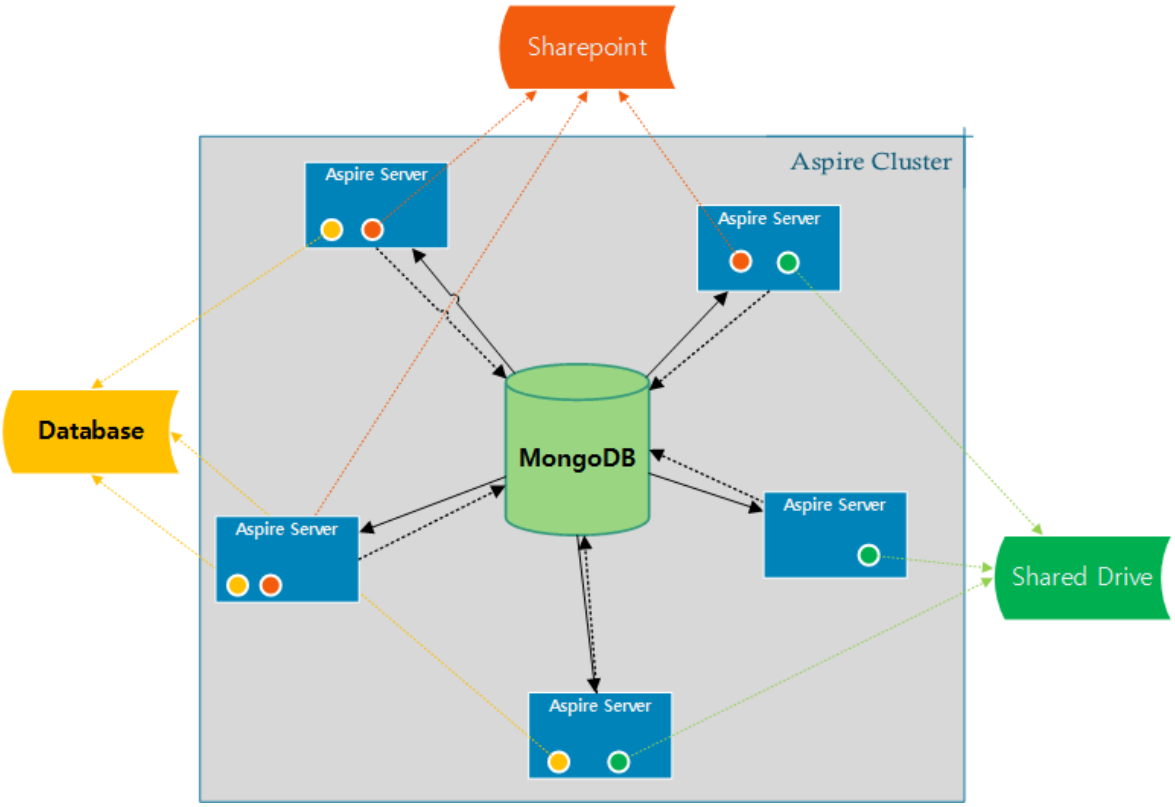
Distributed Processing

Overview

Since the 3.1 release, Aspire connectors are able to crawl in distributed mode automatically. Since all the crawl control data is stored in MongoDB, by just adding more Aspire servers configured to use the same MongoDB, the common connectors are going to crawl distributively.

Each connector is responsible for talking to the repositories, scanning through all the items to fetch and store its IDs to MongoDB for being processed later by any other server or itself.

On this page:



Configuration

In order to setup an Aspire Cluster for Distributed Processing, you need to do the following steps:

1. **Setup MongoDB**

   You need to configure all Aspire servers to use the same MongoDB Installation, configure all the Aspire Servers **config/settings.xml** file

MongoDB Settings

```
  <!-- noSql database provider for the 3.1 connector framework -->
  <noSQLConnectionProvider connectionsPerHost="10" sslEnabled="false" sslInvalidHostNameAllowed="
false">
    <implementation>com.searchtechnologies.aspire:aspire-mongodb-provider</implementation>
    <dropOnClear>false</dropOnClear>
    <servers>mongodb-host:27017</servers>
  </noSQLConnecitonProvider>
```

If you need to connect to a multi node MongoDB installation, check: Connect to a Multi-node MongoDB Installation

2. ## Setup Zookeeper

More details for Zookeeper installation and settings at Failover Settings (Zookeeper)

Failover Settings (Zookeeper)

For each Aspire Server make the following change to the <configAdministration> section of the **settings.xml** file

```
<zookeeper enabled="false" libraryFolder="config/workflow-libraries" root="/aspire" updatesEnabled="
false">
```

to

```
<zookeeper enabled="true" libraryFolder="config/workflow-libraries" root="/aspire" updatesEnabled="
true">
```

Uncomment the line with:

```
<!-- <externalServer>127.0.0.1:2182,127.0.0.1:2183,127.0.0.1:2181</externalServer> -->
```

And write the zookeeper server that you have installed as follows:

```
<externalServer> host:port </externalServer>
```

If you are using a cluster of zookeeper servers separate each server with a comma:

```
<externalServer> host1:port1, host2:port2, host3:port3, ... </externalServer>
```

By default if no external server is specified, Aspire will start an embedded ZooKeeper server on the port specified in the **<clientPort>** tag, and it will not be connected to any other ZooKeeper. This is the default for non-failover installations.

3. ## Install the content sources to distribute

Now it is time to think about which content sources you want to crawl distributively, and from what Aspire Servers, according to your solution architecture.

For this, configure the content sources in one of the servers and once you have them correctly configured **export** the content source and **import** it into the Aspire Servers you want to crawl this content source in parallel.

Crawl Control

Controlling distributed processing is very simple, all you need to know is that if you start the crawl from any of the Aspire Servers, the crawl will **start** from all the servers, the same applies if you **pause**, **stop** or **resume** a crawl.

> You are safe to shutdown all aspire servers when you do a Pause.

> If you need to shutdown all the servers in the cluster, it is mandatory to pause the crawl first.