Migration guide

Before starting to migrate your Aspire deployments to Aspire 5.0, it is strongly advised to understand the architectural change: Aspire 5.0 Architecture

Migrating to Aspire 5.0 is a process that not only changes how the configuration for the crawls are done, but also changes to the hardware architecture must be considered.

The current guide describes the typical journey a migration from Aspire 3/4 would look like.

Step 1. Resource allocation considerations

Aspire 3 and 4 had a horizontal distributed architecture, where all the Aspire nodes executed the exact same software and configuration. All nodes were equal, which meant more complex synchronization, and hard to balance throughput and resource utilization.

Aspire 5.0 consists of two distinct types of nodes: **Manager** and **Workers.** More Manager nodes means more simultaneous crawls. The more worker nodes higher the throughput, but you can have an heterogeneous set of worker nodes, where some would run certain crawls, and the others would run other types of crawls.

For production deployments, where high availability is required, it is recommended to have at least 2 **manager** nodes, as if one fails, the other one can assume the work from the failed one, while the failed one recovers and re-claims work.

Resource requirements:

Node	Minimum nodes	Recommended nodes	Minimum	Recommended
Manager	1	2	2 GB RAM 2 CPU cores	4 GB RAM 4 CPU cores
Worker	1	2	8 GB RAM 4 CPU cores	16 GB RAM 4 CPU cores

Java version

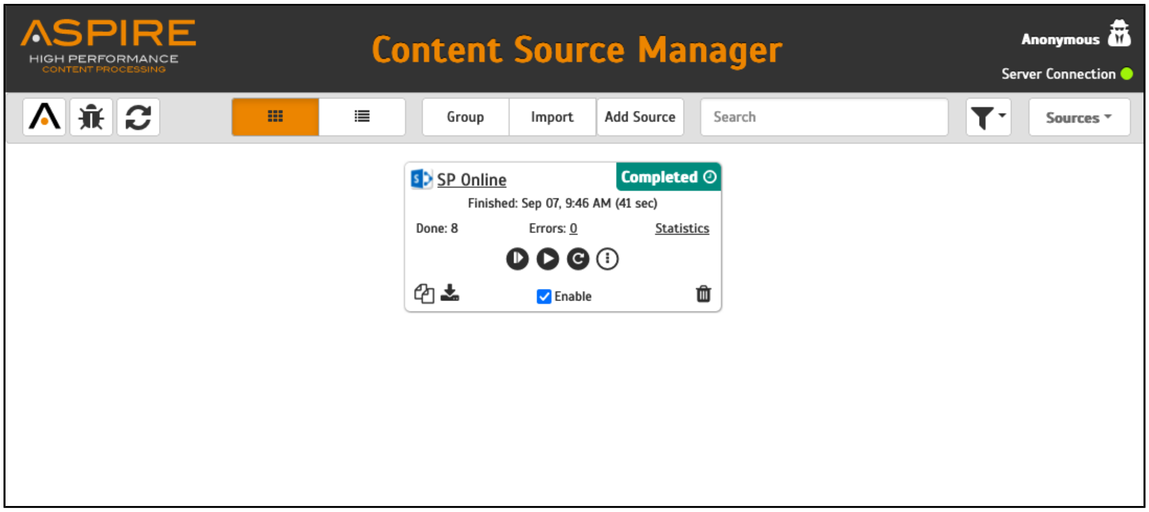
Aspire 5 was developed and tested using OpenJDK 11

Step 2. Deploy your Aspire 5 cluster

There are several options on deploying Aspire 5, from on-premise installations both Windows or Linux based, up to container based deployments using Kubernetes. Choose your preferred deployment option and follow the instructions at How to Install Aspire.

Step 3. Migrate a content source crawl configuration

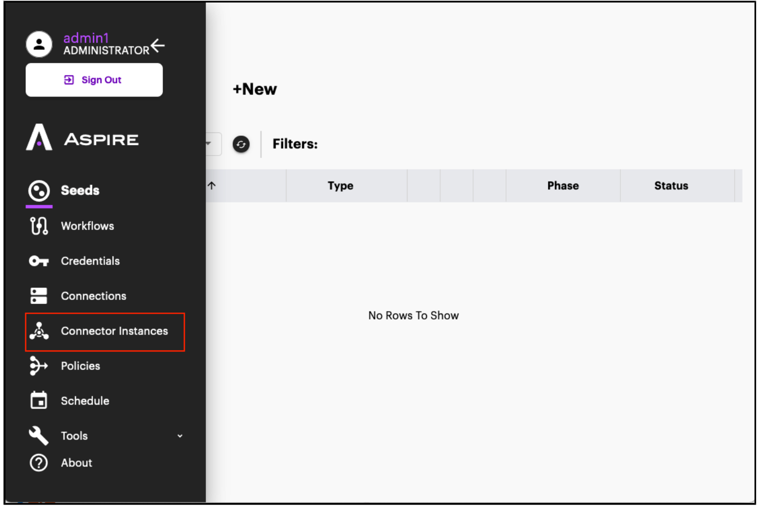
Choose a content source on Aspire 3/4 you want to migrate to Aspire 5. Verify the availability of the connector in Aspire 5 at Connectors.

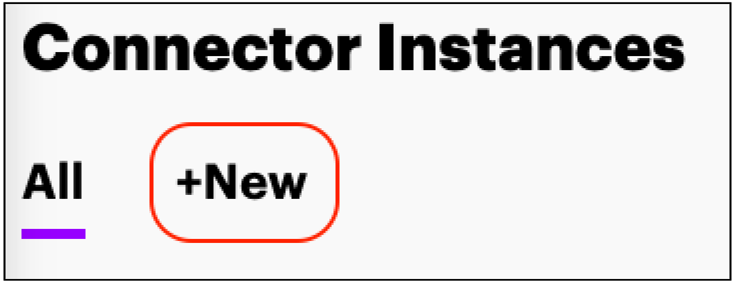


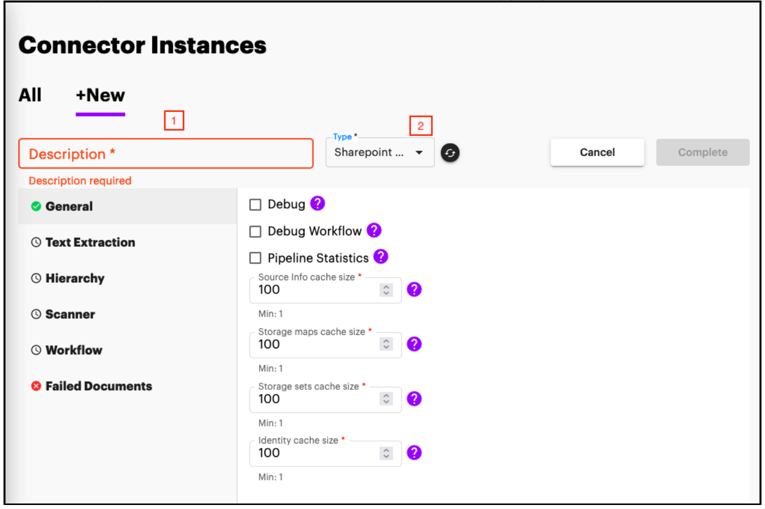
Finished: Sep 07, 9:46 AM (41 sec)

Done: 8 Errors: 0

We'll use a SharePoint Online content source in Aspire 4 as an example

1. Create a **Connector Instance** in Aspire 5 for the connector you'll use. Note that this connector instance can be shared across multiple crawl configurations, so you may only create one connector instance per content source connector type in Aspire 4.
    a. Open Aspire 5 Admin UI and click on the "**Connector Instances**" Section on the left side menu
        i. 

    b. On the Connector Instances Section, click on **"New"**
        i. 

    c. Enter the name for your new Connector instance, and select its Type (in this case SharePoint online)
        i. 

Description *

Description required

Type * Sharepoint ...

□ Debug

□ Debug Workflow

□ Pipeline Statistics

Source Info cache size * 100

Min: 1

Storage maps cache size * 100

Min: 1

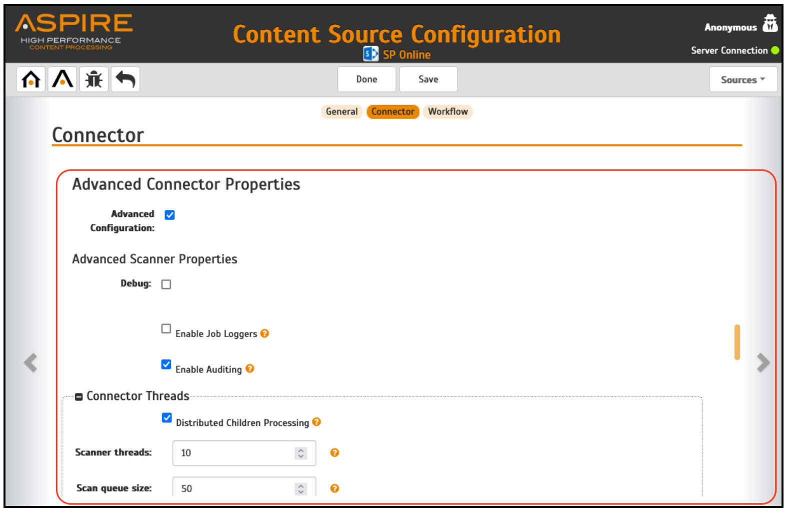
Storage sets cache size * 100

Min: 1

Identity cache size * 100

Min: 1

    d. Configure the properties you need for this connector instance, all the properties you can select here can be found in **Aspire 3/4** at the "**Connector**" section of the content-source configuration, at the "**Advanced Connector Properties**" sub-section.

i. 

Connector

Advanced Connector Properties

Advanced Configuration: ☑

Advanced Scanner Properties

Debug: □

□ Enable Job Loggers

☑ Enable Auditing

Connector Threads

☑ Distributed Children Processing

Scanner threads: 10

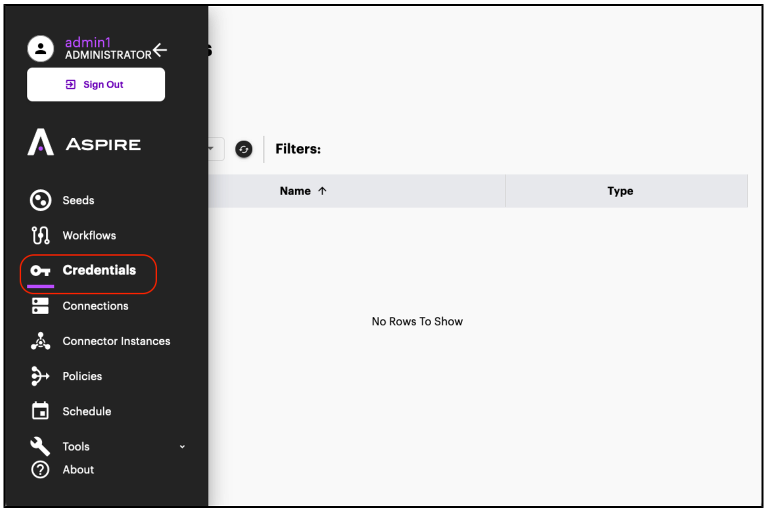
Scan queue size: 50

e. Click on **"Complete"** on the Aspire 5, connector instance creation, once all the properties have been set-up.

---

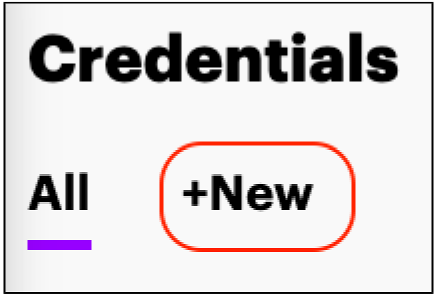
2. Once a connector instance is created, now create a **Credential** configuration object. We'll use this to create our connection at step 3.

   a. Open the **"Credentials"** section on the left side menu

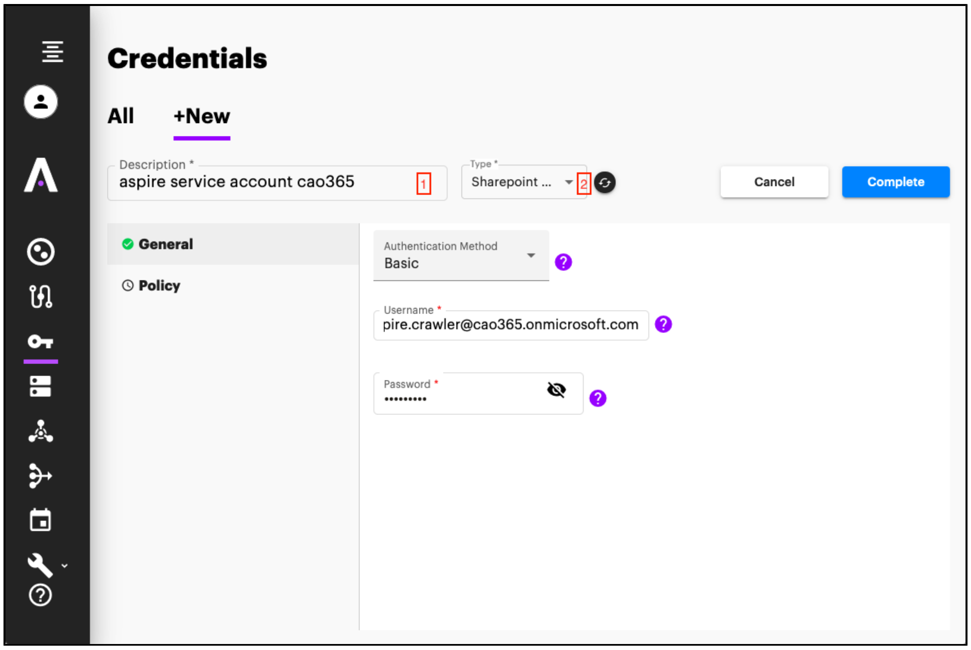
   i. 

No Rows To Show

   b. On the **Credentials** section, Click on **"New"**

   i. 

   c. Enter a name for your **Credentials** object and the type of source (in this case SharePoint online)

i. 

Credentials

Description * aspire service account cao365

Type * Sharepoint ...

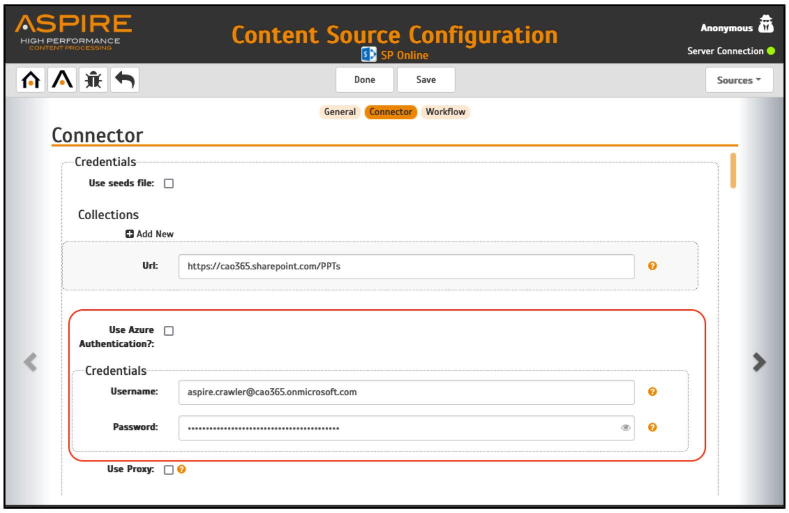
General

Authentication Method Basic

Username * pire.crawler@cao365.onmicrosoft.com

Password * •••••••••

d. Choose and fill the right credentials properties, this can typically be found on the **"Connector"** section on the **Aspire 3/4** content source

i. 

Content Source Configuration

SP Online

Connector

Credentials

Use seeds file: □

Collections

Url: https://cao365.sharepoint.com/PPTs

Use Azure Authentication?: □

Credentials

Username: aspire.crawler@cao365.onmicrosoft.com

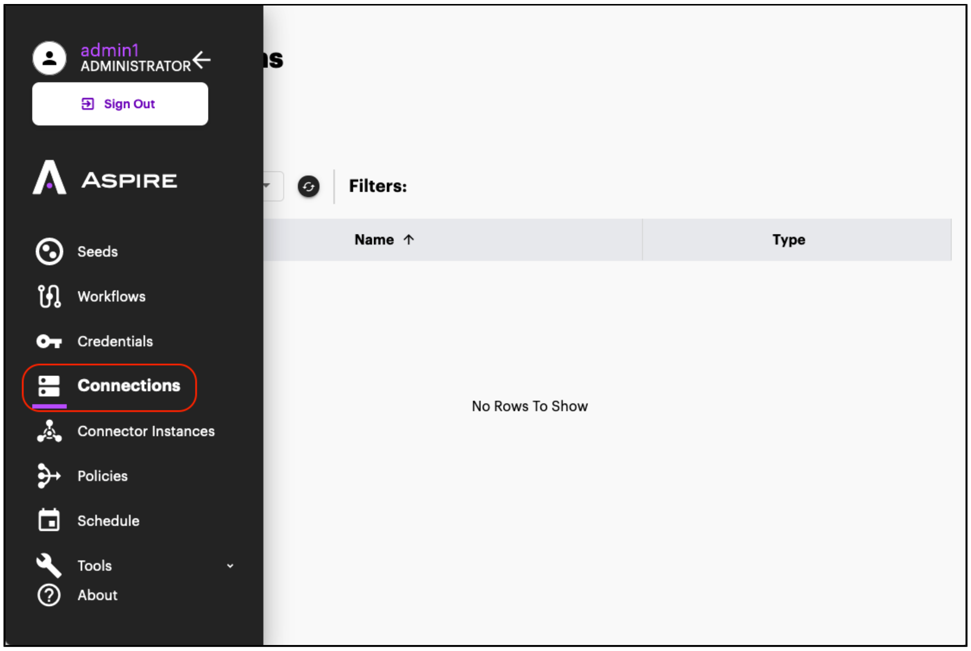
Password: ••

Use Proxy: □

e. Once all the properties have been set, Click on **"Complete"** to create the **Credentials** object

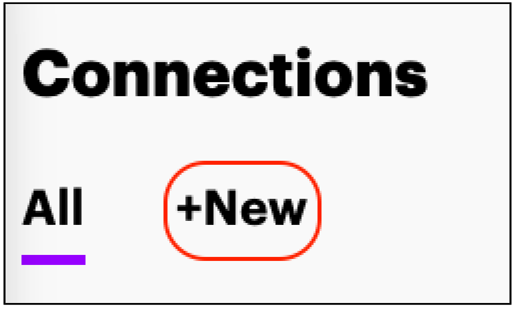
---

3. Once the Credentials object is created, now create a **"Connection"** configuration object. This will point to your SharePoint online instance, without indicating the site collections to crawl (each site collection or list would be a different configuration object).
   a. Open the **"Connection"** section on the left side menu

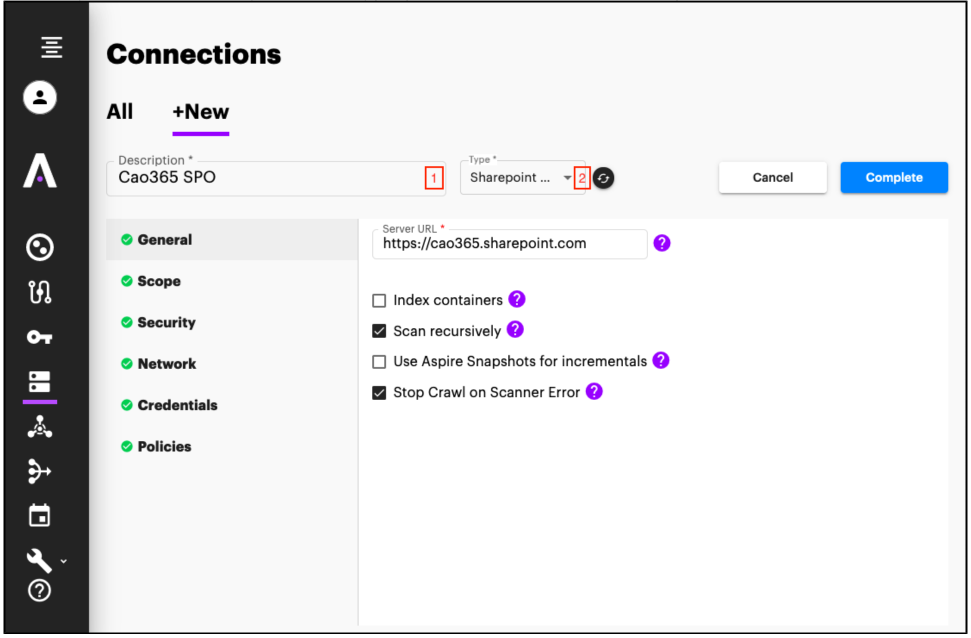
i. 

No Rows To Show

b. On the Connections section, Click on **"New"**

i. 

c. Enter a name for your Connection object, and its type (in this case SharePoint Online)

i. 

Description * Cao365 SPO

Type * Sharepoint ...

Server URL * https://cao365.sharepoint.com

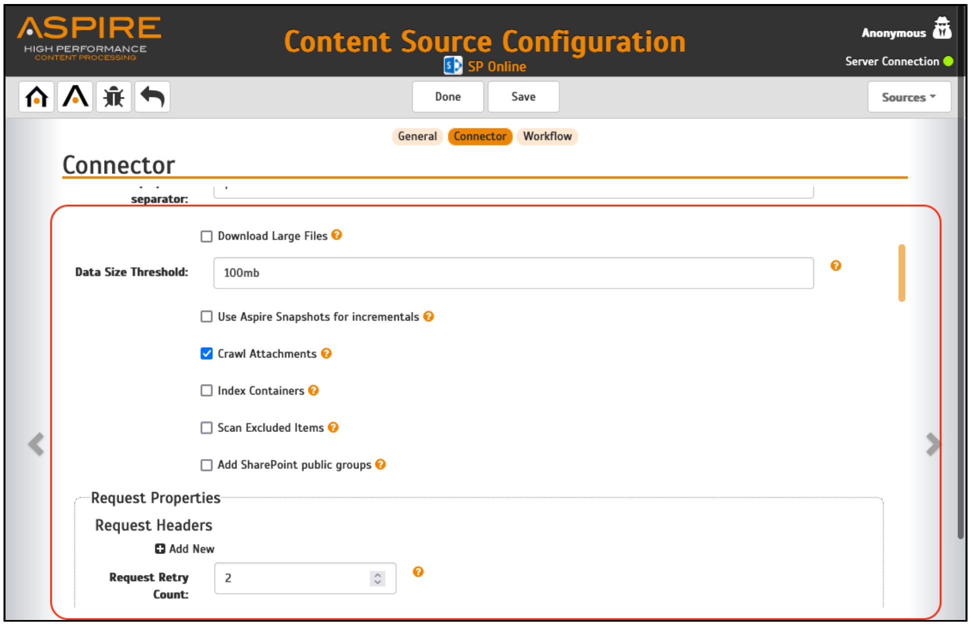
☐ Index containers

☑ Scan recursively

☐ Use Aspire Snapshots for incrementals

☑ Stop Crawl on Scanner Error

d. Enter the properties required for your connection object, these properties can be found in **Aspire 3/4** at the **"Connector"** section of the **content-source.** Notice that the **Server URL** is NOT the crawl path, but rather, a base URL to use for the crawls. All paths configured in section 4 seeds, will be relative to this Server URL. What's considered a Base URL changes from connector to connector. Please check each connector documentation for more details.

i. 

separator:

☐ Download Large Files

Data Size Threshold: 100mb

☐ Use Aspire Snapshots for incrementals

☑ Crawl Attachments

☐ Index Containers

☐ Scan Excluded Items

☐ Add SharePoint public groups

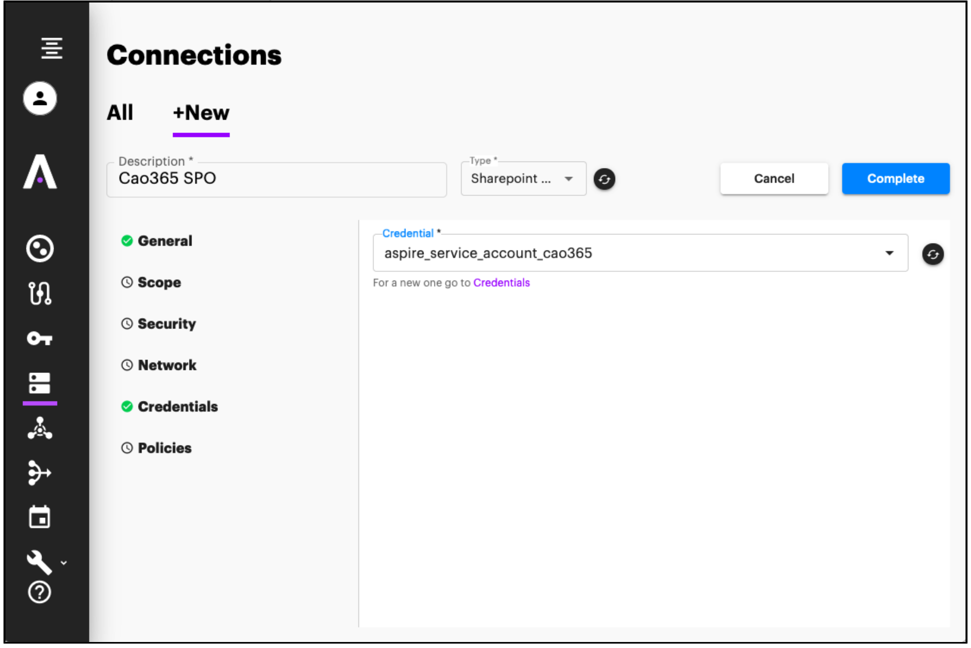
Request Properties

Request Headers

Add New

Request Retry Count: 2

e. Select the credential you created in step 2

i. 

Description * Cao365 SPO

Type * Sharepoint ...

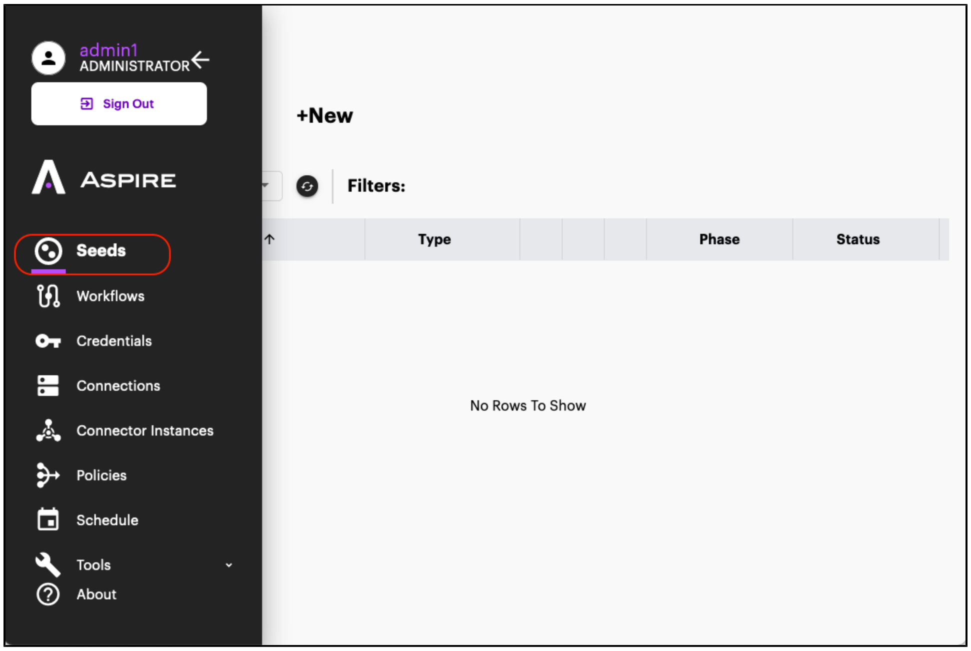
Credential * aspire_service_account_cao365

For a new one go to Credentials

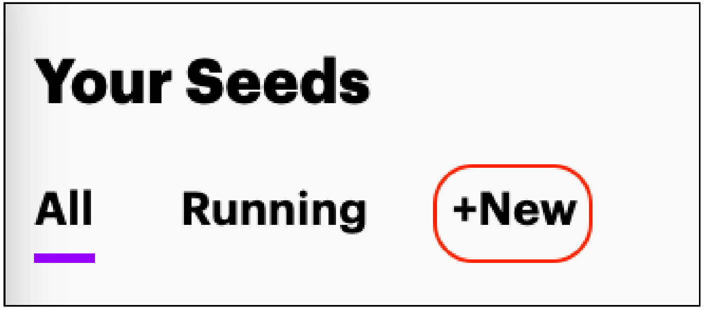
f. Once everything is configured, Click on **"Complete"** to create the **Connection** object

---

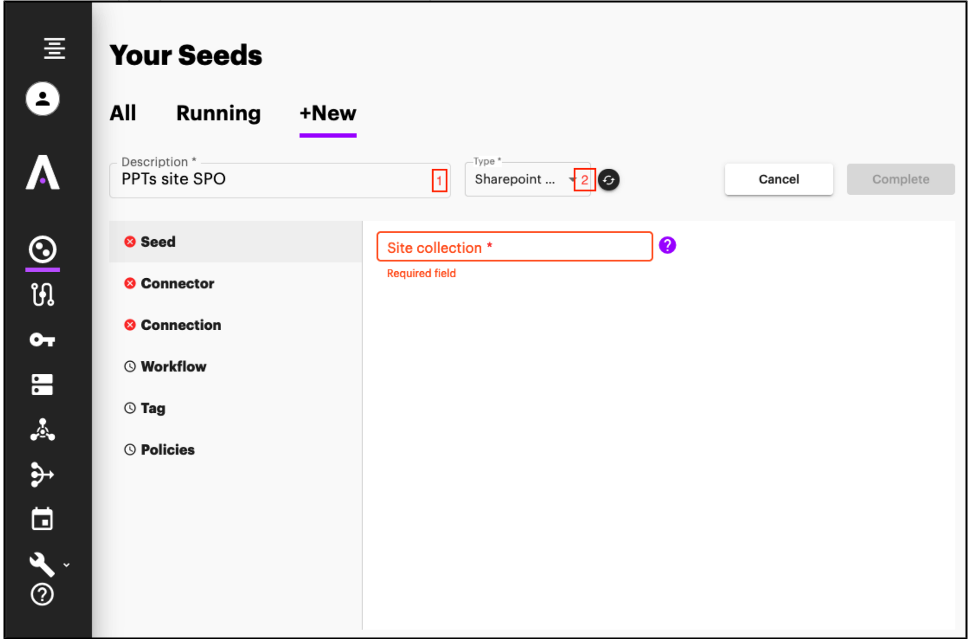
4. Once the Connection object is created, now the **Seed** objects can be created. **Seeds** are the starting points for the crawls, they represent specific locations to start the crawls from. They are configured relative to **"Connection"** and **Connector Instance** objects.
   a. Open the **"Seeds"** section on the left side menu

i. 

b. Once on the Seeds section, Click on "New"

i. 

c. Enter a name and a Type (in this case SharePoint Online)

i. 

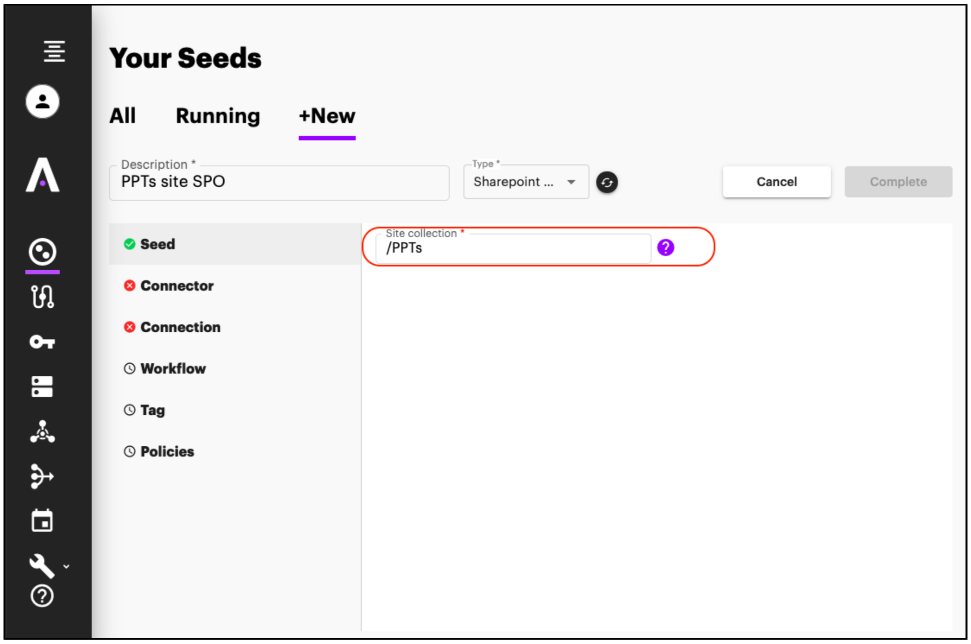
Description * PPTs site SPO

Type * Sharepoint ...

Site collection *

Required field

d. Enter the **relative** path to your start site collection or list (do not include the Server URL)

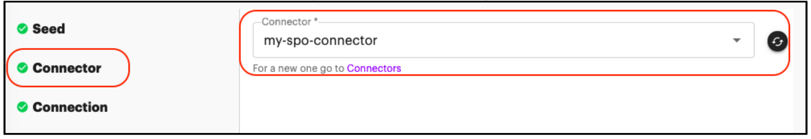
i. 

Description * PPTs site SPO

Type * Sharepoint ...

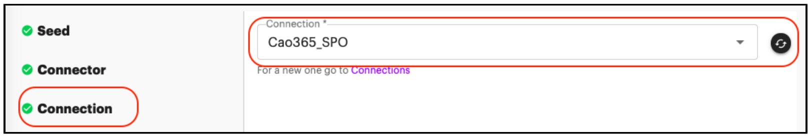
Site collection * /PPTs

e. Choose the **Connector** and **Connection** objects created in Steps 1 and 3

   i. 

   Connector * my-spo-connector

   For a new one go to Connectors

   ii. 

   Connection * Cao365_SPO

   For a new one go to Connections

f. On the **Workflow** section, select the **workflows** that the documents generated by the crawls will execute.
   i. Follow Workflows - Migration Guide for details on migrating your existing Aspire 3/4 workflow configurations
   ii. If you don't have a workflow, you can leave it Empty.

g. Click on Complete to create the new **Seed**

Congratulations you have successfully migrated a content source from Aspire 3/4 into Aspire 5.0! Now you can Start crawling! Now let's configure scheduling and policies.

Step 4. Schedule your crawls

Scheduling a Crawl in Aspire 5 differs from Aspire 3/4 on where it is configured. Like you have seen so far, everything is configured in separate configuration objects, and scheduling is no different: it must be configured via its own configuration object.

A Schedule Configuration object specifies when to start an Action (Start, Stop, Pause, Resume) on a set of Seed objects.

Follow instructions at Schedules for more information on how to set them up.

Step 5. Throttling and Routing Policies

Applying policies to crawls is a new feature of Aspire 5.0, which limits where and when documents can be processed during a crawl.

There are two types of policies: **throttling** which limits the crawl rate and **routing** which limits the Aspire worker nodes on which documents can be processed.

Follow the documentation at Policies to learn more about how they work and how to set them up.

Reference Summary

Here is a summary on each configuration objects we have covered.

- **Connector Instances**
  - All properties under "Advanced Configuration" in Aspire 3/4.
  - REST Endpoint documentation Connectors API

- **Credentials**
    - All access related properties, account names, passwords, authentication type, etc
    - A single credential instance can be reused for many different **connections instances.**
    - REST Endpoint documentation Credentials API
- **Connection**
    - Everything that has to do with the actual connection to the repository like: server URL, connection timeouts, proxies, etc.
    - Can be associated with 1 credential object and 1 connector object.
    - REST Endpoint documentation Connections API
- **Workflow**
    - Same old workflow, but on Aspire 5 this must be configured from scratch on the UI or via REST commands, as this is no longer an xml file.
    - Details on migrating workflows at Workflows - Migration Guide
    - REST Endpoint documentation Workflow API
- **Schedule**
    - Similar to the "content-source" schedules in Aspire 4.0, it supports time schedules, but also supports the new "sequence" schedules which can trigger crawls after another schedule has been completed.
    - REST Endpoint documentation Schedules API
- **Policies**
    - New to Aspire 5.0, there are two types of policies
        - routing
        - throttle
    - REST Endpoint documentation Policies API.
- **Seed**
    - Starting point of a crawl.
    - REST Endpoint documentation Seeds API

What's next?

- Workflows - Migration Guide