Kafka How to Configure

Use the following steps to configure your Kafka Connector in Aspire.

On this page

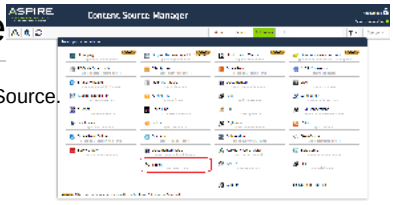
Step 1. Launch Aspire / Open the Content Source Management Page

Launch Aspire (if it's not already running). See:

- Launch Control
- Browse to: http://localhost:50505.

For details on using the Aspire Content Source Management page, please refer to the Admin UI.

Step 2. Add a New Kafka Content Source



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

To create a new content source:

1. From the *Content Source Manager*, click **Add Source**.
2. Select the **Kafka Connector**

Step 2a. Specify Basic Information

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 on the **Scheduled** pull-down 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 on the **Action** pull-down 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 on the **Crawl** pull-down 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 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 rearrange the order of the schedules.

If you want to disable the content source just clear the 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 crawls will be available depending on the connector.

Step 2b. Specify the Connector Information

Kafka Servers: The servers(s) to connect to, in <host>:<port> format. This can be a comma separated list of servers

e.g. first.server:9092,second.server:9092

Topic: The message stream to subscribe to.

Starting Offset for Full Crawls: Which messages to start fetching when starting a full crawl

- **Earliest:** Start from the earliest message available in the stream.
- **Manually Specify Offsets**: Manually specify the offset of the message to start from
  - **Partition:** Partition number
  - **Offset:** The message offset to start from

Note: *If certain partitions are not specified, Kafka will default to fetching the earliest message.*

Step 2c. Specify Workflow Information

In the *Workflow* tab, specify the workflow steps for the jobs that come out of the crawl. Drag and drop rules determine which steps an item should follow after being crawled. These rules could include 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**.

After completing this step, click **Save** then **Done** and you'll be sent back to the *Home* page.

Step 3: Initiate a 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 on the Start button.

During the Crawl

During the crawl, you can do the following:

- 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.
- 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.

Note: *For the Kafka connector, the behavior of the crawl is different between Incremental and Full crawls*

- **Full:** Obtains all messages that are available during crawl start time, stop when these messages have been processed.
- **Incremental:** Obtain messages continuously, stop only when the content source is manually stopped or paused.