File System Scanner

The *File System Scanner* component performs full and incremental scans over a file system folder, maintaining a snapshot of the filesystem and comparing it with the current content to establish what content has been updated. 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 into three types -add, update and delete-. Each type of content is published as 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* or *resume*. 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 completing itself.

File System Scanner	
Factory Name	com.searchtechnologies.aspire:aspire-filesystem-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 File System Scanner component.

General Scanner Component Configuration

Basic Scanner 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.

File System Specific Configuration

Element	Type	Default	Description
maxBytes	long	unlimited	The maximum file size in bytes. Files whose size is greater than this parameter will not be sent to the pipeline.

Configuration Example

```
  <component name="Scanner" subType="default" factoryName="aspire-filesystem-connector">
    <debug>true</debug>
    <snapshotDir>${aspire.home}/data/snapshots</snapshotDir>
    <fileNamePatterns>
      <include pattern=".*" />
      <exclude pattern=".*tmp$" />
    </fileNamePatterns>
    <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 file URL to crawl. Use the default Windows/Linux file path formats depending on the platform you are running on. I.e. Linux: /home/user/folder1/Windows: C:\folder1\
partialScan	boolean	false	To run a partial scan – i.e. to only scan a portion of the larger directory. This is useful to re-process portions of your system without having to process the entire content source.
subDirUrl	string		Configurable when partialScan is set to *true*. The sub-directory which contains the documents to be processed for this partial scan. This directory must be a relative path to the parent directory. Only the documents in this sub-directory will be scanned. This is useful to re-process portions of your system without having to process the entire content source. For Windows use \ as folder separator, for linux use /.
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.

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.
acl/user	string	none	Optional. A list of users that can be added as fixed ACLs. See valid attributes below.
acl/user/@domain	string	none	Domain of the user added.
acl/user/@name	string	none	Name of the user added.
acl/user/@type	string	none	Type of the user acl added: **allow** or **deny** are the valid types.
acl/group	string	none	Optional. A list of groups that can be added as fixed ACLs. See valid attributes below.
acl/group/@name	string	none	Name of the group added.
acl/group/@type	string	none	Type of the group acl added: **allow** or **deny** are the valid types.
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.

Scanner Configuration Example

```
<doc action="start" actionProperties="full" normalizedCSName="testFile" scheduleId="1">
  <connectorSource>
    <url>D:\AspireTesting\</url>
    <partialScan>true</partialScan>
    <subDirUrl>LSA</subDirUrl>
    <indexContainers>true</indexContainers>
    <scanRecursively>true</scanRecursively>
    <useACLs>true</useACLs>
    <acl>
      <user domain="search" name="user1">
        <type>allow</type>
      </user>
      <user domain="search" name="user2">
        <type>deny</type>
      </user>
      <group name="wikiusers">
        <type>allow</type>
      </group>
    </acl>
    <fileNamePatterns>
      <include pattern=".*LSA.*"/>
      <exclude pattern=".*\.bak$"/>
    </fileNamePatterns>
  </connectorSource>
  <displayName>testFile</displayName>
</doc>
```

Output

```
<doc>
  <url>D:\AspireTesting\LSA\Videos youtube.txt</url>
  <snapshotUrl>003 D:\AspireTesting\LSA\Videos youtube.txt</snapshotUrl>
  <docType>item</docType>
  <repItemType>aspire/file</repItemType>
  <fetchUrl>file:/D:/AspireTesting/LSA/Videos%20youtube.txt</fetchUrl>
  <displayUrl>D:\AspireTesting\LSA\Videos youtube.txt</displayUrl>
  <id>D:\AspireTesting\LSA\Videos youtube.txt</id>
  <lastModified>2012-07-25T05:57:30Z</lastModified>
  <dataSize>111</dataSize>
  <sourceName>testFile</sourceName>
  <sourceType>filesystem</sourceType>
  <acls>
    <acl access="allow" domain="search" entity="user" fullname="search\user1" name="user1" scope="global"/>
    <acl access="allow" entity="group" fullname="wikiusers" name="wikiusers" scope="global"/>
    <acl access="deny" domain="search" entity="user" fullname="search\user2" name="user2" scope="global"/>
  </acls>
  <connectorSource>
    <url>D:\AspireTesting\</url>
    <partialScan>true</partialScan>
    <subDirUrl>LSA</subDirUrl>
    <indexContainers>true</indexContainers>
    <scanRecursively>true</scanRecursively>
    <useACLs>true</useACLs>
    <acl>
      <user domain="search" name="user1">
        <type>allow</type>
      </user>
      <user domain="search" name="user2">
        <type>deny</type>
      </user>
      <group name="wikiusers">
        <type>allow</type>
      </group>
    </acl>
    <fileNamePatterns>
      <include pattern=".*LSA.*"/>
      <exclude pattern=".*\.bak$"/>
    </fileNamePatterns>
    <displayName>testFile</displayName>
    <partialScanUrl>LSA</partialScanUrl>
  </connectorSource>
  <action>add</action>
  <hierarchy>
    <item id="4C1AB83A0DB23C7B3F1022F6FB2CBA86" level="3" name="Videos youtube.txt" url="D:
\AspireTesting\LSA\Videos youtube.txt">
      <ancestors>
        <ancestor id="00BE2B03F63AB87312C149D16263A6AB" level="2" name="LSA\" parent="true" type="aspire
/folder" url="D:\AspireTesting\LSA\"/>
        <ancestor id="CC00EC8FA97894C732CF72EF719D703E" level="1" name="testFile" type="aspire/filesystem"
url="D:\AspireTesting\"/>
      </ancestors>
    </item>
  </hierarchy>
</doc>
```