MongoDB Collections Description

Each Connector installed as a Content Source in Aspire will create a new Database in MongoDB with its name, and multiple collections that it uses for crawling and serving security information. The following describes what every collection is used for as well as an explanation of each of the fields of the documents stored in them.

What's next?

- Write Your Own Connector from Scratch

- processQueue
- scanQueue
- hierarchy
- audit
- errors
- statistics
- status
- snapshots

Document Queues and Metadata

• processQueue

Manages the items that needs to be processed by the workflow, these items may or may not be sent to scanned.

Field Name	Example	Description
_id	C:\test-folder\folderA\testDocument.txt	The unique id of the document
metadata	[depends on each connector]	The necessary metadata fields the connector needs to fetch or populate this document
type	[depends on each connector]	The serialized version of the ItemType of the document
status	C, P or A	The document processing status: C: **Completed**, means it have been already processed P: **in Progress**, means it is currently been processed A: **Available**, means it is available for been processed
action	add, update, delete	The action to be performed to the search engine for the document
timestamp	1465334398471	The time-stamp when this document was added to the queue
signature	CBEC1210FE2D51A8166C3E70D38F8A07	An MD5 signature, when a document changes this signature should also change
parentId	C:\test-folder\folderA	The id of the parent document, in other words the document that scanned the current document
processor	File_System-192.168.1.15:50505	The identifier of the Aspire server that processed or is processing the current document
shouldScan	false	Determines whether or not this document should be considered for scanning
shouldProcess	true	Determines whether or not this document should be considered for being processed by the workflow
retries	0	The number of times this document has been retried
name	testDocument.txt	The name of this document
isCrawlRootItem	false	Indicates if this is one of the root crawl items (for internal control)
hierarchyId	C:\test-folder\folderA\testDocument.txt	Unique Id for using to generate the hierarchy for this document, it may be different from the _id field

Example:

```
{
    "_id" : "C:\\test-folder\\folderA\\testDocument.txt",
    "metadata" : {
        "fetchUrl" : "file://C:/test-folder/folderA/testDocument.txt",
        "url" : "file://C:/test-folder/folderA/testDocument.txt"
    },
    "type" :
"vtwqabl6oiadwy3pnuxhgzlbojrwq5dfmnug433mn5twszltfzqxg4djojss4y3pnvyg63tfnz2hglsgnfwgk43zon2gk3kjorsw2vdzobsq
aaaaaaaaaaaaciaaa6dsaahguylwmexgyylom4xek3tvnuaaaaaaaaaaaasaaahq4duaacgm2lmmu",
    "status" : "C",
    "action" : "add",
    "timestamp" : NumberLong(1465334398471),
    "signature" : "CBEC1210FE2D51A8166C3E70D38F8A07",
    "parentId" : "C:\\test-folder\\folderA",
    "processor" : "File_System_Source-192.168.56.1:50505",
    "shouldScan" : false,
    "shouldProcess" : true,
    "retries" : 0,
    "name" : "0.txt",
    "isCrawlRootItem" : false,
    "hiearchyId" : "C:\\test-folder\\folderA\\testDocument.txt"
}
```

- **scanQueue**

Manages the items that needs to be scanned by the connector, these items may or may not be have been sent to process previously.

Field Name	Example	Description
_id	C:\test-folder\folderA	The unique id of the document
metadata	[depends on each connector]	The necessary metadata fields the connector needs to fetch or populate this document
type	[depends on each connector]	The serialized version of the ItemType of the document
status	C, P or A	The document processing status: C: **Completed**, means it have been already processed P: **in Progress**, means it is currently been processed A: **Available**, means it is available for been processed
action	add, update, delete	The action to be performed to the search engine for the document
timestamp	1465334398471	The time-stamp when this document was added to the queue
signature	CBEC1210FE2D51A8166C3E70D38F8A07	An MD5 signature, when a document changes this signature should also change
parentId	C:\test-folder	The id of the parent document, in other words the document that scanned the current document
processor	File_System-192.168.1.15:50505	The identifier of the Aspire server that processed or is processing the current document
shouldScan	false	Determines whether or not this document should be considered for scanning
shouldProcess	true	Determines whether or not this document was considered for being processed by the workflow
retries	0	The number of times this document has been retried
name	folderA	The name of this document
isCrawlRootItem	false	Indicates if this is one of the root crawl items (for internal control)
hierarchyId	C:\test-folder\folderA\testDocument.txt	Unique Id for using to generate the hierarchy for this document, it may be different from the _id field

Example:

```
{
    "_id" : "C:\\test-folder\\folderA",
    "metadata" : {
        "fetchUrl" : "file://C:/test-folder/folderA",
        "url" : "file://C:/test-folder/folderA",
        "displayUrl" : "C:\\test-folder\\folderA",
        "lastModified" : "2016-02-23T17:08:55Z",
        "dataSize" : 0,
        "acls" : null
    },
    "type" :
"vtwqabl6oiadwy3pnuxhgzlbojrwq5dfmnug433mn5twszltfzqxg4djojss4y3pnvyg63tfnz2hglsgnfwgk43zon2gk3kjorsw2vdzobsq
aaaaaaaaaaaaciaaa6dsaahguylwmexgyylom4xek3tvnuaaaaaaaaaaaaasaaahq4duaadgm33mmrsxe",
    "status" : "C",
    "action" : "add",
    "timestamp" : NumberLong(1465334398103),
    "signature" : "CD2C65824E45BFE94C71970EEEA18A8C",
    "parentId" : "C:\\test-folder",
    "processor" : "File_System_Source-192.168.56.1:50505",
    "shouldScan" : true,
    "shouldProcess" : true,
    "retries" : 0,
    "name" : "folderA",
    "isCrawlRootItem" : false,
    "hiearchyId" : "C:\\test-folder\\folderA"
}
```

- ### hierarchy

  Holds the hierarchy information about every single parent document scanned by the connector, each parent contains the information about all its parents all the way up to the root document.

Field Name	Example	Description
_id	C:\test-folder\folderA	Unique id of the parent document
name	folderA	Name to be used in the hierarchy metadata
ancestors	[parent hierarchy info]	Holds the same information but for the parent of document, or **null** if this is a root document

Example:

```
{
    "_id" : "C:\\test-folder\\folderA",
    "name" : "folderA",
    "ancestors" : {
        "_id" : "C:\\test-folder",
        "name" : "test-folder",
        "ancestors" : null
    }
}
```

Statistics and Logging

- ### audit

  Holds the actions done by the content source for each of the documents.

Field Name	Example	Description
_id	ObjectId("5750bfa610163e3f58fd7019")	Mongo Internal ID
id	C:\\test-folder\\folderA\\testDocument.txt	Unique Id of the document
crawlStart	1464909728339	Crawl identifier, each crawl has a different crawlStart time
url	file://C:/test-folder/folderA/testDocument.txt	URL of the document
type	**job** or **batch**	Specifies what type of audit log is the current object
action	ADD, UPDATE, NOCHANGE, DELETE, BATCH_COMPLETED, BATCH_ERROR, WORKFLOW_COMPLETE, WORKFLOW_TERMINATED, WORKFLOW_ERROR or EXCLUDED	**ADD:** Discovered as new document to be added **UPDATE:** Discovered document with a change **NOCHANGE:** Found no change in document **DELETE:** Document was found to be deleted **BATCH_COMPLETED:** The current batch finished **BATCH_ERROR:** There was an error closing the batch **WORKFLOW_COMPLETE:** The document completed the workflow without errors **WORKFLOW_TERMINATED:** The document was terminated during the workflow **WORKFLOW_ERROR:** The document had an error executing the workflow **EXCLUDED:** The document was excluded by the include/exclude patterns
batch	10.10.20.203:50506/2016-06-03T16:04:59Z/batch-0	If any, contains the id of the batch of the current document
ts	1464970015441	The time this entry was added to the log

Example:

```
{
    "_id" : ObjectId("5751ab210afca2469094bb23"),
    "id" : "C:\\test-folder\\folderA\\testDocument.txt",
    "crawlStart" : NumberLong(1464970009642),
    "url" : "file://C:/test-folder/folderA/testDocument.txt",
    "type" : "job",
    "action" : "WORKFLOW_COMPLETE",
    "batch" : "10.10.20.203:50506/2016-06-03T16:04:59Z/batch-0",
    "ts" : NumberLong(1464970015441)
}
```

- ## errors

Holds the possible document errors that occurs either in the scanning or workflow processing.

Field Name	Example	Description
_id	ObjectId("576844914b4ae74664a414bd")	Mongo's internal id
error /@time	1466451089287	Time when this error entry was logged
error /@crawl Time	1466451085183	Identifier of the crawl
error /@cs	File_System_Source	Identifier of the content source
error /@proc essor	File_System_Source-192.168.56.1:50505	The server that processed and reported this error
error /@type	S, D, B, F or U	**S:** Scanner errors relates to errors caused in the connector scanning stages **D:** Document errors relates to fetch, text extraction or workflow processing errors **B:** Batch errors relates to failed batches of Aspire jobs **F:** Failed errors are not currently being used but they could be later **U:** Unknown errors relates to errors where the source is unknown
error/_$	Error processing: C:\\test-folder/folderA/testDocument2.txt\ncom.searchtechnologies.aspire.services.AspireException: Exception whilst running script: Rule: 1\r\n\tat..... (more)	The error message

Example:

```
{
    "_id" : ObjectId("576844914b4ae74664a414bd"),
    "error" : {
        "@time" : NumberLong(1466451089287),
        "@crawlTime" : NumberLong(1466451085183),
        "@cs" : "File_System_Source",
        "@processor" : "File_System_Source-192.168.56.1:50505",
        "@type" : "D",
        "_$" : "Error processing: C:\\test-folder/folderA/testDocument2.txt\ncom.searchtechnologies.aspire.
services.AspireException: Exception whilst running script: Rule: 1\r\n\tat ... (more)"
    }
}
```

- ## statistics

  Holds the crawl statistics per server, what you see in the Administration UI is the sum of all the server statistics associated with the same crawl identified.

FieldName	Example	Description
_id	1466450887680-File_System_Source-192.168.56.1:50505	Unique identifier of each statistics object
statistics/@processor	File_System_Source-192.168.56.1:50505	The server+content source name
statistics/@server	192.168.56.1:50505	The server identifier

statistics/@status	A, S, E, F, L, I, N, IP, IWP, IWR, X, IWS or U	The crawl status: **A:** Aborted **S:** Completed **E:** Errored **F:** Failed **L:** Loading **I:** In-Progress **N:** New **iP:** Paused **IWP:** Pausing **IWR:** Resuming **X:** Stopped **IWS:** Stopping **U:** Unknown
statistics/@mode	F, FR, I, IR, R, T, U	**F:** Full crawl **FR:** Full recovery **I:** Incremental crawl **IR:** Incremental recovery **R:** Real time **T:** Test **U:** Unknown
statistics/@startTime	1466450887680	The time when the crawl started
statistics/@endTime	1466450905466	The time when the crawl ended
statistics/@cs	File_System_Source	The identifier of the content source
statistics/queue/scan/@toScan	0	Number of documents in the scan queue pending to be scanned
statistics/queue/scan/@scanning	0	Number of documents from the scan queue currently being scanned
statistics/queue/scan/@scanned	11	Number of documents from the scan queue already scanned
statistics/queue/scan/@total	11	Total documents in the scan queue
statistics/queue/process/@toProcess	0	Number of documents in the process queue pending to be processed
statistics/queue/process/@processing	0	Number of documents from the process queue currently being processed
statistics/queue/process/@processed	121	Number of documents from the process queue already processed
statistics/queue/process/@total	121	Total documents in the process queue
statistics/nProgress/@adding	0	Number of documents currently being processed as "ADD"
statistics/inProgress/@updating	0	Number of documents currently being processed as "UPDATE"
statistics/inProgress/@deleting	0	Number of documents currently being processed as "DELETE"

statistics/inProgress/@total	0	Total documents currently being processed
statistics/processed/@added	121	Number of documents processed as "ADD"
statistics/processed /@updated	0	Number of documents processed as "UPDATE"
statistics/processed /@deleting	0	Number of documents processed as "DELETE"
statistics/processed /@unchanged	0	Number of documents processed as "NOCHANGE"
statistics/processed /@excluded	0	Number of documents "EXCLUDED" from being processed
statistics/processed /@terminated	0	Number of documents processed but ended as "TERMINATED"
statistics/processed/@errored	0	Number of documents processed with Errors
statistics/processed/@bytes	129470	Total bytes processed so far
statistics/processed/@total	121	Total number of documents processed
statistics/errors/@batch	0	Number of batch errors
statistics/errors/@scan	0	Number of scanner errors (errors that happened while scanning for documents)
statistics/errors/@document	0	Number of document errors (errors that occurred while processing the document)
statistics/errors/@total	0	Total number of errors

Example:

```
{
    "_id" : "1466450887680-File_System_Source-192.168.56.1:50505",
    "statistics" : {
        "@processor" : "File_System_Source-192.168.56.1:50505",
        "@server" : "192.168.56.1:50505",
        "@status" : "S",
        "@mode" : "F",
        "@startTime" : NumberLong(1466450887680),
        "@endTime" : NumberLong(1466450905466),
        "@cs" : "File_System_Source",
        "queue" : {
            "scan" : {
                "@toScan" : 0,
                "@scanning" : 0,
                "@scanned" : 11,
                "@total" : 11
            },
            "process" : {
                "@toProcess" : 0,
                "@processing" : 0,
                "@processed" : 121,
                "@total" : 121
            }
        },
        "inProgress" : {
            "@adding" : 0,
            "@updating" : 0,
            "@deleting" : 0,
            "@total" : 0
        },
        "processed" : {
            "@added" : 121,
            "@updated" : 0,
            "@deleting" : 0,
            "@unchanged" : 0,
            "@excluded" : 0,
            "@terminated" : 0,
            "@errored" : 0,
            "@bytes" : 129470,
            "@total" : 121
        },
        "errors" : {
            "@batch" : 0,
            "@scan" : 0,
            "@document" : 0,
            "@total" : 0
        }
    }
}
```

Controlling and Incremental

- **status**

  Holds all the crawl control information and its status, this determines when a crawl should be started, paused, stopped, or even complete as successful.

Field Name	Example	Description
_id	ObjectId ("5768448d4b4ae74664a41495")	Mongo's internal ID
connectorSource	[depends on specific connector]	Contains the configuration set for running a new crawl, it depends on what each specific connector needs as configuration

@action	start	Property from the scheduler specifying the action to be done for this content source. For crawls it should always be "start"
@actionProperties	full or incremental	Property from the scheduler specifying if the crawl should be either an incremental or a full
@crawlId	0	Aspire's internal ID for the crawl
@normalizedCSName	File_System_Source	Aspire's internal name for the current content source
displayName	File System Source	The content source name as the user entered it
@scheduler	AspireSystemScheduler	Identifies the scheduler that created the crawl request. By default it should be "AspireSystemScheduler"
@scheduleId	0	The schedule id corresponding to the current crawl request.
@jobNumber	5	A sequential counter of how many jobs has the scheduler served.
@sourceId	File_System_Source	The content source identifier
@actionType	manual or scheduled	Determines if the crawl was started by a periodic schedule or a manual request
@dbId	1	Legacy property from the scheduler
crawlStart	1466460900181	The time in milliseconds that this request was created, this will be used as the crawl identifier for rest of the crawl life
crawlStatus	A, S, E, F, L, I, N, IP, IWP, IWR, X, IWS or U	The crawl status: **A:** Aborted **S:** Completed **E:** Errored **F:** Failed **L:** Loading **I:** In-Progress **N:** New **iP:** Paused **IWP:** Pausing **IWR:** Resuming **X:** Stopped **IWS:** Stopping **U:** Unknown
processDeletes	none	If any, holds the ID of the server that is scanning through the snapshots to find the deletes at the end of the crawl
processingDeletesStatus	finished	This flag is only present when the deletes processing is finished
crawlEnd	1466460912343	If any, the time in milliseconds that this crawl finished

Example:

```
{
    "_id" : ObjectId("57686d0d4b4ae74664a417a8"),
    "connectorSource" : {
        "url" : "C:\\test-folder",
        "partialScan" : "false",
        "subDirUrl" : null,
        "indexContainers" : "true",
        "scanRecursively" : "true",
        "scanExcludedItems" : "false",
        "useACLs" : "false",
        "acls" : null,
        "includes" : null,
        "excludes" : null
    },
    "@action" : "start",
    "@actionProperties" : "full",
    "@crawlId" : "0",
    "@normalizedCSName" : "File_System_Source",
    "displayName" : "File System Source",
    "@scheduler" : "AspireSystemScheduler",
    "@scheduleId" : "2",
    "@jobNumber" : "7",
    "@sourceId" : "File_System_Source",
    "@actionType" : "manual",
    "@dbId" : "2",
    "crawlStart" : NumberLong(1466461453589),
    "crawlStatus" : "S",
    "processDeletes" : "none",
    "processingDeletesStatus" : "finished",
    "crawlEnd" : NumberLong(1466461465352)
}
```

- ## snapshots

Holds the incremental information needed for determining when a document has changed, have been added or when the get deleted. This is only used by the connectors where it's repositories APIs doesn't provide a way of getting the updates from a single call without having to scan through all the documents again.

Field Name	**Example**	**Description**
_id	C:\test-folder\folderA\testDocument.txt	The unique ID of each document
container	true or false	**true**: If this document can contain documents **false**: If it doesn't
crawlId	0	The id of the crawl that introduced this entry
signature	CBEC1210FE2D51A8166C3E70D38F8A07	An MD5 digest of the main metadata of each document needed for determine changes
timestamp	1466461845637	The crawlStart time
parentId	C:\test-folder\folderA	The name of the parent of this item
error	true or false	**true**: if this document had an error **false**: otherwise

Example:

```
{
    "_id" : "C:\\test-folder\\folderA\testDocument.txt",
    "container" : false,
    "crawlId" : 0,
    "signature" : "CBEC1210FE2D51A8166C3E70D38F8A07",
    "timestamp" : NumberLong(1466461845637),
    "parentId" : "C:\\test-folder\\folderA",
    "error" : false
}
```