Azure Blob Storage Connector For Developers

Azure Blob Storage

The Azure Blob Storage connector performs full and incremental scans over an Azure Blob Container and will extract security, metadata, and content from each object scanned. Each scanned object will be tagged with one of three possible actions: add, update, or delete, and can be routed to any Aspire pipeline as desired.

The connector, once started, can be stopped, paused or resumed via the Scheduler Component. Typically the start job will contain all information required by the job to perform the scan. When pausing or stopping, the connector will wait until all the jobs it published have completed before updating the statistics and status of the connector.

Azure Blob Storage Connector

AppBundle Name	Azure Blob Storage Connector
Factory Name	com.accenture.aspire:aspire-azureblob-source
Aspire Version	4.0
Inputs	AspireObject from a content source submitter holding all the information required for a crawl
Outputs	An AspireObject containing the URL, content, ACLs and Metadata processed for each file

Configuration

This section lists all configuration parameters available to install the Azure Blob Application Bundle and to execute crawls using the connector.

Property	Type	Default	Description
storageConnectionString	string	DefaultEndpointsProtocol=http;AccountName=myAccount;AccountKey=myKey;	Azure Blob Storage Credentials
useRootURL	boolean	True	Indicates if we should load the information from a configuration provided address
seSeedsFile	boolean	False	Indicates if we should load the information from a user provided text file
seedsFilePath	string	${dist.data.dir}/${app.name}/urls.txt	url for the seeds file path. The file must contain a valid azure blob container url per line. The credentials will be used in all urls
url	string		The address of the Blob Container that we want to crawl, it follows the format: http://hostname/tennant/blobContainer
useExtraBlobContainers	boolean	False	Indicates if we should use more than one container for the crawl
siteCollectionsToCrawl	collection		A list/collection of Azure Blob Container urls that we want to crawl
scanRecursively	boolean	True	Indicates if we want to scan subfolders

⚠ Note, if you are using the Azure Storage Emulator, then for the storageConnectionString property, you can use the next connection string "UseDevelopmentStorage=true;" this will connect to the emulator installed locally in your machine.

Configuration Example

To install the application bundle, add the configuration, as follows, to the <autoStart> section of the Aspire settings.xml.

Configuration

```
<application config="com.accenture.aspire:aspire-azureblob-source">
<properties>
<property name="storageConnectionString"> DefaultEndpointsProtocol=http;AccountName=myAccount;
AccountKey=myKey;</property>
<property name="useRootURL">true</property> <property name="useSeedsFile">false</property>
<property name="seedsFilePath"> ${dist.data.dir}/${app.name}/urls.txt </property>
<property name="url"> http://127.0.0.1:10000/devstoreaccount1/test/ </property>
<property name="useExtraBlobContainers">true</property> <property name="siteCollectionsToCrawl">
<property name=" siteCollectionUrl">http://127.0.0.1:10000/devstoreaccount1/defaultContainer/</property>
<property name="siteCollectionUrl">http://127.0.0.1:10000/devstoreaccount1/noSubs/</property> </property>
<property name=" scanRecursively ">true</property>
</properties>
</application>
```