Aspider How to Configure

On this page

Step 1. Launch Aspire / Go to Content Source Management

? Unknown Attachment

1. Launch Aspire (if it's not already running).
    - See Launch Control
2. Open the **Content Source Management** page.
3. Browse to: http://localhost:50505.
    - For details on using the Aspire Content Source Management page, go to Admin UI.

Step 2. Add an Aspider Web Crawler Content Source

? Unknown Attachment

To specify exactly what 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 ***Add Source***.
2. Click ***Aspider Web Crawler***.

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.
    a. 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 the **Scheduled** pulldown list and select one of the following: *Manually, Periodically*, *Daily, Weekly or Advanced.*
    a. 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 the **Action** pulldown list to select one of the following: *Start, Stop*, *Pause,* or *Resume.*
    a. This is the action that will be performed for that specific schedule.
4. Click the **Crawl** pulldown list and select one of the following: *Incremental, Full*, *Real Time,* or *Cache Groups.*
    a. This will be the type of crawl to execute for that specific schedule.

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.

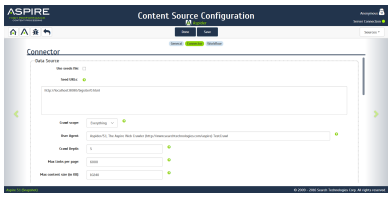
- *Advanced*: Enter a custom CRON Expression (e.g. 0 0 0 ? * *)

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

If you want to disable the content source just unselect the the "Enable" checkbox. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

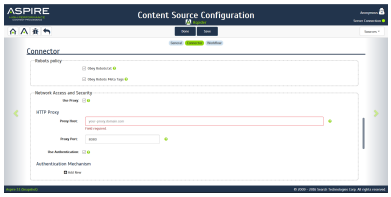
Real Time and Cache Groups crawl will be available depending of the connector.

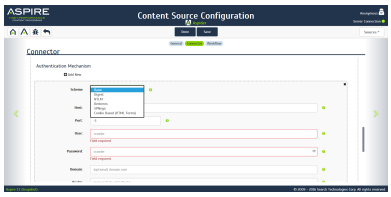
Step 2b. Specify the Connector Information



In the "Connector" tab, specify the connection information to crawl the Aspider Web Crawler.

1. Select "**Use seeds file**" if you want your seed urls to be in a filesystem text file.
    a. If selected, specify the path to the seeds file in the "**Seed URLs**" option.
    b. Otherwise enter the seed urls in the text area "**Seed URLs**"
2. Select the "**Crawl Scope**"
    a. **Everything**: No filter will be applied, every link discovered will be crawled
    b. **Domain Only**: Only links pointing to the same domain will be crawled. For example if the domain of the seed URL is searchtechnologies.com, URLs like aspire.searchtechnologies.com will be crawled.
    c. **Host Only**: Only links pointing to the exact same host will be crawled. For example if the host of the seed URL is aspire.searchtechnologies.com, URLs like XXX.searchtechnologies.com will NOT be crawled.
3. Customize the "**User Agent**" only if your site requires it.
4. Modify the "**Crawl Depth**" depending on how deep you want the Crawler to go.
    a. If the depth is 1, only the seed URL and its children will be crawled
    b. If the depth is 2, the crawl will end when the seed's grandchildren are processed.
    c. Etc.
5. Modify the "**Max Links per page**" to allow more links to be discovered in each page. This option is there to prevent spider traps from generating lots of dummy links.
6. Specify the "**Max content size (in bytes)**" option for limiting how big the pages pages can be to be processed. This is also an option to prevent spider traps from generating lots of dummy content.
    a. Any URL bigger than this limit will be logged as an Error.
7. Select the "**Deletes Policy**"
    a. "**Immediate**" means the URLs are deleted as soon as they are not found in the latest incremental crawl. You have to be careful with this option since some networking errors can cause big chunks of your pages not to be crawled.
    b. "**Time based**" will only delete the URLs not found in the latest crawls when certain (configurable) days have passed without seeing those URLs in the crawl.
        i. The deletes are sent on the next incremental crawl after the condition is met.
    c. "**After X Incrementals**" will only delete the URLs not found in the latest crawls when certain (configurable) incremental crawls have passed without seeing those URLs in the crawl.
8. Select "**Customize Connection Timeouts**" if your crawl has timeouts issues because of the network you are in. All timeouts must be specified in milliseconds.
    a. **Connection Timeout** is how long will the crawler wait for the connections to be established with the web servers.
    b. **Connection Request Timeout** is how long will the crawler wait for a connection to be pulled from the connection manager.
    c. **Socket Timeout** is how long the crawler will wait between two consecutive data packages.
9. Modify the **Connection Throttling** options as required
    a. **Max Urls Per Hostname per minute** this option specify the maximum URLs to retrieve from a single host each minute. If your crawl only has one host and the quota is reached, the crawler will wait for the next minute so the quota will be reset.





Be extra careful when configuring this option, since the web servers might be heavily impacted by the crawler requests. Consult with the Web Server Administrator before setting this option.

    b. **Hosts processors per Aspire node** a host processor is a thread that send URLs for a single host to be processed to the Aspire pipelines. Having only one processor means only URLs from one host at the time will be processed at the time in each server.
10. Select "**Obey Robots.txt**" if you want to respect the rules set by the servers in the robots.txt files. This is the recommended option.
    a. More information about the robots.txt rules at: http://www.robotstxt.org/robotstxt.html
11. Select "**Obey Robots Meta Tags**" if you want the HTML Robots Meta tags to be respected. This is the recommended option.
    a. More information on these meta tags at: http://www.robotstxt.org/meta.html
12. Select "**Use Proxy**" if your server requires a proxy to access the URLs to be crawled.
    a. Specify the "**Proxy Host**" and "**Proxy Port**"
13. Select "**Use Authentication**" if the URLs to crawl require it. If selected add the required Authentication Mechanisms required along with the credentials. The available mechanism are: Basic, Digest, NTLM, SPNego, Kerberos and Cookie Based (HTML Forms).
    a. For "**Kerberos**" and "**SPNego**"
        i. **Key Distribution Center**: The network service that supplies session tickets and temporary session keys to users and computers within an Active Directory domain. Visit Prerequisites for more information on how to get this address.
        ii. Select "**Verbose Negotiation**" If you want the negotiation handshake to be logged.
    b. **Cookie Based (HTML Forms).** Visit Prerequisites for more information on how to get this information
        i. **Login URL**: This is the URL of the login form you are redirected when you are not logged in.
        ii. **Form Element Path**: The CSS Selector of the HTML form inside the login page from above. (You can use the inspection on Google Chrome browser to generate this for you)
        iii. **Username Field**: The "name" attribute of the HTML "input" element where the username should be filled.
        iv. **Password Field**: The "name" attribute of the HTML "input" element where the password should be filled.
        v. If you need extra fields to be sent as constants along with the form select "**Add New**" (i.e. the login submit button is sometimes required)
            1. **Name**: the id of the field to send
            2. **Value**: the value to send
    c. Common options
        i. **Host**: The host where the authentication mechanism will be used.
        ii. **Port**: The port of the host where where the authentication mechanism will be used.
        iii. **User**: The username to authenticate with
        iv. **Password**: The password to authenticate with
        v. **Domain**: If the user belongs to a domain, and the authentication scheme (mechanism) requires it, you must specify it here. If this is not empty the final username will be sent as "DOMAIN\username"
        vi. **Realm**: If a realm is required by the authentication scheme (mechanism)
14. **Use Fixed ACLs**: Check if you want to use a fixed acl that will be attached to all the documents fetched
    a. **Users**: Fixed user acls
        i. **Domain**: User's domain
        ii. **Name**: Username
        iii. **Type**: Select between Allow and Deny
    b. **Groups**: Fixed group acls
        i. **Name**: Group's name
        ii. **Type**: Select between Allow and Deny
15. **Include/Exclude patterns**: Enter regex patterns to include or exclude documents based on URL matches.
16. Select "**Scan Excluded Items**" if you want the excluded URLs to be "scanned" for new links.
17. Select "**Reject images / videos / javascript / css**" if you want those items to be excluded from the crawl.

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. After completing these steps click ***Save*** and ***Done.***
    - 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. Click on the crawl type option to set it as "Full" (is set as "Incremental" by default and the first time it'll work like a full crawl.
    - After the first crawl, set it to "Incremental" to crawl for any changes done in the repository).
2. Click ***Start***.

During the Crawl

During the crawl, you can do the following:

1. Click on the "Refresh" button 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.
2. Click on "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, you will get a clickable "Error" flag that 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 Aspider Web Crawler (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button.

- The Aspider Web Crawler 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.