

HBase Connector How to Configure

Before Launching Aspire

Change the felix.properties file and add this lines if the Kerberos authentication is going to be used:

```
# To append packages to the default set of exported system packages,
# set this value.
org.osgi.framework.system.packages.extra=\
...
sun.security.krb5, \
com.sun.security.auth.callback

# The following property makes specified packages from the class path
# available to all bundles. You should avoid using this property.
org.osgi.framework.bootdelegation=\
...
javax.security.sasl, \
sun.security.krb5
```

On this page:

- [Before Launching Aspire](#)
- [Step 2. Add a new HBase Content Source](#)
 - [Step 2a. Specify Basic Information](#)
 - [Step 2b. Specify the Connector Information](#)
 - [Step 2c. Specify Workflow Information](#)
 - [Step 2d. HBase settings file](#)
- [Step 3: Initiate a Full Crawl](#)
 - [During the Crawl](#)
- [Step 4: Initiate an Incremental Crawl](#)
 - [Group Expansion](#)

Step 1. Launch Aspire and 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 [Admin UI](#)

Before launching Aspire, you need to change the felix.properties file and add these lines if the Kerberos authentication is going to be used:

? Unknown Attachment

Step 2. Add a new HBase Content Source

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 **HBase Connector**.

? Unknown Attachment

Step 2a. Specify Basic Information


In the **General** tab in the *Content Source Configuration* window, specify basic information for the content source:


1. Name: Enter a content source name
 - 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**
 - 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**
 - 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**
 - This will be the type of crawl to execute for that specific schedule.


? Unknown Attachment

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.

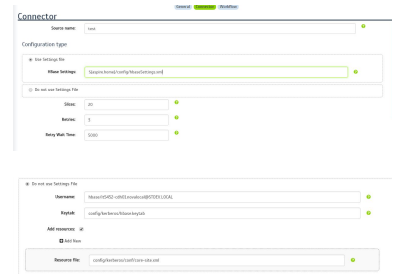
 If you want to disable the content source, clear the Enable check box. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

 Real Time and Cache Groups 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 HBase.

1. **Source Name:** Enter the source name to use for publishing the document.
2. **Namespace prefix:** Enter the prefix of the name space.
3. **Create Namespaces:** Select this option if the publisher should attempt to create the namespaces
4. **Configuration type**
 - a. **Use Settings file:**
 - i. HBase configuration file path: Enter the path that contains the HBase configuration file.
 - b. **Do not use Settings File**
 - i. **Username:** Kerberos User with the permissions to crawl from HBase.
 - ii. **Keytab:** Path to the Keytab file to use.
 - iii. **Add resources:** Check if you are going to pass Hadoop resources files (hbase-site.xml) to the publisher.
 1. **Resource file:** Hadoop resource file path.
 - iv. **Add properties:** Check if you are going to pass specific Hadoop properties to the publisher.
 1. **Name:** Hadoop property name.
 2. **Value:** Hadoop property value.
5. **Slices:** The number of slices to use for the crawl.
6. **Retries:** The number of times to retry a HBase scan call.
7. **Retry Wait Time:** The time in milliseconds to wait before each retry



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 to 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*.
2. Specify a *Name* and *Description* for the Publisher.
3. Click **Add**.

After completing these steps, click **Save** and **Done**. You'll be sent back to the *Home* page.

Step 2d. HBase settings file

If used, the HBase settings file has the following structure:

 Unknown Attachment

```
<settings>
  <properties>
    <property name="hbase.zookeeper.quorum">10.0.0.114</property>
  </properties>
  <configDir>config\kerberos\conf</configDir>
  <security>
    <kerberos >
      <user>hbase/it5452-cdh01.novalocal@STDEV.LOCAL</user>
      <path>config\kerberos\hbase.keytab</path>
    </kerberos>
  </security>
</settings>
```

1. **Properties:** Put any Hadoop property required
2. **Config Dir:** Path where the Hadoop Resources files are located
3. **User:** Kerberos User with the permissions to crawl from HBase
4. **Kerberos Keytab:** Path to the Keytab file to use

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 to **Full**. (**Incremental** is the default, and the first time it will 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:

- 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, 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 HBase (documents which are added, modified, or removed), then click **Incremental** instead of **Full**.

- The HBase 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.

Group Expansion

Group expansion configuration is done on the "Advanced Connector Properties" of the Connector tab.

1. Select the **Advanced Configuration** check box to enable the advanced properties section.
2. Select the **Group Expansion** check box.

3. Add a new source for each repository you want to expand groups from. (You'll need administrator rights on all of them to be able to do this.)
4. Set the default domain, user name and password of the crawl account.
5. Set a schedule for group expansion refresh and cleanup.
6. (Optional) Select the **Use external Group Expansion** check box to select an LDAP Cache component for LDAP group expansion.
See more info on the LDAP Cache component on [LDAP Cache](#).