SharePoint 2013 Scanner

he *SharePoint 2013 Scanner* component performs full and incremental scans over a SharePoint 2013 repository, maintaining the last SharePoint change token of the repository to get updates next time an incremental crawl is executed. Updated content is then submitted to the configured pipeline in AspireObjects attached to Jobs. As well as the URL of the changed item, the AspireObject will also contain metadata extracted from the repository. Updated content is split in to three types: add, update, and delete. Each type of content is published on a different event so that it may be handled by different Aspire pipelines.

The scanner reacts to an incoming job. This job may instruct the scanner to *start*, *stop*, *pause*, *resume* or *cacheGroups*. Typically the *start* job will contain all information required by the job to perform the crawl. However, the scanner can be configured with default values via application.xml file. When pausing or stopping, the scanner will wait until all the jobs it published have completed before itself completing.

SharePoint 2013 Scanner	
Factory Name	com.searchtechnologies.aspire:aspire-sharepoint2013-connector
subType	default
Inputs	AspireObject from a content source submitter holding all the information required for a crawl
Outputs	Jobs from the crawl

Configuration

This section lists all configuration parameters available to configure the SharePoint 2013 Scanner component.

General Scanner Component Configuration

Basic Configuration

Element	Type	Default	Description
snapshotDir	String	snapshots	The directory for snapshot files.
numOfSnapshotBackups	int	2	The number of snapshots to keep after processing.
waitForSubJobsTimeout	long	600000 (=10 mins)	Scanner timeout while waiting for published jobs to complete.
maxOutstandingTimeStatistics	long	1m	The max about of time to wait before updating the statistics file. Whichever happens first between this property and maxOutstandingUpdatesStatistics will trigger an update to the statistics file.
maxOutstandingUpdatesStatistics	long	1000	The max number of files to process before updating the statistics file. Whichever happens first between this property and maxOutstandingTimeStatistics will trigger an update to the statistics file.
usesDomain	boolean	true	Indicates if the group expansion request will use a domain\user format (useful for connectors that does not support domain in the group expander).

Branch Handler Configuration

This component publishes to the *onAdd*, *onDelete* and *onUpdate*, so a branch must be configured for each of these three events.

Element	Type	Description
branches/branch/@event	string	The event to configure - *onAdd*, *onDelete* or *onUpdate*.
branches/branch/@pipelineManager	string	The name of the pipeline manager to publish to. Can be relative.
branches/branch/@pipeline	string	The name of the pipeline to publish to. If missing, publishes to the default pipeline for the pipeline manager.
branches/branch/@allowRemote	boolean	Indicates if this pipeline can be found on remote servers (see Distributed Processing for details).
branches/branch/@batching	boolean	Indicates if the jobs processed by this pipeline should be marked for batch processing (useful for publishers or other components that support batch processing).
branches/branch/@batchSize	int	The max size of the batches that the branch handler will created.
branches/branch/@batchTimeout	long	Time to wait before the batch is closed if the batchSize hasn't been reached.

branches/branch /@simultaneousBatches	int	The max number of simultanous batches that will be handled by the branch handler.

SharePoint Scanner Configuration

Element	Type	Default	Description
userName	String	username	The user name to connect to SharePoint with, if one is not given in the control job.
password	String	secretpassword	The password to connect to SharePoint with, if one is not given in the control job.
domain	string		Domain used to authenticate against SharePoint.
defaultDisplayName	String	SharePoint2013	The *name* of the crawl, if one is not given in the control job.
groupPrefixSeparator	String	|	The separator inserted between the site URL and group name when extracting groups from sites.
snapshotDir	String	.	The directory for snapshot files.
waitForSubJobsTimeout	long	600000 (=10 mins)	Scanner time out while waiting for published jobs to complete.
scanRecursively	boolean	false	Indicates whether the child containers should be scanned or not.
indexContainers	boolean	false	Indicates whether the container items should be indexed or not.
crawlAttachments	boolean	false	Crawl attachments from list items. E.g. documents attached to an Event.
crawlExtraSiteCollections	boolean	false	Indicates if the user will crawl more than one site collection.
subSiteCollections/siteCollectionUrl	string	empty	List of sub site collections to crawl. More than one allowed.
useLDAPCache	boolean	false	Check for an installed "Aspire LDAP Cache" component for group expansion.
externalGroupServerPath	string	empty	List of installed "Aspire LDAP Cache" components.

Example Configuration

```
 <component name="Scanner" subType="default" factoryName="aspire-sharepoint2013-connector">
  <debug>${debug}</debug>
  <groupPrefixSeparator>${groupPrefixSeparator}</groupPrefixSeparator>
  <snapshotDir>${snapshotDir}</snapshotDir>
  <scanRecursively>${scanRecursively}</scanRecursively>
  <indexContainers>${indexContainers}</indexContainers>
  <crawlAttachments>${crawlAttachments}</crawlAttachments>
  <useLDAPCache>${useLDAPCache}</useLDAPCache>
  <externalGroupServerPath>${externalGroupServerPath}</externalGroupServerPath>
  <crawlExtraSiteCollections>${crawlExtraSiteCollections}</crawlExtraSiteCollections>
  <subSiteCollections>
    <siteCollectionUrl>${siteCollectionUrl}</siteCollectionUrl>
  </subSiteCollections>
  <branches>
    <branch event="onAdd" pipelineManager="../ProcessPipelineManager"
     pipeline="addUpdatePipeline" allowRemote="true" batching="true"
     batchSize="50" batchTimeout="60000" simultaneousBatches="2" />
    <branch event="onUpdate" pipelineManager="../ProcessPipelineManager"
     pipeline="addUpdatePipeline" allowRemote="true" batching="true"
     batchSize="50" batchTimeout="60000" simultaneousBatches="2" />
    <branch event="onDelete" pipelineManager="../ProcessPipelineManager"
     pipeline="deletePipeline" allowRemote="true" batching="true"
     batchSize="50" batchTimeout="60000" simultaneousBatches="2" />
  </branches>
</component>
```

Source Configuration

Scanner Control Configuration

The following table describes the list of attributes that the AspireObject of the incoming scanner job requires to correctly execute and control the flow of a scan process.

Element	Type	Options	Description
@action	string	start, stop, pause, resume, abort	Control command to tell the scanner which operation to perform. Use **start** option to launch a new crawl.
@actionProperties	string	full, incremental	When a **start** @action is received, it will tell the scanner to either run a **full** or an **incremental** crawl.
@normalizedCSName	string		Unique identifier name for the content source that will be crawled.
displayName	string		Display or friendly name for the content source that will be crawled.

Header Example

```
  <doc action="start" actionProperties="full" actionType="manual" crawlId="0" dbId="0" jobNumber="0"
normalizedCSName="FeedOne_Connector"
   scheduleId="0" scheduler="##AspireSystemScheduler##" sourceName="ContentSourceName">
    ...
    <displayName>testSource</displayName>
    ...
  </doc>
```

All configuration properties described in this section are relative to /doc/connectorSource of the AspireObject of the incoming Job.

Element	Type	Default	Description
url	string		The URL to scan (allowed http or https).
username	string		The username to connect to SharePoint with.
password	string		The password to connect to SharePoint with.
domain	string		Domain used to authenticate against SharePoint.
indexContainers	boolean	false	*true* if folders (as well as files) should be indexed.
scanRecursively	boolean	false	*true* if subfolders of the given URL should be scanned.
indexContainers	boolean	false	Indicates whether the container items should be indexed or not.
crawlAttachments	boolean	false	Crawl attachments from list items. E.g. documents attached to an Event.
crawlExtraSiteCollections	boolean	false	Indicates if the user will crawl more than one site collection.
subSiteCollections /siteCollectionUrl	string	empty	List of sub site collections to crawl. More than one allowed.
fileNamePatterns /include/@pattern	regex	none	Optional. A regular expression pattern to evaluate file urls against; if the file name matches the pattern, the file is included by the scanner. Multiple include nodes can be added.
fileNamePatterns /include/@pattern	regex	none	Optional. A regular expression pattern to evaluate file urls against; if the file name matches the pattern, the file is excluded by the scanner. Multiple exclude nodes can be added.

Content Source Configuration Example


```
<doc action="start" actionProperties="full" actionType="manual" crawlId="0" dbId="2" jobNumber="5"
normalizedCSName="SharePoint2013" scheduleId="2" scheduler="AspireScheduler" sourceName="SharePoint2013">
  <connectorSource>
    <url>http://10.10.21.127/sites/aspire</url>
    <crawlExtraSiteCollections>true</crawlExtraSiteCollections>
    <subSiteCollections>
      <siteCollectionUrl>http://10.10.21.127/sites/mysite</siteCollectionUrl>
    </subSiteCollections>
    <domain>qa</domain>
    <username>sp_farm</username>
    <password>encrypted:562E81591F85B858E5A5D3876F9C9FDB</password>
    <scanRecursively>true</scanRecursively>
    <indexContainers>true</indexContainers>
    <crawlAttachments>true</crawlAttachments>
    <fileNamePatterns/>
  </connectorSource>
  <displayName>SharePoint2013</displayName>
</doc>
```


Output


```
<doc>
  <url>http://10.10.21.127/sites/aspire/_api/Web</url>
  <snapshotUrl>001 http://10.10.21.127/sites/aspire/_api/Web</snapshotUrl>
  <repItemType>aspire/sharePoint</repItemType>
  <docType>container</docType>
  <GUID>0f7bc97f-ac37-40b7-89be-5c009d79173b</GUID>
  <description/>
  <title>Aspire</title>
  <lastModified>2014-08-22T15:30:01Z</lastModified>
  <dateCreated>2014-01-15T23:38:39Z</dateCreated>
  <dataSize>0</dataSize>
  <displayUrl>http://10.10.21.127/sites/aspire</displayUrl>
  <id>http://10.10.21.127/sites/aspire/_api/Web</id>
  <fetchUrl>http://10.10.21.127/sites/aspire</fetchUrl>
  <sourceName>SharePoint2013</sourceName>
  <sourceType>sp2013</sourceType>
  <connectorSpecific type="sp2013">
    <field name="AllowRssFeeds">true</field>
    <field name="AppInstanceId">00000000-0000-0000-0000-000000000000</field>
    <field name="Configuration">0</field>
    <field name="Created">2014-01-15T23:38:39</field>
    <field name="CustomMasterUrl">/sites/aspire/_catalogs/masterpage/seattle.master</field>
    <field name="DocumentLibraryCalloutOfficeWebAppPreviewersDisabled">false</field>
    <field name="EnableMinimalDownload">true</field>
    <field name="Id">0f7bc97f-ac37-40b7-89be-5c009d79173b</field>
    <field name="Language">1033</field>
    <field name="LastItemModifiedDate">2014-08-22T15:30:01Z</field>
    <field name="MasterUrl">/sites/aspire/_catalogs/masterpage/seattle.master</field>
    <field name="QuickLaunchEnabled">true</field>
    <field name="RecycleBinEnabled">true</field>
    <field name="ServerRelativeUrl">/sites/aspire</field>
    <field name="SyndicationEnabled">true</field>
    <field name="Title">Aspire</field>
    <field name="TreeViewEnabled">false</field>
    <field name="UIVersion">15</field>
    <field name="UIVersionConfigurationEnabled">false</field>
    <field name="Url">http://10.10.21.127/sites/aspire</field>
    <field name="WebTemplate">STS</field>
  </connectorSpecific>
  <acls>
```

```
    <acl Permissions="Read, Limited Access, " Sid="s-1-5-21-3023650700-3092893521-1383129343-1112" access="
allow" domain="qa" entity="user" fullname="qa\spadmin" name="spadmin" scope="global"/>
    <acl Permissions="Full Control, " access="allow" domain="" entity="group" fullname="
902A40CF40D9CD503EE3199DA5D7F113|Aspire Owners" name="Aspire Owners" scope="machine"/>
    <acl Permissions="Read, Limited Access, " Sid="s-1-5-21-3023650700-3092893521-1383129343-1107" access="
allow" domain="qa" entity="user" fullname="qa\sp_farm" name="sp_farm" scope="global"/>
    <acl Permissions="Edit, " access="allow" domain="" entity="group" fullname="
902A40CF40D9CD503EE3199DA5D7F113|Aspire Members" name="Aspire Members" scope="machine"/>
    <acl Permissions="Read, " access="allow" domain="" entity="group" fullname="
902A40CF40D9CD503EE3199DA5D7F113|Aspire Visitors" name="Aspire Visitors" scope="machine"/>
    <acl Permissions="View Only, " access="deny" domain="" entity="group" fullname="
902A40CF40D9CD503EE3199DA5D7F113|Excel Services Viewers" name="Excel Services Viewers" scope="machine"/>
    <acl Permissions="Limited Access, " Sid="S-1-0-0" access="deny" domain="" entity="user" fullname="
SHAREPOINT\system" name="System Account" scope="global"/>
    <acl Permissions="Limited Access, View Only, " access="deny" domain="" entity="group" fullname="
Everyone" name="Everyone" scope="global"/>
  </acls>
  <hierarchy>
    <item id="73886DD2E0982FC87918CD9006F4456A" level="1" name="Aspire" type="aspire/sharePoint" url="
http://10.10.21.127/sites/aspire"/>
  </hierarchy>
  <connectorSource>
    <url>http://10.10.21.127/sites/aspire</url>
    <crawlExtraSiteCollections>true</crawlExtraSiteCollections>
    <subSiteCollections>
      <siteCollectionUrl>http://10.10.21.127/sites/mysite</siteCollectionUrl>
    </subSiteCollections>
    <domain>qa</domain>
    <username>sp_farm</username>
    <password>encrypted:562E81591F85B858E5A5D3876F9C9FDB</password>
    <scanRecursively>true</scanRecursively>
    <indexContainers>true</indexContainers>
    <crawlAttachments>true</crawlAttachments>
    <fileNamePatterns/>
    <displayName>SharePoint2013</displayName>
  </connectorSource>
  <action>add</action>
  <content/>
</doc>
```