

Azure Data Lake For Developers

Azure Data Lake Source

The Azure Data Lake connector will crawl files and folders (configuration-dependent). Execution will result in populating the following fields:

Property	Type	Description
Fullname	String	Full path of the directory or file (from root path "/")
Name	String	File name (minus the path) of the directory or file
Length	Long	Length of a file (does not apply for directories)
Group	String	ID of the group that owns this file/directory
User	String	ID of the user that owns this file/directory
Permission	String	Unix-style permission string for this file or directory
Last Access Time	Date	Date and time of when the file was last accessed
AclBit	Boolean	Flag indicating if the file has ACLs set on it
Block Size	Long	Block size reported by server
Expiry Time	Date	Date and time when the file expires, as UTC time
ReplicationFactor	Int	Replication factor reported by server
isContainer	Boolean	Indicates "true" if is a directory, otherwise File
Fetch Url	String	Azure Data Lake full Absolute Path including FQDN. <code>adl://[yourdomain].azuredatalakestore.net/full/path/to.file</code>
Last Modified Date	Date	Date and time of when the file was last modified
Acls	ACL Array	List of access for file or folder

Example Output

The following code block shows the console output of crawling of a folder called `/test` located at root of testing Data Lake Storage `ad1://dlsjose.azuredatalakestore.net`

2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Received job - action: start
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Initializing crawl:
1528133426127
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Initializing statistics for
crawl: 1528133426127
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Clearing queues, snapshot,
hierarchy and intersection acls - please wait...
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Clearing queues, snapshot,
hierarchy and intersection acls took 200 ms
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Offering crawl root
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Sending start job for crawl:
1528133426127 (status: I)
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ScanQueueLoader]: QueueLoader
(scan) crawl status checker thread started
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ProcessQueueLoader]: QueueLoader
(process) item claim thread started
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ScanQueueLoader]: QueueLoader
(scan) item claim thread started
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ProcessQueueLoader]: QueueLoader
(process) crawl status checker thread started
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Published crawl start job
2018-06-04T17:30:26Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store: [
/test]
2018-06-04T17:30:28Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test
2018-06-04T17:30:28Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager/ProcessCrawlRoot]: Added root
item: /test
2018-06-04T17:30:28Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test
2018-06-04T17:30:29Z INFO [/aspire_azuredatalakestore/ScanPipelineManager/Scan]: Scanning: /test
2018-06-04T17:30:30Z INFO [/aspire_azuredatalakestore/RAP]: >>> Scan Item - Azure DataLake Store: /test
2018-06-04T17:30:31Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test/NOACCESS
2018-06-04T17:30:31Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test/subtest
2018-06-04T17:30:31Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test/NOACCESS
2018-06-04T17:30:31Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test/subtest
2018-06-04T17:30:31Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test/test4.txt
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test/test5.txt
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/RAP]: >>> Processing crawl - Azure DataLake Store:
adl://dlsjose.azuredatalakestore.net/test/test6.txt
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/ScanPipelineManager/Scan]: Item /test scanned 5
subitems
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test/test4.txt
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test/test5.txt
2018-06-04T17:30:32Z INFO [/aspire_azuredatalakestore/ProcessPipelineManager]: Processing: adl://dlsjose.
azuredatalakestore.net/test/test6.txt
2018-06-04T17:30:34Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ScanQueueLoader]: QueueLoader
(scan) crawl status thread stopped
2018-06-04T17:30:34Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ProcessQueueLoader]: QueueLoader
(process) crawl status thread stopped
2018-06-04T17:30:34Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Published crawl end job
2018-06-04T17:30:34Z INFO [/aspire_azuredatalakestore/Main/CrawlController]: Crawl ended with status: S
2018-06-04T17:30:34Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ProcessQueueLoader]: QueueLoader
(process) item claim thread stopped
2018-06-04T17:30:35Z INFO [/aspire_azuredatalakestore/QueuePipelineManager/ScanQueueLoader]: QueueLoader
(scan) item claim thread stopped