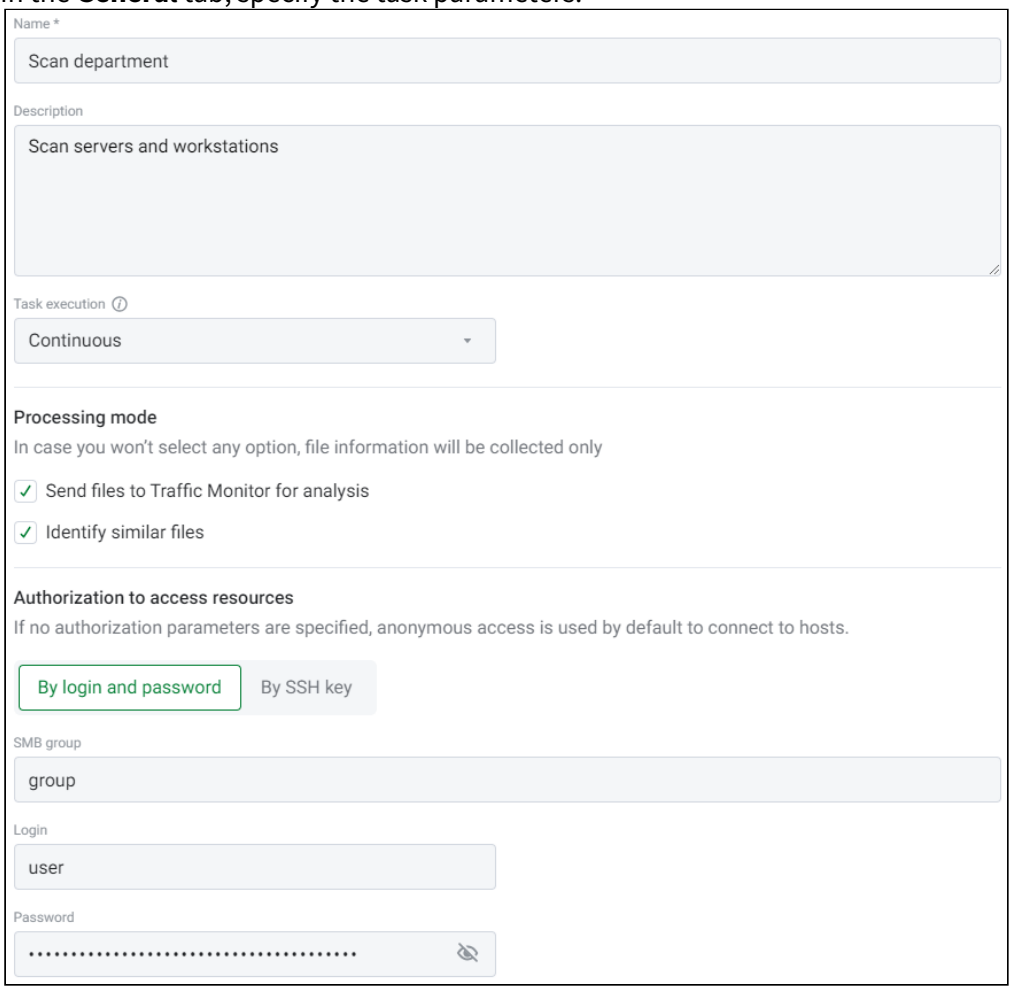
Note:

If it is time for the scheduled task to start automatically and the task has the **Detecting scanning paths** status, then the task will not start. In this case, we recommend starting the task manually when identifying scanning paths is completed.

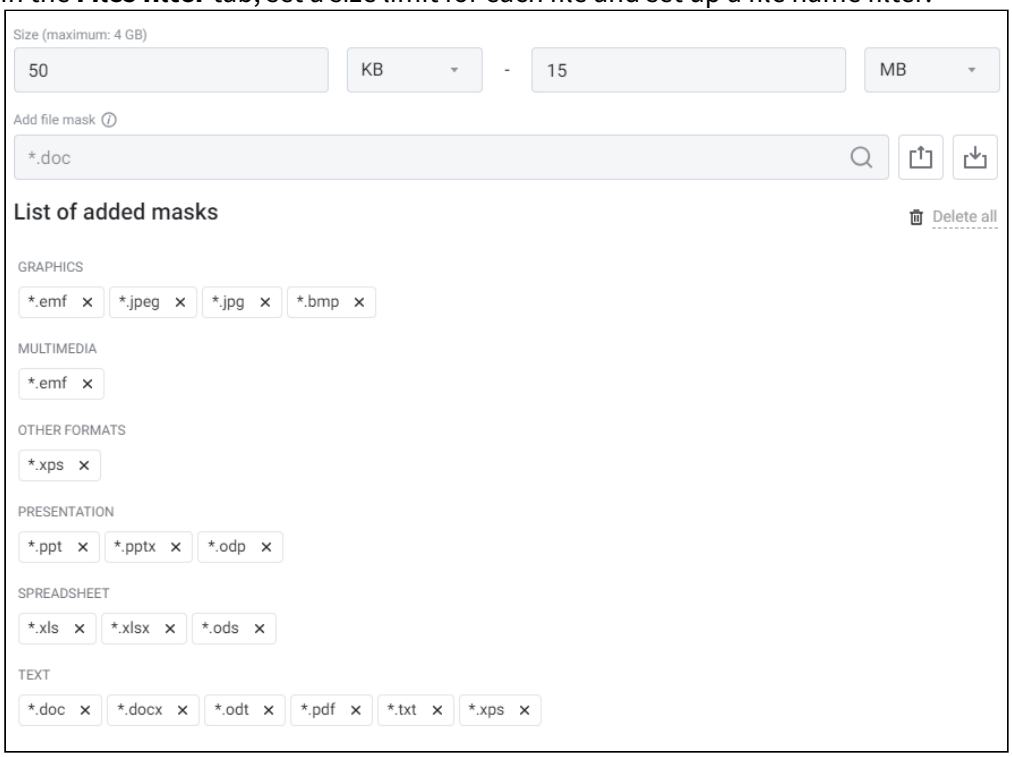
Examples of Task Settings, Host Settings and Adding Scanning Paths

Configure a task:

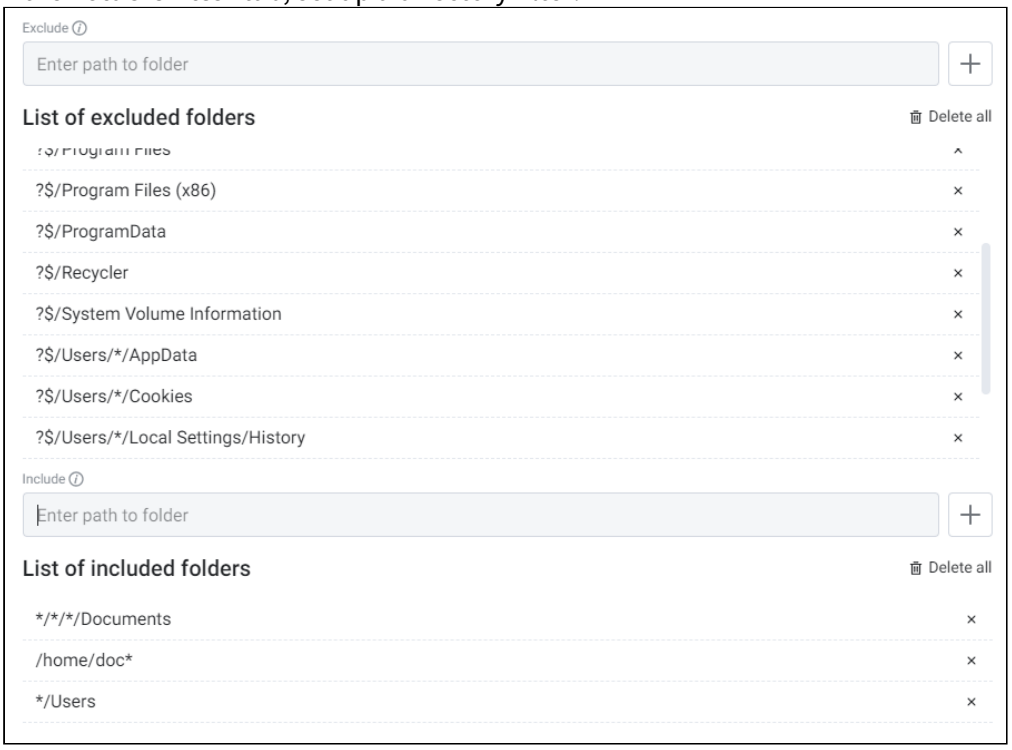
1. In the **General** tab, specify the task parameters:



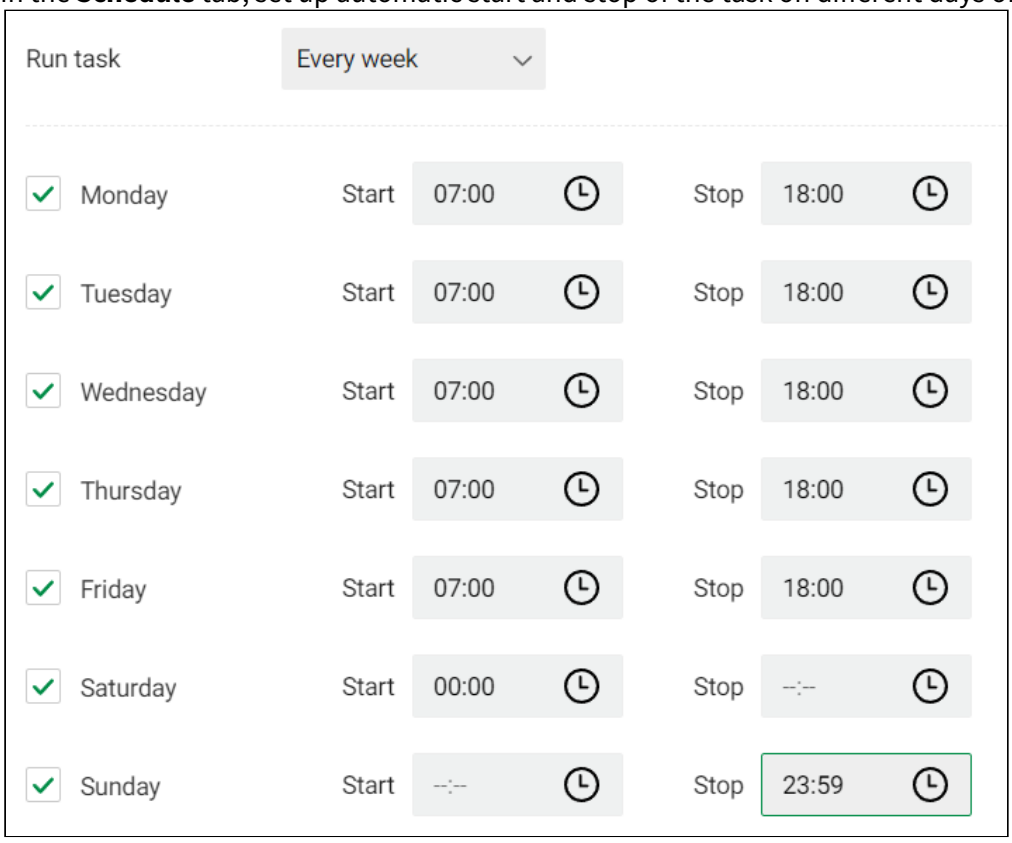
2. In the **Files filter** tab, set a size limit for each file and set up a file name filter:



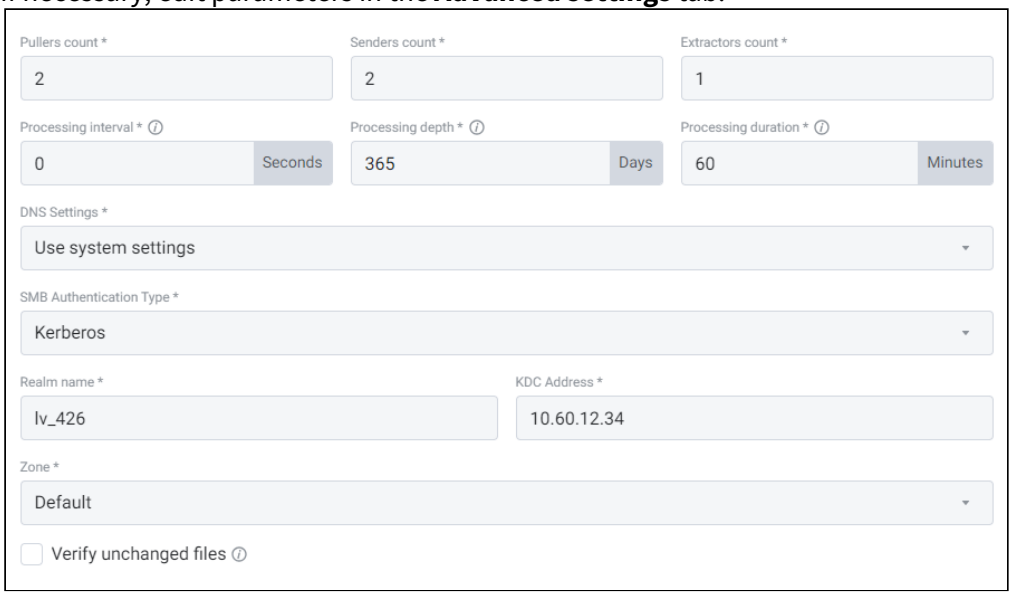
3. In the **Folders filter** tab, set up a directory filter:



4. In the **Schedule** tab, set up automatic start and stop of the task on different days of the week:



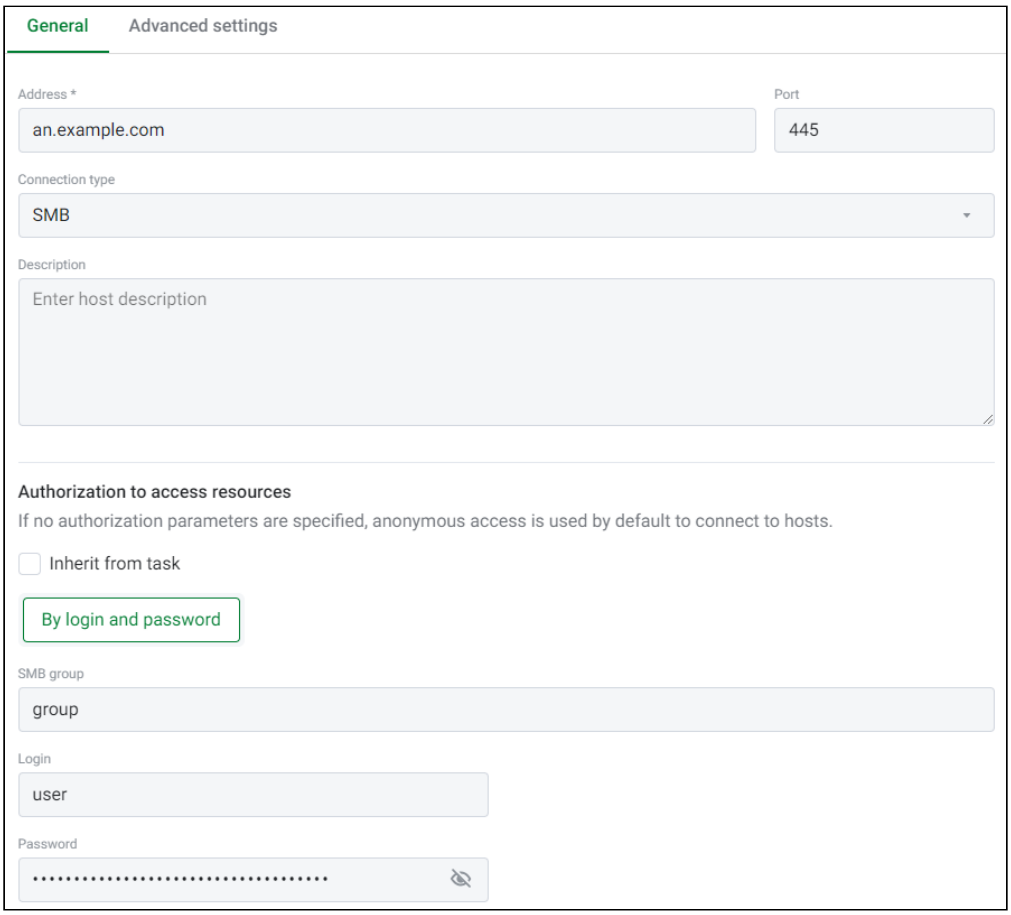
5. If necessary, edit parameters in the **Advanced settings** tab:



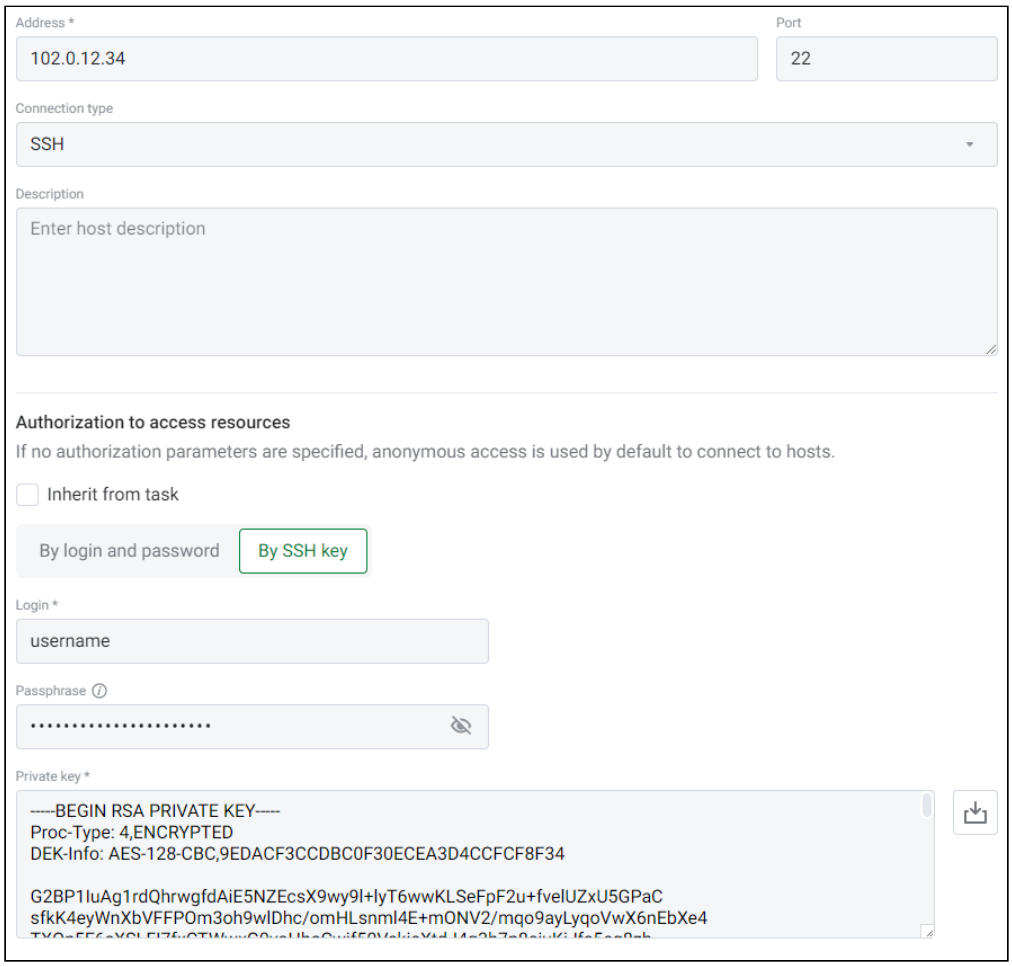
Configure a host:

1. In the **General** tab, specify the host connection parameters:

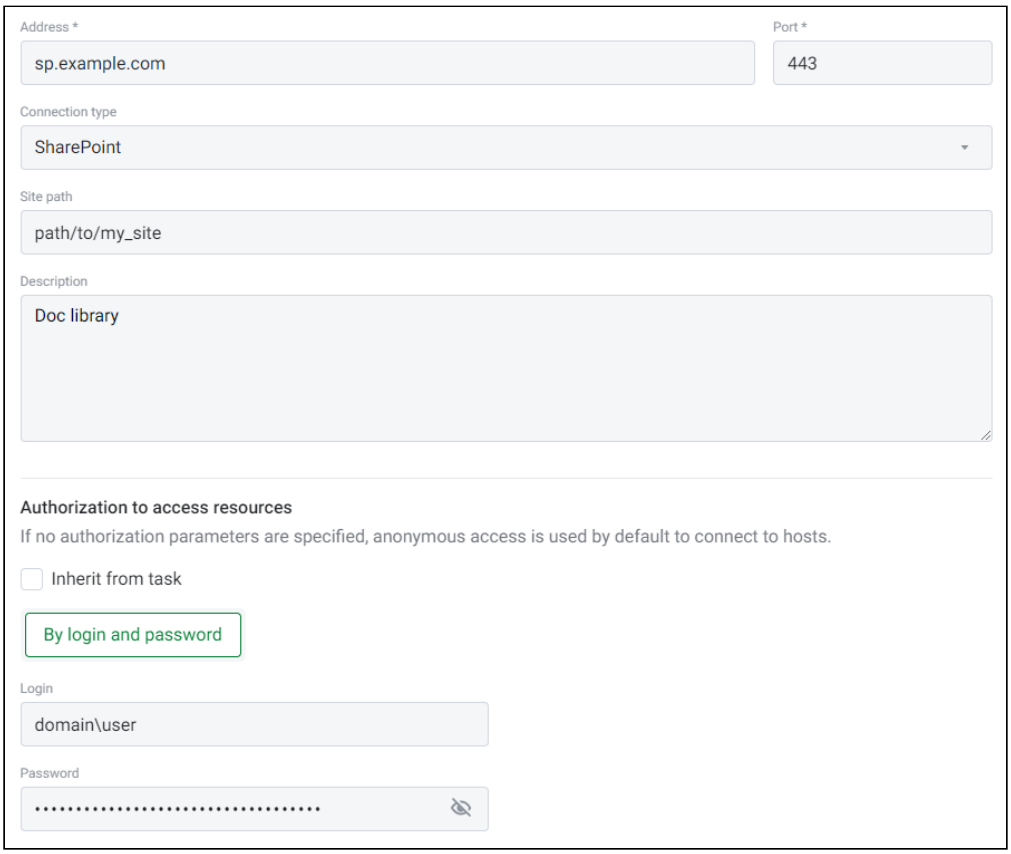
- A host with **SMB** connection:



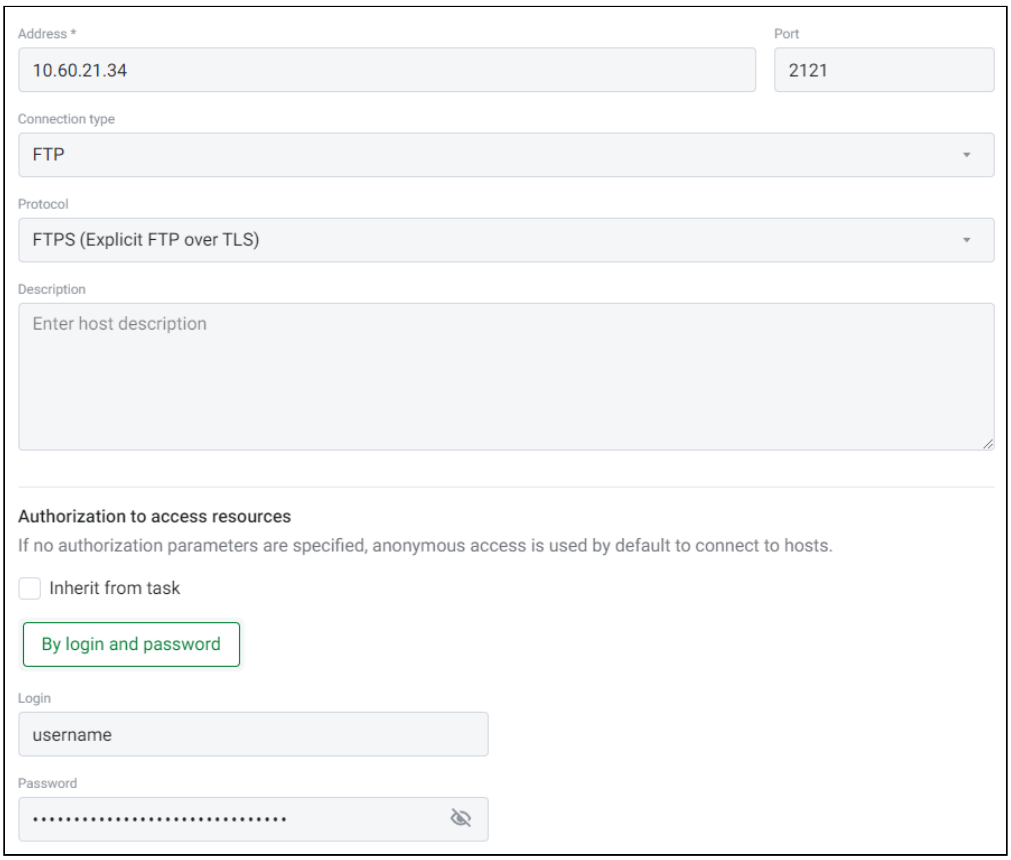
- A host with **SSH** connection:



- A host with **SharePoint** connection:



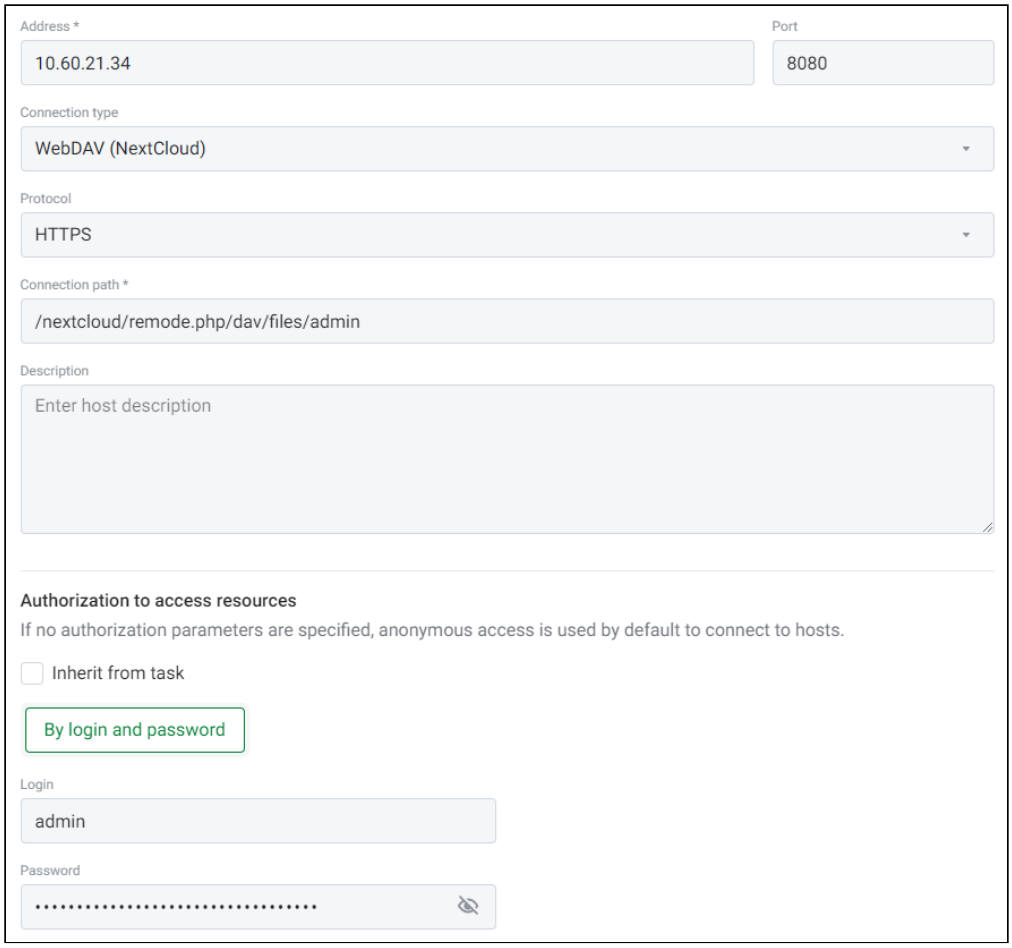
- A host with **FTP** connection:



Note:

When connecting via the FTPS (Explicit FTP over TLS) protocol, certificates of authenticity are not checked.

- A host with **WebDAV** connection:

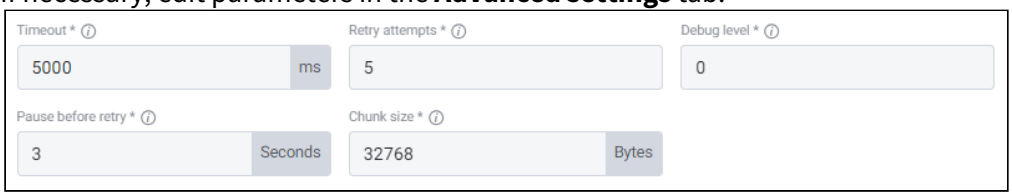


Note:

To speed up scanning of Nextcloud storage, use the application password to connect. For details, see the note for the paragraph Third-party WebDAV clients of Nextcloud official documentation.

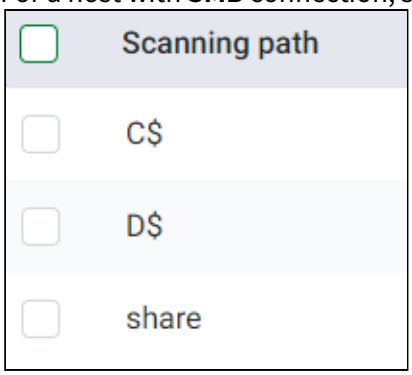
When connecting via WebDAV protocol, TLS/SSL certificates are not checked.

2. If necessary, edit parameters in the **Advanced settings** tab:



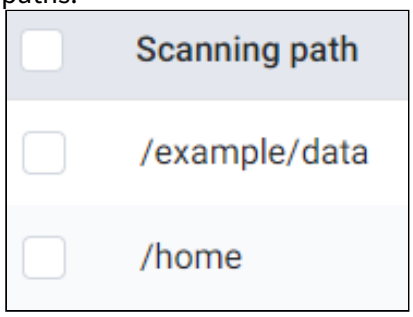
Add scanning paths:

- For a host with **SMB** connection, specify names of network resources as the scanning paths:

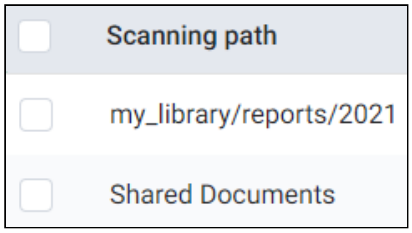


> **Note:**
>
> It is not possible to specify a subdirectory of a network resource (for example: share/folder) as a scanning path.
>
> To scan a subdirectory of an SMB share, configure filters for directories in the **Folders filter** tab in the task settings.

- For a host with **SSH**, **FTP** or **WebDAV** connection, specify full paths to directories as scanning paths:



- For a host with **SharePoint** connection, specify library names and/or directory paths within libraries:



9.1.4 Editing Tasks and Hosts

Editing task and host settings is available only for tasks with the following statuses: **New**, **Awaiting scheduled launch**, **Stopped**, **Completed**. You can click  to view current task/host settings even if editing is unavailable.

To edit a task:

1. Click the **Tasks** section to go to the task list.
2. On the right in the line of the task you want to edit, click .
3. Edit the task settings (for information on task settings see "Creating, Configuring and Running Tasks").
4. Click **Save**.

To edit a host in the task:

1. Click on the task to go to the host list.
2. On the right in the line of the host you want to edit, click .
3. Edit the host settings (for information on host settings see "Adding and Configuring Hosts").
4. Click **Save**.

9.1.5 Copying Tasks

To copy a task:

1. Open the **Tasks** tab to go to the scanning task list.
2. On the right, in the line of the task you want to copy, click .

In the task list, you will see a new task. The name of the new task consists of the original task name and the word "Copy". Its settings, hosts and scanning paths are the same as in the original task.

Warning!

Task files and file metadata will not be copied.

Within the new task, the System will completely re-scan the resources.

9.1.6 Deleting Tasks, Hosts and Scanning Paths

Deleting tasks, hosts and scanning paths is available only for tasks with the following statuses: **New**, **Awaiting scheduled launch**, **Stopped**, **Completed**.

Warning!

If you delete a scanning task, the following data **will be deleted**:

- copies of scanned files which are not sent to Traffic Monitor
- metadata of scanned files

- extracted text contents

For tasks created and run before Data Discovery updating to version 1.10, only copies of scanned files will be deleted.

Note:

If you delete a task, task files will be deleted from storage and will not be sent to Traffic Monitor.

We recommend that you send files to Traffic Monitor before task deletion:

1. In the task settings:
   - Enable the **Send files to Traffic Monitor for analysis** parameter.
   - Specify the following values:
     - **0** for the **Pullers count** parameter.
     - **N** for the **Senders count** parameter, where **N** is greater than or equal to the number of Data Discovery servers in the task zone.
2. Start the task. The System sends files to Traffic Monitor and deletes them from storages, but does not download new files to zone storages.
3. Wait until there are no task files left in the zone storages. To get the number of task files on a Data Discovery server:
   a. Connect to the server ssh console.
   b. Run the `ls PATH/NUMBER -1 | wc -l` command, where:
      - `PATH` is the server storage path, by default: `/mnt/disctps`;
      - `NUMBER` is the task number.

      Example: `ls /mnt/disctps/7 -1 | wc -l`
4. Stop the task.

To delete a task:

1. Click the **Tasks** section, to go to the task list.
2. In the **Actions** column of the task you want to delete, click .
3. Click **Delete** to confirm.

To delete hosts:

1. Click on the task to go to the host list.
2. To delete a single host, in the **Actions** column of the host, click .

   To delete multiple hosts, select  next to required hosts and click  in the top menu.

   To delete all hosts on the page, select  next to the **Host** column name and click  in the top menu.
3. Click **Delete** to confirm.

Warning!

If you add the host again, the System will rescan its resources.

To delete scanning paths:

1. Click on the host to go to the scanning path list.
2. To delete a single path, click  to the right of the table entry.

   To delete multiple paths, select  next to required paths and click  in the top menu.

   To delete all paths on the page, select  next to the **Scanning path** column name and click  in the top menu.
3. Click **Delete** to confirm.

> **Warning!**
>
> If you add the path again, the System will rescan it.

9.1.7 Task Execution Statistics

Data Discovery contains statistics: information about execution of each scanning task. Statistics on hosts and scanning paths is available as well.

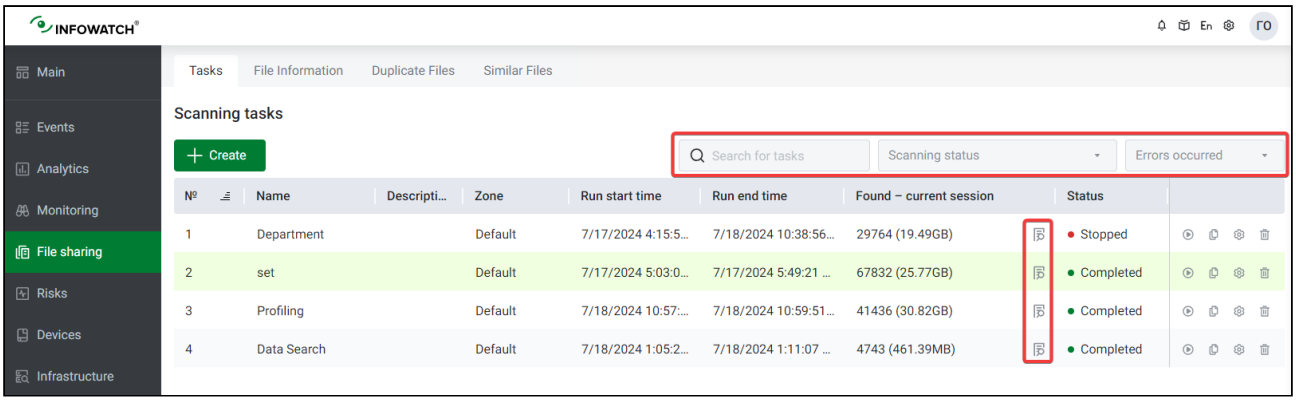
Statistics on scanning tasks

Statistics are available on the **Tasks** tab of the **File sharing** section. To navigate through pages, use the panel at the bottom of the screen. Statistics on tasks include the following data:

Parameter	Description
№	Task number. The number is unique
Name	Task name
Description	Short description of the task
Zone	Task zone
Run start time	Date and time the task was started. The value is displayed in the following format: M/d/yyyy h:mm:ss tt
Run end time	Date and time the task was stopped. The value is displayed in the following format: M/d/yyyy h:mm:ss tt. For tasks with the *Running* status, the data is not available
Found – current session	The number and total size of files found on all task resources, with the filters applied. The System gets the number of files on a resource at each connection (in each session). The number of found files can decrease

Parameter	Description
	if files were deleted from the resource or if the filter settings in the task were changed
Status	Current status of the task: • New • Awaiting scheduled launch • Starting • Running • Stopping • Stopped • Detecting scanning paths • Completed If an error occurred when working with a resource,  is displayed next to the status. To view information about the error, hover the cursor over .

Managing task list



You can sort tasks by the following columns: **№**, **Name**, **Run start time**, **Run end time**, **Status**. To change the sort order of rows in the task list, click on the heading of the column, by which you want to sort tasks.

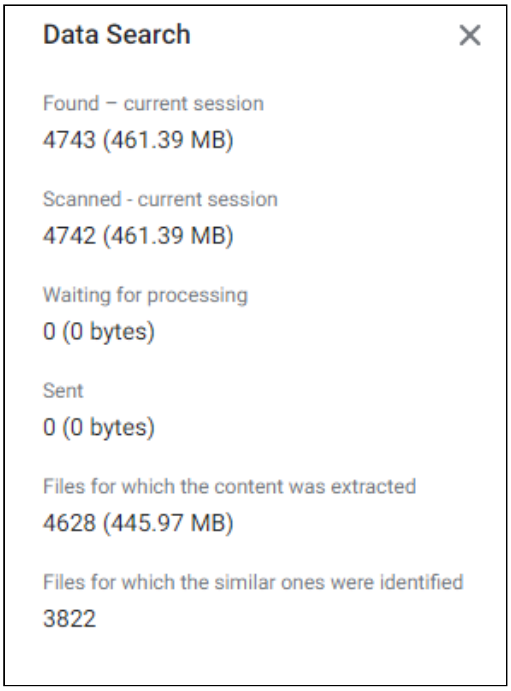
To search for tasks by name, in the **Search for tasks** field, enter the first characters of the task name. The **Search for tasks** field is located at the right above the table. You will see a list of tasks whose name starts with the characters entered in the **Search for tasks** field. Search is case-insensitive.

To filter a task list, select values in the drop-down lists at the right above the table. You can filter a task list:

- **by Scanning status** – you can select one of statuses;
- **by Errors occurred** – you can select one of two values: with **Errors** or with **No errors**.

To reset the filter value, click  in the field with the selected value.

To open the detailed Task performance statistics, in the **Found – current session** column of the task row, click . On the right, the window with the task name and its execution statistics will open:



If the task is running when viewing, the values of the fields will change during the execution. Statistics contain the following information:

- **Found – current session** – the number and total size of files found in the current session, with the filters applied. Data Discovery gets the number of files on the resource at each connection (in each session). The number of found files can decrease, if files were deleted from the resource or if the filter settings in the task were changed;
- **Scanned – current session** – the number and total size of files scanned in the current session. In most cases, the number of files scanned increases with each session. At the end of the pass, the number of files is reset, in the next pass the scanner checks the files and processes new ones
- **Waiting for processing** – the number and total size of files waiting for processing at the moment;
- **Sent** – the number and total size of files sent to Traffic Monitor over the time the task was running;
- **Files for which the content was extracted** – the number and total size of files, from which text data was extracted for searching the documents with the same terms (similar ones);
- **Files for which the similar ones were identified** – the number of files, which have **passed the processing** for detecting the similar ones. You can see the value in the field only after Data Analysis Service has processed all data. By the time, the scanning task has the **Finished** status. Until the end of processing, the field has the 0 value.

To copy the task number or the task name to the clipboard, hover your mouse over the table cell and click .

To move a column, left-click and hold the column heading, then move it along the heading row.

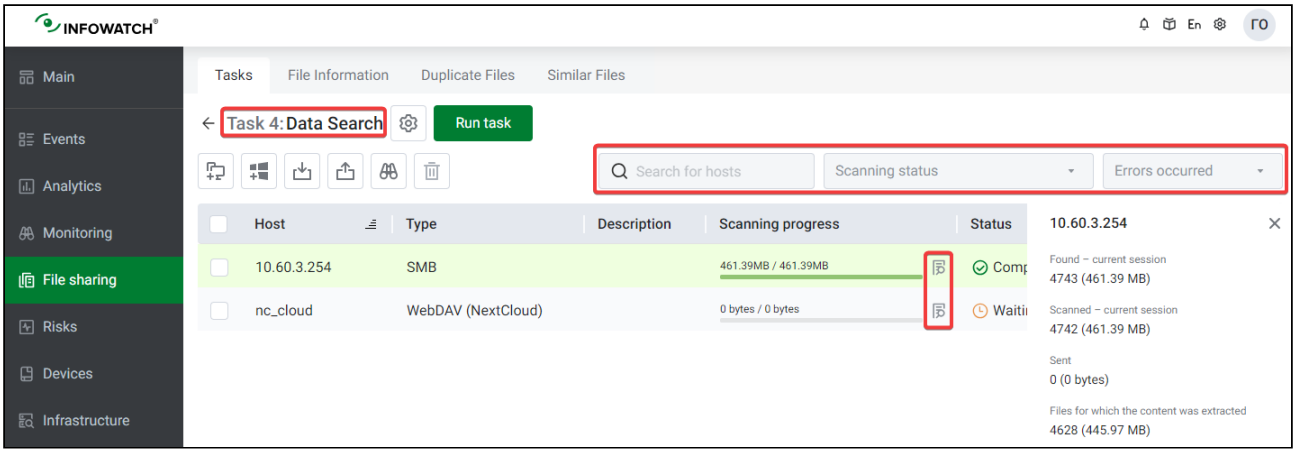
To change a column width, left-click and hold the right border of the column heading, then move the column border along the heading row.

Statistics on hosts

Click on the task to go to the host list. At the left top, you can see the number and the name of the scanning task. To navigate through pages of the host list, use the navigation panel at the bottom of the screen. Statistics include the following data:

Parameter	Description
Host	Host address
Type	Host connection type
Description	Host description
Scanning progress	The total size of scanned files / the total size of found files
Status	**Note:** The host status depends on statuses of the host resources. The current status of the host: • *Scanning* – at least one resource has one of the following statuses: *Connecting*, *Getting info* or *Processing files* • *Completed* – all host resources have the *Completed* status, except for those where an error occurred • *Waiting for scan* – the status is displayed if the conditions for statuses listed above are not met If an error occurred when working with a resource, is displayed next to the status. To view information about the error, hover the cursor over .

Managing host list



You can sort hosts by the following columns: **Host**, **Type**. To change the sort order of rows in the host list, click on the heading of the column, by which you want to sort hosts.

To search for hosts by address, in the **Search for hosts** field at the right above the table, enter the first characters of the IP address/domain name. The **Search for hosts** field is located at the top right of the screen. You will see a list of hosts whose address starts with the characters entered in the **Search for hosts** field. Search is case-insensitive.

Note:

To search for hosts by address, the System uses values specified when creating/editing a host. For example, if you specified the host address as a domain name, the host can't be found by IP address.

To filter a host list, select values in the drop-down lists at the right above the table. You can filter a host list:

- **by Scanning status** – you can select one of statuses;
- **by Errors occurred** – you can select one of two values: with **Errors** or with **No errors**.

To reset the filter value, click  in the field with the selected value.

To open the detailed Host scanning statistics, in the **Scanning progress** column of the host row, click . On the right, the window with the host address and scanning statistics will open.

If the task is running when viewing, the values of the fields will change during the execution. Statistics contain the following information:

- **Found – current session** – the number and total size of files found on the host, with the filters applied. Data Discovery gets the number of files on the resource at each connection (in each session). The number of found files can decrease, if files were deleted from the resource or if the filter settings in the task were changed;
- **Scanned – current session** – the number and total size of files scanned on the resource in the current session. In most cases, the number of files scanned increases with each session. At the end of the pass, the number of files is reset, in the next pass the scanner checks the files and processes new ones;
- **Sent** – the number and total size of host files sent to Traffic Monitor over the time the task was running;
- **Files for which the content was extracted** – the number and total size of files, from which text data was extracted for searching the documents with the same terms (similar ones).

To move a column, left-click and hold the column heading, then move it along the heading row.

To change a column width, left-click and hold the right border of the column heading, then move the column border along the heading row.

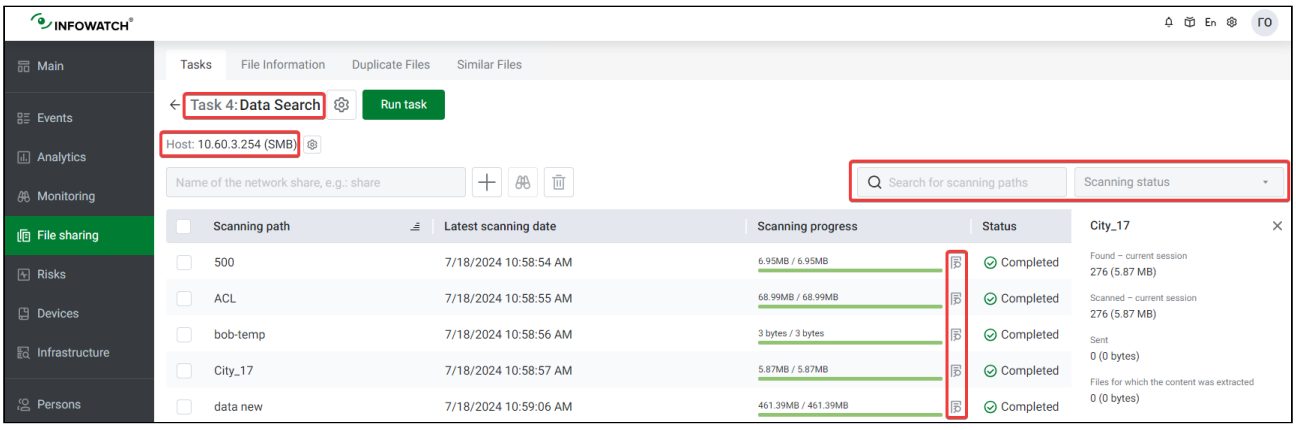
Statistics on scanning paths

Click on the host to go to the scanning path list. At the left top, you can see the number and the name of the scanning task as well as the host address. To navigate through pages of the scanning path list use the navigation panel at the bottom of the screen. Statistics include the following data:

Parameter	Description
Scanning path	Scanning path (resource) of the host
Latest scanning date	Date and time of the last connection to the resource. The value is displayed in the following format: M/d/yyyy h:mm:ss tt
Scanning progress	The total size of scanned files / the total size of found files
Status	Current status of the scanning path:

Parameter	Description
	• *Waiting for scan* – the System has not yet connected to the resource in this task • *Connecting* – the System is connecting to the resource • *Getting info* – the System is getting a list of files and information about files • *Processing files* – the System is saving file information and downloading file copies to a storage • *Stopped* – the resource scanning session ended **or** the scanning task is stopped • *Completed* – the System has finished traversing through the resource and all files found on the resource are processed • *Error* – an error has occurred when working with the resource (to view information about the error, hover the cursor over the status of the resource) **Error examples** In some cases, error messages may contain the following text: • Operation not permitted • Permission denied • Authentication failed Such error messages may indicate that incorrect authorization parameters are used or the account has insufficient access rights to scan the resource.

Managing scanning path list



You can sort scanning paths by the **Scanning path** column. To change the sort order of rows in the scanning path list, click on the heading of the **Scanning path** column.

To search for paths, in the **Search for scanning paths** field, enter the first characters of the path. The **Search for scanning paths** field is located at the right above the table. You will see a list of paths that start with the characters entered in the **Search for scanning paths** field. Search is case-insensitive.

To filter a scanning path list, select values in the drop-down lists at the right above the table. You can filter a path list **by Scanning status** – you can select one of statuses.

To reset the filter value, click  in the field with the selected value.

To open the detailed Resource scanning statistics, in the **Scanning progress** column of the host row, click . On the right, the window with the path name and scanning statistics will open.

If the task is running when viewing, the values of the fields will change during the execution. Statistics contain the following information:

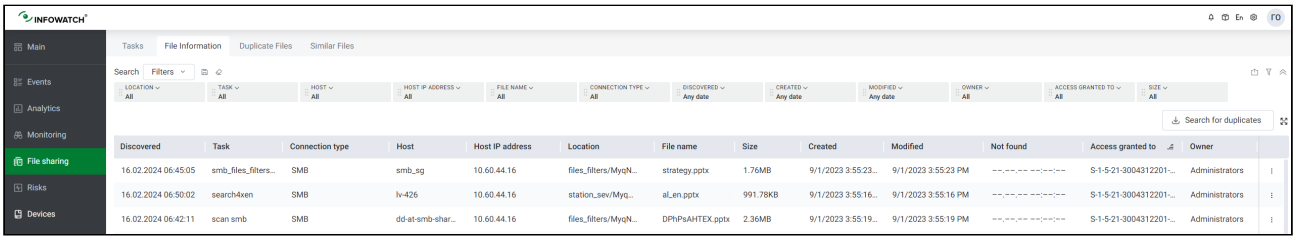
- **Found – current session** – the number and total size of files found on the resource in the current session, with the filters applied. Data Discovery gets the number of files on the resource at each connection (in each session). The number of found files can decrease, if files were deleted from the resource or if the filter settings in the task were changed;
- **Scanned – current session** – the number and total size of files scanned on the resource in the current session. In most cases, the number of files scanned increases with each session. At the end of the pass, the number of files is reset, in the next pass the scanner checks the files and processes new ones;
- **Sent** – the number and total size of resource files sent to Traffic Monitor over the time the task was running;
- **Files for which the content was extracted** – the number and total size of files, from which text data was extracted for searching the documents with the same terms (similar ones).

To move a column, left-click and hold the column heading, then move it along the heading row.

To change a column width, left-click and hold the right border of the column heading, then move the column border along the heading row.

9.2 "File Information" Tab

On this tab, you can find information about files that have been scanned within tasks



.

Detailed file information is presented in the form of a table; results of grouping – in the form of a list of groups which can be disclosed:

- for structure of the file information table, for information on how entries are added to the table and for actions with files, see "File Information Table";
- for configuration of table and groups display, see "Table Display Configuration";
- for filtering of table entries, see "Single Filter";
- for searching for duplicate files by template, see "Search for Duplicates";
- for table data export, see "Exporting Reports with File Information".

See also "Example of Managing the Table".

9.2.1 File Information Table

On the **File Information** tab, Data Discovery displays the results of scanning resources. Information about files is represented as a table. At the end of each resource processing session, information about files scanned during the session is added to the table.

How entries are added to the detailed file information table within a task

A *pass* is a complete scanning of a single resource. A pass may consist of several sessions.

A *session* is a connection to a resource while the System passes through it.

During the first pass through the resource, Data Discovery collects metadata of all files and adds an entry for each file to the table.

When the System connects to the resource again, it processes only new and changed files. To detect changed files, the System checks the file metadata for changes:

- Path to the file
- File name
- File size
- Date and time the file was modified
- File owner
- Users/groups that have access to the file

> **Warning!**
>
> Checking the following metadata for changes is supported only when scanning SMB shares on hosts running Windows:
>
> - File owner
> - Users/groups that have access to the file

If the file is new or changed, the System adds a new entry to the table at the end of the session.

Structure of the detailed file information table

Each table entry contains information about one scanned file. Each entry contains the following fields:

Field	Description
Discovered	Date and time when the file was first found on the resource by the System within a scanning task. The date and time are in the following format: `M/d/yyyy h:mm:ss tt` **Note**: If the unchanged file is found again in the following sessions within the task: • no new entries are added to the table; • **Discovered** field is not updated in the file entry.
Task	Name of the scanning task within which the file was found on the resource **Note**: If a file on the resource is found within several scanning tasks, an entry is added to the table about the file for each such task

Field	Description
Host	Domain name of the host where the file was found
Host IP address	IP address of the host where the file was found. If the address for a SharePoint host is specified in the domain name format in the host settings, the field is not filled.
Location	Location is a full path to the directory where the file was found **Note**: Full path includes the scanning path of the resource. For files from SharePoint hosts, the full path also includes the site path.
File name	File name The System retrieves the value from the file system of the resource
Size	File size The System retrieves the value from the file system of the resource
Created	Date and time the file was created. The date and time are specified in the following format: `M/d/yyyy h:mm:ss tt` The System retrieves the value from the file system of the resource. **Note**: Determining the file creation date is not supported if: • SMB shares are scanned on hosts running Linux • hosts running Linux are scanned via SSH protocol In this case, the **Modified** field value is used as the value of **Created** field
Modified	Date and time when the file was modified. The date and time are in the following format: `M/d/yyyy h:mm:ss tt` The System retrieves the value from the file system of the resource. **Note**: Date and time of the file modification specified in this field indicate the following changes: • When scanning SMB shares on hosts running Windows OS: • Last change to file contents • Last file name change • Last file path change • Last change of the file owner or the list of users and groups who have access to the file • When scanning SMB shares on hosts running Linux / hosts running Linux via SSH protocol: • Last change to file contents • When scanning document libraries in SharePoint Server 2016: • Last change to file contents • Last file name change

Field	Description
Not found	The date and time of a pass completion within a task, if the file was not found on the resource again according to the results of the pass. The date and time are specified for the first pass in which the file was not found. The date and time are in the following format: `M/d/yyyy h:mm:ss tt` The system will not find the file again within the task if: • The file has been deleted from the resource. • The file metadata has changed. In this case, a new entry is added to the table. • File/folder filters were changed in task settings and the file did not match new filters conditions.
Owner	In the **Owner** field, Data Discovery specifies the name of a person or a group that owns the file.
Access granted to	In the **Access granted to** field, Data Discovery specifies the list of names of users and groups that have access to the file. Data Discovery identifies persons and groups by the SID values it receives from the resource file system. To identify names by SID, Data Discovery uses data from Active Directory (see "Identification of persons and groups"). If there is no data to identify a person or group, Data Discovery would specify SID instead of the name, e.g.: `S-1-5-21-1922741149-3613679056-2300624754-1000`

Warning!

Obtaining the SID of the owner of a file and the SID of users/groups that have access to a file is supported for:

- SMB shares on hosts running Windows operating systems
- SharePoint Server 2016 document libraries

The **Owner** and **Access granted to** fields can be filled incorrectly if:

- An SMB share is scanned on a Linux host
- An SMB share being scanned has been created on an SMB server that does not support file attributes to determine the owner of a file and users/groups that have access to a file
- An SMB being scanned has been created on a file system that does not support file attributes to determine the owner of a file and users/groups that have access to a file

Getting the SID of the file owner and SID of users/groups that have access to a file is not supported when scanning hosts:

- Running Linux via SSH protocol
- Via FTP protocol
- Via WebDAV protocol

In this case, the *Senders* and *Recipients* parameters have the "`iw data discovery`" value.

Note:

Date and time in the **Discovered**, **Created**, **Modified**, **Not found** fields are displayed in the time zone of the workstation from which you connect to the Data Discovery web interface.

Identification of persons and groups

Data Discovery can identify persons and groups by SID.

If identification is successful:

- Instead of the **person**'s SID, in the **Owner** and **Access granted to** fields, Data Discovery will display the name specified in the dossier
- Instead of the **group** SID, in the **Owner** and **Access granted to** field, Data Discovery will display the group name received from Active Directory

Identification of Windows OS accounts by well-known SIDs

The System identifies the following Windows OS accounts (for more information about well-known SIDs, see Microsoft Windows documentation):

SID	Account	Description
S-1-5-18	System	A special account used by the operating system
S-1-1-0	Everyone	All users

Data about the specified SIDs is added to the System by default. The System performs identification in all table entries.

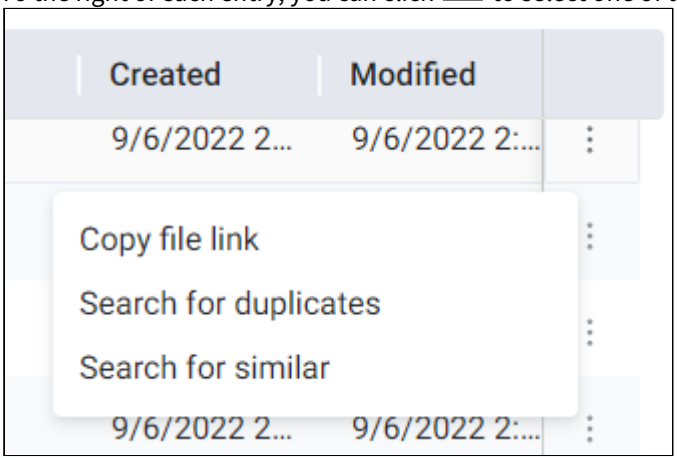
Identification of persons and groups from Active Directory

All persons added to the System as a result of synchronization with Active Directory have a SID contact type which makes it possible to identify them. If groups from Active Directory have a SID, they will be identified. For more information about adding data about persons and groups to the System, see "InfoWatch Data Discovery. Administration Guide", "Integrations".

The System identifies persons and groups in all entries of the table by the current data of each integration. Identification that uses data of the last synchronization may take some time, up to 15 minutes from the moment the synchronization is completed.

Actions with files

To the right of each entry, you can click ⋮ to select one of the available actions with file:



- **Copy file link**;
- **Search for duplicates** – action is similar to searching for duplicates by template. The difference is that here a file from the table will be used as a template;
- **Search for similar** – If a file has passed the processing for detecting the similar ones, you can find files similar to it in meaning. So Data Discovery displays files included in one thematic group according to the results of DAS analysis.

Note:

Searching for duplicates or similar ones by a file from the table will be unavailable if any of the following conditions is met:

- File size exceeds 4 GB
- File copies were not created when scanning - none of processing modes is selected in scanning task settings.

For file to be processed for identifying the similar ones, the extracted **text content** must be at least 200 bytes. Content with size less than 200 bytes will not be sent for analysis to DAS but will be available for viewing in Investigation Centre web interface, in the **File sharing** section.

If files were scanned with Data Discovery version **earlier than 1.8**, there is no information about their contents. Such files are not searched.
In this case:

1. Copy the task in which the files were scanned.
2. Rescan the files in a new task.

Viewing extracted text contents

To view text content extracted from the file, select a row in the table. On the right, the window with the file name and extracted text data will open. If there is no text data extracted from the file, the window contains only the file name.

Note:

Several text contents can be extracted from one file. But for analysis and displaying, they are merged into the one content.

Example: The extractor processes archive file containing several documents. Text contents may be extracted from each document. But, as a result, they will be merged into one common content related to archive file.

9.2.2 Table Display Configuration

To change sorting order of the rows, left-click the column title you want to sort by. Table entries will be sorted by the column values in ascending/descending order.

To move a column, left-click and hold the column title, then move it along the title row.

To change a column width, left-click and hold the right border of the column title, then move the column border along the title row.

To navigate through pages of the file information table, use the navigation panel at the bottom of the screen.

9.2.3 Search for Duplicates

In addition to grouping files by hash and size, Data Discovery allows you to search for file duplicates by specifying a file template. When searching, files are considered duplicates if they have the same content as the template.

The System gets information about file contents within a scanning task if at least one of the processing modes is selected.

Note:

Search for duplicates is unavailable if the scanned file is larger than 4 GB.

Search by template

To search for duplicates:

1. Click **Search for duplicates**.
2. In the window that opens, upload a file.
3. Click **Search**.
4. Wait for the search to finish.

The table will display entries for files that are duplicates of the uploaded file. The Single Filter conditions will apply to the entries (see "Single Filter").

Warning!

If files were scanned by the System version 1.7 or lower, there is no information about their contents. Such files are not searchable.

In this case:

1. Copy the task in which the files were scanned.
2. Scan the files again in the new task.

Resetting search results

To reset the search results, click  to the right of the uploaded template name.

9.2.4 Exporting Reports with File Information

In the **File Sharing** section, on the **File information**, **Duplicate Files** and **Similar Files** tabs, you can export presented information to a report. The report file format is XLSX.

To export the data:

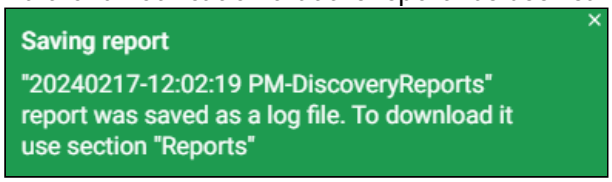
1. Configure the filter (see "Single Filter"), the sorting order (see "Table Display Configuration") and/or search duplicates by template (see "Search for Duplicates").

   > **Warning!**
   >
   > If the table contains more than 10,000 entries after filtering, only the first 10,000 entries are added to the report.

   The data in the configured view will be exported to the report.
2. Click  in upper right part of the screen.
3. Select **Unload report**.
4. In the window that opens, specify the report content:
   a. To add a sheet with additional information to the report, select **Additional information** in the list. The sheet contains:
      - the date and time of the report generation;
      - the filter settings;
      - the link to the Management Console.

   b. Depending on selected tab, you can select:
      - **Files information**
      - **Duplicate files**
      - **Similar files**

   c. Click **Save**.
5. Wait for a notification that the report has been saved.

6. To download the file go to the **Reports** → **Reports archive** section.
7. Click the report name to save the file on your computer.

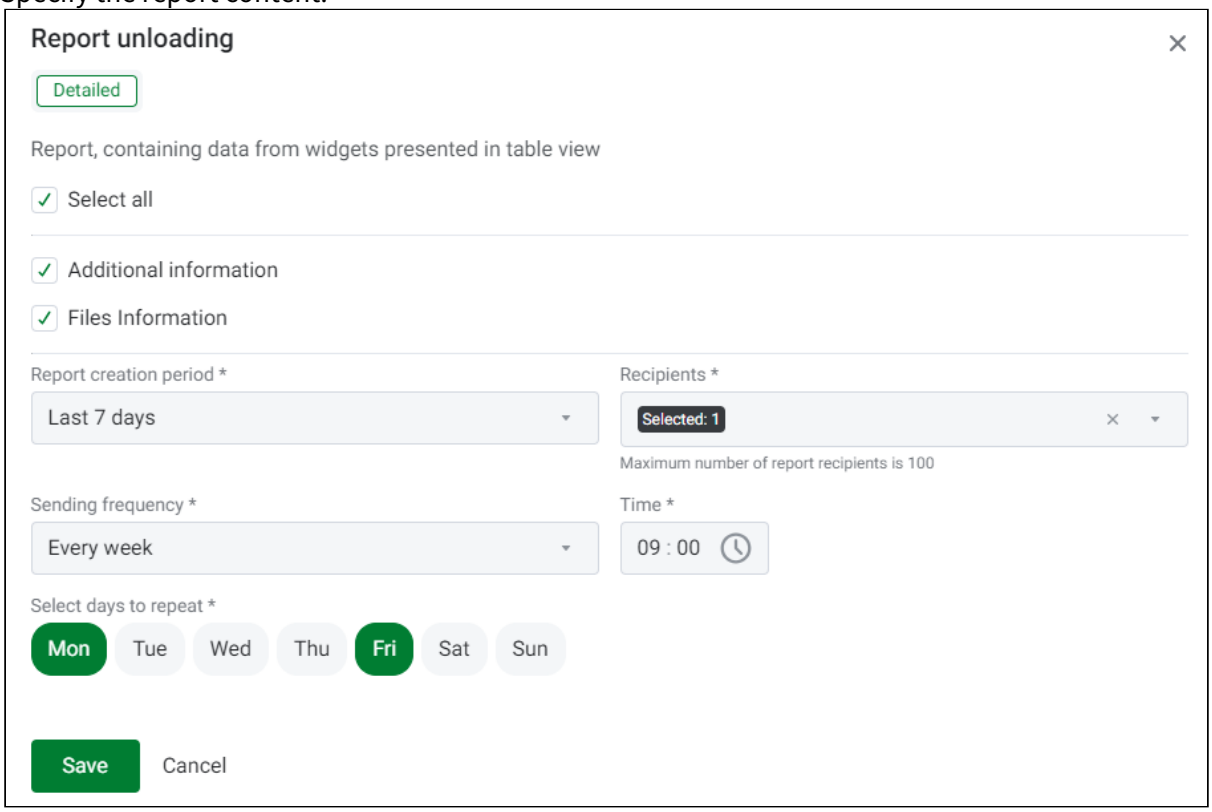
On the **File information** tab, you can **set up regular sending**. For this:

1. Configure the filter (see "Single Filter"), the sorting order (see "Table Display Configuration") and/or search duplicates by template (see "Search for Duplicates").

> **Warning!**
>
> If the table contains more than 10,000 entries after filtering, only the first 10,000 entries are added to the report.

The data in the configured view will be exported to the report.

2. Click  in upper right part of the screen.
3. Select **Set up regular sending**.
4. Specify the report content:



5. Also specify the sending parameters:

Parameter	Description
Report creation period	Specify the period for which to create a report: • **Last 24 hours** • **Current day** • **Last 3 days** • **Last 7 days** • **Last 30 days**
Recipients	Specify the report recipients
Sending frequency	Specify, with what frequency should Officer receive the report: • **Every day**

Parameter	Description
	• **Every week** You can select days of the week when Data Discovery will send the reports. • **Every month** Report will be sent on the 1st day of the month following the reporting one
Time	Specify the time of sending report

6. Click **Save**. A task for regular sending will be created. You can view the task list in the **Reports** section, on the **Tasks for regular sending** tab. For details, see "Tasks for Regular Sending".

Important!

For regular sending reports, you should configure the connection with the mail server in Investigation Centre (for details, see "*InfoWatch Data Discovery. Administration Guide*", the "Notification Services" article).

9.2.5 Example of Managing the Table

The table contains information about files scanned on different resources within several scanning tasks.

You need to determine:

- on what workstations a file with financial information was located;
- in what time period the file was located on the workstations.

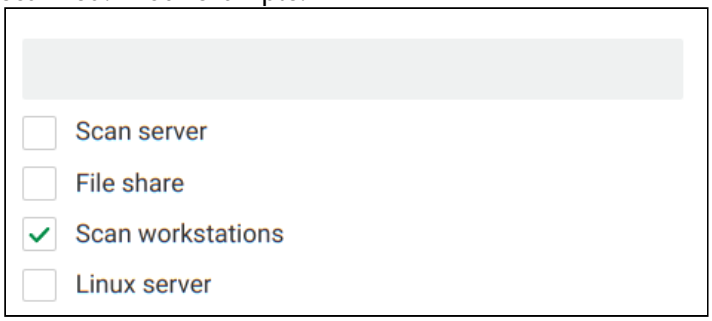
It is known about the file with financial information that:

- the file name includes the string "financ";
- the file was created in the current month;
- file size does not exceed 5 megabytes.

Configuring the filters

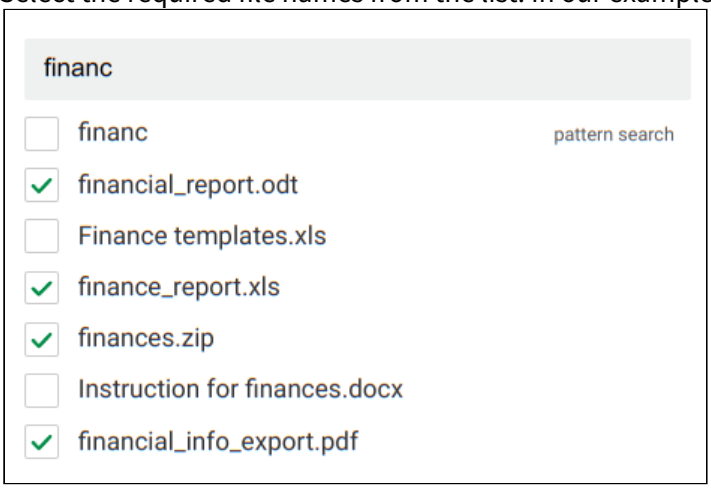
Configure filtering for the table entries:

1. In the **Task** filter, select from the list the name of the task within which the workstations were scanned. In our example:



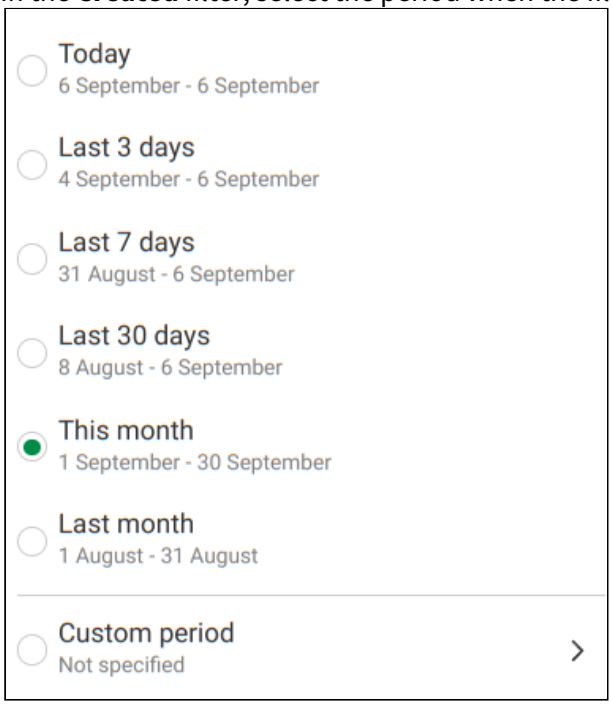
2. Select the values for the **File name** filter:

a. In the filter enter "financ" in the search bar;
b. Select the required file names from the list. In our example:



> **Note:**
>
> The display of items in the list does not depend on the conditions set in the other filters. For example, the list may contain files from the scan task that does not match the conditions of the **Task** filter.

3. In the **Created** filter, select the period when the file was created. In our example:



4. In the **Size** filter, set the file size limit. In our example:



Results of filtering

In the file information table, only entries that match all the specified conditions at the same time are displayed:

- name of the scan task within which the file was found is "Scan workstations";
- file name is "financial_report.odt", "finance_report.xls", "finances.zip" or "financial_info_export.pdf";
- file was created in the current month;
- file size does not exceed 5 megabytes.

Sort the table by the **Host** column:

Discovered	Task	Host	File name	Size	Created	Modified	Not found
9/2/2022 3:22:17 PM	Scan workstations	user1.example.com	finance_report.xls	4.21MB	9/2/2022 3:15:17 PM	9/2/2022 3:15:17 PM	--.--.-- --:--:--
9/4/2022 11:22:17 AM	Scan workstations	user2.example.com	finance_report.xls	4.21MB	9/4/2022 11:07:38 AM	9/4/2022 11:07:38 AM	9/4/2022 2:59:10 PM
9/4/2022 1:52:11 PM	Scan workstations	user2.example.com	financial_info_export.pdf	2.34MB	9/4/2022 1:42:36 PM	9/4/2022 1:42:36 PM	--.--.-- --:--:--
9/5/2022 6:01:29 PM	Scan workstations	user3.example.com	finance_report.xls	4.21MB	9/5/2022 5:59:20 PM	9/5/2022 5:59:20 PM	9/6/2022 1:01:29 PM
9/6/2022 1:01:15 PM	Scan workstations	user3.example.com	finance_report.xls	4.17MB	9/5/2022 5:59:20 PM	9/6/2022 1:00:01 PM	--.--.-- --:--:--

Files matching the conditions were found on 3 workstations:

Workstation name	Description of file entries
user1.example.com	The System found a **finance_report.xls** file on the host. The file was first found on 9/2/2022 at 3:22:17 PM. The **Not found** field is not filled, which is possible in the following cases: • The unmodified **finance_report.xls** file is still on the resource. • The file was modified/deleted, but the resource was not scanned again after the file was modified/deleted.
user2.example.com	1. The System found a **finance_report.xls** file on the host. The file was first found on 9/4/2022 at 11:22:17 AM. The **Not found** field is filled – the System did not find the file when the resource was scanned again. It is possible in the following cases: • The file has been deleted from the resource. • The file was modified. If the file matches the filter conditions in the task settings, a new entry about the modified file is added to the table when the resource is scanned again. **Note**: The entry about the modified file may not be displayed in the table if the file does not match the current table filters conditions. • The filter conditions in the task settings were changed and the file did not match the new conditions when the resource was scanned again.

Workstation name	Description of file entries
	2. The System found a **financial_info_export.pdf** file on the host. The file was first found on 9/4/2022 at 1:52:11 PM. The **Not found** field is not filled.
user3.example.com	1. The System found a **finance_report.xls** file on the host. The file was first found on 9/5/2022 at 6:01:29 PM. The **Not found** field is filled. 2. The System found a **finance_report.xls** file on the host. The file was first found on 9/6/2022 at 1:01:15 PM. The **Not found** field is not filled. The creation date and the file name are the same in both entries. The entries differ in file size and file modification date. The **finance_report.xls** file may have been edited on the host, then when scanning the resource again: a. The System found a modified **finance_report.xls** file and added a new entry to the table. b. The System did not find a **finance_report.xls** file with unchanged metadata, the System filled in the date and time in the existing entry when the file was not found again.

9.3 "Duplicate Files" Tab

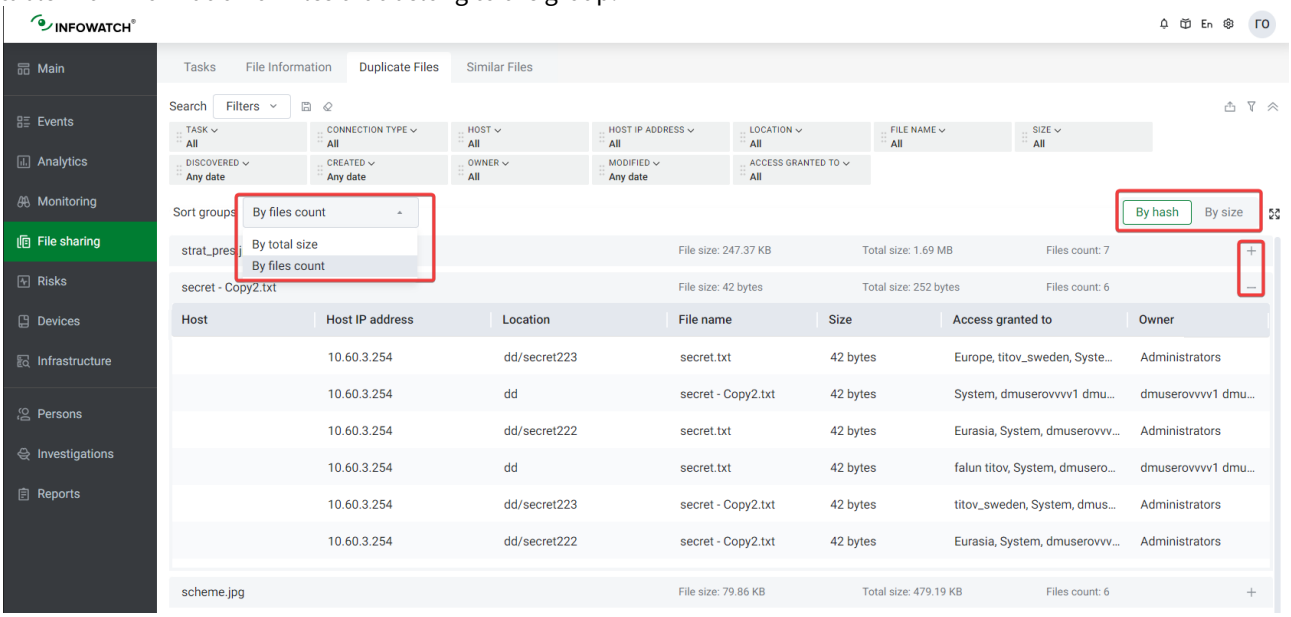
In addition to search for duplicates by template, Data Discovery is automatically comparing scanned files:

- ***by hash*** - if during scanning Data Discovery created file copies. For this, **you should select at least one of processing modes** in the scanning task. Otherwise Data Discovery can search for duplicates only by size because file copies for analysis are not created;
- ***by size*** (to byte) - if during scanning Data Discovery did not create file copies and collected only metadata.

Data Discovery considers files with the same hash or size as duplicates and groups them.

> **Note:**
>
> Precision of search by hash is significantly higher.

Results of search for duplicates are presented on the **Duplicate Files** tab as a list of groups where every entry is a table with information on files that belong to the group.



A group name is a name of the first file that belongs to it.
For each group, you can see:

- file size
- total group size
- number of files in the group

You can sort groups:

- **by total size**
- **by files count**

To open the group, click  on the right.

For each file in the group, the following information is displayed:

Field	Description
Host	Domain name of the host on which the file was detected
Host IP address	IP address of the host on which the file was detected If the address for SharePoint host is specified as a domain name in the host settings, do not fill this field
Location	File location – a full path to the directory where the file was detected **Note**: The full path contains the scanning resource path. For files from SharePoint hosts, the full path also contains path to the site.
File name	Name of the file. The System gets the value from the resource file system.

Field	Description
Size	Size of the file. The System gets the value from the resource file system.
Owner	In the **Owner** field, Data Discovery specifies the name of a person or a group that owns the file.
Access granted to	In the **Access granted to** field, Data Discovery specifies the list of names of users and groups that have access to the file. Data Discovery identifies persons and groups by the SID values it receives from the resource file system. To identify names by SID, Data Discovery uses data from Active Directory (see *Identification of persons and groups* in the "File Information Table" article). If there is no data to identify a person or group, Data Discovery would specify SID instead of the name, e.g.: `S-1-5-21-1922741149-3613679056-2300624754-1000`

To change the sorting order of file entries in the table, click on the heading of the column by which you want to sort. All entries in the table will be sorted in ascending/descending order of selected column values.

To move a column, left-click and hold the column heading, then move it along the heading row.

To change a column width, left-click and hold the right border of the column heading, then move the column border along the heading row.

To navigate through pages, use the bar in the bottom of the screen.

The listed fields and the way of filling them in are the same as in detailed file information table.

On the page with duplicates, you also can **view text contents extracted from the file.** For this, select a row in the table. On the right, the window with the file name and extracted text data will open. If there is no text data extracted from the file, the window contains only the file name.

> **Note:**
>
> Several text contents can be extracted from one file. But for analysis and displaying, they are merged into one content.
>
> **Example**: The extractor processes archive file containing several documents. Text contents may be extracted from each document. But, as a result, they will be merged into one common content related to archive file.

You can filter the data and **export the report**. For details, see "Exporting Reports with File Information".

9.4 "Similar Files" Tab

Data Discovery allows you to detect documents with the same terms, that are similar in meaning. The functionality is designed to detect storage locations and identify access rights to documents of the same subject. To search for similar files, Data Discovery interacts with InfoWatch Data Analysis Service (DAS). When scanning resources, Data Discovery extracts text contents from file copies, saves these contents in storage and notifies DAS that there is data

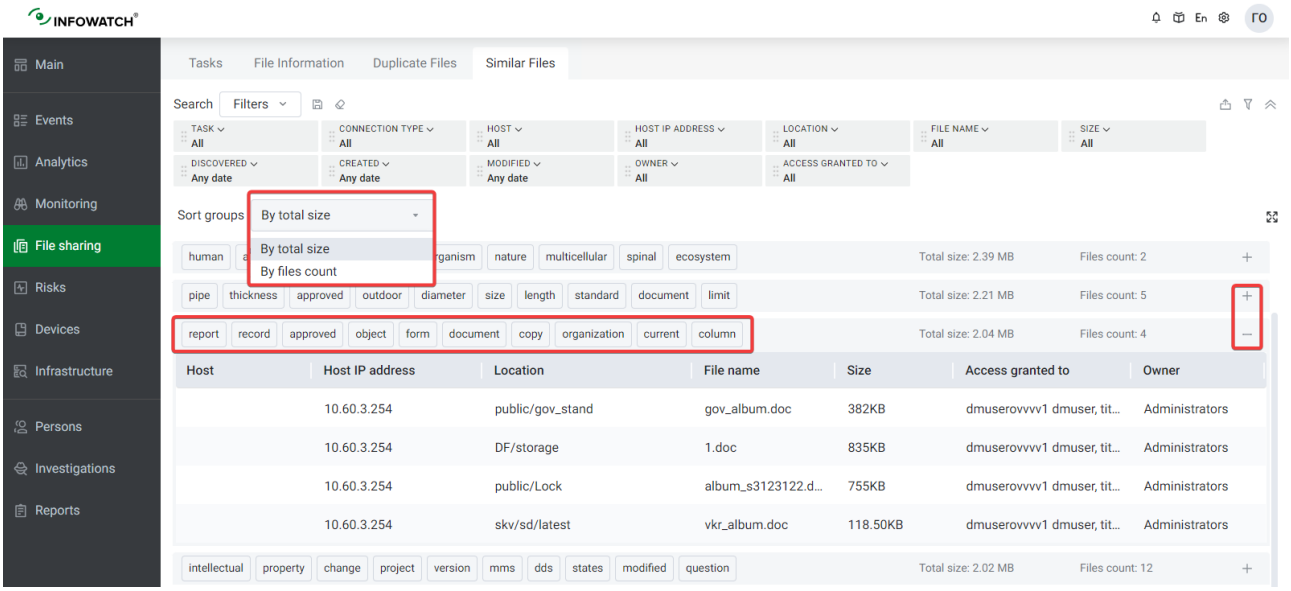
for processing. To find information about installing and configuring DAS, see "*InfoWatch Data Analysis Service. Installation, Configuration and Administration Guide*".

To search for similar files:

- DAS version 1.2.1 or later must be installed in the same cluster with Data Discovery;
- Protoclusterization functionality must be enabled in DAS;
- DAS must be configured for processing the node, on which Data Discovery is installed;
- The **Identify similar files** processing mode must be selected in the scanning task.

> **Note:**
>
> For a file to be processed for identifying the similar ones, the extracted **text content** must be **at least 200 bytes**. Content with size less than 200 bytes will not be sent for analysis to DAS but will be available for viewing in Investigation Centre web interface, in the **File sharing** section.
>
> Several text contents can be extracted from one file. But for analysis and displaying, they are merged into one content.
>
> **Example**: The extractor processes an archive file with several documents. Text contents may be extracted from each document. But, as a result, they will be merged into one common content related to the archive file.

Results of search for similar files are presented on the **Similar Files** tab as a list of thematic groups where every entry is a table with information on files that belong to the group. Grouped files are similar in meaning.



As a result of content analysis, groups containing the only one file can be created. It means that there are no files similar by meaning among scanned ones. Such uniqueness may also matter for Security Officer when analyzing file topic. You can also search for similar files on the File Information tab by selecting a file in table for comparison.

For each group, you can see:

- **tags defining the common meaning** of documents most closely
- total group size
- number of files in the group

You can sort groups:

- **by total size**

- **by files count**

To open the group, click  on the right.

For each file in the group, the following information is displayed:

Field	Description
Host	Domain name of the host on which the file was detected
Host IP address	IP address of the host on which the file was detected If the address for SharePoint host is specified as a domain name in the host settings, do not fill this field
Location	File location – a full path to the directory where the file was detected **Note**: The full path contains the scanning resource path. For files from SharePoint hosts, the full path also contains path to the site.
File name	Name of the file. The System gets the value from the resource file system.
Size	Size of the file. The System gets the value from the resource file system.
Owner	In the **Owner** field, Data Discovery specifies the name of a person or a group that owns the file.
Access granted to	In the **Access granted to** field, Data Discovery specifies the list of names of users and groups that have access to the file. Data Discovery identifies persons and groups by the SID values it receives from the resource file system. To identify names by SID, Data Discovery uses data from Active Directory (see *Identification of persons and groups* in the "File Information Table" article). If there is no data to identify a person or group, Data Discovery would specify SID instead of the name, e.g.: `S-1-5-21-1922741149-3613679056-2300624754-1000`

To view text content extracted from the file, select a row in the table. On the right, the window with the file name and extracted text data will open. If there is no text data extracted from the file, the window contains only the file name.

To change the sorting order of file entries in the table, click on the heading of the column by which you want to sort. All entries in the table will be sorted in ascending/descending order of selected column values.

To move a column, left-click and hold the column heading, then move it along the heading row.

To change a column width, left-click and hold the right border of the column heading, then move the column border along the heading row.

To navigate through pages, use the bar in the bottom of the screen.

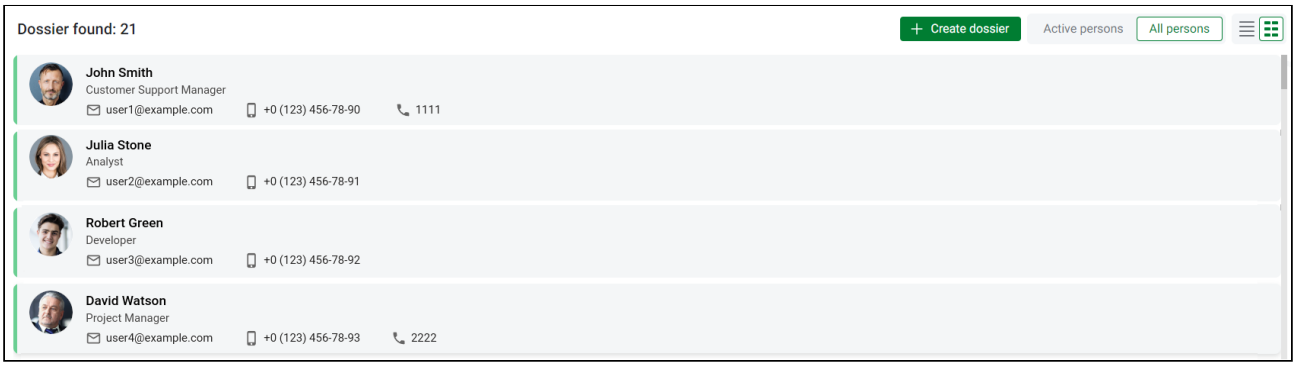
The listed fields and the way of filling them in are the same as in detailed file information table.

You can filter the data and **export the report**. For details, see "Exporting Reports with File Information".

10 Persons Section

The section contains the list of persons whose data has been loaded into the System during synchronization with Active Directory or Traffic Monitor and persons whose data has been added manually.

To view the person's dossier, click the desired person. To return to the list of persons, click  in the upper left corner of the working area.



The list of persons displays only those persons who satisfy conditions of the Single Filter. To specify what data should be displayed in the dossier, use buttons in the right part of the working area:

- **Active persons** – to display dossiers of employees with an active Active Directory account and at least one event

  > **Note:**
  >
  > If a person has no events or has been blocked/removed from AD, the person is considered inactive and is displayed in the **All persons** list.
  >
  > To view active persons, you can use filters. If no value has been selected in any filter, the System will not be able to display the list of active persons.

- **All persons** – to display dossiers of employees who have ever worked for the company
-  – to display the dossier as a table
-  – to display the dossier as a tile

In the section, the following actions are available:

- Viewing Person's Dossier
- Personal Information Tab
- File Sharing Tab
- Creating, Editing and Deleting Dossier
- Adding Contacts to Person
- Changing Person's Manager
- Merging Dossiers

10.1 Viewing Person's Dossier

A dossier contains information about a specific employee in the organization.

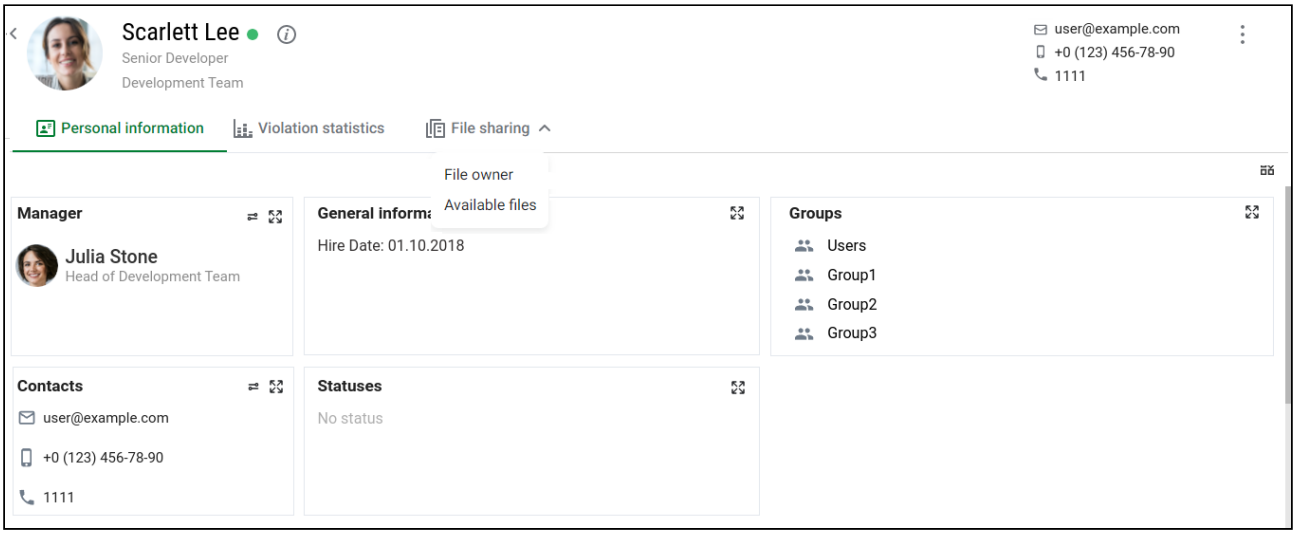
The person's dossier includes:

- Employee's full name

- Position
- Department
- Contacts: an account from Active Directory of the form *login@company.com*, personal or work contacts of the person, including email address, phone number, mobile phone number

> **Note:**
>
> All persons added during integration with Active Directory have a SID contact type used to identify them.

> **Important!**
>
> Do not create persons with the same SID contact value. If there are two or more persons with the same SID, the file owner may be determined incorrectly.



To the right of the person's name, you can see a color status indicator:

- – Active employees including manually added persons and persons whose AD account has expired, but their AD status remains *Active*
- – Employees disabled in AD or with undefined status

Click on the icon to the right of the person's name to view:

- Dossier creation date
- Dossier update date
- Dossier owner – either the System or a Security Officer (if the dossier was added manually)

To edit data in the dossier, use the button in the right part of the working area (for details, see "Creating, Editing and Deleting Dossier").

Information in the person's dossier is represented on the following tabs:

- **Personal information** – dossier widgets specific to a person: **General information**, **Manager**, **Contacts**, **Groups**, **Description** (see "Personal Information Tab")
- **Violation statistics** - statistics widgets related to this person (see "Statistics widgets");
- **File sharing** – click to choose one of the following subsection:
  - **File owner** – a table with files for which the person is considered the owner (see "File Sharing Tab")
  - **Available files** – a table with files the person can access (see "File Sharing Tab")

> **Note:**
>
> Tabs listed above are available if you have installed Data Discovery.
>
> The list of tabs in the person's dossier depends on products installed in the Investigation Centre cluster.

To return to the list of persons, click  in the upper left corner of the working area.

10.2 Personal Information Tab

Widgets on the **Personal information** tab allow the Security Officer to view all data about the person that is stored in the System.

Widget name	Description
Manager	Displays the list of the person's managers and allows you to specify a manager manually
Contacts	Allows you to manage the person's contacts and displays contacts loaded from connected sources
General information	Displays the person's state in different sources: employment date, length of service, workplace, date of dismissal
Groups	Displays the list of ActiveDirectory or InfoWatch Traffic Monitor groups the person belongs to
Description	Displays description of the person, allows you to specify and edit the description manually

To expand the widget to full screen, click .

To edit the widget data, click .

10.3 File Sharing Tab

File Sharing tab contains two subsections:

- **File Owner**

- **Available Files**

In both subsections, the structure of file tables is similar to the **File sharing** section (for details, see "File Information Table"), but here all entries refer to a selected person. On how to filter entries in tables, see "Single Filter".

You can also export a report with available data in XLSX format (for details, see "Exporting Reports with File Information").

Important!

Data Discovery uses SID to determine the file owner and whether a person has access to the file. If there is no SID in the person's contacts, then information in the **File owner** and **Available files** subsections will be unavailable.

Persons added as a result of integration with Active Directory (see "InfoWatch Data Discovery. Руководство администратора", "Integration with Active Directory") have a contact of SID type.

If a person is added manually, add a contact of SID type to the person (see "Adding Contacts to Person"). In this case, Data Discovery will be able to determine if the person owns a file or has access to files.

10.4 Creating, Editing and Deleting Dossier

To create a dossier for a new person or create an additional dossier for a person who already exists in the System:

1. Go to the **Persons** section.
2. Click + Create dossier.
3. Enter personal information:

Field	Description
Person name	Name of the person
Position	Position in the company
Department	Department where the person works
Description	Additional description

4. Add a photo.
5. Add the person's contact data: personal or work phone number, email address, social media accounts.
6. Click + Create dossier. The new dossier will be available in the list of employees' dossiers.

To edit a dossier:

1. Open the required dossier.,
2. Click ⋮ and select **Edit**.
3. Make changes and click **Save**.

To delete a dossier:

1. Open the required dossier.
2. Click  and select **Delete**.
3. Click **OK**. The dossier will be permanently deleted.

> **Important!**
>
> If you delete a dossier with a SID type contact, then as a result of identifying the person by this SID, the name of the person from the deleted dossier will be displayed in the **Owner** and **Access granted to** fields. To fix this, create a new dossier and add this SID to the new dossier.
>
> To update values of the **Owner** and **Access granted to** fields, run scanning again.

10.5 Adding Contacts to Person

During synchronization with Active Directory and InfoWatch Traffic Monitor, persons' contacts become available. These are phone numbers, email addresses, social network accounts (when synchronizing with TM), etc. The user can also add additional contacts manually. For this:

1. Open the person's dossier.
2. On the **Contacts** widget of the **Personal information** tab, click .
3. In the window that opens, you will see all person's contacts that were automatically added from other systems during synchronization. For the following contact types, only one contact is displayed by default: domain login, phone number, email, mobile phone number.
   Click **Add contact**.
4. Specify the contact type (e.g., personal phone number) and enter its value. You can add several values and add a comment to the contact. For this, click .

   > **Note:**
   >
   > To identify a person who owns a file or has access to a file, add a SID type contact for the person.

   > **Important!**
   >
   > Do not create persons with the same value of the SID contact. If there are two or more persons in the System with the same SID, the file owner may be determined incorrectly.

5. Click **Save**. The new contact is now available in the dossier. If the contact has a comment, the contact is marked with . To read the comment, click this icon.

10.6 Changing Person's Manager

The person's manager is displayed on the **Manager** widget of the **Personal information** tab. This information is added during synchronization with Active Directory and InfoWatch Traffic Monitor.

The Security Officer can change a person's manager if the manager was added manually. If the person's manager was loaded from a source system, changing the manager is not allowed.

To change the manager:

1. On the widget, click .
2. The **Changing manager** window will open. In the *Added manually* field, enter the name of a new manager. You can select a manager from the list of persons who are registered in the System. You can add multiple values.
3. Click **Save**. The selected managers will be displayed on the **Manager** widget.

To open a manager's dossier, click the manager name on the **Manager** tab.

10.7 Merging Dossiers

You can merge two or more dossiers into one. The new dossier will contain all information from dossiers that have been merged. To merge dossiers:

1. Open a dossier you want to merge with other dossiers.
2. Click  and select **Merge dossiers**.
3. A dialog box will open. In the **Merge with** field, select one or more dossiers from the list.
4. Click **Merge**.

11 Investigation Section

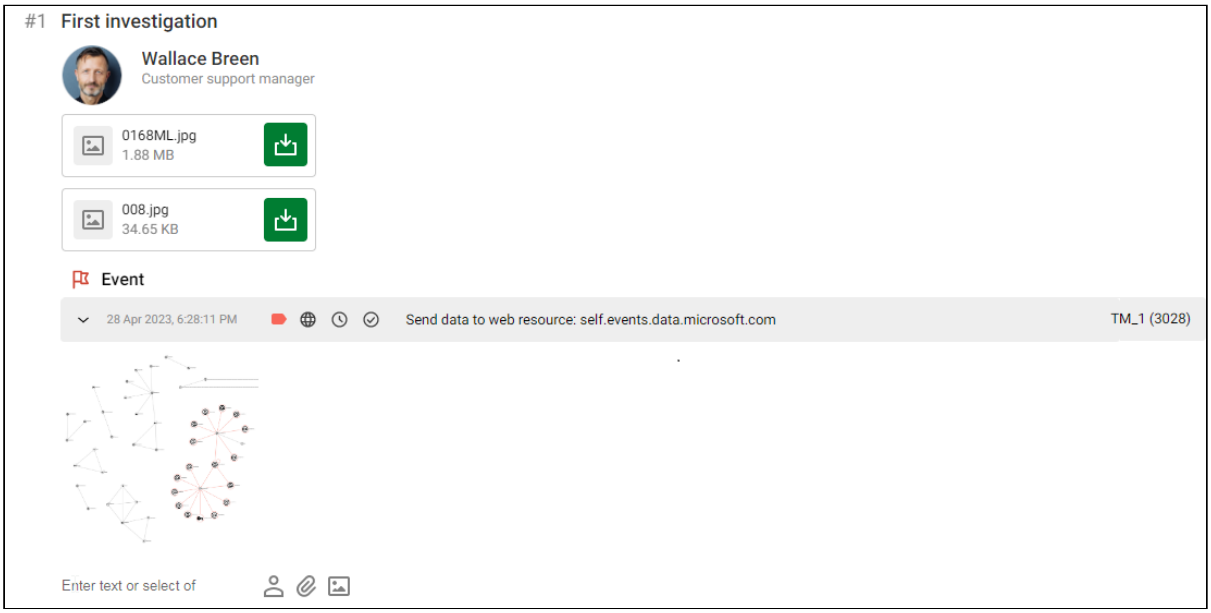
In this section, you can accumulate data collected from different sources and combine it into investigations. You can use this data for making decisions and creating reports. The order of collecting materials and conducting investigations is as follows:

1. The System represents facts of policy violations in the form of events.
2. A Security Officer manually adds such events to an investigation. If necessary, the Security Officer adds descriptions, attaches files, screenshots and dossiers of persons that are suspected of violating.
3. Other Security Officers join the investigation, add their comments and attach other materials. Thus, they can promptly exchange information.
4. The Security Officers represent the accumulated data as a report.

On the navigation bar, you can find a list of all open investigations and an archive – this is a base of investigations with which Security Officers are not currently working.

In the working area, you can see a card of the selected investigation. This card may include:

- Links to dossiers of persons involved in the investigation,
- Cards of the Traffic Monitor event from the **Events** section,
- Attachments in different formats,
- Graphic files,
- A free-form text comment.



To learn more about the attached materials, you can:

- Click on the dossier link and go to the **Persons** section,
- Click  and download the attached file to your hard drive,
- Open the event card and view detailed information on the event,
- View graphic files.

For your convenience, you can reorder materials in the investigation card. For this, hold  and drag the item to the right place. To delete an obsolete item, click  next to it and select **Delete**.

To update the state of the investigation card, click .

User's actions:

- Creating and Archiving Investigations
- Adding Event to Investigation
- Adding Materials to Investigation
- Collaborative Investigations
- Editing, Exporting, Printing and Deleting Investigation

11.1 Creating and Archiving Investigations

To create an investigation:

1. On the navigation bar on the left, click .
2. In the window that will open, enter the investigation name.
3. Click **Save**. A card of the new investigation will open and you will be able to add materials to it.

While you are working on the investigation, other users can accidentally edit the investigation data. To prevent it, we recommend archiving the investigation. We also recommend that you should archive the investigation when it is over.

To move the investigation to an archive:

1. On the navigation bar, open the required investigation.
2. Click .
3. In the window that will open, click **Move to archive**. Archieved investigations can be viewed, but they cannot be editied.

To get back to working on an archived investigation:

1. Select the required archive.
2. Click .
3. In the window that will open, click **Unarchive**. Now you can edit the investigation.

11.2 Adding Materials to Investigation

To conduct an investigation, you should collect materials as fully as possible. For this:

1. On the navigation bar, open the investigation card.
2. In the working area, you can add different kinds of materials to the investigation:
    - **Text comment** – for this:
        a. Enter a free-form comment.
        b. If necessary, format the comment text using built-in standard tools. Double-click the comment text to access the tools.
        c. Click **Enter** to save the comment.
    - **A link to the person's dossier** – for this:
        a. Click .
        b. In the window that will open, choose the required dossier and click **Accept**.
    - **Attachments** – for this:
        a. Click .
        b. On your computer's hard drive, select one or more files of any format and click **Open**. The process may take some time and depends on the size of the files. You can upload

up to 10 files at a time with a total size of up to 100 MB. Selected files will be attached to the investigation. These files can be viewed after a Security Officer who is admitted to the investigation downloads them to a hard drive.

- **Graphic files** – for this:
  a. Click  .
  b. On your computer's hard drive, select one or more graphic files with the following extenstions: jpg, jpeg, png, webp, gif and click **Open**. These files will be displayed in the card in preview mode.
- **A link to the Traffic Monitor event** (for details, see "Adding Event to Investigation")

11.3 Collaborative Investigations

Collaborative work on investigations is one of the main ways to achieve high efficiency. For this purpose, we have implemented co-editing of investigations. Features of co-editing investigations:

- If a Security Officer opens an investigation for viewing and at the same time another Security Officer adds new material, then the first Security Officer gets a notification that the investigation has been modified. The first Security Officer will be able to save their changes and changes made by both Security Officers will be merged.
- If a Security Officer is trying to save changes to the investigation that has already been deleted by another Security Officer, then the first Security Officer will get a notification saying that the removed material is unavailable and changes cannot be saved.

11.4 Editing, Exporting, Printing and Deleting Investigation

To change the investigation name:

1. Open the investigation card.
2. Click .
3. Enter a new name and click **Save**.

To export or print the investigation card:

1. Open the investigation card.
2. Click .
3. Choose either of the following:
   - Print PDF document. Information from the card will be printed using standard browser tools
   - Save an archive to a hard drive. The archive includes a PDF document with text data on the investigation and attachments

To delete the investigation:

1. Open the investigation tab.
2. Click .
3. In the window that opens, click **Delete**. The investigation can be deleted only by its author.

12 Reports Section

In this section, the following tabs are available:

- Reports Archive – where you can view the reports list and download the one you need;
- Tasks for Regular Sending – where you can view the list of tasks for regular sending, suspend sending or delete a task.

12.1 Reports Archive

In this section, you can view a table containing files of all reports created in the Investigation Centre.

For each report, you can view:

- Name
- Size
- Generation date and time

You can change the width of table columns.

To download the report, click on its name. The report file will be saved on your computer.

The table contains last 50 reports, which are stored for 30 days. If number of files exceeds 50, then the System will delete the earliest ones.

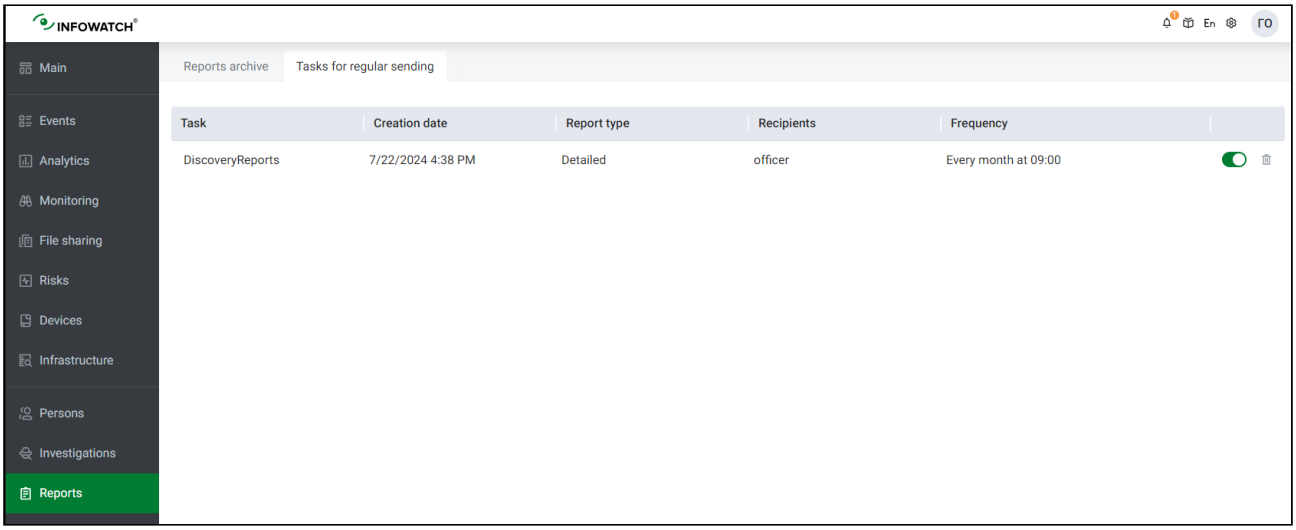
12.2 Tasks for Regular Sending

List of tasks for regular sending contains the fields:

- **Task** – the task name. By default, the name depends on the section or tab, from which the report is loaded. You can change the task name.
- **Creation date** – date and time, when the task is created
- **Report type** – detailed, presentational

> **Note:**
>
> Data Discovery generates only detailed reports.

- **Recipients** – users who are receiving the report
- **Frequency** – days and time when users need to receive the report



In the section you can:

- disable regular report sending;
- delete task from the list;
- change the columns width in the table header;
- change the task name.

To edit the task name, click on the current task name and enter a new one.

To disable a report sending for some time, use the switcher .

To delete the task:

1. In the task row, click .
2. In the window that opens, press OK. The task will be deleted.

13 End-user license agreement

ATTENTION! Please read the terms and conditions of this End-User License Agreement (hereinafter referred to as the "Agreement") carefully before using the Software. Installation and/or use of the Software shall constitute your full and unconditional acceptance of all the terms and conditions of this Agreement.

If you do not accept to the terms and conditions of this Agreement, or if you enter into this Agreement on behalf of a legal entity, but you do not have such authority, you must interrupt the installation and/or stop using the Software and delete any copies of the same.

1. Grant of License

1.1. This End-User License Agreement shall constitute a standard form contract entered into by and between the End User and Laboratory InfoWatch LLC, Taxpayer Identification Number (INN) 7734583888 (hereinafter referred to as the "Copyright Holder").

1.2. You shall be granted hereunder a personal, non-transferable and non-exclusive license granting you the rights to use the Copyright Holder's software (hereinafter referred to as the "Software"), namely the rights to download, install, backup and use the Software for its intended purpose, as set forth in the Software Documentation (User's Guide, Administrator's Guide, Installation Guide), this Agreement and your Licensing Agreement. The Software Documentation shall be an integral part hereof.

1.3. If you have received, downloaded and/or installed the Software for information, you may use the Software only for informational purposes and only during the trial period. Any use of the Software for any other purposes or after the trial period shall be prohibited.

1.4. If you use different Software versions or Software versions for different languages, or if you receive the Software on several media, or if you otherwise receive several copies of the Software or receive the Software as part of a package of other software, the total number of your licenses shall not exceed the number specified in your Licensing Agreement.

1.5. You may make a copy of the Software, provided that such a copy is intended only for archival purposes or to replace a duly purchased copy in cases where the original is lost, destroyed or becomes unusable. This copy may not be used for any other purposes and shall be destroyed, if the possession of the Software copy ceases to be legitimate.

1.6. You shall be solely responsible and ensure compliance with the applicable export and import laws, as well as applicable trade sanctions and embargo with regard to the transfer of rights and use of the Software.

2. Terms and Conditions of Use

2.1. The Software is intended solely for use in analyzing and checking corporate information, and specifically for detecting and preventing unauthorized use of confidential data.

2.2. You shall use the Software exclusively in accordance with the applicable statutory requirements of the Russian Federation or any other countries, including but not limited to:

- legislation on personal data;
- legislation on trade secrets;
- legislation on combating the misuse of insider information;
- anti-corruption legislation;
- legislation on critical information infrastructure;
- legislation on state information systems;
- National Payment System data protection legislation;
- statutory provisions of the Labor Code of the Russian Federation;
- and any other applicable sources of law.

2.3. You may not use the Software for any purpose or in any manner that violates applicable law, this Agreement and/or any third-party rights. The Copyright holder shall not be liable for your breach of this Clause.

2.4. If you are not a resident of the Russian Federation, you shall use the Software in full compliance with applicable law, including laws and regulations on personal data and trade secrets.

2.5 In order to avoid your violation of Articles 138, 138.1 of the Criminal Code of the Russian Federation, Article 23 of the Constitution of the Russian Federation or any other legal regulations, before using the Software, you are recommended to inform your employees in writing that:

- this Software is used on the workstation computer, and you should inform your employees about the features of this Software and data that may be obtained as a result of using the same;
- the workstation computer and the communication channels used (e-mail, local corporate network, telephony and IP-telephony, etc.) are your property and are provided to such employees for temporary use so that they could perform their official duties;
- all information created on the workstation computer is your property and is subject to inspection and logging on a legal basis.

2.6. Please note that you need to obtain the consent of the persons whose data will be processed when using the Software to the relevant collection, storage, use, processing, distribution or any other form processing of such data.

2.7.We remind you that installing a program on another person's computer without prior authorization from such a person, obtaining personal data of any third party without his/her consent, violating privacy, breaching confidentiality of correspondence, telephone conversations, telegraphic or any other types of messages shall constitute a criminal offense and shall be prosecuted, as prescribed by the current legislation of the Russian Federation.

2.8. The Copyright Holder shall not have access to the data created and/or obtained as part of your use of the Software. You shall have exclusive access to such data. In view of the foregoing, the Copyright Holder shall not collect, or store, or use, or distribute such data.

2.9. You shall be held fully and solely liable if the information created and/or obtained as part of your use of the Software contains unethical and/or illegal data, computer viruses or programs (or links thereto) that may interrupt or disrupt the standard features of computer equipment, software or telecommunication facilities of any persons.

2.10. You shall not:

- edit, change, adapt, translate, decompile, disassemble, reverse engineer, create or recreate source code, modify the Software, whether in full or in part, or attempt to do so in respect to the Software, or create derivative works based on the Software and/or Documentation;
- integrate the Software into any third-party website or software product, except where such integration is expressly authorized by the Copyright Holder and the copyright holder of such a third-party website or software product. The Software may be integrated into a third-party software product only through the built-in API module within the features of such a module;
- remove or modify any notice of copyright, proprietary rights to the Software and/or its elements;
- let on lease, rent out or assign for temporary use, sublicense, distribute the Software or otherwise grant any rights with respect to the Software and/or Documentation to any person or entity, or disclose the results of development testing of the Software, except as expressly permitted by this Agreement;
- allow or provide access to the license key to any third party in violation of this Agreement and your Licensing Agreement;
- collect, systematize, store or disseminate confidential and/or personal information in violation of law;
- use the Software for unauthorized remote access to any third-party computer, or changing the configuration of the computer so that such features can appear in third-party software;
- use the Software to collect, store, use, process or distribute any information or data that violates the applicable laws or the rights and legitimate interests of any third party.

2.11. The license key shall constitute confidential information. The Copyright Holder shall reserve the right to use the means to verify the authenticity of the license key you have installed.

3. Disclaimer of Warranties and Liability of the Copyright Holder

3.1. The Software shall be provided "AS IS" in accordance with generally accepted international business practice. The Copyright Holder shall not be responsible for any issues arising during the operation of the Software, including those related to compatibility with other software products (packages, drivers, etc.) or those arising from ambiguous interpretation of the Software Documentation, or for failure of the Software to meet your expectations, etc.).

3.2. You confirm that you are aware of the features of the Software licensed hereunder, and you shall be held fully and solely liable if the Software or the results of using the same do not meet your individual needs or expectations.

3.3. The Copyright Holder makes no warranties, whether expressed or implied, and hereby disclaims all warranties.

The Copyright Holder makes no representations or warranties regarding the fact that the Software is flawless, uninterrupted, fast or accurate, and shall not be resposible for any direct or indirect consequences of using the Software, including those arising from potential errors in the Software or typos in the Software and/or the Documentation attached thereto or further provided.

The Copyright Holder makes no representations or warranties and shall not be held liable in the event of any changes in a third-party software that occurred after the installation/deployment of the Software and resulted in a loss of the Software features (including, but not limited to, a change in the data transfer protocol, data storage format, logic of a third-party software, an update of a software that ceases to support the Software). The Copyright Holder makes no warranties or representations and sets no conditions or provisions (whether expressed or implied) in respect of any matters, including but not limited to infringement of third-party rights, merchantability, integration or fitness for a particular purpose.

3.4. Any data that becomes available in the course of using the Software, including any file, link, image or text, shall be the sole responsibility of a disclosing person or entity and shall be the property of such a person or entity. This Agreement shall not grant any rights to such data. You accept responsibility for any potential consequences, security risks or any losses caused by such data viewed or obtained by using the Software.

3.5. The Copyright Holder shall not be liable for any loss or damage, regardless of the reasons for such a loss or damage (including, but not limited to, direct or indirect damage, lost profits, interruption of commercial or business activities, loss of business information, negligence, or any other losses) arising from the use or failure to use the Software, even if the Copyright Holder has been notified of eventual losses. The Copyright Holder's liability shall be at any time limited to reimbursement for documented actual damage of up to Three thousand rubles (RUB 3,000). The Copyright Holder shall be held liable only for actions or omissions for which the Copyright Holder is at fault.

4. Intellectual Property Rights

4.1. You hereby agree that the exclusive rights to any intellectual property embodied in the Software and /or any documentation provided to you shall be vested in the Copyright Holder. The Software shall not be assigned or sold hereunder, but shall be granted under a limited non-exclusive license. Nothing in this Agreement grants you any rights to the said intellectual property other than those conferred under your Licensing Agreement.

4.2. You agree that the source code and the license key for the Software shall be the property of the Copyright Holder.

5. Rights to Information Accessed in the Course of Performance of this Agreement

5.1. You hereby agree that you do not have any rights to any information that is not subject to intellectual property, as set forth in Section 4, accessed by you in the course of performance hereof.

5.2. Such information may include, but shall not be limited to, any system, method of work or any other information.

5.3. You may use the above information only for the purpose of exercising the rights to the Software granted under this Agreement and the Licensing Agreement. You may not use this information for your own benefit or beyond the scope of this Agreement or the Licensing Agreement.

6. Open Source Software

You are aware that the Software contains open source software distributed under certain licenses, available in the licenses.inf license.info file, provided together with the Software as part of the distribution kit.

7. Miscellaneous

7.1. This Agreement shall come into force immediately upon installing the Software and shall be valid throughout the period of use.

7.2. In the event that you violate any of the provisions of this Agreement and/or your Licensing Agreement, or in the event that the performance of this Agreement and/or your Licensing Agreement violates or may result in violation of the applicable laws, the Copyright Holder shall have the right to repudiate and terminate this Agreement (including, but not limited to, by blocking the license key, terminating the grant of rights) at any time without prior notice. To the extent set forth above in this Clause, you shall not be reimbursed for the cost of the Software/the amount of the Software license fee, or any eventual losses arising under this Agreement and/or your Licensing Agreement.

7.3. Upon termination or expiration hereof, the End User shall destroy all copies of the Software installed.

7.4. You are hereby notified that when the usage statistics collection function is enabled in the Software, the Software may collect and record anonymized information on:

a) hardware, operating system of the device on which the Software is installed and operated, as well as other technical data containing information on the use of hardware, hardware load and processing speed of the Software;

b) end user's activities in the user interface of the Software.

The Copyright Holder may use the specified data to make decisions on the development of features and improvement of the operational characteristics of the Software.

The Copyright Holder shall ensure the confidentiality of the data obtained and shall not disclose or transfer such data to any third party.

7.5. This Agreement, as well as any disputes arising out of or in connection with this Agreement, shall be governed by and construed in accordance with the laws of the Russian Federation.

7.6. All disputes and disagreements arising during the performance hereof shall be resolved through negotiations, if possible. In case of failure to reach an agreement, disputes shall be settled in court in the Arbitration Court of Moscow.

7.7. This Agreement may be amended by the Copyright Holder at any time without prior notice. The amended Agreement shall enter into force once it is posted at the address specified below in this Clause. The current version of the Agreement is always available at https://kb.infowatch.com/display/EULA/. By continuing to use the Software, you accept the amendments and/or additions made to this Agreement. If you do not accept the terms and conditions of such amended Agreement, you shall immediately discontinue the use of the Software.

8. Contact Details of the Copyright Holder

Tel./Fax: +7(495)229-00-22

Sales Department: sales@infowatch.com

Technical Support: support@infowatch.com

Web site: www.infowatch.ru

14 Glossary

Term	Definition
Administrator	Administrator is a user of the System who installs, configures and maintains the System. See also: Security Officer, User
SMB group	A parameter that makes it possible to use the System as part of a specific workgroup for network management of folders located on workstations, terminal servers, file storages, etc. (for example, a domain)
Scanning task	An operation of checking resources (scanning paths) for sensitive data. A task contains a description of scanning resources, downloading files from the resources and sending those files to Traffic Monitor. The task includes a list of hosts, a list of scanning paths, a filter by names of scanned files, a filter by size of scanned files, a filter by directories and a schedule to start and stop the task automatically
Zone	A group of Data Discovery servers. A zone can include several servers from the System cluster. Each zone has its own settings for data export to Traffic Monitor
User interface	A set of means and methods by which the user interacts with the System
Information about file	Information about a file includes file metadata and information about scanning the file/resource. Information about scanning a file/resource includes: • task name • file discovery date • IP address of the host, etc. See also: File metadata
Management Console	Graphical user interface for managing Data Discovery (administrating and configuring the System, working with scanning tasks)
License	Right to use the System. The right is obtained by purchasing the System and defines allowed actions with the System
Mask	A search template. This is a method of describing a search query using metacharacters. Masks are used in file name filters and in folder filters
File metadata	

Term	Definition
	Information about a file Data Discovery gets from a file system of the resource: • Full path to the file • File name • File size • Date and time the file was created • Date and time the file was modified • File owner • User/groups that can access the file See also: Information about file
Security Officer	The main user of the Management Console. Also, a preset user role in the Management Console with privileges for all actions in the System, except for administrative ones
Task execution plan	The order of actions for the scanner service (Puller) generated by the Execution Controller service. Defines the sequence of processing resources
User	A user of Data Discovery: an Administrator, a Security Officer, etc. See also: Administrator, Security Officer
Platform	Principal design of the product. This is a set of parts, subsystems, interfaces and production processes. The product Platform is the basis for deploying a product family and assembling finished products from standardized components
Pass	Complete scanning of a single resource (all files on the resource must be processed by the scanner service (Puller)). Pass may consist of several sessions
Work station/Computer	In terms of the System, a controlled workstation or terminal device
Schedule	Schedule for automatic start and stop of the task. Both daily and weekly schedule can be used
Resource	A host resource which contains files. The System scans files on resources and downloads files to storages. To connect to a resource, you need to specify the scanning path for the host
Session	Connecting to a resource while the System passes through it. Duration of a scanning session (the **Processing duration** parameter) is specified in Advanced Settings of the task

Term	Definition
ClickHouse service	Distributed analytical columnar DBMS designed for real-time analysis of stored data. Product Platform component
PostgreSQL service	DBMS that is used in the Data Discovery to store settings, configuration task, and statistics on task. Product Platform component
Folder filters	A set of conditions for processing directories on resources. Is implemented as a list of directories included in processing and a list of directories excluded from processing
File filters	A set of conditions in task settings for scanning and downloading files from resources. Is implemented as a list of masks for filtering by file name and filters by maximum and minimum file size
Host	Workstation/server to which the System connects to scan files and download them for further analysis. Contains a list of resources for scanning
Storage	Temporary storage for downloaded files. Files are stored there until they are sent to Traffic Monitor
Data export to Traffic Monitor	Description of connection to Traffic Monitor. Data export allows files downloaded from resources to be sent to Traffic Monitor and then analysed. Indicates a specific installation of Traffic Monitor, an authorization token and a list of XAPI addresses are required. In the Management Console, data export settings are in the **Zones and servers** section
Contents Processor	The service transfers data to Ripper and notifies DAS. A pair of instances of Ripper and Contents Processor services is an extractor. It extracts text data from files and notifies DAS that there is data for processing
Discoperator	The service is a backend and it is used for the Management Console operation
FTP	Network data transfer protocol
InfoWatch Data Analysis Service (DAS)	Data Analysis Service is designed for flow post-processing and analysis of data provided by InfoWatch company products. Module of text data protoclusterization serves Data Discovery
InfoWatch Traffic Monitor	Traffic Monitor is a software package designed to control different types of traffic (SMTP, IMAP, POP3, HTTP, HTTPS, IMAP, XMPP, ICQ, NRPC) and shadow copies of data that are copied to removable media and sent for printing

Term	Definition
Kubernetes	Containerized application orchestration software. Orchestration includes automation of application deployment, scaling and coordination in a cluster environment
NATS	Messaging system for components
Nextcloud	A set of client-server programs for creating and using data storage
Puller	Service for scanning resources and downloading files from them. Scanners are instances of the Puller service. They collect files metadata and download files from resources to a storage
Ripper	Service for extracting text data from files. A pair of instances of Ripper and Contents Processor services is an extractor. It extracts text data from files and notifies DAS that there is data for processing
Sender	Service for sending downloaded files to Traffic Monitor. Senders are instances of the Sender service. Senders send data from a storage to Traffic Monitor for analysis. After a file is sent to Traffic Monitor, the sender removes it from the storage
SMB (Server Message Block)	Network protocol for remote access to files, printers and other network resources. Is also used for interprocess communication
SSH (Secure Shell)	A network protocol that allows remote control of the operating system and tunneling of TCP connections (for example, for sending files)
Task Execution Controller	Service for task execution management. Controls how other services work on the task
TPS Agent	Storage status monitoring service. Helps Puller and Sender services to work with the storage, provides statistics on the storage
WebDAV	HTTP protocol extension that allows users to manage files stored on remote servers
XAPI	Private API for sending files to Traffic Monitor