Rest API - Connectors Configuration

Each Seed entity requires a reference to a Connector in order to be created. This page details how to create a Connector using the Rest API

Create Connector

Field	Required	Default	Multiple	Notes	Example
type	Yes	-	No	The value must be the same as the type of the seed that will use this connector.	"filesystem"
description	Yes	-	No	Name of the connector object.	"MyFileSystemConnector"
artifact	Yes	-	No	The mvn coordinates of the connector.	"com.accenture.aspire:aspire-filesystem-source"
properties	**Yes**	**-**	**No**	**Configuration object**	
debug	No	false	No	Set to true to enable the debug messages.	true / false
wDebug	No	false	No	Set to true to enable job logging.	true / false
enableStatistics	No	true	No	Set to true to gather pipeline job statistics in the debug console.	true / false
infoCacheSize	Yes	-	No	The size of the Source Info cache used by the connector.	200
mapCacheSize	No	100	No	The number of Storage maps kept in memory per seed.	200
setCacheSize	No	100	No	The number of Sets kept in memory per seed.	200
identityCacheSize	No	100	No	The number of identities kept in memory per seed.	200
enableFetcher	Yes	-	No	Set to true to enable document fetching for the seeds that use this connector.	true / false
enableTextExtract	Yes	-	No	Set to true to enable text extraction. By default, connectors use Apache Tika to extract text from downloaded documents. To apply special text processing to a downloaded document in the workflow, disable text extraction. The downloaded document is then available as a content stream.	true / false
extractTextConfiguration	No	false	No	Set to true to override default text extraction settings.	true / false
extractTextMaxSize	No	20971520	No	Maximum extract text size in number of characters or \"unlimited\". Doesn't apply if HTML Output option is enabled.	10000
extractTimeout	No	180000	No	Maximum time (in ms) to wait for the text extraction.	18000
xmlMaxDepth	No	2147483647	No	The max depth level for a file inner structure. Can be used to block denial of service attacks or corrupted files.	2147483647
structuredText	No	false	No	Set to true to include formatting in output (in HTML) instead of plain text.	true / false
tikaConfig	No	-	No	Path for Apache Tika configuration file. It can be passed as empty to use the default configuration.	"/path/to/tikaConfig.xml" / ""
pdfParserProperties	No	false	No	Set to true to enable changing the default PDFBox properties.	true / false
enableAutoSpace	No	true	No	If set to true, the parser should estimate where spaces should be inserted between words. For many PDFs this is necessary as they do not include explicit whitespace characters.	true / false
suppressDuplicateOverlappingText	No	false	No	If set to true the parser should try to remove duplicated text over the same region. This is needed for some PDFs that achieve bolding by re-writing the same text in the same area. Note that this can slow down extraction substantially (PDFBOX-956) and sometimes remove characters that were not in fact duplicated (PDFBOX-1155).	true / false
extractAnnotationText	No	true	No	If set to true, text in annotations will be extracted.	true / false
sortByPosition	No	false	No	If set to true, sort text tokens by their x/y position before extracting text. This may be necessary for some PDFs (if the text tokens are not rendered \"in order\"), while for other PDFs it can produce the wrong result (for example if there are 2 columns, the text will be interleaved).	true / false

extractAcroFormContent	No	true	No	If set to true, extract content from AcroForms at the end of the document.	true / false
extractInlineImages	No	false	No	If set to true, extract inline embedded OBXImages. Beware: some PDF documents of modest size (~4MB) can contain thousands of embedded images totaling > 2.5 GB. Also, at least as of PDFBox 1.8.5, there can be surprisingly large memory consumption and/or out of memory errors. Set to true with caution.	true / false
extractUniqueInlineImagesOnly	No	true	No	Multiple pages within a PDF file might refer to the same underlying image. If extractUniqueInlineImagesOnly is set to false, the parser will call the EmbeddedExtractor each time the image appears on a page. This might be desired for some use cases. However, to avoid duplication of extracted images, set this to true.	true / false
enable-non-text-filter	No	false	No	Set to true to filter non text documents.	true / false
enableFetchForNonText	No	true	No	Set to true if the workflow needs to stream the non-text documents.	true / false
non-text-document	No	false	No	Set to true to filter using document extensions. Set to false to use a file to match non-text documents.	true / false
nonTextDocumentsExtensions	No	-	No	Comma separated list of non-text document extensions. Used based on the non-text-document value.	"jpg,jpeg,gif,png"
nonTextDocuments	No	-	No	Path to a file containing a list of regex that matches the non-text documents, one regex expression per line. Used based on the non-text-document value.	"config/nonTextDocuments.txt"
metadataMap	**No**	**[ ]**	**Yes**	**Settings for mapping extracted fields to a destination field.**	
from	No	-	No	Field to be mapped.	"fieldA"
to	No	-	No	Field where the value will be mapped.	"fieldB"
addHierarchy	Yes	-	No	Set to true to add hierarchy information to the documents.	true / false
hierarchyCacheSize	No	500	No	The hierarchy in memory cache size. Reducing this value may increase the number of requests to the NoSQL database.	5000
scanThreads	Yes	-	No	The maximum number of threads that will scan the repository at any one time.	10
scanQueue	Yes	-	No	The size of the in memory queue for items that need scanning in the repository. The recommended queue size is at least as large as the number of threads, if not two to three times larger. Larger queue sizes allow database access to be performed farther in advance, and smooth fluctuations in the time it takes to claim items from NoSql.	50
processThreads	Yes	-	No	The maximum number of threads that will process items from the repository at any one time.	20
processQueue	Yes	-	No	The size of the in memory queue for items that need to be processed. The recommended queue size is at least as large as the number of threads, if not two to three times larger. Larger queue sizes will allow database access to be performed further in advance and smooth fluctuations in the time it takes to claim items from NoSql.	200
deleteCompleteQueueEntries	No	false	No	Set to true if completed queue entries should be deleted (or just marked as complete).	true / false
flushSyncTime	No	"30ms"	No	Time to wait for all servers to finish their flushes to the snapshot at the end of each incremental crawl.	"15s"
deleteCheckAfterErrors	Yes	-	No	Checks if "delete" candidates still exist after an incremental when they are part of scan error.	"ALWAYS" / "NEVER"
maxIdentitiesTimestamp	Yes	-	No	Number of crawls to execute before removing the oldest identity items.	3
workflowErrorTolerant	No	false	No	If set to true, exceptions in workflow rules will only affect the execution of the rule in which the exception occurs. Subsequent rules will be executed and the job will complete the workflow successfully. Otherwise, exceptions in workflow rules will be re-thrown and the job will be moved to the error workflow.	true / false
retriesEnabled	No	true	No	If set to true, failed documents will be reprocessed at the end of the crawl and in the following incremental crawls.	true / false
removeFailedFromSnapshot	No	false	No	Check to remove the snapshot entry for each failed document. This makes the retries to be performed on all next incremental crawls. This overrides the \"Maximum crawls to retry\" option.	true / false
useRetryPattern	Yes	false	No	Whether the failed nodes processing should check the exception for regex pattern matches.	true/false
retryPatterns	No	[ ]	Yes	A regex pattern to match against any exception raised by either the individual document or publisher. If matched, the document or documents will be retried using the limits configured below.	[".*\\.pdf"]
maxInCrawlRetries	No	3	No	Maximum number of retries per crawl for a failed document.	3

maxCrawls	No	3	No	Maximum number of incremental crawls in which a failed document will be retried.	3

Example

POST aspire/_api/connectors

```
{
    "type": "filesystem",
    "description": "Test Description",
    "artifact": "com.accenture.aspire:aspire-filesystem-source",
    "properties": {
        "debug": false,
        "wDebug": false,
        "enableStatistics": false,
        "infoCacheSize": 100,
        "mapCacheSize": 100,
        "setCacheSize": 100,
        "identityCacheSize": 100,
        "enableFetcher": true,
        "enableTextExtract": true,
        "extractTextConfiguration": true,
        "extractTextMaxSize": "20971520",
        "extractTimeout": 180000,
        "xmlMaxDepth": 100,
        "structuredText": false,
        "tikaConfig": "",
        "pdfParserProperties": true,
        "enableAutoSpace": true,
        "suppressDuplicateOverlappingText": false,
        "extractAnnotationText": true,
        "sortByPosition": false,
        "extractAcroFormContent": true,
        "extractInlineImages": false,
        "extractUniqueInlineImagesOnly": true,
        "enable-non-text-filter": true,
        "enableFetchForNonText": true,
        "non-text-document": true,
        "nonTextDocumentsExtensions": "jpg,jpeg,gif,png,tif,mp3,mp4,mpg,mpeg,avi,mkv,wav,bmp,swf,war,rar,tgz,
dll,exe,class",
        "metadataMap": [{
                "from": "fieldA",
                "to": "destA"
            }, {
                "from": "fieldB",
                "to": "destB"
            }
        ],
        "addHierarchy": true,
        "hierarchyCacheSize": 5000,
        "scanThreads": 10,
        "scanQueue": 50,
        "processThreads": 20,
        "processQueue": 200,
        "deleteCompleteQueueEntries": false,
        "flushSyncTime": "15s",
        "deleteCheckAfterErrors": "ALWAYS",
        "maxIdentitiesTimestamp": 3,
        "workflowErrorTolerant": true,
        "retriesEnabled": true,
        "removeFailedFromSnapshot": true,
        "retryPattern": [".*tika.*", ".*png.*"],
        "maxInCrawlRetries": 3,
        "maxCrawls": 3
    }
}
```

Update Connector

Field	Required	Default	Multiple	Notes	Example
id	Yes	-	No	Id of the connector to update.	"e3fc3f4b-2784-4e5a-b27e-87a8f9a726a9"
type	No	-	No	The value must be the same as the type of the seed that will use this connector.	"filesystem"
description	No	-	No	Name of the connector object.	"MyFileSystemConnector"
artifact	No	-	No	The mvn coordinates of the connector.	"com.accenture.aspire:aspire-filesystem-source"
properties	**No**	**-**	**No**	**Configuration object**	
debug	No	false	No	Set to true to enable the debug messages.	true / false
wDebug	No	false	No	Set to true to enable job logging.	true / false
enableStatistics	No	false	No	Set to true to gather pipeline job statistics in the debug console.	true / false
infoCacheSize	No	100	No	The size of the Source Info cache used by the connector.	200
mapCacheSize	No	100	No	The number of Storage maps kept in memory per seed.	200
setCacheSize	No	100	No	The number of Sets kept in memory per seed.	200
identityCacheSize	No	100	No	The number of identities kept in memory per seed.	200
enableFetcher	No	true	No	Set to true to enable document fetching for the seeds that use this connector.	true / false
enableTextExtract	No	true	No	Set to true to enable text extraction. By default, connectors use Apache Tika to extract text from downloaded documents. To apply special text processing to a downloaded document in the workflow, disable text extraction. The downloaded document is then available as a content stream.	true / false
extractTextMaxSize	No	20971520	No	Set to true to override default text extraction settings.	10000
extractTimeout	No	180000	No	Maximum extract text size in number of characters or \"unlimited\". Doesn't apply if HTML Output option is enabled.	18000
xmlMaxDepth	No	2147483647	No	The max depth level for a file inner structure. Can be used to block denial of service attacks or corrupted files.	2147483647
structuredText	No	false	No	Include formatting in output (in HTML) instead of plain text.	true / false
tikaConfig	No	-	No	Path for Apache Tika configuration file. It can be passed as empty to use the default configuration.	"/path/to/tikaConfig.xml" / ""
pdfParserProperties	No	false	No	Set to true to enable changing the default PDFBox properties.	true / false
enableAutoSpace	No	true	No	If set to true, the parser should estimate where spaces should be inserted between words. For many PDFs this is necessary as they do not include explicit whitespace characters.	true / false
suppressDuplicateOverlappingText	No	false	No	If set to true the parser should try to remove duplicated text over the same region. This is needed for some PDFs that achieve bolding by re-writing the same text in the same area. Note that this can slow down extraction substantially (PDFBOX-956) and sometimes remove characters that were not in fact duplicated (PDFBOX-1155).	true / false
extractAnnotationText	No	true	No	If set to true, text in annotations will be extracted.	true / false
sortByPosition	No	false	No	If set to true, sort text tokens by their x/y position before extracting text. This may be necessary for some PDFs (if the text tokens are not rendered \"in order\"), while for other PDFs it can produce the wrong result (for example if there are 2 columns, the text will be interleaved).	true / false
extractAcroFormContent	No	true	No	If set to true, extract content from AcroForms at the end of the document.	true / false

extractInlineImages	No	false	No	If set to true, extract inline embedded OBXImages. Beware: some PDF documents of modest size (~4MB) can contain thousands of embedded images totaling > 2.5 GB. Also, at least as of PDFBox 1.8.5, there can be surprisingly large memory consumption and/or out of memory errors. Set to true with caution.	true / false
extractUniqueInlineImagesOnly	No	true	No	Multiple pages within a PDF file might refer to the same underlying image. If extractUniqueInlineImagesOnly is set to false, the parser will call the EmbeddedExtractor each time the image appears on a page. This might be desired for some use cases. However, to avoid duplication of extracted images, set this to true.	true / false
enable-non-text-filter	No	false	No	Set to true to filter non text documents.	true / false
enableFetchForNonText	No	true	No	Set to true if the workflow needs to stream the non-text documents.	true / false
non-text-document	No	false	No	Set to true to filter using document extensions. Set to false to use a file to match non-text documents.	true / false
nonTextDocumentsExtensions	No	jpg,jpeg, gif,png,tif, mp3,mp4, mpg, mpeg, avi,mkv,w av,bmp, swf, war,rar, tgz,dll, exe,class	No	Comma separated list of non-text document extensions. Used based on the non-text-document value.	"jpg,jpeg, gif,png"
nonTextDocuments	No	-	No	Path to a file containing a list of regex that matches the non-text documents, one regex expression per line. Used based on the non-text-document value.	"config /nonTextDocuments. txt"
metadataMap	**No**	**[ ]**	**Yes**	**Settings for mapping extracted fields to a destination field.**	
from	No	-	No	Field to be mapped.	"fieldA"
to	No	-	No	Field where the value will be mapped.	"fieldB"
addHierarchy	No	true	No	Set to true to add hierarchy information to the documents.	true / false
hierarchyCacheSize	No	500	No	The hierarchy in memory cache size. Reducing this value may increase the number of requests to the NoSQL database.	5000
scanThreads	No	10	No	The maximum number of threads that will scan the repository at any one time.	10
scanQueue	No	50	No	The size of the in memory queue for items that need scanning in the repository. The recommended queue size is at least as large as the number of threads, if not two to three times larger. Larger queue sizes allow database access to be performed farther in advance, and smooth fluctuations in the time it takes to claim items from NoSql.	50
processThreads	No	20	No	The maximum number of threads that will process items from the repository at any one time.	20
processQueue	No	200	No	The size of the in memory queue for items that need to be processed. The recommended queue size is at least as large as the number of threads, if not two to three times larger. Larger queue sizes will allow database access to be performed further in advance and smooth fluctuations in the time it takes to claim items from NoSql.	200
deleteCompleteQueueEntries	No	false	No	Set to true if completed queue entries should be deleted (or just marked as complete).	true / false
flushSyncTime	No	"15s"	No	Time to wait for all servers to finish their flushes to the snapshot at the end of each incremental crawl.	"30s"
deleteCheckAfterErrors	No	"ALWAYS"	No	Checks if "delete" candidates still exist after an incremental when they are part of scan error.	"ALWAYS" / "NEVER"
maxIdentitiesTimestamp	No	3	No	Number of crawls to execute before removing the oldest identity items.	3
workflowErrorTolerant	No	false	No	If set to true, exceptions in workflow rules will only affect the execution of the rule in which the exception occurs. Subsequent rules will be executed and the job will complete the workflow successfully. Otherwise, exceptions in workflow rules will be re-thrown and the job will be moved to the error workflow.	true / false
retriesEnabled	No	true	No	If set to true, failed documents will be reprocessed at the end of the crawl and in the following incremental crawls.	true / false
removeFailedFromSnapshot	No	false	No	Check to remove the snapshot entry for each failed document. This makes the retries to be performed on all next incremental crawls. This overrides the \"Maximum crawls to retry\" option.	true / false

retryPattern	Yes	[ ]	Yes	A regex pattern to match against any exception raised by either the individual document or publisher. If matched, the document or documents will be retried using the limits configured below.	[".*\\.pdf"]
maxInCrawlRetries	No	3	No	Maximum number of retries per crawl for a failed document.	3
maxCrawls	No	3	No	Maximum number of incremental crawls in which a failed document will be retried.	3

Example

PUT aspire/_api/connectors/e3fc3f4b-2784-4e5a-b27e-87a8f9a726a9


```
{
    "id": "e3fc3f4b-2784-4e5a-b27e-87a8f9a726a9",
    "type": "filesystem",
    "description": "Test Description",
    "artifact": "com.accenture.aspire:aspire-filesystem-source",
    "properties": {
        "debug": false,
        "wDebug": false,
        "enableStatistics": false,
        "infoCacheSize": 100,
        "mapCacheSize": 100,
        "setCacheSize": 100,
        "identityCacheSize": 100,
        "enableFetcher": true,
        "enableTextExtract": true,
        "extractTextConfiguration": true,
        "extractTextMaxSize": "20971520",
        "extractTimeout": 180000,
        "xmlMaxDepth": 100,
        "structuredText": false,
        "tikaConfig": "",
        "pdfParserProperties": true,
        "enableAutoSpace": true,
        "suppressDuplicateOverlappingText": false,
        "extractAnnotationText": true,
        "sortByPosition": false,
        "extractAcroFormContent": true,
        "extractInlineImages": false,
        "extractUniqueInlineImagesOnly": true,
        "enable-non-text-filter": true,
        "enableFetchForNonText": true,
        "non-text-document": true,
        "nonTextDocumentsExtensions": "jpg,jpeg,gif,png,tif,mp3,mp4,mpg,mpeg,avi,mkv,wav,bmp,swf,war,rar,tgz,
dll,exe,class",
        "metadataMap": [{
                "from": "fieldA",
                "to": "destA"
            }, {
                "from": "fieldB",
                "to": "destB"
            }
        ],
        "addHierarchy": true,
        "hierarchyCacheSize": 5000,
        "scanThreads": 10,
        "scanQueue": 50,
        "processThreads": 20,
        "processQueue": 200,
        "deleteCompleteQueueEntries": false,
        "flushSyncTime": "15s",
        "deleteCheckAfterErrors": "ALWAYS",
        "maxIdentitiesTimestamp": 3,
        "workflowErrorTolerant": true,
        "retriesEnabled": true,
        "removeFailedFromSnapshot": true,
        "retryPattern": [".*tika.*", ".*png.*"],
        "maxInCrawlRetries": 3,
        "maxCrawls": 3
    }
}
```