



User Requirements Document

Project identification

Project Name: deDup for SharePoint

Date: 10.05.2024

Owner: Nonlinear SRL

Project Status: Implemented

Document History

Version	Date	Author(s)	Summary
1.0	12.01.2024	Nonlinear SRL	first version
1.2	10.05.2024	Nonlinear SRL	revised version

The contents of this document remain the property of and may not be reproduced in whole or in part without the express permission of Nonlinear.

Contents

Project identification.....	1
Contents.....	2
1. Overview.....	3
1.1 Purpose.....	3
2. Authentication.....	3
2.1 Microsoft Authentication.....	3
3. Deduplication Process.....	3
3.1 File Hashing.....	3
4. User Interface.....	3
4.1 Intuitive Design.....	3
4.2 Dashboard.....	3
5. Functionality.....	4
5.1 Duplicate Detection.....	4
5.2 Manual Deduplication.....	4
5.3 Reporting.....	4
6. Compliance.....	4
6.1 GDPR Compliance.....	4
6.2 Data Retention.....	4
7. Maintenance.....	4
7.1 Regular Updates.....	4
7.2 Update procedure.....	5
7.3 Support.....	5
8. Technical requirements.....	5
8.1 Compute.....	5
8.2 Database.....	5
Application deployment.....	6
User manual.....	13

1. Overview

1.1 Purpose

deDup for SharePoint is designed to identify and assist in the removal of duplicate files within SharePoint sites. It operates as a multi-tenant application, utilizing Microsoft authentication for secure access to SharePoint sites.

2. Authentication

2.1 Microsoft Authentication

- The app employs Microsoft authentication for user access.
- Required permissions include:
 - **openid**
 - **profile**
 - **offline_access**
 - **email**
 - **User.Read**
 - **Sites.Read.All**
 - **Files.ReadWrite.All** (Note: Specifically requested for the first-time the deletion process is triggered)

3. Deduplication Process

3.1 File Hashing

- The app utilizes file hashes directly from Microsoft Graph for comparison.
- No actual files are stored on the Sharepoint Deduplicator servers, ensuring data privacy and GDPR compliance.

4. User Interface

4.1 Intuitive Design

- The user interface is designed to be user-friendly, allowing easy navigation and understanding of the deduplication process.

4.2 Dashboard

- A central dashboard provides an overview of the deduplication process, including progress and any identified duplicates.

5. Functionality

5.1 Duplicate Detection

- The primary function is to detect duplicate files within SharePoint sites using file hashes obtained from Microsoft Graph.

5.2 Manual Deduplication

- The app facilitates a manual deduplication process, allowing users to review and selectively remove duplicate files, ensuring sensitive data remains under user control.
- Deletion Process:
 - Deleting files from SharePoint requires the additional permission of *Files.ReadWrite.All*.
 - This permission will be specifically requested the first time the app initiates the deletion process.
 - Users will be prompted to grant this permission during the initial deletion attempt.

5.3 Reporting

- Users can generate reports detailing the identified duplicates and actions taken during the manual deduplication process.

6. Compliance

6.1 GDPR Compliance

- The Sharepoint Deduplicator app adheres to GDPR guidelines by not storing any actual files and maintaining a focus on data privacy.

6.2 Data Retention

- The app does not retain any sensitive data beyond caching deduplication process, ensuring compliance with data retention policies.

7. Maintenance

7.1 Regular Updates

- The Sharepoint Deduplicator app will receive regular updates to enhance performance, security, and compatibility with the latest SharePoint features.

7.2 Update procedure

- At present, there is no automatic update feature in place. To obtain the latest version, the application must be redeployed. Alternatively, you may contact support to manually update the version using the Just In Time (JIT) access feature.

7.3 Support

- A dedicated support system is in place to promptly address user queries and concerns.

8. Technical requirements

8.1 Compute

- The application requires an app service to run, this gets deployed automatically on application creation. On deployment, the user has the opportunity to select which machine size the app will use, based on their requirements. Note that for larger sites, a machine type with a higher memory/disk storage is recommended.
- The VM sizes available are as follows:
 - B1 - Suitable for light testing purposes
 - B2 - Suitable for small amount of sharepoint sites, up to 3 mil files (recommended)
 - P1 - Suitable for large sharepoint sites (P1v3 on azure)
 - P2 - Suitable for maximum performance (P2v3 on azure)
- Pricing for different VM sizes can be found in the official azure documentation [here](#).

8.2 Database

- The application will store metadata information about files in a user-provided database. A mongodb connection string is required for this. The string needs to contain a database name.
- Example: `mongodb://<user>:<pass>@<db-url>:<port>/<db-name>?ssl=true....`
- For cost efficiency, it is recommended to use the serverless mongodb option provided by azure. This requires the user to deploy it separately.
- If the user chooses to deploy the database on azure, there is a guide provided within the application deployment section of this document. The official pricing documentation on azure can be found [here](#).

Application deployment

- In order to access the user's account, a single tenant app registration on azure is needed. A new one can be created using the settings below.

Register an application ...

*** Name**
The user-facing display name for this application (this can be changed later).

dedup-testing ✓

Supported account types

Who can use this application or access this API?

☒ Accounts in this organizational directory only (nonlinear.ro only - Single tenant)

☐ Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant)

☐ Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant) and personal Microsoft accounts (e.g. Skype, Xbox)

☐ Personal Microsoft accounts only

[Help me choose...](#)

Redirect URI (optional)
We'll return the authentication response to this URI after successfully authenticating the user. Providing this now is optional and it can be changed later, but a value is required for most authentication scenarios.

Select a platform ▼ e.g. https://example.com/auth

- Ensure that the Sites.Read.All application permission is granted, along with admin consent by navigating to the API permissions section.

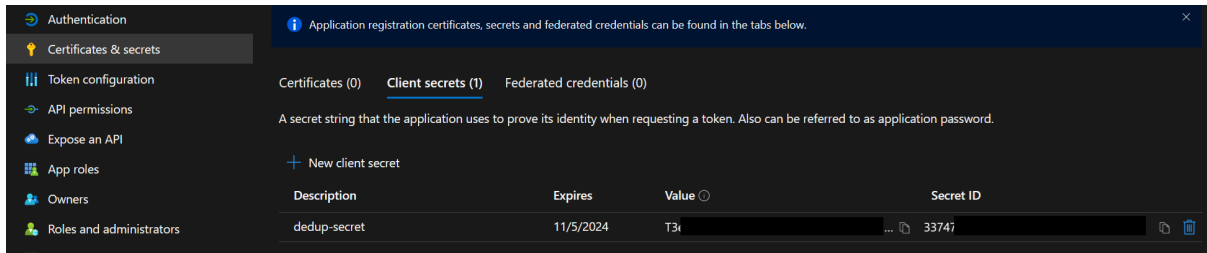
Configured permissions

Applications are authorized to call APIs when they are granted permissions by users/admins as part of the consent process. The list of configured permissions should include all the permissions the application needs. [Learn more about permissions and consent](#)

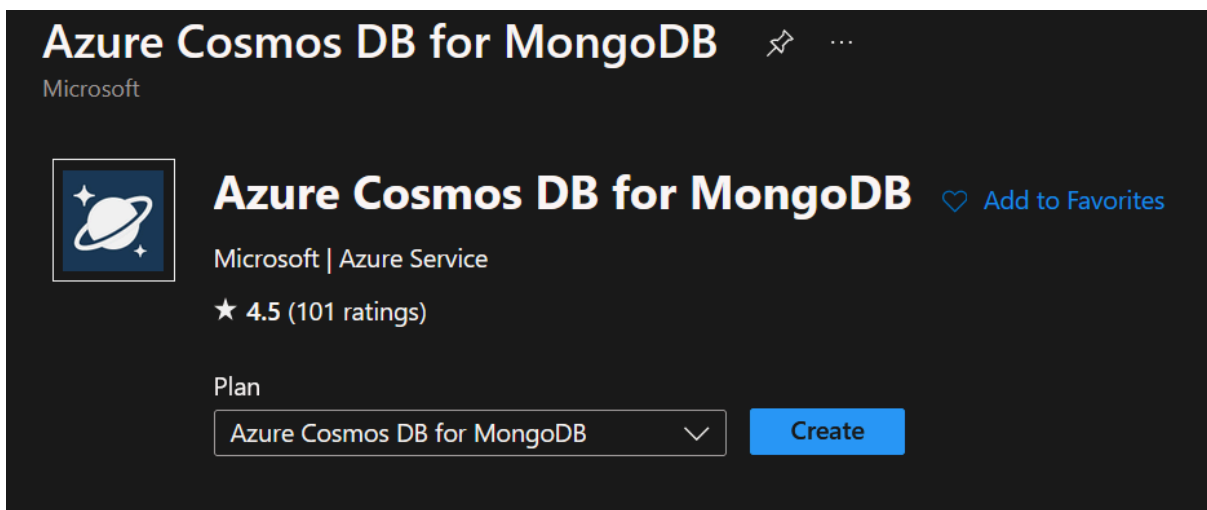
+ Add a permission ✓ Grant admin consent for MSFT

API / Permissions name	Type	Description	Admin consent requ...	Status
▼ Microsoft Graph (6)				
email	Delegated	View users' email address	No	...
offline_access	Delegated	Maintain access to data you have given it access to	No	✓ Granted for MSFT ...
openid	Delegated	Sign users in	No	...
profile	Delegated	View users' basic profile	No	...
Sites.Read.All	Delegated	Read items in all site collections	No	✓ Granted for MSFT ...
User.Read	Delegated	Sign in and read user profile	No	✓ Granted for MSFT ...

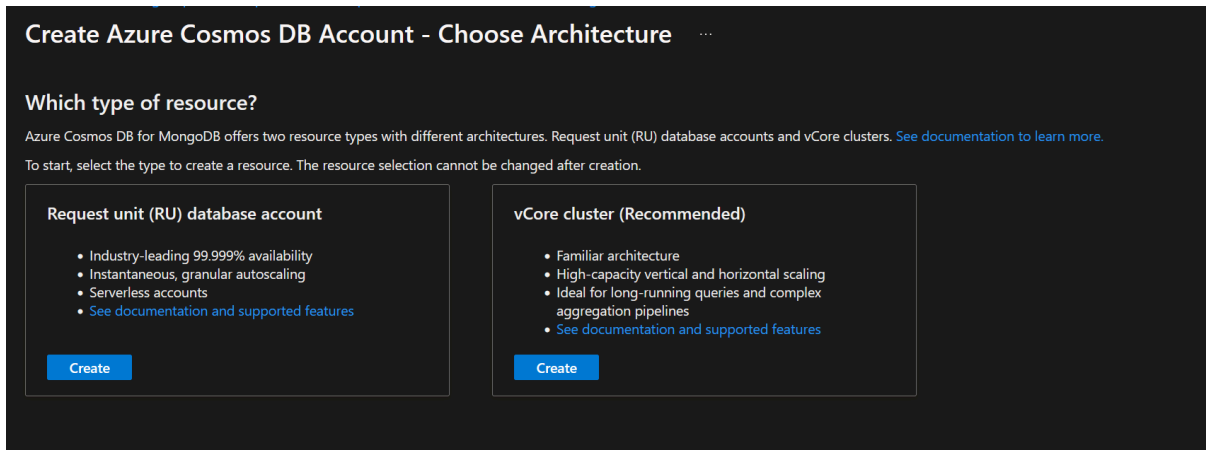
- The user then has to create a secret for the application by navigating to the Certificates and secrets section.



- Before the application can be deployed, it requires a database. This step can be skipped if the user already has a mongodb instance deployed and accessible.
- If not, the user can deploy a mongodb database on azure by creating an instance of the "Azure Cosmos DB for MongoDB" service.



- At the next screen, it is recommended to choose the Request unit option, due to the improved cost savings features.
- For serverless pricing details, the user can consult the Azure documentation [here](#).



Create Azure Cosmos DB Account - Choose Architecture

Which type of resource?

Azure Cosmos DB for MongoDB offers two resource types with different architectures. Request unit (RU) database accounts and vCore clusters. [See documentation to learn more.](#)

To start, select the type to create a resource. The resource selection cannot be changed after creation.

Request unit (RU) database account

- Industry-leading 99.999% availability
- Instantaneous, granular autoscaling
- Serverless accounts
- [See documentation and supported features](#)

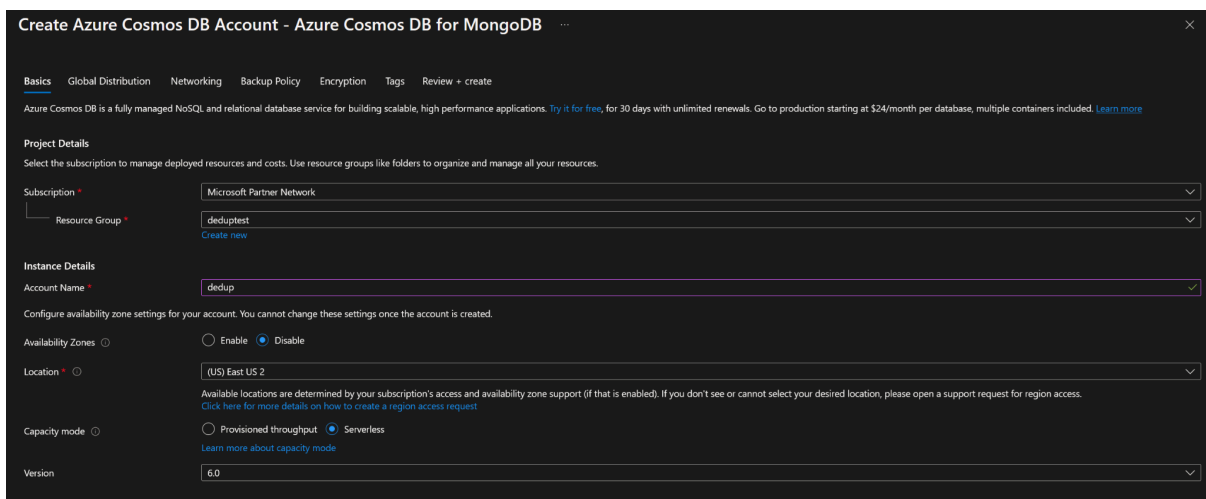
Create

vCore cluster (Recommended)

- Familiar architecture
- High-capacity vertical and horizontal scaling
- Ideal for long-running queries and complex aggregation pipelines
- [See documentation and supported features](#)

Create

- When creating the database account, the user can choose the serverless option.



Create Azure Cosmos DB Account - Azure Cosmos DB for MongoDB

Basics Global Distribution Networking Backup Policy Encryption Tags Review + create

Azure Cosmos DB is a fully managed NoSQL and relational database service for building scalable, high performance applications. [Try it for free](#), for 30 days with unlimited renewals. Go to production starting at \$24/month per database, multiple containers included. [Learn more](#)

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * Microsoft Partner Network

Resource Group * dedup test [Create new](#)

Instance Details

Account Name * dedup

Configure availability zone settings for your account. You cannot change these settings once the account is created.

Availability Zones ☐ Enable ☒ Disable

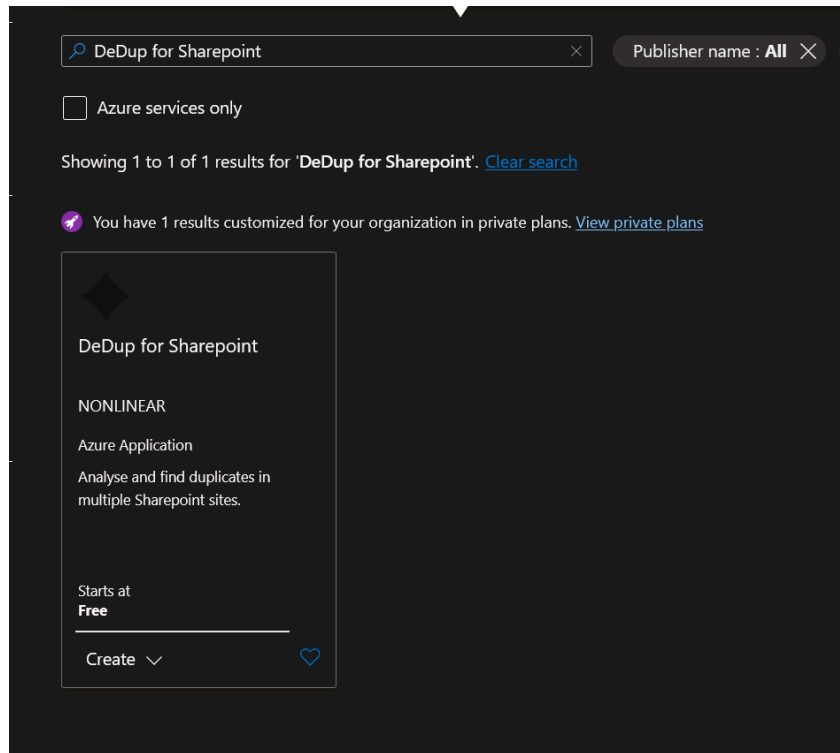
Location * (US) East US 2

Available locations are determined by your subscription's access and availability zone support (if that is enabled). If you don't see or cannot select your desired location, please open a support request for region access. [Click here for more details on how to create a region access request.](#)

Capacity mode ☐ Provisioned throughput ☒ Serverless [Learn more about capacity mode](#)

Version 6.0

- The user can deploy the application by navigating to the microsoft marketplace, and searching for "DeDup for sharepoint" and clicking on the create button.



- After choosing a plan, the next multi-stage menu allows the user to configure the details about the applications like subscription and resource group target, as well as its name.

Basics Application Settings SKU Configuration JIT Configuration Review + create

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ Microsoft Azure Sponsorship ▼

Resource group * ⓘ dedupctest ▼
[Create new](#)

Instance details

Region * ⓘ East US ▼

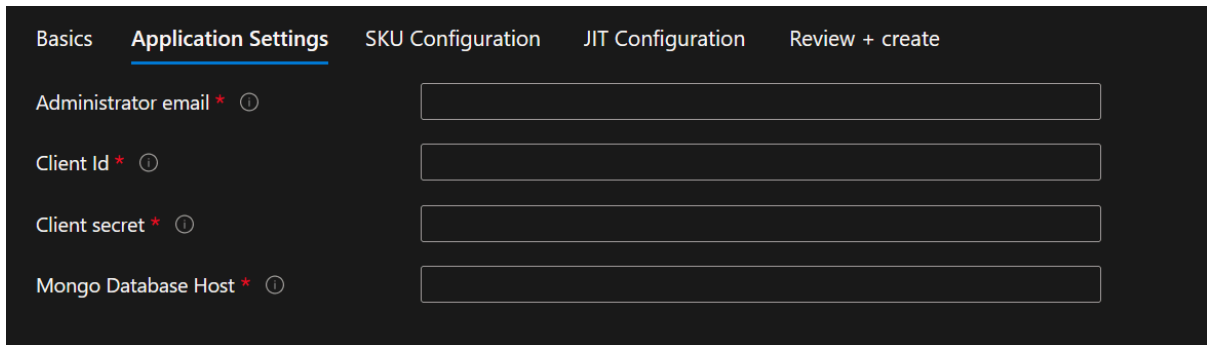
Managed Application Details

Provide a name for your managed application, and its managed resource group. Your application's managed resource group holds all the resources that are required by the managed application which the consumer has limited access to.

Application Name * dedupctesting ✓

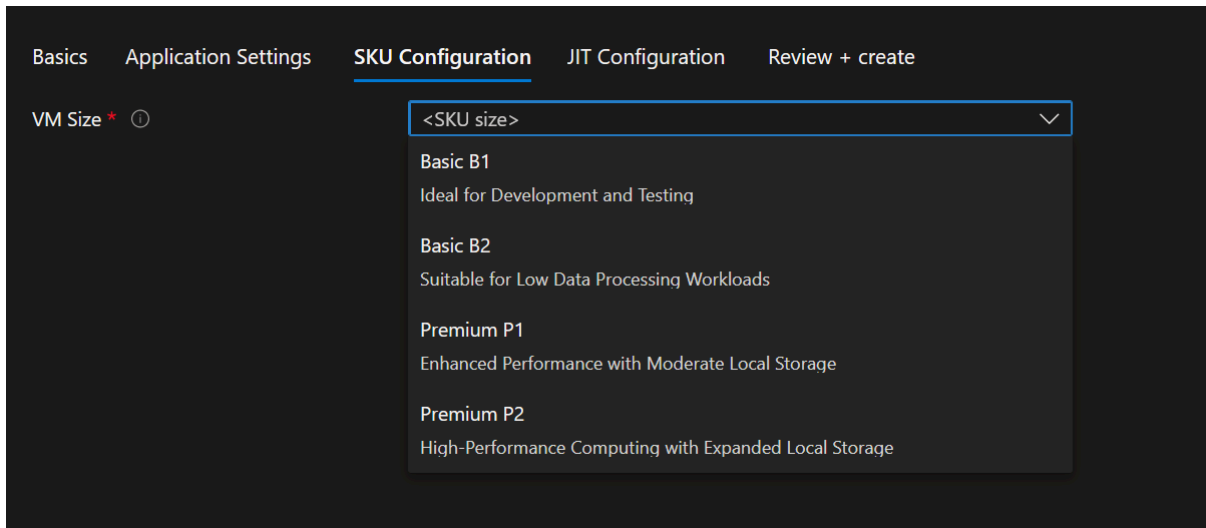
Managed Resource Group * ⓘ mrg-sharepoint_deduplicator-20240509104000 ✓

- Next, the user is asked to enter their email, client id and secret obtained from the app registration, and the MongoDB database connection string obtained after database creation.
- Note: it is important to include a database name in the connection string. eg: `mongodb://<user>:<pass>@<db-url>:<port>/dedup?ssl=true...`

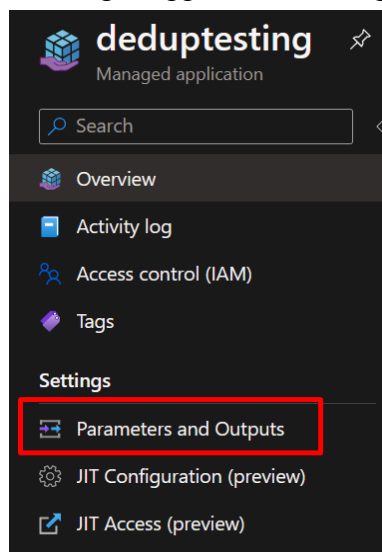


The screenshot shows a configuration interface with a dark background. At the top, there are five tabs: 'Basics', 'Application Settings' (which is selected and underlined in blue), 'SKU Configuration', 'JIT Configuration', and 'Review + create'. Below the tabs, there are four input fields, each with a label, a red asterisk indicating a required field, and a help icon (a circle with an 'i'). The labels and their corresponding input fields are: 'Administrator email' with a text box, 'Client Id' with a text box, 'Client secret' with a text box, and 'Mongo Database Host' with a text box.

- In the next step, the user needs to choose the virtual machine configuration used for deploying the application
- Size recommendation:
 - B1: Light testing purposes
 - B2: Suitable for small amount of sharepoint sites, up to 3 mil files (recommended)
 - P1: For large sharepoint sites (P1v3 on azure)
 - P2: For maximum performance (P2v3 on azure)
- For pricing details please check the official azure documentation [here](#).

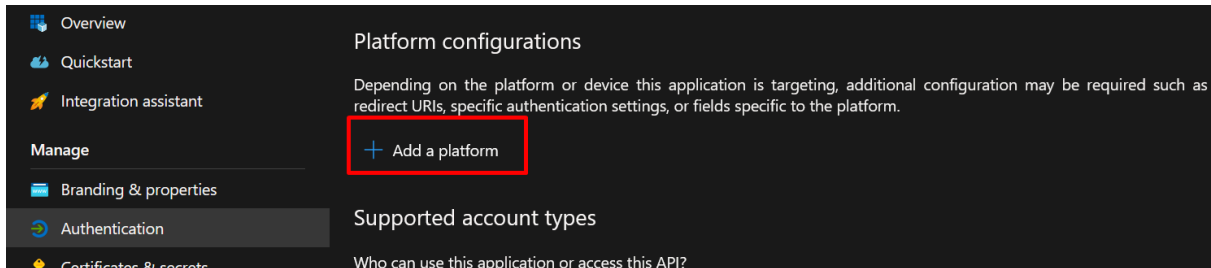


- After this, the user is asked to agree to terms and conditions and to click on the create button. The application is now being deployed.
- After the deployment is finished, the user must register the deployment url as a valid redirect url in the app registration. The deployment url can be found under parameters and outputs section in the managed application settings.



1 output		
Name	Type	output
appEndpoint	String	https://dedup testing.azurewebsites.net/

- Then, by navigating to the app registration settings, the Web platform can be added from within the Authentication section. The redirect url must follow the following format: {appEndpoint}/.auth/login/aad/callback



* Redirect URIs

The URIs we will accept as destinations when returning authentication responses (tokens) after successfully authenticating or signing out users. The redirect URI you send in the request to the login server should match one listed here. Also referred to as reply URLs. [Learn more about Redirect URIs and their restrictions](#)

✓

Front-channel logout URL

This is where we send a request to have the application clear the user's session data. This is required for single sign-out to work correctly.

Implicit grant and hybrid flows

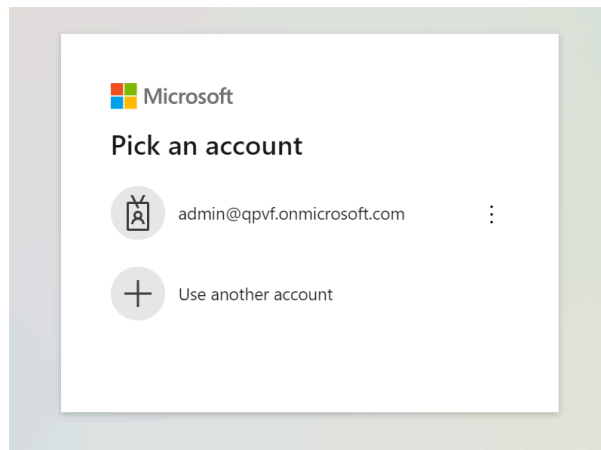
Request a token directly from the authorization endpoint. If the application has a single-page architecture (SPA) and doesn't use the authorization code flow, or if it invokes a web API via JavaScript, select both access tokens and ID tokens. For ASP.NET Core web apps and other web apps that use hybrid authentication, select only ID tokens. [Learn more about tokens.](#)

Select the tokens you would like to be issued by the authorization endpoint:

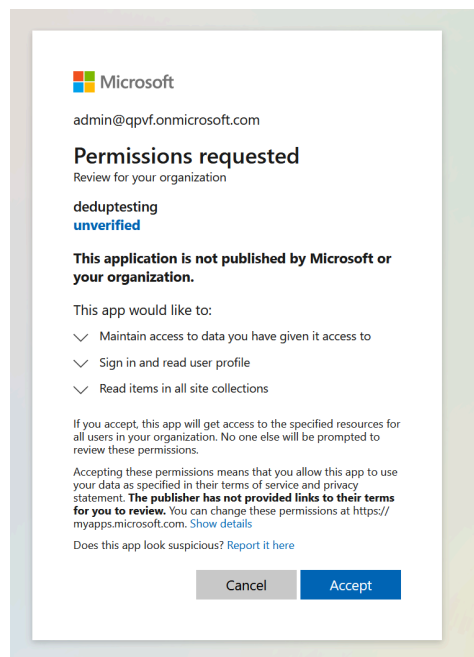
- ☐ Access tokens (used for implicit flows)
- ☒ ID tokens (used for implicit and hybrid flows)

User manual

- Initially the user will be prompted to login and redirected to microsoft where they will be asked to login using their account.



- On successful login, the user is presented with a consent screen, listing the permission grants the application needs.



- In order to begin a scanning session, the user must add the list of sites they want to target by navigating to the sites page accessible from the user menu..

Media type

Site

Refresh page

Rescan

Upload Report

Download Page Report

13.88 GB

out of

37.86 GB

is duplicated storage

Archive

Unique data

More details

Duplicate files	Nr. of duplicates	Disk size	Total size
marketing_resources_2024.zip	2	6.68 GB	26.71 GB
vecteezy_waterfall-in-a-wild-nature_2140013.mov	2	377.09 MB	754.18 MB
vecteezy_morning-sunrise-reflections-natural-lagoon-krabi-thailand_2375835.mp4	2	149.39 MB	298.78 MB
INNT82-8f0e32a7-dc13-4861-bfa9-5d3da8153195_br_backup.jpeg	3	653.96 KB	1.92 MB
NMNM24-3f4280aa-c6d9-47ad-b016-b8919257493a_br_backup.jpeg	2	956.65 KB	1.87 MB
cropped-large-1140x1140.png	2	659.57 KB	1.29 MB
NINV8-2bdbbd09-8367-4c87-889a-93df3d1f39b0_br_backup.jpeg	2	644.88 KB	1.26 MB
NINV8-2bdbbd09-8367-4c87-889a-93df3d1f39b0.jpeg	2	619.70 KB	1.21 MB
INNT81-29acc001-2e5a-468e-b319-clf436df49719_br_backup.jpeg	2	565.96 KB	1.11 MB
BECB63-51a2666f-c305-424f-bf53-9fdeea6c1ae-1152x1536.jpeg	2	401.16 KB	802.32 KB

Jump to page

1

2

3

4

5

...

8

10

Sites

Audit

Users

Change subscription

Switch account

About

- The sites page allows the user to add and delete sites. It also provides a summary of duplicates for each site.

Add sites

Delete sites

Retry

<input type="checkbox"/>	Site <input type="checkbox"/> Show invalid sites only	Last Scanned At	Total Files	Total Size	Amount Of Duplicated Files	Estimated Duplicated Files	
<input type="checkbox"/>	https://qpvf.sharepoint.com/sites/testsite/	07/05/2024 17:40:15	5 385	37.86 GB	75	13.88 GB	
<input type="checkbox"/>	https://qpvf.sharepoint.com/sites/test.site/	07/05/2024 17:40:15	2	7.94 KB	1	3.97 KB	

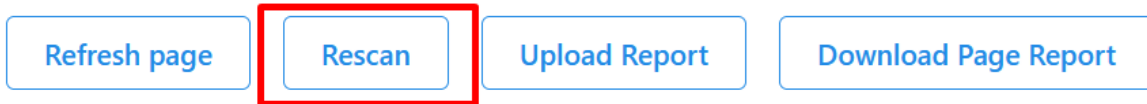
Jump to page

1

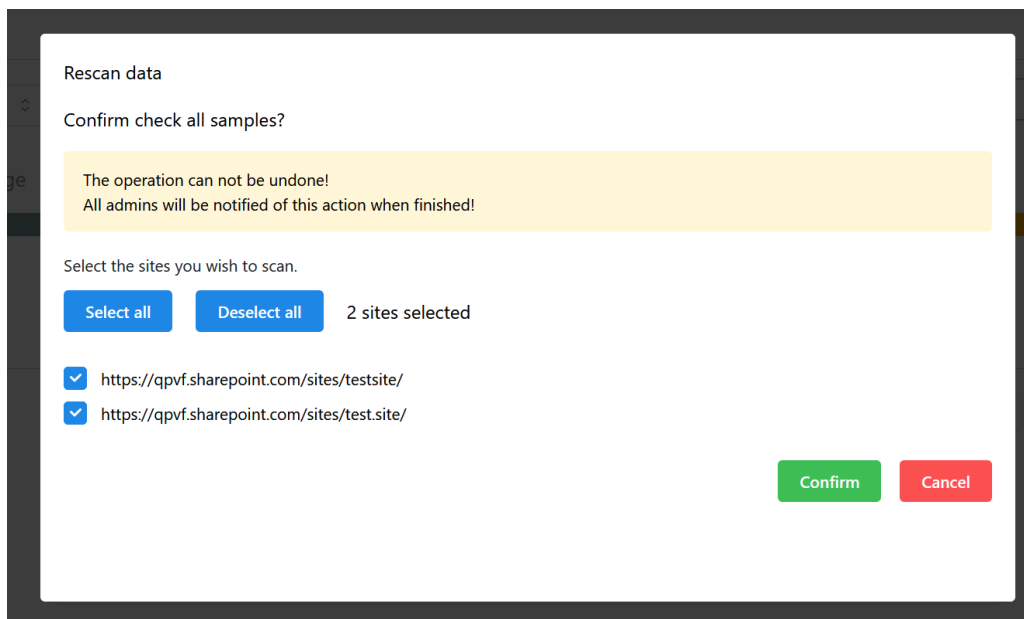
100

- When removing a site from the list, the user has the option to also delete any cached metadata associated with the site's files, if such data exists.
- Upon adding a site, it undergoes validation. Any invalid sites are flagged within the table, allowing the user to initiate a re-test by clicking the "Retry" button.

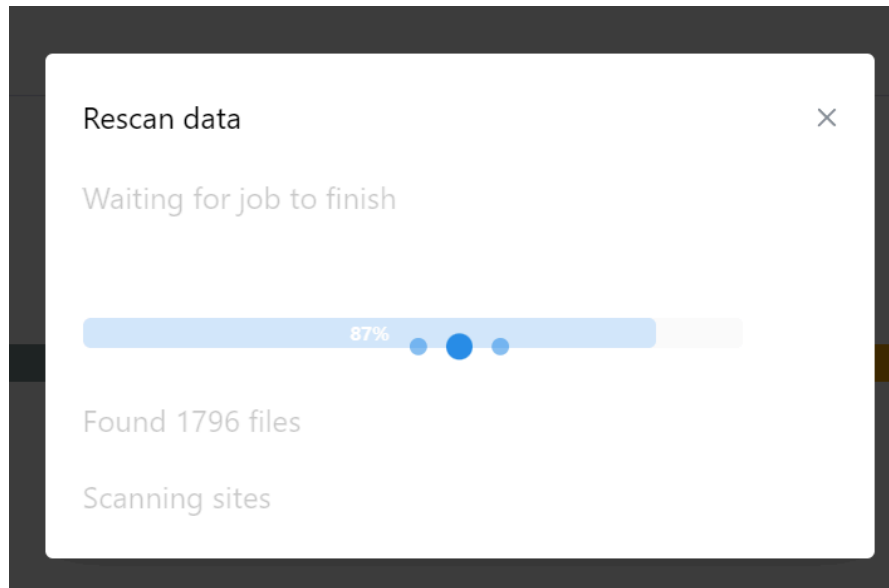
- To start an actual scanning session, the user must press the scan sites button located on the toolbar



- On pressing the button, a modal will be shown allowing the user to select the sites to be scanned. At the end of the scan, an email will be sent to all accounts marked as admin.



- On confirmation, the user will be presented with a modal screen informing them of the status of the scan, along with progress information.



- When the scan has finished, a list of found duplicate files is presented, along with a file type breakdown.. The shown table entries represent a duplicate file group.

deDup Duplicate file finder

Media type Site Refresh page Rescan Upload Report Download Page Report

13.88 GB out of 37.86 GB is duplicated storage

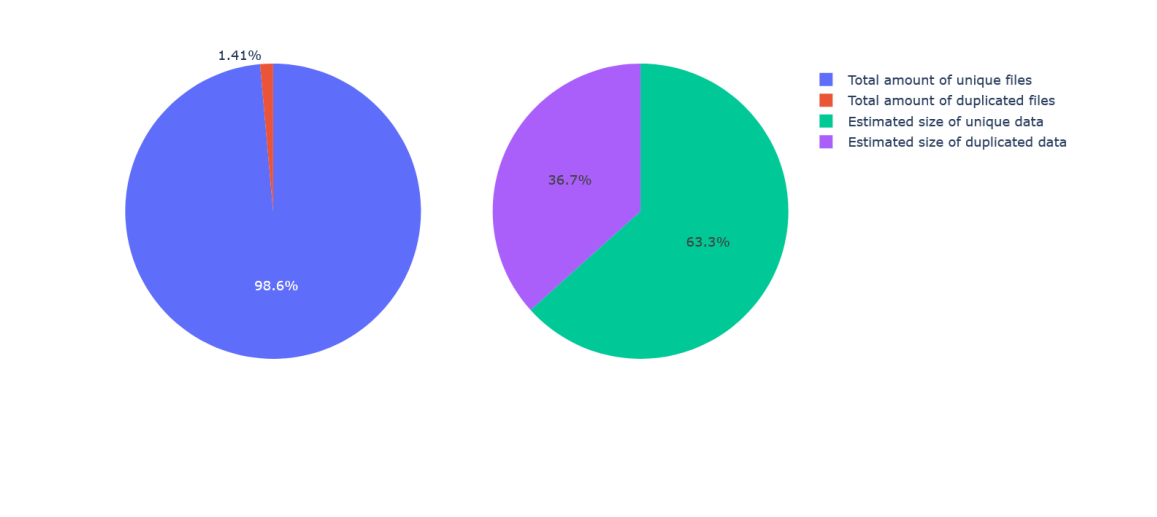
Archive Unique data More details

Duplicate files	Nr. of duplicates	Disk size	Total size
marketing_resources_2024.zip	2	6.68 GB	26.71 GB
vecteezy_waterfall-in-a-wild-nature_2140013.mov	2	377.09 MB	754.18 MB
vecteezy_morning-sunrise-reflections-natural-lagoon-krabi-thailand_2375835.mp4	2	149.39 MB	298.78 MB
INNT82-8f0e32a7-dc13-4861-bfa9-5d3da8153195_br_backup.jpeg	3	653.96 KB	1.92 MB
NMNM24-3f4280aa-c6d9-47ad-b016-b8919257493a_br_backup.jpeg	2	956.65 KB	1.87 MB
cropped-large-1140x1140.png	2	659.57 KB	1.29 MB
NINV8-2bdbbd09-8367-4c87-889a-93df3d1f39b0_br_backup.jpeg	2	644.88 KB	1.26 MB
NINV8-2bdbbd09-8367-4c87-889a-93df3d1f39b0.jpeg	2	619.70 KB	1.21 MB
INNT81-29acc001-2e5a-468e-b319-cf436df49719_br_backup.jpeg	2	565.96 KB	1.11 MB
BECB63-51a2666f-c305-424f-bf53-9fdeaaa6c1ae-1152x1536.jpeg	2	401.16 KB	802.32 KB

Jump to page 1 2 3 4 5 ... 8 10

- The user has access to a data summary statistics page by clicking on the More Details button. In this modal, a table detailing the number of files, and sizes is displayed. Further information about the amount of unique/duplicate files along with unique/duplicated sizes is broken down using two pie charts.

Total amount of scanned files	5387
Size of all scanned files	37.86 GB
Total amount of duplicated files	76
Estimated size of duplicated data	13.88 GB



- When clicking on a duplicate file group, the user will be redirected to a page listing the duplicate locations. From this page, the user is able to clean up duplicate files by deleting them individually, or in bulk. The option of generating a report is also available.

Size	49.92 KB	Estimated duplicate size	33.28 KB	Hash	33	<a>Delete selected <a>Download group report	
<input type="checkbox"/>	Name	Link	Site	Total size			
<input type="checkbox"/>	IMG_4701-300X300.JPEG	/2023/media/08/IMG_4701-300x300.jpeg	https://qpvf.sharepoint.com/sites/testsite/	16.64 KB		<input type="checkbox"/>	
<input type="checkbox"/>	INNT82-8f0e32a7-dc13-4861-bfa9-5d3da8153195-300X300.JPEG	/2023/media/08/INNT82-8f0e32a7-dc13-4861-bfa9-5d3da8153195-300x300.jpeg	https://qpvf.sharepoint.com/sites/testsite/	16.64 KB		<input type="checkbox"/>	
<input type="checkbox"/>	NCFD2-09D9064E-30B8-4147-84F8-52DDF85A15A-300X300.JPEG	/2023/media/08/NCFD2-09d9064e-30b8-4147-84f8-52ddf85a15a-300x300.jpeg	https://qpvf.sharepoint.com/sites/testsite/	16.64 KB		<input type="checkbox"/>	

Jump to page 1 100

- Within the main page, the user can download a report reflecting the current duplicate items listed in the table. The report allows the user to delete duplicates in bulk by providing a boolean column. By selecting the yes option, it allows the user to later upload the report and automatically remove the selected duplicate entries.

[Refresh page](#)[Rescan](#)[Upload Report](#)[Download Page Report](#)

	I	J
	Delete	
aRCQ7	y	
aRCQ7		
aRCQ7		
aRCQ7		
aRCQ7		

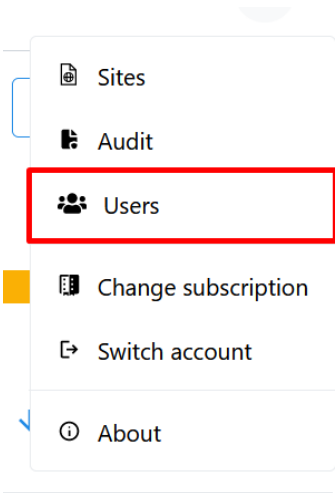
Upload deletion report

Drag and Drop or Select Files

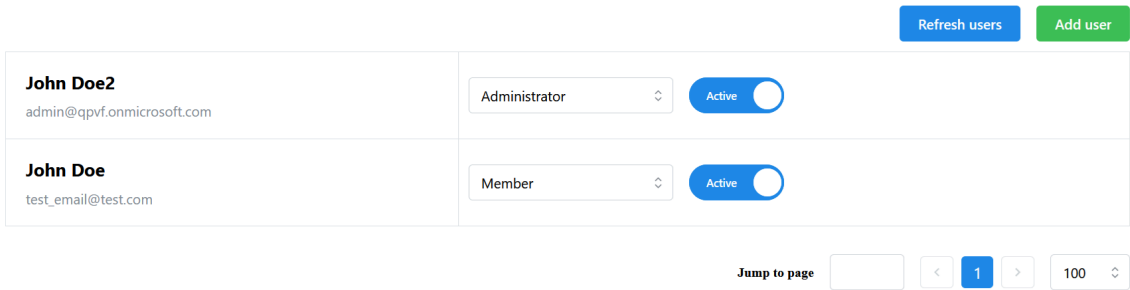
Uploaded file page_files_report.xlsx

Confirm

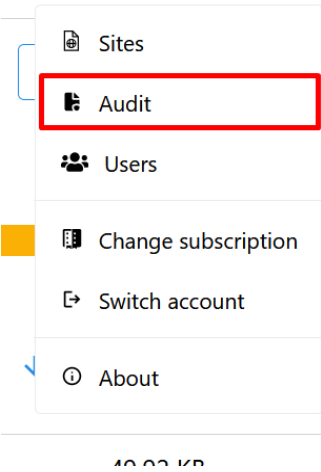
- The application allows for multiple users to have access at once, based on access control rules. More users can be added by navigating to the users page, accessed from the user menu.



- In this page, users can be added, deactivated, or their roles modified.



- In order to be able to monitor the actions taken by other users using the application, the user can inspect the audit page, located in the user menu.



- The audit page lists and details every action executed by all users. The information can be filtered, and exported to an excel report by clicking on the "Download Report" button.

User

Action type

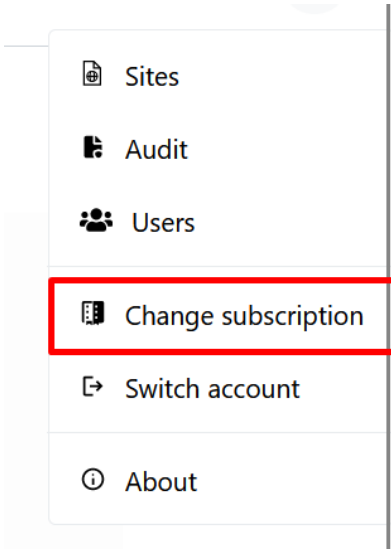
Filter dates

Refresh page

Download report

User	Time	Action	Properties
admin@qpvf.onmicrosoft.com	2024-05-08T12:52:16.444000	Delete	<div>sourcereport</div>
			<div>file_id014763CL2PYNPOW5R25ZA3JHAWMFVTS6AT</div>
			<div>file_urlhttps://qpvf.sharepoint.com/sites/testsite/Shared%20Documents/2023/media/08/NCFD1-bd66f51c-4696-4e67-9191-41ec9a4686b9-100x100_br_backup.jpg</div>
admin@qpvf.onmicrosoft.com	2024-05-07T14:40:15.336000	Rescan	
admin@qpvf.onmicrosoft.com	2024-04-29T09:37:09.796000	Rescan	
admin@qpvf.onmicrosoft.com	2024-04-29T09:36:30.978000	Delete	<div>sourceui</div>
			<div>file_id014763CL2TDHAA3OJKZFH2WARWPKQE3O2J</div>
			<div>file_urlhttps://qpvf.sharepoint.com/sites/testsite/Shared%20Documents/2023/media/11/IMG_9530-1_br_backup.jpeg</div>
admin@qpvf.onmicrosoft.com	2024-04-29T09:35:37.263000	Rescan	

- The subscription plan can be updated by navigating to the "Change subscription" option from the user menu.



- This page allows the user to update their personal information and to request a different plan. The selected plan’s benefits are outlined on the page.

Change your subscription plan to fit your needs!

Billing information

Customer name

Billing email

Billing address

Country

City Postal code

Street address

Company name

Registration number Tax ID

Free **Basic** Professional Enterprise

For small businesses that need regular cleanup.

This plan adds a quota of 500 GB of cleaned data.

- ✓ Ability to visualize duplicate groups
- ✓ Manually clean data
- ✓ Users management
- ✓ Auditing
- ✗ Download group reports
- ✗ Bulk cleaning

Basic Plan (1 Year) 299.00\$

Billing period will start once the plan is activated.

[Request plan](#)

- When the "Change subscription" button is clicked, a support email is issued, and the user will be redirected to stripe to enter payment information. After the payment is complete, the user's subscription is updated and will be redirected back to the application. If the subscription does not seem to update immediately, please allow a few minutes.

deDup **TEST MODE**

Pay Nonlinear

€499.00

deDup for SharePoint €499.00
Effortlessly analyze and eliminate duplicates across multiple SharePoint sites with our advanced...

Subtotal €499.00

[Add promotion code](#)

Total due €499.00



Powered by stripe | [Terms](#) | [Privacy](#)

PayPal

Or pay another way

Email test.email@example.com

Payment method

Card  

Card information

1234 1234 1234 1234

MM / YY

CVC

Cardholder name

Full name on card

Country or region

Romania

Securely save my information for 1-click checkout

Enter your phone number to create a Link account and pay faster on Nonlinear and everywhere Link is accepted.

 0712 034 567

[Optional](#)

[link](#) · [More info](#)

Pay

- In case a referral coupon exists, it can be added to this page also.

deDup TEST MODE

Pay Nonlinear

€499.00

deDup for SharePoint €499.00
Effortlessly analyze and eliminate duplicates across multiple SharePoint sites with our advanced...

Subtotal €499.00

[Add promotion code](#)

Total due €499.00

Powered by stripe | Terms | Privacy

PayPal

Or pay another way

Email eusebiu.fatu@gmail.com

Payment method

Card giropay EPS

Card information

1234 1234 1234 1234 VISA MasterCard American Express

MM / YY CVC

Cardholder name

Full name on card

Country or region

Romania

Securely save my information for 1-click checkout

Enter your phone number to create a Link account and pay faster on Nonlinear and everywhere Link is accepted.

0712 034 567 Optional

link More info

Pay

- In case the user has any questions, suggestions or issues, they can send feedback using the button next to the user menu.

bug icon

user icon

[Download Page Report](#)

- When clicked, a modal will open which will allow the user to type in their feedback message, along with an option to send the application logs to allow for enhanced support.