# Getting Started Tutorial

## Welcome to the Getting Started Tutorial

- Install Aspire, an Aspire connector application, and an Aspire publisher application (that writes to a file instead of a search engine for indexing).
- Get an idea of how to install Aspire applications using the System Admin UI (no programming knowledge needed).

To use Aspire as a componentized pipeline processing system (without the Connector Framework), see the Aspire Framework Pages.



There are other tutorials for each type of Connector application you might want to install located under the section for each Connector.

Make sure that the Aspire version you download matches your tutorial version. The released binaries must match the training versions.

**On this page**


- Welcome to the Getting Started Tutorial
- Prerequisites
- Step 1: Download with Pre-Built Binaries
- Step 2: Start MongoDB
- Step 3: Edit the Aspire config/settings.xml File
- Step 4: Start Aspire
- Step 5: Launch the Aspire Content Source Manager
- Step 6: Add a New Content Source
- Step 7: Initiate the full crawl
- Step 8: Shut Down Aspire


## Prerequisites

1. Before you begin, you must register to use Aspire (go to How to Access Aspire for more details).

- You will need your user registration **name** and **password** in order to complete this tutorial.



Review the Prerequisites for Connectors and Content Processing.

2. After registering, you'll require a valid License file. See Aspire Licensing for more details.

3. You will need to install a crawl state database. MongoDB, Apache HBase, and Elasticsearch are supported. These software can hold the crawl state for the Aspire Connector framework. Please see their corresponding wiki pages for more details:

- MongoDB - Installing and Configuring a Crawl Status Database
- Apache HBase - Installing and Configuring a Crawl Status Database
- Elasticsearch - Installing and Configuring a Crawl Status Database



Aspire versions lower than 4.0 only allow for MongoDB as their crawl state database

## Step 1: Download with Pre-Built Binaries

You must be registered to access the binaries download page. Go to How to Access Aspire , to check the Aspire Registration process.

A pre-built distribution of Aspire can be downloaded from Aspire Binaries .


This distribution contains pre-built binaries for a functioning Aspire installation which works for both Windows and Linux.

After downloading Aspire, extract the contents of the zip or tar file into a directory.

## Step 2: Start MongoDB

1. If you haven't installed MongoDB, do it now: MongoDB - Installing and Configuring a Crawl Status Database
2. Make sure your MongoDB server is running by executing:

```
> mongo
MongoDB shell version: 4.0.4
connecting to: test
>
```

**Error**

If a **"NETWORK  Failed to connect to 127.0.0.1:27017"** error occurs, then it is not running.

Start it by executing:

***For Windows:***

Open a terminal and CD to your MongoDB installation folder and execute:

```
C:\Program Files\MongoDB\Server\4.0\bin> mongod
```

***For Linux:***

```
$ sudo service mongod start
```

# Step 3: Edit the Aspire config/settings.xml File

1. Go to the directory where you unpacked Aspire (such as "aspire-quick-start").
2. Go to the configuration directory **/config**.
3. Open the **settings.xml** file with a text or XML editor.
4. Look for the  clusterId tag.
5. You need to specify a cluster Id, we recommend to replace the **dev** in the settings.xml with your unique cluster id

```
<!-- By default all Aspire servers start in their own cluster. To make servers work together, set a
common
       cluster id across multiple instances that are connected to a common zooKeeper instance and
database
       provider (for example "dev" or "prod") -->
<clusterId>dev</clusterId>
```

If you encounter Parsing Error:  org.xml.sax.SAXParseException: Content is not allowed in prolog

This error is caused by the presence of a byte-order mark (BOM) at the start of your XML file. You can see the BOM using "od -c <filename>". Remove the BOM by copying and pasting the entire file contents to a new file in a text editor (since the BOM does not copy to the clipboard).

You may also need to edit the settings if you are running MongoDB on a separate machine.

- If this is the case, Modify the following section and point to the server where you installed MongoDB. In this case, MongoDB is running in the host called **mongodb-host**:

**MongoDB Settings**

```
<!-- noSql database provider for the 4.0 connector framework -->
<noSQLConnectionProvider sslEnabled="false" sslInvalidHostNameAllowed="false">
  <implementation>com.accenture.aspire:aspire-mongodb-provider</implementation>
  <servers>mongodb-host:27017</servers>
</noSQLConnectionProvider>
```

# Step 4: Start Aspire

First, make sure you have access to the internet so that Aspire can download components from our Maven repository.

To start Aspire, follow these steps:

1. Open a terminal and cd to the **bin/** directory inside the downloaded Aspire Distribution.
2. Execute the startup script

- *Windows*

```
> aspire.bat
```

***Linux***

```
$ aspire.sh
```

***OS X*** (this command must be executed from Aspire home directory)

```
$ bin/aspire.sh
```

**Error**

If you are unable to start Aspire and an error occurs, try accessing a URL via the browser and logging in with your registered username/password. For example:

- Can you get to https://repository.searchtechnologies.com/artifactory/public/com/accenture/aspire/aspire-application/4.0/aspire-application-4.0.jar in a browser?
  Enter the username/password that you entered inyour settings.xml file.
- Do you use a proxy?
  If so, modify settings.xml to include information as described in “Proxy Settings” on the General Settings page.

The "aspire" batch script (on Windows) or shell script (on Unix) can be modified as necessary if you want to assign more memory or need to set other JVM system properties.

**Example output:**

```
Removing Felix-Cache and AppBundle-Cache directories



********************************************************************************
*
* ASPIRE BOOTLOADER
*
* Bundle id : 10
*
* Location  : file:bundles/boot/aspire-bootloader-4.0.jar
*
.
.
.
.
2014-02-04T20:38:43Z INFO [/Workflow]: Installed component: /Workflow/WfManager
2014-02-04T20:38:43Z INFO [aspire]: Successfully started appBundle: /Workflow (location: com.accenture.
aspire:app-workflow-manager)
AUTOSTART: No applications to start
```

Aspire may take a few minutes to load all of the necessary components. You will see feedback to the command prompt during the startup.

3. Use the standard administration interface (http://localhost:50505) to install pre-packaged applications such as connectors and search engine publishers.

Go to Admin UI for more details.

### Errors

4. If you see an error like this:

```
2014-11-19T10:20:35Z INFO [BOOTLOADER]: Fetching: com.accenture.aspire:aspire-application:4.0
2014-11-19T10:20:43Z ERROR [BOOTLOADER]: Cannot get file from repository (https://repository.
searchtechnologies.com/artifactory/public/) - check the component is properly deployed. File:
https://repository.searchtechnologies.com/artifactory/public/com/accenture/aspire/aspire-application/4.0
/aspire-application-4.0.jar
org.apache.maven.wagon.authorization.AuthorizationException: Access denied to: https://repository.
searchtechnologies.com/artifactory/public/com/accenture/aspire/aspire-application/4.0/aspire-application-4.0.
jar
       at org.apache.maven.wagon.providers.http.LightweightHttpWagon.fillInputData(LightweightHttpWagon.java:
119)
       at org.apache.maven.wagon.StreamWagon.getInputStream(StreamWagon.java:116)
       at org.apache.maven.wagon.StreamWagon.getIfNewer(StreamWagon.java:88)
       at org.apache.maven.wagon.StreamWagon.get(StreamWagon.java:61)
       at org.sonatype.aether.connector.wagon.WagonRepositoryConnector$GetTask.run(WagonRepositoryConnector.
java:511)
       at java.util.concurrent.ThreadPoolExecutor$Worker.runTask(ThreadPoolExecutor.java:895)
       at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:918)
       at java.lang.Thread.run(Thread.java:662)
```

and Aspire exits:

```
ERROR: Failed to load Aspire application
Shutting down OSGI
```

5. Check that you have the correct username and password in the *config/settings.xml* file. See General Settings for details.

### More Information

For more information on how to start, stop, install as service, see Launch Control

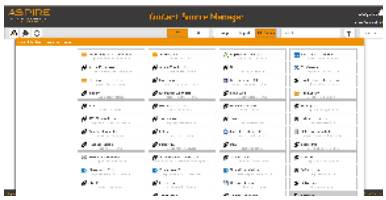
# Step 5: Launch the Aspire Content Source Manager

1. Browse to: http://localhost:50505 if you are running Aspire locally or change the host to the server hosting Aspire.

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

# Step 6: Add a New Content Source

## Step 6a: Add a New File System Content Source



We would like to crawl a specific folder in our file system. We need to create a "Content Source" using the "File System Connector".

To create a new content source:

1. From the Aspire Content Source Manager Home page, ***Add Source***.

2. Click ***File System Connector.***

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

## Step 6b: 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.
    - 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.

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 in the ***Add New*** option. Note that you can separately schedule full crawls, incremental crawls, and user-group cache (used for group expansion with document-level security) downloads.

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

## Step 6c: Specify the Connector Information

In the **Connector** tab, specify the connection information to crawl the File System folder.

1. Enter the folder path you want to crawl.
    - *For Windows*: Use the following format D:\folder\folder1\
    - *For Linux*: Use the following format /home/user/folder/folder1/
2. Check on the other options as needed:
    a. Index Containers?: Index containers as items. If unchecked, only files will be indexed.
    b. Scan Recursively?: Scan through subfolder's child nodes.
    c. Scan Excluded Item: It will scan sub items of container that have been excluded
    d. Include/Exclude patterns: Enter regex patterns to include or exclude files/folders based on URL matches.

## Step 6c: Specify the Connector Information

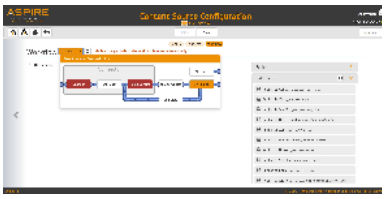
In the **Connector** tab, specify the connection information to crawl the File System folder.

1. Enter the folder path you want to crawl.
    - *For Windows*: Use the following format D:\folder\folder1\

- *For Linux*: Use the following format /home/user/folder/folder1/

2. Check on the other options as needed:
    a. Index Containers?: Index containers as items. If unchecked, only files will be indexed.
    b. Scan Recursively?: Scan through subfolder's child nodes.
    c. Scan Excluded Item: It will scan sub items of container that have been excluded
    d. Include/Exclude patterns: Enter regex patterns to include or exclude files/folders based on URL matches.

## Step 6d: 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 should an item follow after being crawled. These 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 **Publisher s** tab to the **onPublish** Workflow tree.
    a. Specify a *Name* and *Description* for the Publisher.
    b. Click ***Add***.

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

# Step 7: Initiate the full crawl

Now that the content source is set up, you will see a box (called a 'card') on the main page which shows your new content source. You can use the buttons in this box to initiate a crawl.

1. Click ***Full crawl*** (left most black circle) to initiate the crawl.

## During the Crawl

During the crawl, you can do the following:

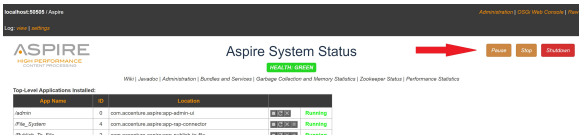
- Click ***Refresh*** on the **Content Sources** page to view the latest status of the crawl.
  The status will show **RUNNING** while the crawl is going, and **CRAWLED** when it is finished.
- Click ***Complete*** to view the number of documents crawled so far, the number of documents submitted, and the number of documents with errors.

**Important**: If there are errors, you will get a clickable **Error** that will take you to a detailed error message page.

## Checking the Output

Because you are publishing the results to a file, you can see the published jobs in the file you specified when installing the "Publish To File" component. Typically this will be in the "logs" directory, in the "*publis hToFile.jobs*" file (which can be usually be found under the "Publish_To_File" sub-directory).

# Step 8: Shut Down Aspire




Congratulations! You have completed the 20-minute quick start.

1. To shut down Aspire, go to the **Debug Console** page (http://localhost:50505/aspire).
2. Click ***Shutdown*** (the red button to the right of the server name).

Or, go to the Aspire console window (where you started Aspire with "bin\startup") and

1. Type **shutdown.**
2. Press ***Return*** or ***Enter***.