Netwrix Data Classification for Google Drive

Quick-Start Guide

Version: 5.6.1
8/12/2021

Table of Contents

Legal Notice

The information in this publication is furnished for information use only, and does not constitute a commitment from Netwrix Corporation of any features or functions, as this publication may describe features or functionality not applicable to the product release or version you are using. Netwrix makes no representations or warranties about the Software beyond what is provided in the License Agreement. Netwrix Corporation assumes no responsibility or liability for the accuracy of the information presented, which is subject to change without notice. If you believe there is an error in this publication, please report it to us in writing.

Netwrix is a registered trademark of Netwrix Corporation. The Netwrix logo and all other Netwrix product or service names and slogans are registered trademarks or trademarks of Netwrix Corporation. Microsoft, Active Directory, Exchange, Exchange Online, Office 365, SharePoint, SQL Server, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. All other trademarks and registered trademarks are property of their respective owners.

Disclaimers

This document may contain information regarding the use and installation of non-Netwrix products. Please note that this information is provided as a courtesy to assist you. While Netwrix tries to ensure that this information accurately reflects the information provided by the supplier, please refer to the materials provided with any non-Netwrix product and contact the supplier for confirmation. Netwrix Corporation assumes no responsibility or liability for incorrect or incomplete information provided about non-Netwrix products.

1. Introduction

This guide is intended for the first-time users of Netwrix Data Classification for Google Drive. It can be used for evaluation purposes, therefore, it is recommended to read it sequentially, and follow the instructions in the order they are provided. After reading this guide you will be able to:

- Prepare your IT infrastructure for scanning
- Install and configure Netwrix Data Classification
- Add a source to start crawling Google Drive
- Review classification results
- Leverage reporting capabilities and export results for custom reports

NOTE: This guide only covers the basic configuration and usage options for crawling Google Drive with Netwrix Data Classification. For advanced installation scenarios and configuration options, as well as for information on various reporting possibilities and other product features, refer to Netwrix Data Classification Online Help Center.

2. Configure G Suite for Crawling

Netwrix Data Classification for Google Drive uses the **OAuth 2.0** protocol to authenticate to your G Suite domain. You will need to create a service account and authorize it to access data in individual and shared Drives on behalf of users using the Google Drive API. Do the following:

In Google Cloud Platform web console:

1. Create a new project
2. Select Application type
3. Create a new service account
4. Create a service account key (JSON, save a copy for the data source configuration)
5. Enable G Suite domain-wide delegation for the service account (write down the Client ID)
6. Enable Google Drive API

In G Suite Admin Console:

1. Authorize service account to access the Google Drive API

To configure G Suite for crawling

IMPORTANT! Google administrative interfaces tend to change over time, so refer to the following guide for up-to-data instructions on creating OAuth 2.0 service accounts: Using OAuth 2.0 for Server to Server Applications.

Review the following for additional information:

To...	Do...
Create a new project	1. Navigate to https://console.developers.google.com (Google Cloud Platform web console) while logged in as a G-Suite administrator within the domain to be crawled (if the user is not added within the correct domain then the correct data will not be identified). 2. Create a new project.
Select Application type	1. Once a new project has been created, navigate to **APIs&Services → OAuth consent screen**. 2. Set **User type** to "*Internal*". 3. Provide the name for new application. 4. Click **Save**.

To...	Do...
Create a new service account	1. In **Google Cloud Platform** web console, navigate to **Credentials** and click **Create Credentials**. 2. Then, click **Service account**. 3. Create service account as described in Google official article. 4. On the **Grant this service account access to project (optional)** step, do not select any roles. 5. On the **Grant users access to this service account (optional)** step, do not grant any user access. Click **Done**.
Create a service account key	1. On the **Service accounts** section, click edit on the account you want to create a key for. 2. Click ⋮ icon under **Actions** and select **Create key**. 3. In the **Create private key for <Service account name>** dialog, select **JSON** format, and download the file to a known location as it will be required later. **NOTE:** Your new public / private keypair is generated and downloaded to your machine; it serves as the only copy of this key. You are responsible for storing it securely. If you lose this keypair, you will need to generate a new one.
Delegate domain-wide authority to the service account	1. On the **Service accounts** section, select your service account and click **Edit**. 2. Click the **Show Domain- Wide Delegation** link and tick the **Enable G Suite Domain-wide Delegation** checkbox. 3. Click **Save**. 4. Once completed, review the "*Domain wide delegation*" column for this account and make sure that the delegation enabled. 5. Click the **View Client ID** link. 6. Copy your Client ID, you will need it later.
Enable Google Drive API	1. In **Google Cloud Platform** web console, navigate to the **API Dashboard** and select **Enable APIs and Services** (if APIs have not previously been enabled). 2. Search for **Google Drive API** and click **Enable** (or **Manage**). 3. Search for **Admin SDK API** and click **Enable** (or **Manage**).

To...	Do...
	4. Switch to **G Suite Admin Console**. 5. Navigate to **Security** → **API Controls** → **Manage Domain-wide Delegation** within the Google admin portal. 6. Set the client name to the **Client ID** you copied on the previous step. 7. Set the API scopes and select **Authorize**: • https://www.googleapis.com/auth/drive • https://www.googleapis.com/auth/admin.directory.user

3. Install Netwrix Data Classification

1. Run **Netwrix_Data_Classification.exe**.
2. Review minimum system requirements and then read the License Agreement. Click **Next**.
3. Follow the instructions of the setup wizard. When prompted, accept the license agreement.
4. On the **Product Settings** step, specify path to install Netwrix Data Classification. For example, *C:\Program Files\NDC\*.
5. On the **Configuration** step, specify the directory where **Index files** reside. For example, *C:\Program Files\NDC\Index*.
6. On the **SQL Database** step, provide SQL Server database connection details.

   Complete the following fields:

Option	Description
Server Name	Provide the name of the SQL Server instance that hosts your NDC SQL database. For example, *"WORKSTATIONSQL\SQLSERVER"*.
Authentication Method	Select Windows or SQL Server authentication method.
Username	Specify the account name.
Password	Provide your password.
Database Name	Enter the name of the SQL Server database. Netwrix recommends using **NDC_database** name.

7. On the **Licensing** step, add license. You can add license as follows:
   - Click the **Import** button and browse for your license file

     OR
   - Open your license file with any text editor, e.g., **Notepad** and paste the license text to the **License** field.
8. On the **Administration Web Application** step, review default IIS configuration.
9. On the **Services** step, configure Netwrix Data Classification services:

- Select all services to be installed.
- **File System Path**—Use default path or provide a custom one to store Netwrix Data Classification's Services files. For example, *C:\Program Files\NDC Services.*
- Provide user name and password for the product services service account.

  NOTE: This account is granted the **Logon as a service** privilege automatically on the computer where NDC is going to be installed.
- Select additional service options, if necessary.

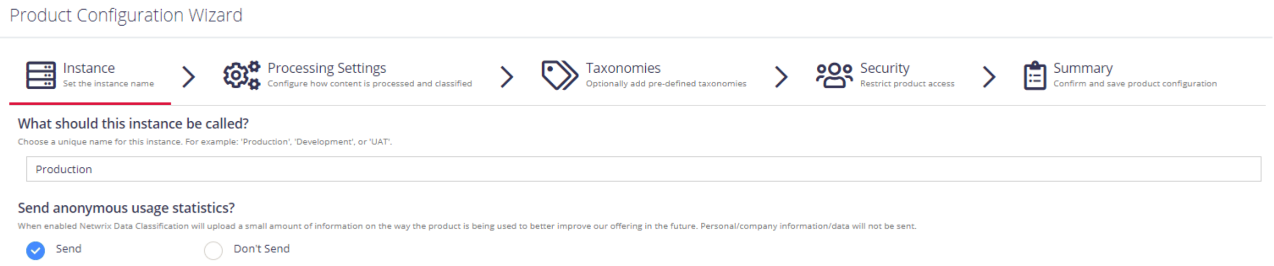
10. On the **Pre-Installation Tasks and Checks** step, review your configuration and select **Install**.
11. When the installation completes, open a web browser and navigate to the following URL: *http://localhost/conceptQS* where **localhost** is the name or IP address of the computer where Netwrix Data Classification is installed. For example, *http://workstationndc/conceptQS*.

4. Initial Product Configuration

The **Product Configuration Wizard** allows you quickly configure basic Netwrix Data Classification settings such as processing mode, taxonomies, etc.

In your web browser, navigate to the following URL: http://hostname/conceptQS where **hostname** is the name or IP address of the computer where Netwrix Data Classification is installed and perform initial configuration steps.

On the **Instance** step, provide the unique name for your Netwrix Data Classification instance. For example, *"Production"*.


Product Configuration Wizard
Instance
Set the instance name
Processing Settings
Configure how content is processed and classified
Taxonomies
Optionally add pre-defined taxonomies
Security
Restrict product access
Summary
Confirm and save product configuration

What should this instance be called?

Choose a unique name for this instance. For example: 'Production', 'Development', or 'UAT'.

Production

Send anonymous usage statistics?

When enabled Netwrix Data Classification will upload a small amount of information on the way the product is being used to better improve our offering in the future. Personal/company information/data will not be sent.

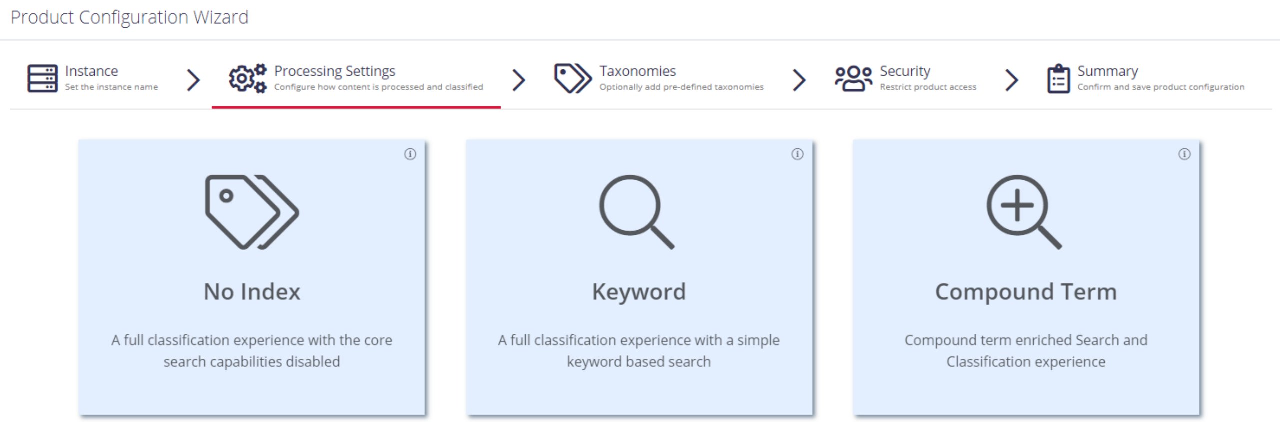
● Send ○ Don't Send

Click **Next** to proceed. See also:

- Select Processing Mode
- Processing Settings
- Add Taxonomy
- Security
- Configure Health Alerting
- Review Your Configuration

4.1. Select Processing Mode

At this step of the wizard, select processing (indexing) mode for your environment.



Product Configuration Wizard

Instance
Set the instance name
Processing Settings
Configure how content is processed and classified
Taxonomies
Optionally add pre-defined taxonomies
Security
Restrict product access
Summary
Confirm and save product configuration

No Index
A full classification experience with the core search capabilities disabled

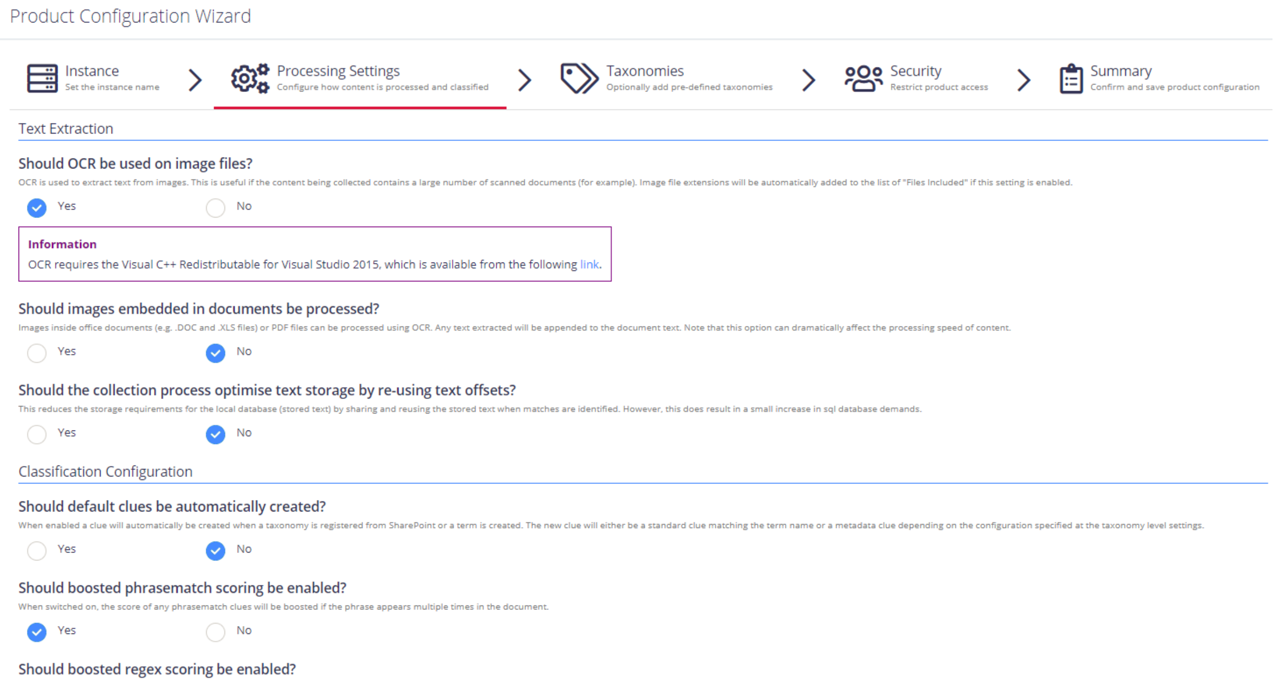
Keyword
A full classification experience with a simple keyword based search

Compound Term
Compound term enriched Search and Classification experience

For starter and evaluation purposes, select **Keyword** mode.

4.2. Processing Settings

On the **Processing Settings** step, review options for data processing and classification. For test and evaluation purposes, Netwrix recommends use default values.



Product Configuration Wizard

Instance
Set the instance name
Processing Settings
Configure how content is processed and classified
Taxonomies
Optionally add pre-defined taxonomies
Security
Restrict product access
Summary
Confirm and save product configuration

Text Extraction

Should OCR be used on image files?
OCR is used to extract text from images. This is useful if the content being collected contains a large number of scanned documents (for example). Image file extensions will be automatically added to the list of "Files Included" if this setting is enabled.
● Yes ○ No

Information
OCR requires the Visual C++ Redistributable for Visual Studio 2015, which is available from the following link.

Should images embedded in documents be processed?
Images inside office documents (e.g. .DOC and .XLS files) or PDF files can be processed using OCR. Any text extracted will be appended to the document text. Note that this option can dramatically affect the processing speed of content.
○ Yes ● No

Should the collection process optimise text storage by re-using text offsets?
This reduces the storage requirements for the local database (stored text) by sharing and reusing the stored text when matches are identified. However, this does result in a small increase in sql database demands.
○ Yes ● No

Classification Configuration

Should default clues be automatically created?
When enabled a clue will automatically be created when a taxonomy is registered from SharePoint or a term is created. The new clue will either be a standard clue matching the term name or a metadata clue depending on the configuration specified at the taxonomy level settings.
○ Yes ● No

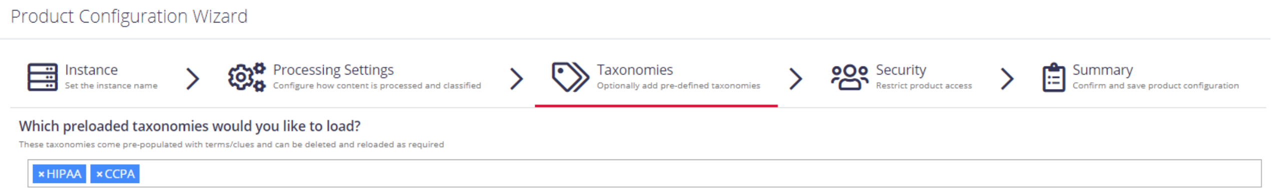
Should boosted phrasematch scoring be enabled?
When switched on, the score of any phrasematch clues will be boosted if the phrase appears multiple times in the document.
● Yes ○ No

Should boosted regex scoring be enabled?

Proceed with adding taxonomies.

4.3. Add Taxonomy

On this step, you are prompted to load predefined taxonomies.



Product Configuration Wizard

Instance
Set the instance name
Processing Settings
Configure how content is processed and classified
Taxonomies
Optionally add pre-defined taxonomies
Security
Restrict product access
Summary
Confirm and save product configuration

Which preloaded taxonomies would you like to load?

These taxonomies come pre-populated with terms/clues and can be deleted and reloaded as required

×HIPAA ×CCPA

Click the search bar and select one or several taxonomies you want to add. See Built-in Taxonomies Overview for the full list of built-in taxonomies supported by Netwrix Data Classification.

4.4. Review Your Configuration

On this step, review your configuration. Once you complete the wizard, you can:

- Add a Source
- Add a Taxonomy
- Take the Product Tour
- Get Help

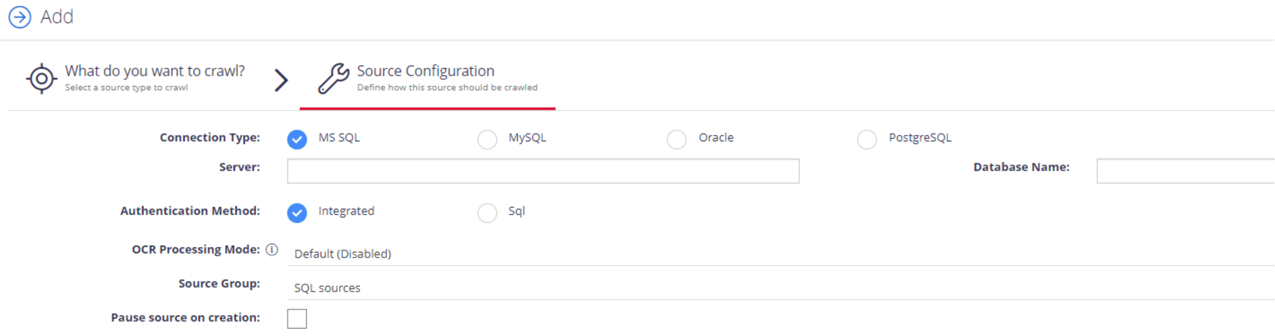
5. Add Database Source

The **Database** source configuration screen allows you to enable the crawling and classification of content stored in your Microsoft SQL Server, MySQL, and Oracle databases.

Content must either be configured / crawled using the configured service accounts (IIS Application Pool User, Windows Services) or by using specific connection details.

Once connected it is possible to create an intelligent content mapping, crawling certain fields as unstructured index text, and other fields as mapped metadata. For more information please see the Database Configuration Wizard section.

If you wish to make other configuration changes before collection of the source occurs ensure you tick the checkbox "*Pause source on creation*".


Add
What do you want to crawl?
Select a source type to crawl
Source Configuration
Define how this source should be crawled

Connection Type: ● MS SQL ○ MySQL ○ Oracle ○ PostgreSQL

Server:

Database Name:

Authentication Method: ● Integrated ○ Sql

OCR Processing Mode: Default (Disabled)

Source Group: SQL sources

Pause source on creation: □

Complete the following fields:

Option	Description
Connection Type	Select your connection type: MS SQL, MySQL, or Oracle.
Server	Specify the server name of the database system to be crawled ("." can be used to indicate the local server).
Database Name	Specify the database that will be crawled. It is possible to configure multiple databases from the same server.
Authentication Method	Select authentication method: **Integrated** or **SQL**. • With **Integration** option selected, database will be accessed under the account currently logged on. • With **SQL** option selected, specify user name and password to be used when accessing the database.
OCR Processing Mode	Select processing mode for images in the documents:

Option	Description
	• **Disabled** – documents' images will not be processed. • **Default** – defaults to the source settings if configuring a path or the global setting if configured on a source. • **Normal** – images are processed with normal quality settings. • **Enhanced** – upscale images further to allow more accurate results. This will provide better accuracy but can lead to longer processing time if the images do not contain text.
Source Group	If you want to add database to a source group, select existing, or create a new one.
Pause source on creation	Select to make other configuration changes before the initial data collection starts.

After the source configuration is completed, you will be prompted to lauch SQL crawling configuration wizard. See Database Configuration Wizard for more information.

5.1. Database Configuration Wizard

For the database sources, you can enable security-based crawling, that is, finding sensitive data (which logically will either be stored in text or binary-based columns). It is possible to create an intelligent content mapping, crawling certain fields as unstructured index text, and other fields — as mapped metadata.

This section explains how to use the **Database Configuration Wizard** for configuring the crawling process. You can run this wizard when adding the data source, or you can later open the **Source** tab, select your database source and click **Launch Wizard**.

IMPORTANT! If you want to crawl a target database in your MS SQL replication model, you must backup your database before running the configuration wizard.

See next:

- Introduction
- Tables
- Exceptions
- Summary

5.1.1. Introduction

On this step, provide matching rules to search in the database for data that match exactly or are similar to a specific pattern. You can indicate both: exact or partial matches over the database strings.

5.1.2. Tables

On this step, review the grid of the tables in the database that are not currently enabled for crawling (if already enabled then don't show in this grid) and have at least one text/binary column. Configure your crawling scope considering the following:

Column	Description
Table	Contains the list of all tables in the database, followed by alphabetically.
Text Columns	Contains the number of text/binary columns for each table. Click the number link to review the full list.
Metadata Columns	Contains the number of non-text/binary columns for each table. Click the number link to review the full list.
Primary Key	Contains the primary key for each table. Review the following Microsoft article for more information on SQL Server primary keys: Primary Keys Constraints.
Modified Filter	To improve performance the product performs automatic re-indexing against a field in each table that indicates the last modified date of the row. Where possible, the product will automatically map this based upon the exact match or inclusion of one of the below values within the field name. Additional values can be added below in order to support other naming conventions for modified fields (different language or internal convention).
Include?	Select if you want to disable crawling for this table. **NOTE:** You can disable crawling for all listed tables using the **Include none** option in the upper right corner of the wizard or enable crawling accordingly with the **Include all**.
View Sample	Shows a table of the top 15 rows allowing to view if the table is one to exclude.

5.1.3. Exceptions

On this step, review tables with missing primary keys and/or missing modified filters.

- **Missing primary keys** – only shows if users have tables that are missing primary keys where the user can select the primary key from a dropdown of all the columns. This step does not show if there are no missing primary keys.
- **Missing modified filters** – only shows if there are tables missing modified filters. Here tables are shown that are missing a modified and that have a datetime (or equivalent) typed column to select. If there are none this stage is skipped.

5.1.4. Summary

At this step, review your database configuration.

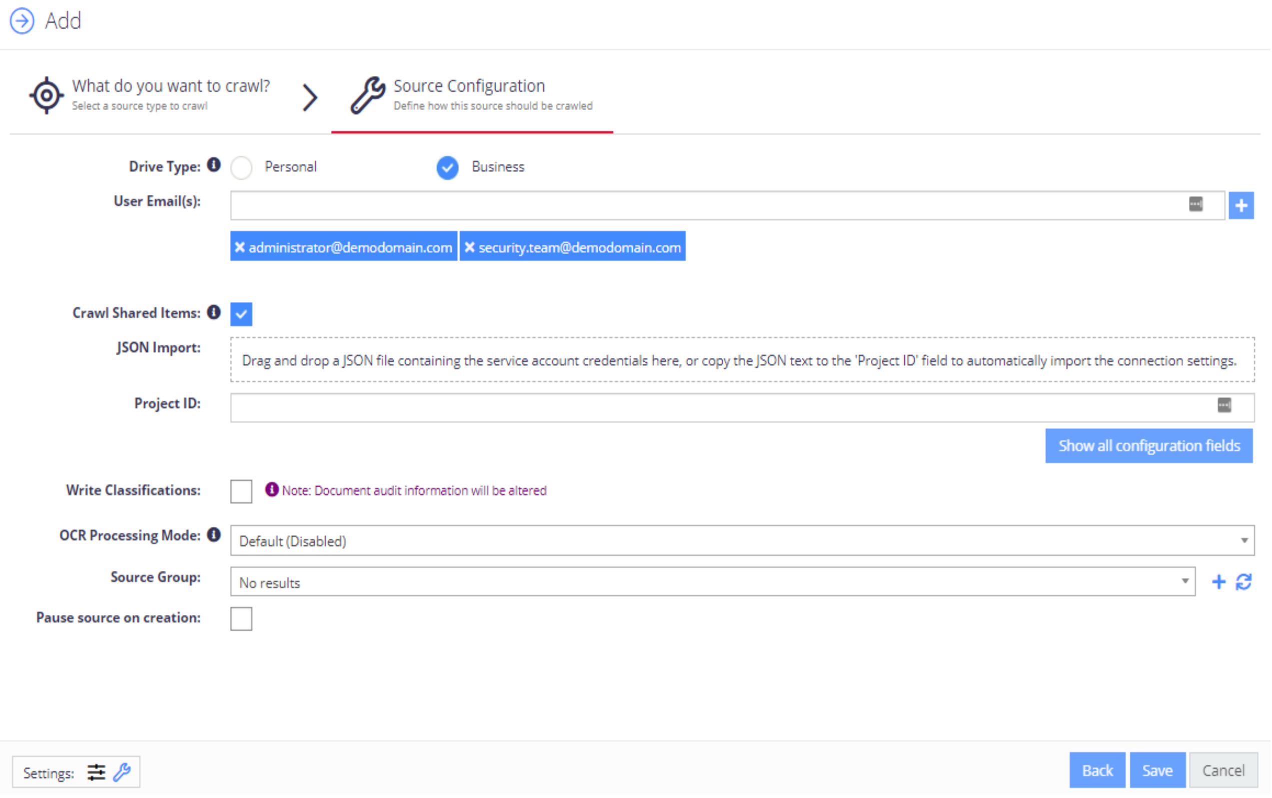
- **Overview** – review a high-level overview of the number of configured tables and excluded tables with their details.
- **Configured Tables** – double-check the configuration of tables to be crawled.
- **Excluded Tables** – review the full list of the tables to be excluded from classification scope with exclusion reason.

When the database configuration has been completed you will be redirected to the **Advanced Source Configuration**, this allows you to define how the database will be crawled. It is possible to crawl either specific tables, or crawl custom queries (defined select statements, which may use JOIN statements across multiple tables). See Database for more information.

6. Add Google Drive Source

The **Google Drive** source configuration screen allows you to enable the crawling and classification of content stored in both G-Suite repositories and Google Drive personal accounts.

IMPORTANT! Make sure you created App for GDrive crawling prior to start adding the source. See the Installation and Configuration Guide for more information.


Add
What do you want to crawl?
Select a source type to crawl
Source Configuration
Define how this source should be crawled
Show all configuration fields
Settings:
Back
Save
Cancel

Drive Type: ○ Personal ● Business

User Email(s): administrator@demodomain.com security.team@demodomain.com

Crawl Shared Items: ☑

JSON Import: Drag and drop a JSON file containing the service account credentials here, or copy the JSON text to the 'Project ID' field to automatically import the connection settings.

Project ID:

Write Classifications: □ Note: Document audit information will be altered

OCR Processing Mode: Default (Disabled)

Source Group: No results

Pause source on creation: □

Complete the following fields:

Option	Description
	Basic settings
Drive Type	Select *Business*.
User Email(s)	When adding a G-Suite source, enter the email address of the user's drive that you wish to crawl (via impersonation).
Crawl Shared Items	Select to crawl all files shared with the specified user in addition to any team drives shared with the user.

Option	Description
Crawl Shared Items	Select to enable crawling of any types of documents shared with the specified user.
JSON Import	Drag the JSON connection file you downloaded while creating Google service account in the form.
Project ID	Open the JSON connection file and copy file contents to **Project ID** field.
Write Classifications	Leave this checkbox empty.
OCR Processing Mode	(missing or bad snippet)
Source Group	Netwrix recommends creating a dedicated source group for Google Drive.
Pause source on creation	Select if you want to make other configuration changes before collection of the source occurs.

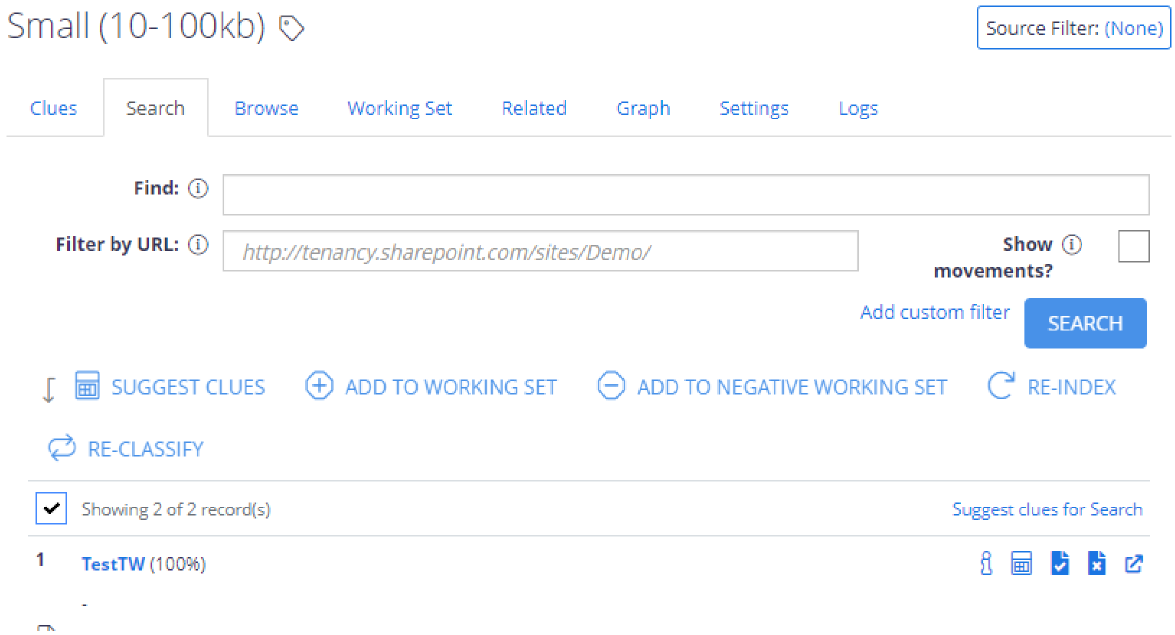
7. Review Reports and Browse Classified Documents

Once your documents are classified, you can identify sensitive information and reduce its exposure. Netwrix recommends starting with the **Document Tagging** report to see automatic and manual classifications of the documents within the reporting set. Further, you can browse your documents to see a list of documents achieving the minimum score set for classification in the term. Review the following for additional information:

- To browse classification results
- To review the Document Tagging report

To browse classification results

1. In administrative web console, navigate to **Taxonomies** → **Term Management**.
2. Select **Taxonomy** in the dropdown on the left and then expand specific term you are interested in.
3. Switch to **Browse** tab:



Small (10-100kb)

Source Filter: (None)

Find:

Filter by URL: http://tenancy.sharepoint.com/sites/Demo/

Show movements? □

☑ Showing 2 of 2 record(s)

1 TestTW (100%)

4. Click **Filter** to start browsing your documents.

To review the Document Tagging report

1. In administrative web console, navigate to **Reports** and expand the **Document Reports** set.
2. Select the **Document Tagging** report and click **Show filters** to narrow report scope.

Filter	Description
Taxonomy	By default, the report shows results for all taxonomies. Select the taxonomy you are interested in to restrict report scope.
Score Range	Select the score. Review Scoring for more information.
Classification	By default, the report shows results for all terms within a taxonomy. Limit your results by selected term.
Page URL	Filter your results by selected page URL.
Source	Select source group you created for Google Drive.

3. Click **Generate** and review results.
4. You can also export displayed page to .csv and .xlsx table or download the whole results.

TIP: Upon export, you will be prompted to include any associated document metadata to the report. It can be useful if you want to generate custom security reports. Specify metadata fields and click **Export** to download report.