

Azure Blob Storage Connector 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 1. Launch Aspire and open the Content Source Management Page](#)
- [Step 2. Add an Azure Blob Storage Content Source](#)
 - [Step 2a. Specify Basic Information](#)
 - [Step 2b. Specify Connector Information](#)
 - [Step 2c. Specify Workflow Information](#)
- [Step 3: Initiate a Full Crawl](#)
 - [During the Crawl](#)
- [Step 4: Initiate an Incremental Crawl](#)

Step 2. Add an Azure Blob Storage Content Source

To specify exactly which shared folder to crawl, first create a new Content Source.

1. From the Content Source , click **Add Source**.
2. Click **Azure Blob Storage 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 ? * *)

Step 2b. Specify Connector Information

In the Connector tab, specify the connection information to crawl an Azure Blob container.

Scan Recursively: Select if sub-folders 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 Storage Blob (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The Azure Blob Storage connector will automatically identify only changes that 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.