Parquet Extractor How to configure

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

? Unknown Attachment

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 Admin UI

Step 2. Add a new Content Source

- For this step please follow the step from the Configuration Tutorial of the connector of you choice, please refer to Connector list

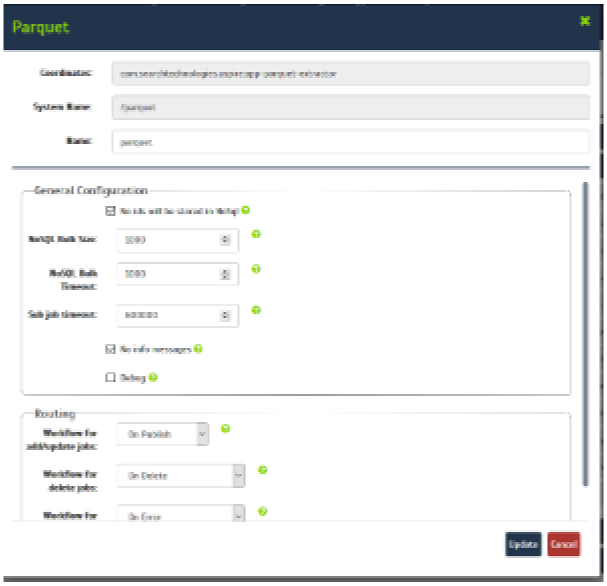
Step 3. Add a new Parquet Extractor to the Workflow

To add a Parquet Extractor application drag from the **Parquet Extractor** rule from the *Workflow Library* and drop to the *Workflow Tree* where you want to add it. This will automatically open the Parquet Extractor window for the configuration of the application.

Step 3a. Specify Application Information

In the Parquet Extractor window, specify the configuration information of the application.

1. General configuration
   a. No ids will be stored in NoSql - Check if you do not want to store ids in NoSql. Warning! Aspire will not be able to delete items in the index in future crawls.
   b. NoSQL Bulk Size - The size of the bulk write operations done to the NoSql Database
   c. NoSQL Bulk Timeout - The amount of time to wait before flushing the bulk operations after the last insert
   d. No info messages - Check if you want info messages disabled
   e. Sub Job timeout.
      i. Time in milliseconds to wait before the current job is killed for inactivity.
      ii. **Example:** 60000
   f. Debug:
      i. Enable debug messages.
2. Routing
   a. Workflow for add/update jobs:
      i. Workflow to send the generated add or update jobs.
   b. Workflow for delete jobs:
      i. Workflow to send the generated delete jobs.
   c. Workflow for error jobs:
      i. Workflow to send the generated error jobs.



Parquet

Coordinates: com.searchtechnologies.aspire:app-parquet-extractor

System Name: /parquet

Name: parquet

General Configuration

☑ No ids will be stored in NoSql

NoSQL Bulk Size: 1000

NoSQL Bulk Timeout: 1000

Sub job timeout: 600000

☑ No info messages

☐ Debug

Routing

Workflow for add/update jobs: On Publish

Workflow for delete jobs: On Delete

Workflow for: On Error

Step 3b. Share rule to a library

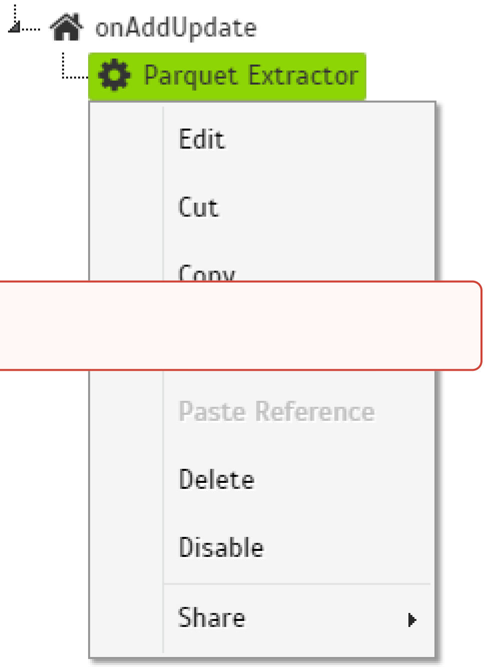
Once the application is set, it must be shared to a (new or existing) library.

Note: This is a required step.

Step 3c. Copy the rule from the shared library

Add the shared application from the library to the Delete workflow.

Note: This is a required step.



In order to work, the application requires to disable the ExtractText stage performed by the connector on the connector's Advanced Properties.

Once you've clicked on the *Add* button, it will take a moment for Aspire to download all of the necessary components (the Jar files) from the Maven repository and load them into Aspire. Once that's done, the application will appear in the Workflow Tree.

For details on using the Workflow section, please refer to Workflow introduction.