Azure Data Lake How to Configure

Step 1. Launch Aspire and open the Content Source Management Page

1. Launch Control
2. Browse to: http://localhost:50505. For details on using the Aspire Content Source Management page, please refer to Admin UI

On this page

Step 2. Add a new Azure Data Lake Content Source

To specify exactly which shared folder to crawl, we will need to create a new "Content Source".

To create a new content source:

1. From the Content Source , click on "Add Source" button.
2. Click on "Azure Data Lake Connector".

Step 2a. Specify Basic Information

? Unknown Attachment

In the "General" tab in the Content Source Configuration window, specify basic information for the content source:

1. Enter a content source name in the "Name" field.
   - This is any useful name which you decide is a good name for the source. It will be displayed in the content source page, in error messages, etc.
2. Click on the **Scheduled** pulldown list and select one of the following: *Manually, Periodically*, *Daily,Weekly or Advanced.*
   - Aspire can automatically schedule content sources to be crawled on a set schedule, such as once a day, several times a week, or periodically (every N minutes or hours).For the purposes of this tutorial, you may want to select Manually and then set up a regular crawling schedule later.
3. Click on the **Action** pulldown list to select one of the following: *Start, Stop*, *Pause,* or *Resume.*
   - This is the action that will be performed for that specific schedule.
4. Click on the **Crawl** pulldown list and select one of the following: *Incremental, Full*, *Real Time,* or *Cache Groups.*
   - This will be the type of crawl to execute for that specific schedule.
5. After selecting a Scheduled, specify the details, if applicable:

- *Manually*: No additional options.
- *Periodically*: Specify the "Run every:" options by entering the number of "hours" and "minutes."
- *Daily*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options.
- *Weekly*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options, then clicking on the day checkboxes to specify days of the week to run the crawl.
- *Advanced*: Enter a custom CRON Expression (e.g. 0 0 0 ? * *)

You can add more schedules by clicking in the **Add New** option, and rearranging the order of the schedules.

If you want to disable the content source, clear the Enable check box. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

⚠ Real Time and Cache Groups crawl will be available depending of the connector.

Step 2b. Specify Connector Information

In the Connector tab, specify the connection information to crawl a Azure Data Lake folder.

1. **Credentials**
    a. *Authorization Token End Point:* OAuth 2 End Point supplied by Azure for your App, format https://login.microsoftonline.com/[mykey]/oauth2/token
    b. *Application ID*: Client ID for your application
    c. *Application Secret:* Key supplied by Azure
    d. *Fully Qualified Domain Name or FQDN:* full path of your Data Lake domain, format [mydomain].azuredatalakestore.net
2. **Source**
    a. Collect from Root: Within this option, connector will crawl from root directory of Azure Data Lake FQDN supplied. Meaning *"/"*
    b. Use Seeds File: This option will allow collect paths from a supplied file location, very useful if paths will be constantly changing and controlled by a 3rd party process. Paths should be listed one per line in a form of */folder/sub-folder*
        i. For Windows: D:\folder\folder1\paths.txt
        ii. For Linux: /home/user/folder/folder1/paths.txt
    c. Specific Paths: This option will allow submit N paths. Admin is able to supply as many paths in a format of */folder/sub-folder*

Note: Permission should be granted to the Application for each folder to crawl, in a form of Read and Execute (r-x). *For more information about Data Lake Access Control, please refer to this link*

Index Containers: Select if folders are to be indexed

Scan Recursively: Select if sub-folder are to be scanned

Scan Excluded Items: If selected, the scanner will scan sub items of container items that have been excluded by a pattern (because it matches an exclude pattern or because it doesn't match an include pattern)

Step 2c. Specify Workflow Information

? Unknown Attachment

In the "Workflow" tab, specify the workflow steps for the jobs that come out of the crawl. Drag and drop rules to determine which steps should an item follow after being crawled. This rules could be where to publish the document or transformations needed on the data before sending it to a search engine. See Workflow for more information.

1. For the purpose of this tutorial, drag and drop the *Publish To File* rule found under the *Publishers* tab to the **onPublish** Workflow tree.
    a. Specify a *Name* and *Description* for the Publisher.
    b. Click *Add*.
2. Click **Save** and **Done** and you'll be sent back to the Home Page.

Step 3: Initiate a Full Crawl

Now that the content source is set up, the crawl can be initiated.

1. Set the Crawl Type option to "Full". (The default is Incremental. After the first crawl, set it back to "Incremental" to crawl for any changes in the repository).
2. Click **Start** .

During the Crawl

During the crawl, you can do the following:

- Click **Refresh** on the Content Sources page to view the latest status of the crawl.
    - The status will show **RUNNING** while the crawl is going, and **CRAWLED** when it is finished.
- Click **Complete** to view the number of documents crawled so far, the number of documents submitted, and the number of documents with errors.

If there are errors, a clickable "Error" flag will take you to a detailed error message page.

Step 4: Initiate an Incremental Crawl

If you only want to process content updates from the Azure Data Lake (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The Azure Data Lake connector will automatically identify only changes which have occurred since the last crawl.

If this is the first time that the connector has crawled, the action of the "Incremental" button depends on the exact method of *change* discovery. It may perform the same action as a "Full" crawl crawling everything, or it may not crawl anything. Thereafter, the Incremental button will only crawl updates.

Statistics are reset for every crawl.

Group Expansion

An Azure Data Lake Store as part of the Azure ecosystem relies on the Azure Active Directory to delimit permissions against Users and Groups. Aspire provides a separate connector to pull those records via service.

Group expansion configuration is performed on the *Azure Active Directory Group Expander.*

Please refer to Azure Active Directory Group Expander for more information.