How to Configure Kinesis

On this page

- Step 1. Launch Aspire / Open the Content Source Management page
- Step 2. Add a new Kinesis Content Source
  - Step 2a. Specify Basic Information
  - Step 2b. Specify the Connector Information
  - Step 2c. Specify Workflow Information
- Step 3: Initiate a Crawl
  - During the Crawl

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 Kinesis Content Source

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

1. From the *Content Source Manager*, click **Add Source**.
2. Select **Kinesis 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:

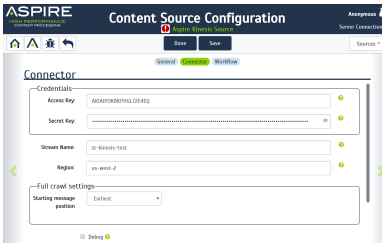
- Manually: No additional options
- Periodically: Specify "Run every:" options by entering the number of "hours" and "minutes."
- Daily: Specify the "Start time:" Click the hours and minutes drop-down lists and select options
- Weekly: Specify the "Start time:" by clicking on the hours and minutes drop-down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

> **Note:** You can add more schedules. Select **Add New** and rearrange the order of the schedules. To disable a content source, clear the the **Enable** check box. This may be useful if the folder is 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

1. **Enter AWS account credentials**
   a. Access Key
   b. Secret Key

2. **Enter the stream details**
    a. Stream Name
    b. Region (e.g. us-west-1, see AWS Regions and Endpoints)
3. **Configure full crawl behavior (set starting point per shard)**
    a. Earliest (default): Start at the first untrimmed record for all shards
    b. Latest: Fetch only records that arrives after crawl start
    c. At Sequence Number: Start from the record starting from the specified sequence number
    d. After Sequence Number: Start from first available record that is immediately after the specified sequence number
    e. At Timestamp: Start from the record at or after the specified timestamp. Timestamp must be specified using the ISO format with milliseconds precision (e.g. "2018-03-13T15:55:15.027Z")



Note: *If using the* At Sequence Number, After Sequence Number *or* At Timestamp *options, unspecified shards will default to "Earliest".*

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.
2. After the first crawl, set it to "Incremental" to crawl for any changes done in the repository).
3. 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. (*see note below)
- 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, you will get a clickable "Error" flag that will take you to a detailed error message page.

Important: *Due to the nature of Kinesis Data Streams, the connector does not actually stop crawling and will keep running as long as errors are not encountered, or the connector is not stopped or paused. This is normal behavior and allows the connector to continuously pick up any incoming new data. The only time it will stop is that if all the shards that were picked up at the start of the crawl end up being closed by a reshard operation.*