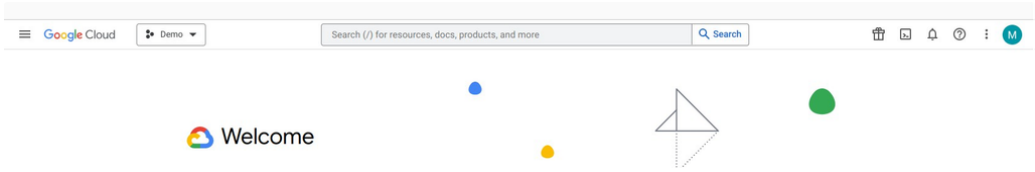
Google Drive Crawler Setup Guide

Introduction

This document will walk you through the steps required to setup and gather the necessary details required to set up a Google drive crawler on SimSage. To follow this guide, you will need access to an administrator account on your Google system.

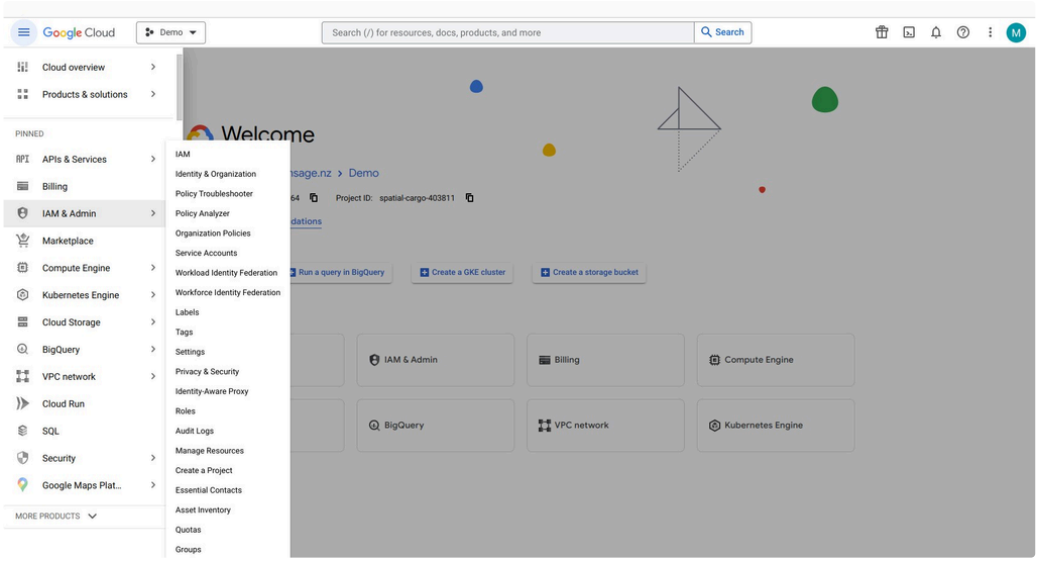
Create Service Account

1. Go to the Google portal and sign into the required user account: https://console.cloud.google.com/
2. Navigate to your desired project using the dropdown project list.

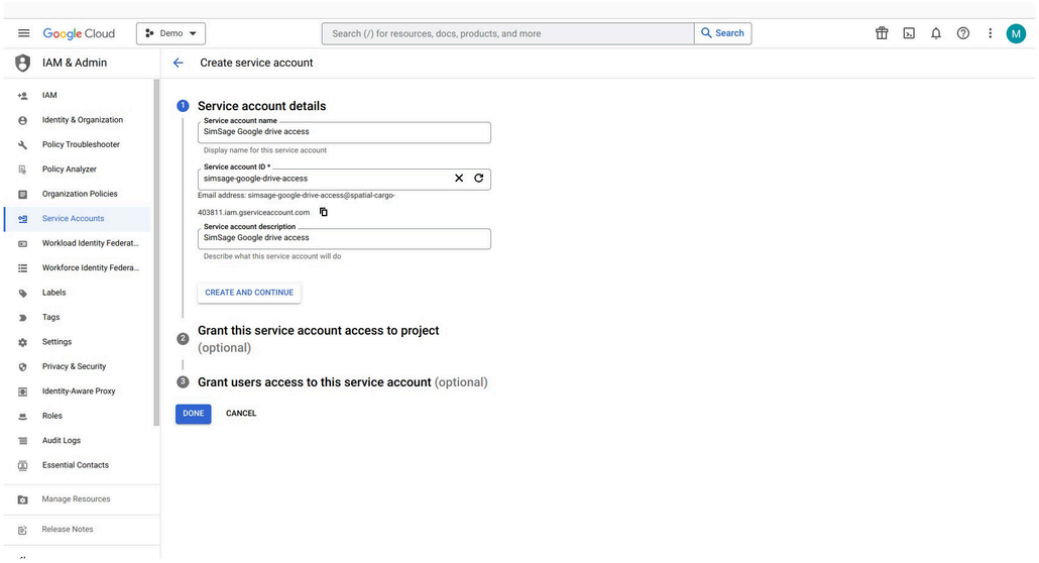


3. Or, create a new project using the new project button after clicking the dropdown project list.

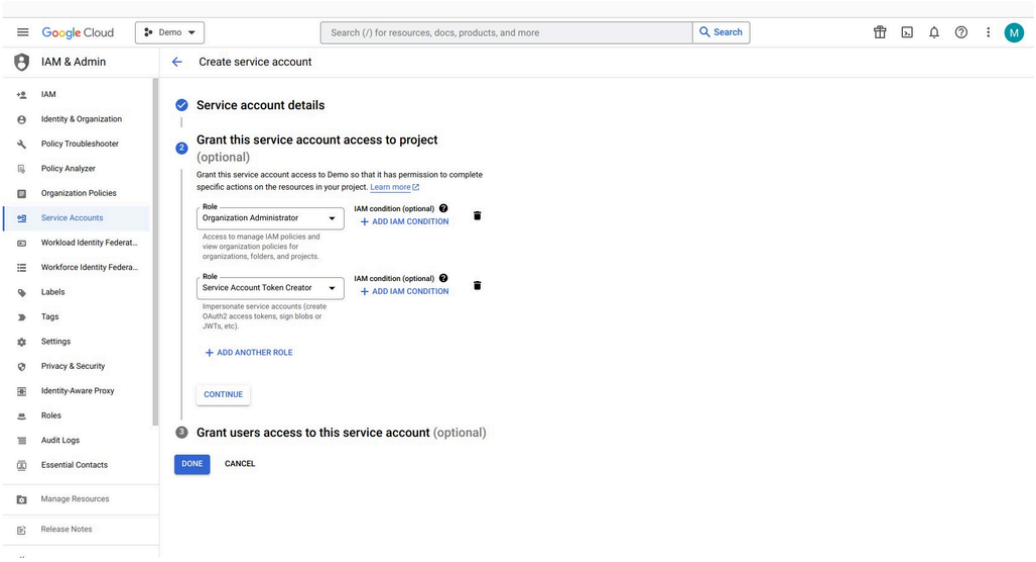
4. Populate the “New Pordject” form with the necessary values.
   In our example we used the following values. You do not need to use these specific values.
   Project name: Demo
   Organization: SimSage | SimSage
   location: SimSage | SimSage
5. Next, using the left-hand side navigation menu. Go to “IAM & Admin” > “Service Accounts”.



6. On the Service accounts screen click “+ CREATE SERVICE ACCOUNT”.
7. The following values are our suggestions. You do not need to use these specific values.
   Service account name: SimSage Google drive access
   Service account ID: simsage-google-drive-access
   Service account description: SimSage Google drive access
8. Click “CREATE AND CONTINUE” to Grant this service account access to the project.



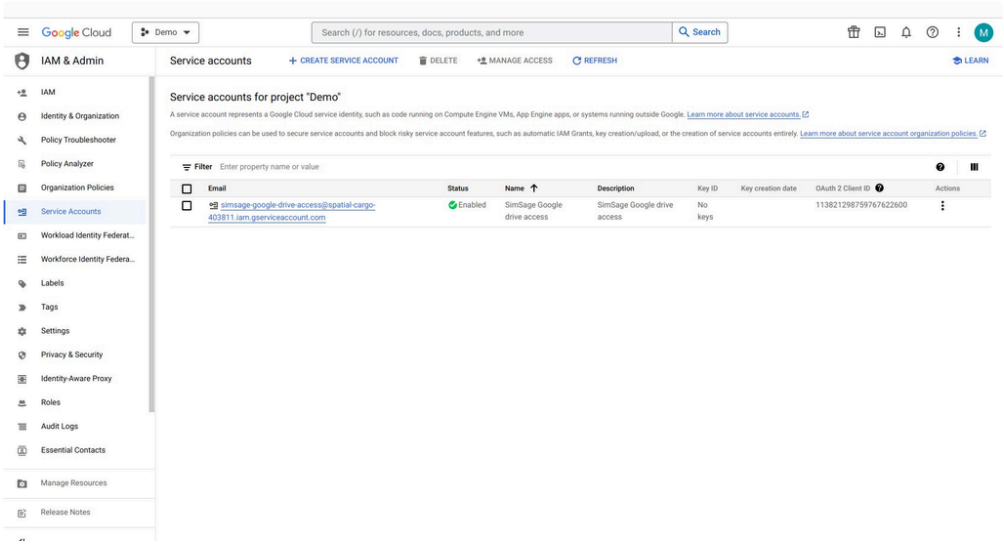
9. Next, use the “Select a role” dropdown to assign the “Organisation Administrator” Role to the service account. NB: This gives SimSage access to all resources in your Organisation.
10. Click “+ ADD ANOTHER ROLE” to add the other required role.
11. Again, use the “Select a role” dropdown to assign the “Service Account Token Creator” Role to the service account. NB: This allows SimSage to impersonate your users and access accounts as required.
12. Click “Done”



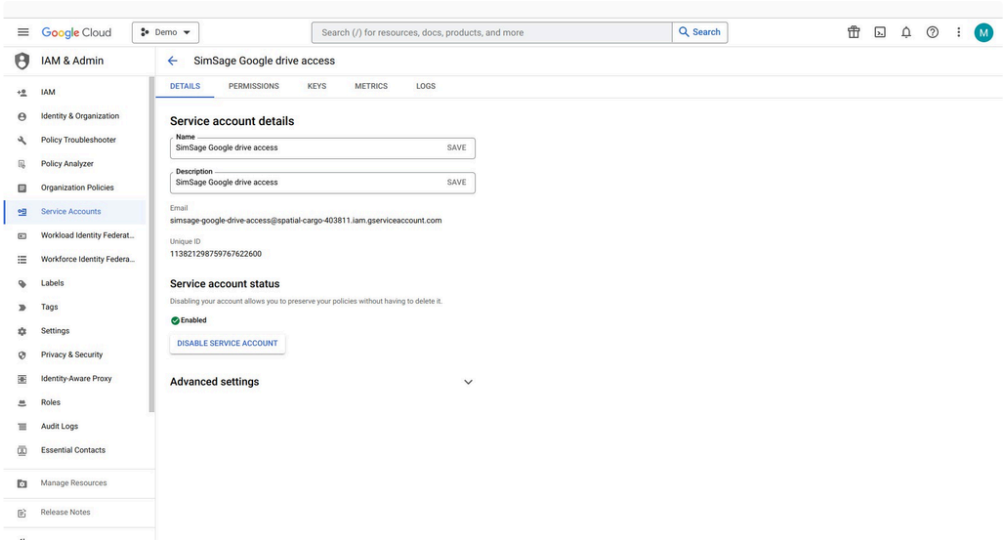
Create JSON Key

Now the service account has been created. In our case this is called “simsage-google-drive-access@spatial-cargo-403811.iam.gserviceaccount.com”. We now need to create a key for the service account to allow SimSage to use their account. NB: The service account name will differ for your organization.

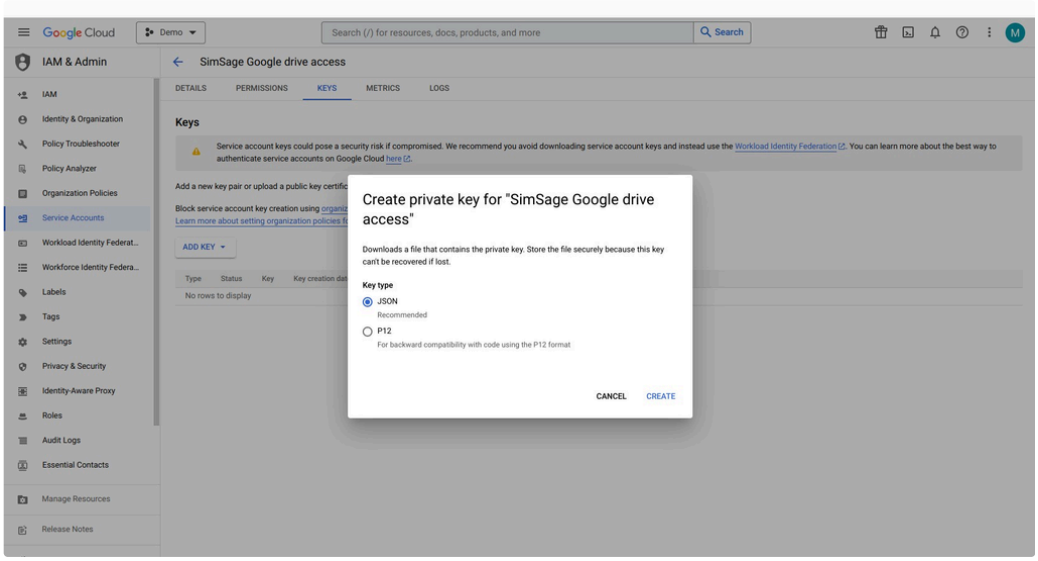
1. Navigate to your new service accounts account details by clicking on its name.



2. Make note of the unique ID for this account.



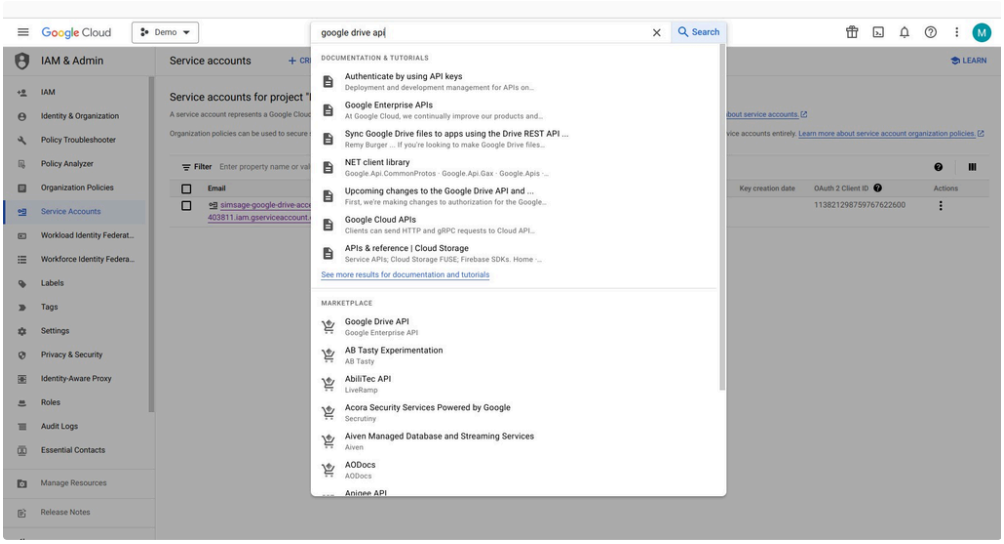
3. Navigate to the “KEYS” tab.
4. Click “ADD KEY” > “Create new key”.
5. Keep the default key type as JSON and click “CREATE”.



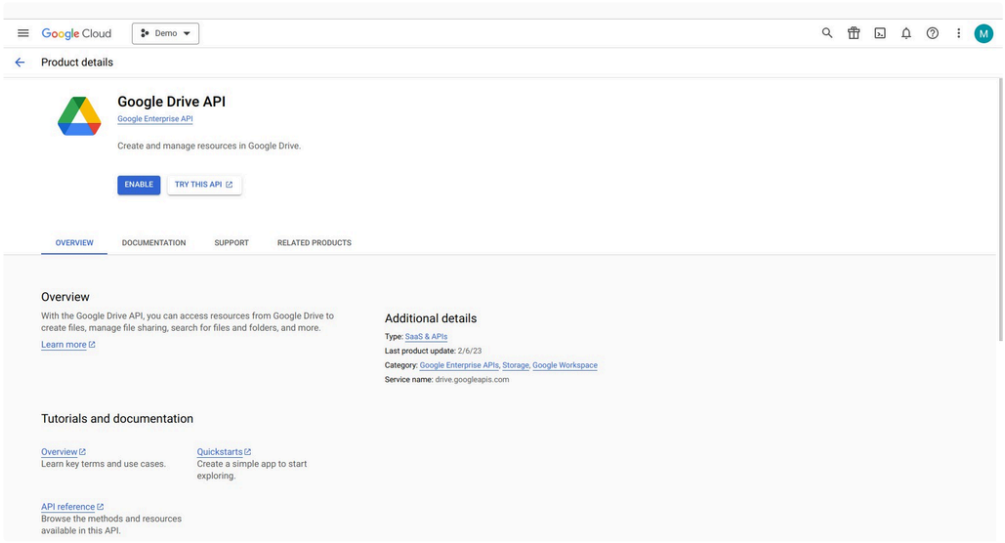
6. The key will be downloaded locally to your machine. SimSage will need the contents of this file to operate the crawler, so we advise you store it securely until you need it.

Enable Google Drive API

1. Type “Google Drive API” into the Search bar at the top of the screen.
2. In the drop down select “Google Drive API” from the marketplace options

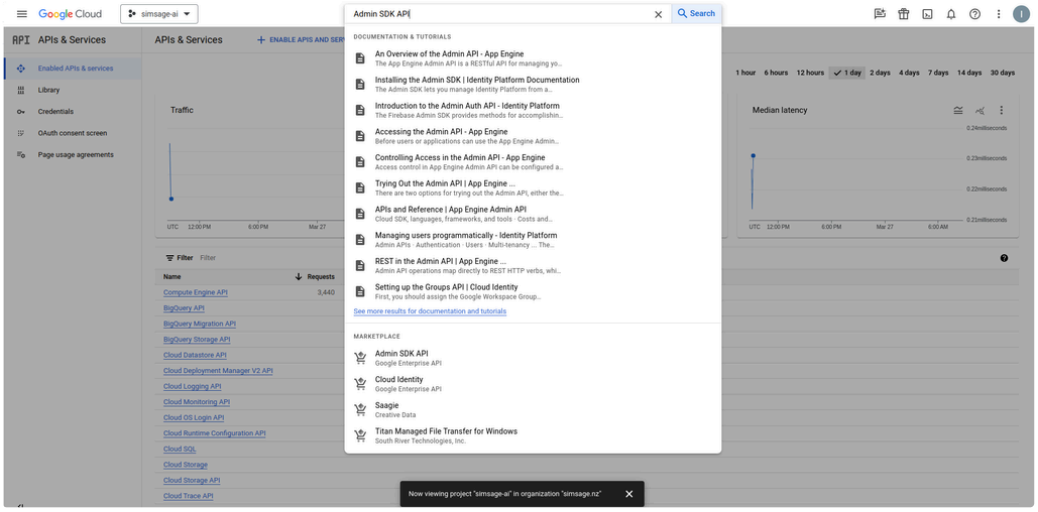


3. Click “Enable”

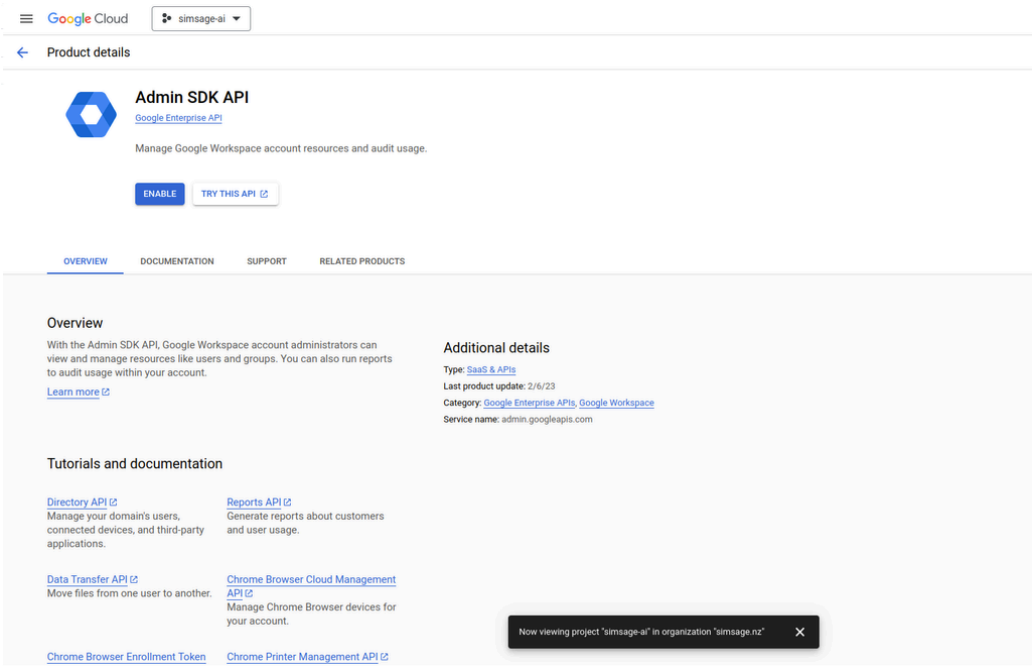


Enable Admin SDK API

1. Type “Admin SDK API” into the Search bar at the top of the screen.
2. In the drop down select “Admin SDK API” from the marketplace options

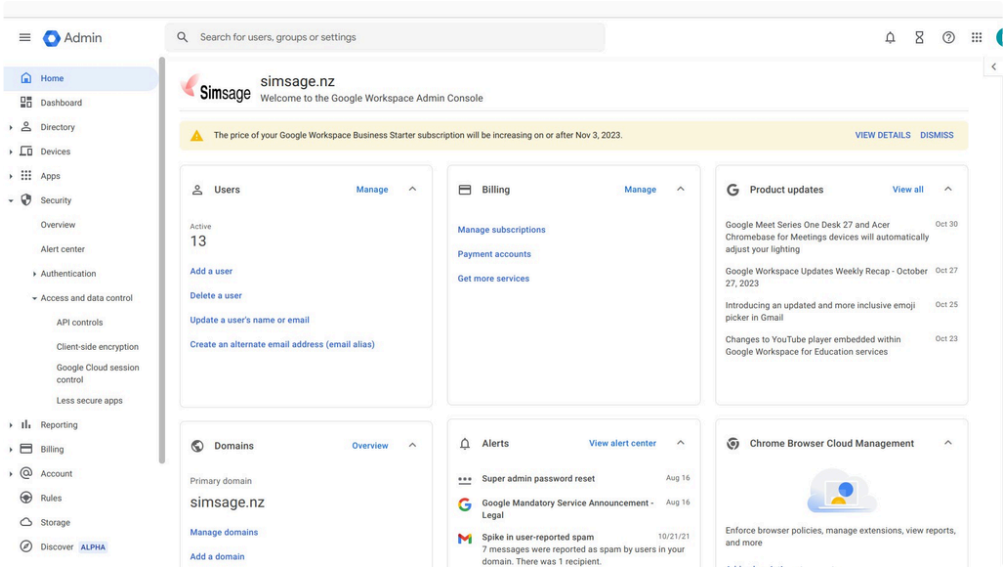


3. Click enable

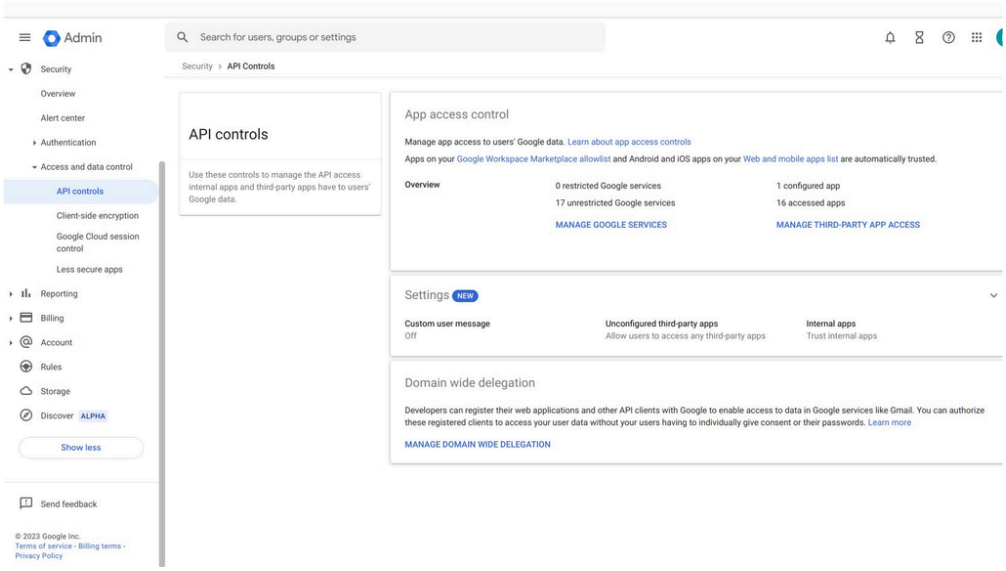


Authorize Domain Wide Delegation

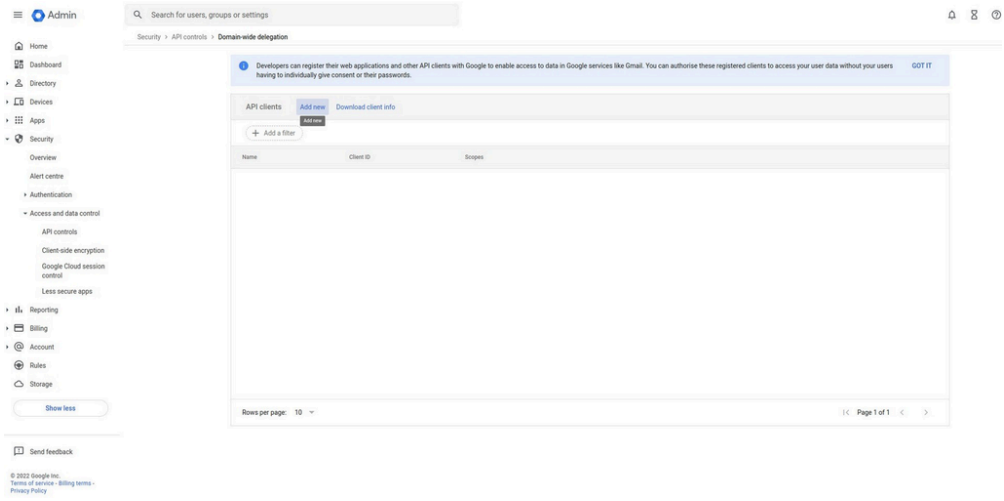
1. Open https://admin.google.com
2. Using the lefthand navigation bar, go to “Security” > “Access and data control” > “API controls”. NB: You may need to click “show more” to reveal the “Security” option.



3. Click “MANAGE DOMAIN-WIDE DELEGATION” to switch to the API clients screen.



4. Click “Add new” to give the service account access to specific items of your Google drive.

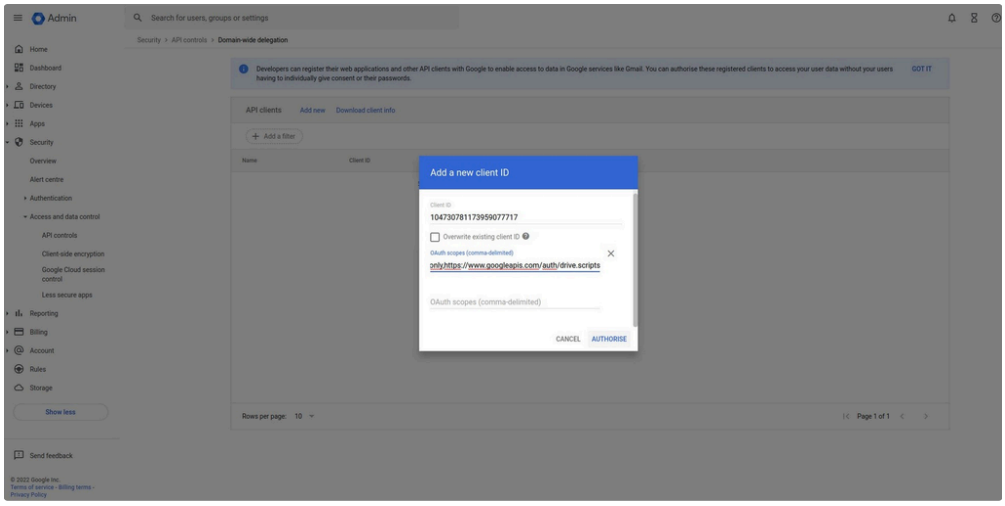


5. Populate the “Client ID” field with your service accounts unique ID.
6. Populate the “Oauth scopes” field with the following strings:

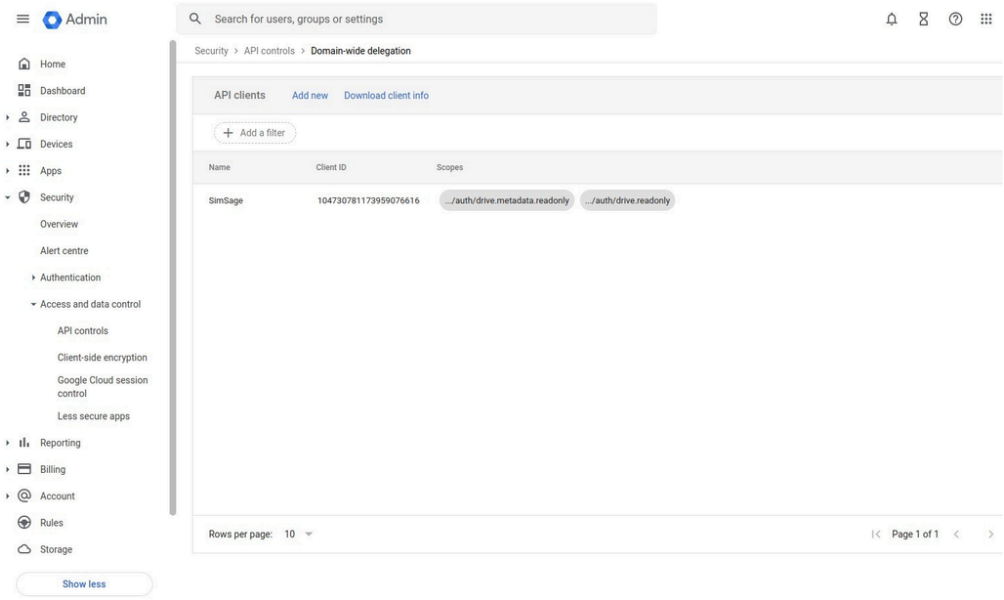
- `https://www.googleapis.com/auth/drive.metadata.readonly`

- `https://www.googleapis.com/auth/drive.readonly`
- `https://www.googleapis.com/auth/admin.directory.group.member.readonly`
- `https://www.googleapis.com/auth/admin.directory.group.readonly`
- `https://www.googleapis.com/auth/admin.directory.domain`
- `https://www.googleapis.com/auth/admin.directory.user.readonly`

7. Click “AUTHORIZE”.



8. Google will now notify you that this client has been added with 6 scopes.

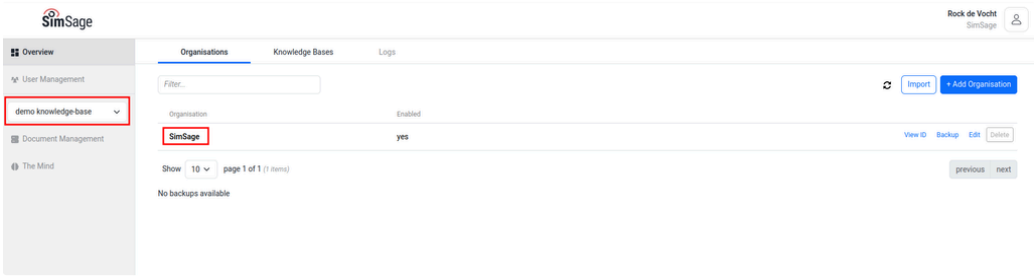


We can now setup a new SimSage crawler to access any drive accounts you want to index.

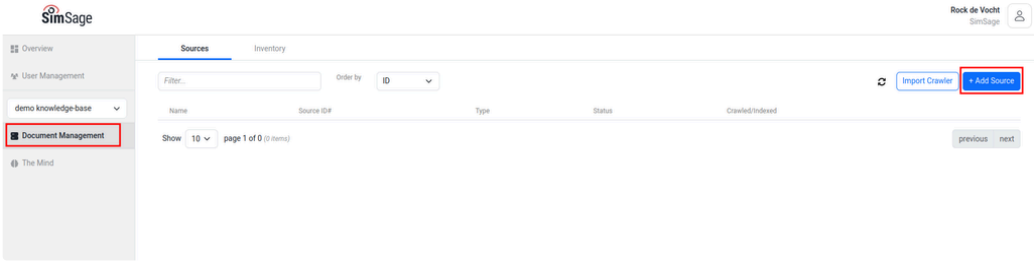
SimSage Source configuration

General Tab

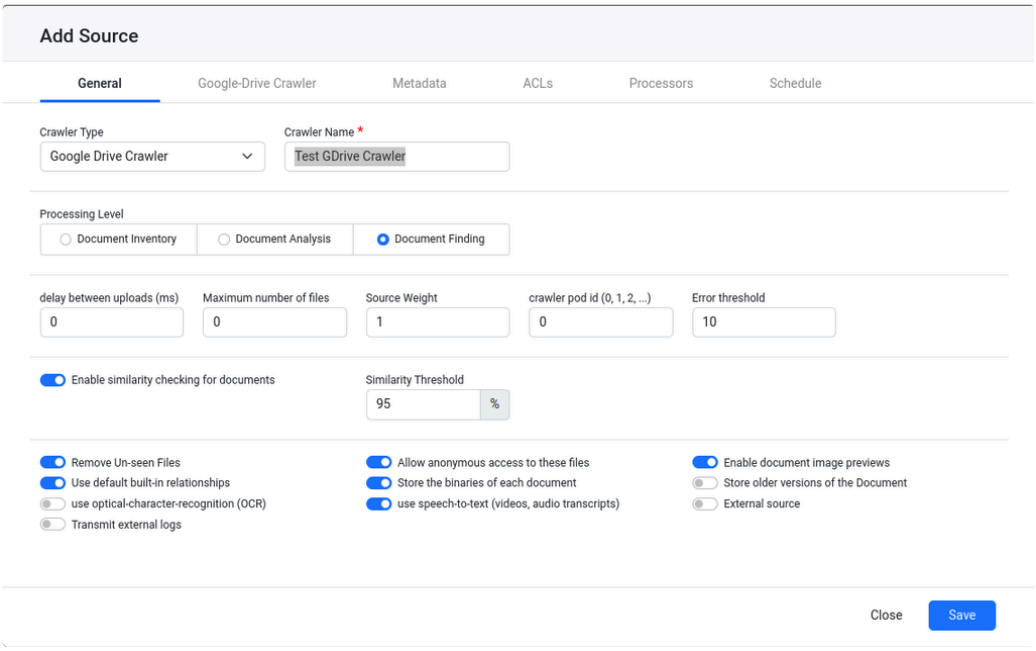
1. Select your organisation (SimSage shown in the image below)
2. Select your knowledge-base from the drop-down box (demo knowledge-base shown below)



3. Click on Document Management link in the left hand side of the menu
4. Click on the + Add Source button



5. In the General tab
   a. set the Crawler Type to “Google Drive Crawler”
   b. give the Crawler a Name (Test GDrive Crawler in the image below)
   c. Select the “Google-Drive Crawler” tab to set up your google drive crawler

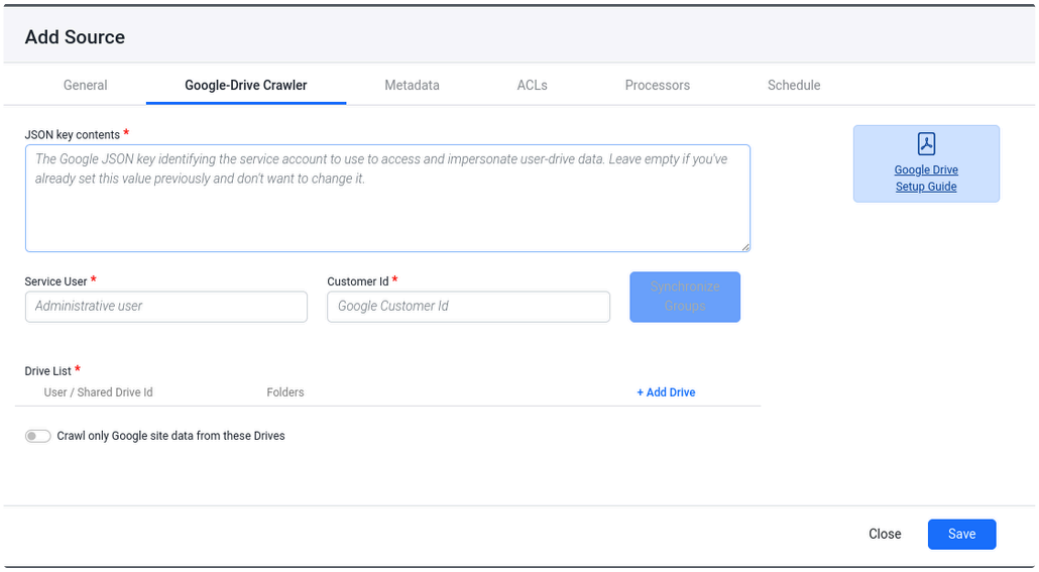


Google-Drive Crawler Tab

Access Details

- Steps
  - Add the Json Key generated earlier into the “JSON key contents” text area

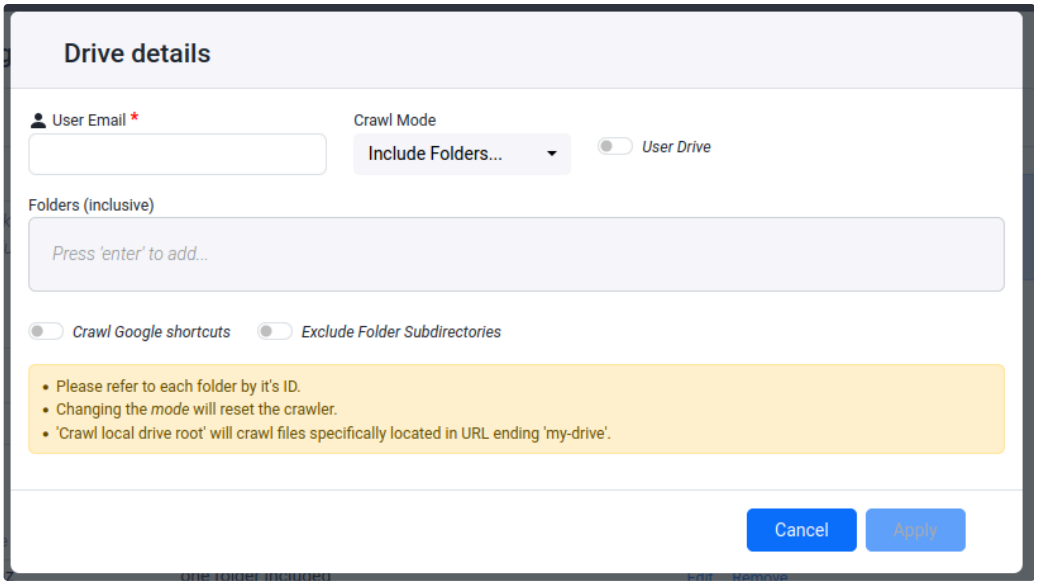
- Add the email of the ‘Service User’ into the ‘Service User’ text box
  N.B. The Service user represents the Google Account used to crawl shared drives as well as to retrieve Domain Group and User information from Google to manage Access restrictions to the crawled files. Please assure the entered user has access to the shared drives required as well as the necessary roles to get User and domain information!
- Add your Organisation’s Google Customer Id in the ‘Customer Id’ text box
  If you do not know your Customer Id:
  - Sign in to your Google Admin console. Sign in using your administrator account (does not end in @gmail.com).
  - In the Admin console, go to Menu Account Account settings. Profile.
  - Next to Customer ID, find your organization's unique ID.
- If this crawler is used to only retrieve Google site data from the Drive select the “Crawl only Google site data from these Drives” toggle



Google-Drive Crawler Tab

Drive Details and Settings

- Add drives to crawl
  - Press the ‘+ Add Drive’ link above the Drive List table
  - Select the kind of Drive you wish to crawl, e.g. Shared Drive or a User’s personal drive
  - Depending on your selection enter either the User’s email or the drive Id for the shared drive
  - If you want to further restrict the files crawled depending on the folders they reside in, add a comma separated list of folders to include into the ‘Folder List’ text area



In the above scenario note there are multiple options to the drive:

- Crawl Google Shortcuts: Crawl shortcuts in specified directories
- Exclude Folder Sub-directories: Only crawl files in folder specified not in child folders
    - Note this option is not available for *exclude* mode

Things to Consider

Be sure to press 'enter' upon typing a folder ID to be sure it is saved. A blue bubble will surround the text as a indicator: