Best Practices for Scanning Files Larger Than 30MB Using Discover

Doc ID	TECH218585
Version:	1.0
Status:	Published
Published date:	12/17/2010
Categories:	How To , 10.0 , 11.0 , 12.0
Available To:	Internal

Author: marie_coon@symantec.com

Problem

How to setup Discover to scan files larger than 30MB.

Solution

NOTE: These settings will slow down processing, so we do not recommend making these changes for non-Discover Detection Servers.

To process files that are larger than the 30MB standard limit, you must modify several settings in Discover. The plan is to use two different Discover Servers. One server is for files 30MB or smaller. The first server should be using default settings. The second server is for files larger than 30MB. The configuration in this KB is for the second server.

The idea is to reduce the number of message chains while increasing the capacity of the chain (max file size, number of tokens, etc.). You must also adjust timeouts.

Regarding the servers

In the example below the modified parameters are based on a max size of 120 MB:

- The fast path Discover Server (which could eventually run on the same machine as the Enforce Server) must have the standard configuration settings for a Discover server.
- The slow path Discover server should be configured with the following Advanced Settings found on the Advanced Settings (Server Settings) Page in the UI.
- BoxMonitor.FileReaderMemory = -Xms1578M -Xmx1578M (default = –Xrs -Xms1200M -Xmx1200M)

  This increases the available FileReader Memory.
- BoxMonitor.HeartbeatGapBeforeRestart = 2100000 (default = 960000)

  The time interval (in milliseconds) that the BoxMonitor waits for a monitor process (for example, FileReader, IncidentWriter) to report the heartbeat. If the heartbeat is not received within this time interval the BoxMonitor restarts the process. Increasing this value gives more time for FileReader to respond to Box Monitor.

- ContentExtraction.LongTimeout = 300000 (default = 120000)

  The time interval (in milliseconds) given to the ContentExtractor to process a document larger than ContentExtraction.LongContentSize. If the document cannot be processed within the specified time it's reported as unprocessed. This value should be greater than ContentExtraction.ShortTimeout and less than ContentExtraction.RunawayTimeout.

- ContentExtraction.MaxContentSize = 120M (default = 30M)

The maximum size (in MB) of the document that can be processed by the ContentExtractor. This increases the maximum file size limitation during Content Extraction.

- ContentExtraction.RunawayTimeout = 600000 (default = 300000)

  The time interval (in milliseconds) given to the ContentExtractor to finish processing of any document. If the ContentExtractor does not finish processing some document within this time it will be considered unstable and it will be restarted. This value should be significantly greater than ContentExtraction.LongTimeout.

- FileReader.MaxFileSize = 125829120 (default = 30000000)

  The maximum size of a message to be processed. Larger messages are truncated to this size. This should match the ContentExtraction.MaxContentSize.

- FileReader.MaxReadGap = 45 (default = 15)

  The time that a child process can have data but not have read anything before it stops sending heartbeats. Increasing this value gives FileReader more time.

- IncidentDetection.MaxContentLength = 20000000 (default = 2000000)

  Applies only to regular expression rules. On a per component basis, only the first MaxContentLength number of characters are scanned for violations. The default (2,000,000) is equivalent to > 1000 pages of typical text. The limiter exists to prevent regular expression rules from taking too long. This allows us to look throughout the document for regular expressions.

- Lexer.MaximumNumberOfTokens = 120000 (default = 30000)

  Maximum number of tokens (including separators) extracted from each message component for detection. Applicable to all detection technologies where tokenization is required, e.g. System patterns, EDM, DGM. Increasing this value may cause the detection to run out of memory and restart.

- MessageChain.CacheSize = 1 (default = 8)

  Limits the number of messages that can be queued in the message chains.

- MessageChain.MaximumComponentTime = 1200000 (default = 600000)

  The time interval (in milliseconds) allowed before any chain component is restarted. Giving more time for processing.

- MessageChain.NumChains = 1 (default = 8)
  Note: For normal usage, it is recommended to set MessageChain.NumChains = # of processors on the Discover box.

  The number of messages, in parallel, that the filereader will process. Setting this number higher than 8 (with the other default settings) is not recommended. A higher setting does not substantially increase performance and there is a much greater risk of running out of memory. Setting this to less than 8 (in some cases 1) helps when processing big files, but it may slow down the system considerably.

Additionally: Add the following line to the \vontu\protect\config\crawler.properties on the Discover server machine:

filesystemcrawler.workqueue.max.memory = 120000000

This value defaults to 60000000, but it must be the same or larger than the maximum message size. All other settings should be standard.

NOTE: As per other settings, changes to the Advanced Server Settings and to properties files require a recycling of the Vontu Monitor in order to take effect.

Targets should be configured for each set of shares to scan: One target is assigned to the slow path server and only scans files larger than 10MB; the other target is assigned to the fast path and scans files smaller than 10MB. This setup allows you to scan all file types up to 120MB.

Text files larger that 120MB will be truncated, but the first 120MB will be processed.

Other file types: *.doc, *.xls, *.ppt, *.pdf, *.zip, etcetera will be **ignored** if they are larger than 120MB because Vontu's Message Cracking technology cannot recognize them.

If you must include .xls files, you must disable formula extraction.

To disable formula extraction:

1. Edit the formats.ini file in the directory, \Vontu\Protect\lib\native\formats.ini.
2. Change "getformulastring=2" to "getformulastring=0."
3. Restart the Monitor Server.
4. Disable formula extraction on all the detection servers.
   Note: If you use index document matching (IDM) on Excel files, disable formula extraction on Vontu Enforce Server for consistency between IDM indexing and detection.

If you are using IDM and it does not work on files that exceed the content extraction limit, this problem has been addressed in the E-track 2229997. This is the workaround:
"You can change the advanced setting 'DDM.MaxBinMatchSize' to 30,000,000 (instead of 300,000,000) on each Detection Server and matching of large binary files will work. This will only fix the issue with files that verity cannot extract any partial text for."

Please note: It would also be advisable to use 64 Bit environments, since you may be required to adjust the overall JVM size of the FileReader JVM beyond the out of the box settings to accomodate the amount of detection chains. Otherwise you may runinto an out of memory error (OOM)

Legacy ID

42415