ACCELERATOR DEDUPLICATION WHITE PAPER

FEBRUARY 2011

PURPOSE

This white paper describes the deduplication feature that Compliance Accelerator and Discovery Accelerator provide.

OVERVIEW

The deduplication feature was introduced in version 9.0 of Compliance Accelerator and Discovery Accelerator. It provides two methods to identify similar and duplicate items:

- METADATA ANALYSIS. This method is quick to run but can give rise to false positives in certain situations. The method is available in both Compliance Accelerator and Discovery Accelerator.
- CONTENT AND METADATA ANALYSIS. This method is only available in Discovery Accelerator implementations where analytics is enabled. The reason for this is that content analysis requires the availability of the full HTML content of each item, which is true of analytics-enabled cases only.

  Analyzing both the content of items and their metadata is better than analyzing the metadata alone.

The following sections describe these two methods in detail and explain the advantages and disadvantages of using them.

GLOSSARY

Term	Description
FPCN	A hash that Enterprise Vault generates of a file or file-based attachment.
GSS	Guaranteed sampling search that Compliance Accelerator uses to obtain a specified percentage of messages for review.

METADATA ANALYSIS (IDENTIFICATION OF "SIMILAR" ITEMS)

For Discovery Accelerator, this method is intended as a quick and quite effective pre-assessment mode. In Compliance Accelerator, the same set of metadata properties is used to generate hashes of items. Similar items are removed from the set, and only one copy is kept for review to increase the quality of the review set.

The results of a Compliance Accelerator or Discovery Accelerator search contain a subset of each item's metadata. Elements of this subset are used to construct a unique hash of each item, and so identify similar items. The list of attributes used is given below. Some items such as files do not have all these attributes, and in these cases the next best set is used.

Field	Document type	Comments
Author Display Name	• Exchange mails • Domino mails • EML mail files • Files • SharePoint documents	
CC Display Name	• Exchange mails • Domino mails • EML mail files	Enterprise Vault truncates this list to 256 characters. The hash-generation algorithm sorts this field into alphabetical order before calculating the fingerprint.
Content	• Exchange mails • Domino mails • EML mail files • Files • SharePoint documents	Enterprise Vault truncates the content to 120 characters, 118 of which are used for fingerprint calculation.
ConversationID	• Exchange mails	
Modified Date	• Files • SharePoint documents	
Number of attachments	• Exchange mails • Domino mails • EML mail files	
Number of CC	• Exchange mails • Domino mails • EML mail files	The number of recipients in the CC field.
Number of To	• Exchange mails • Domino mails • EML mail files	The number of recipients in the To field.
Subject	• Exchange mails • Domino mails • EML mail files • Files • SharePoint documents	For files and SharePoint documents, the subject is the file name.
To Display Name	• Exchange mails • Domino mails • EML mail files	Enterprise Vault truncates this list to 256 characters. The hash-generation algorithm sorts this field into alphabetical order before calculating the fingerprint.
Message Type	• SharePoint documents with a document type of EML	In version 10.0 and later of the Accelerators, the message type is also included for SharePoint items with a document type of EML.

NOTE: For SharePoint Items with a document type of EML, the metadata hash is calculated as for EML items, but with one extra field of MsgType. For SharePoint items that do not have a document type of EML, the metadata hash is calculated as for files.

The metadata analysis method can return false positives for two reasons:

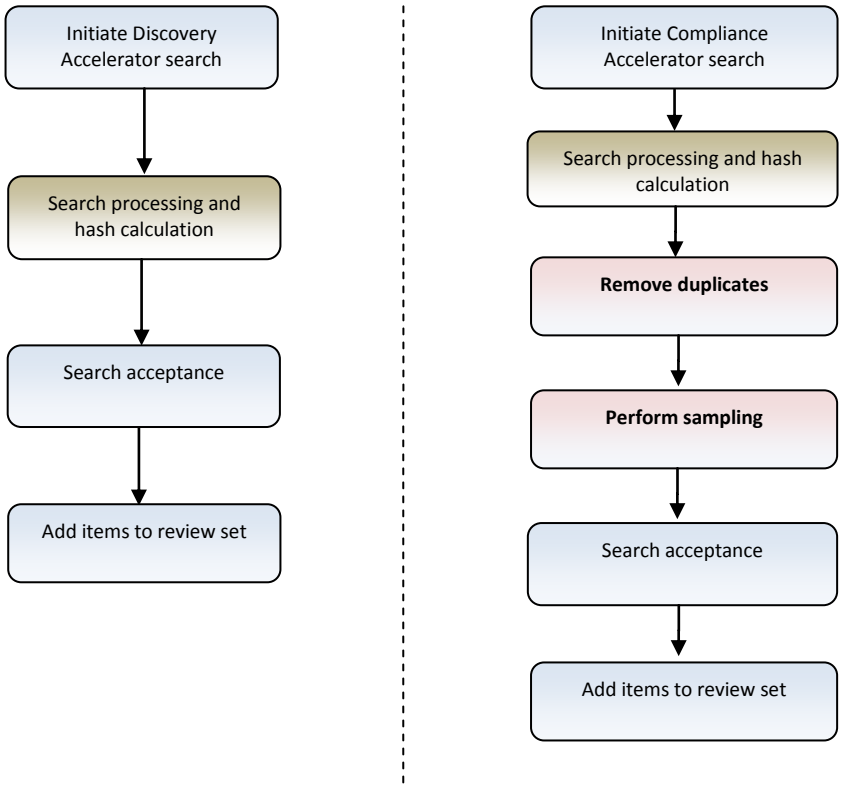
- It cannot evaluate the entire content of each item.
- It does not have access to the full list of recipients.

So, there is a small possibility that items have differing content or recipient lists, but in all other respects are identical. This is why the metadata analysis method is said to identify “similar” items rather than duplicates.

Conversely, there may be cases where FSA items are found to be “not similar”, but content-based deduplication may find the items to be duplicates. This can happen when the items are the same in all respects except for the modified date. This is because content hashing does not consider the modified date.

See page 7 for some examples of the differing results that metadata analysis and content and metadata analysis can produce.

DEDUPLICATION HASH CALCULATION WORKFLOW



Discovery Accelerator calculates and stores the deduplication hashes of search items at the time that it processes them, before the results of the search are accepted. These hashes are used to stack similar items in the review pane after search results are accepted.

Compliance Accelerator automatically deduplicates similar items from the search results. This deduplication is performed for guaranteed sample searches, scheduled searches and ad hoc searches only.

Compliance Accelerator does not deduplicate items when only sampling is performed.

CONTENT AND METADATA ANALYSIS (IDENTIFICATION OF "DUPLICATE" ITEMS)

The second deduplication method analyzes both the full HTML content of items and their metadata. This significantly reduces the chances of false positives occurring.

The metadata and content analysis method can return false positives for two reasons:

- It cannot evaluate the content of item.
- The content of the item is not in an HTML-convertible format. For example, this is true of images.

This method is only available in Discovery Accelerator cases where analytics is enabled. The reason for this is that content analysis requires the availability of the full HTML content of each item, which is true of analytics-enabled cases only.

As part of analytics data collection, Discovery Accelerator gets the complete content and metadata of each item, including the full list of recipients. The data is in full fidelity to generate a more accurate deduplication hash for the item.

Discovery Accelerator uses the following properties to create the hash of an item (it ignores empty properties).

File, SharePoint document, and email properties

Field	Document type	Availability	Comments
$UNID	• Domino emails	Enterprise Vault 2007 and later	This is the unique identifier that Lotus Domino assigns to each email. For more information, see this article: http://doctorapi.blogspot.com/2009/05/noteid-unid-dbid-originatorid.html
Author	• Exchange emails • Domino emails • EML email files • Files • SharePoint documents	All Enterprise Vault versions	
CC	• Exchange emails • Domino emails • EML email files	All Enterprise Vault versions	
Content	• Exchange emails • Domino emails • EML email files • Files • SharePoint documents	All Enterprise Vault versions	Discovery Accelerator considers only the content of the HTML Body element, and not the full HTML. For hash-generation purposes, Discovery Accelerator removes the metadata tags from the HTML.
PR_Conversation_ID	• Exchange emails	All Enterprise Vault versions	Items generated using Outlook 2003 or later have this property embedded to track conversations. When two or more items share the same value for this property, they are likely to be duplicates.

Field	Document type	Availability	Comments
Subject	• Exchange emails • Domino emails • EML email files • Files • SharePoint documents	All Enterprise Vault versions	For files and SharePoint documents, this field is the file name.
To	• Exchange emails • Domino emails • EML email files	All Enterprise Vault versions	
Message Type	• SharePoint documents with a document type of EML	Accelerators 10.0 onwards	In version 10.0 and later of the Accelerator products, the message type is also included for SharePoint items with a document type of EML.

NOTE: For SharePoint Items with a document type of EML, the content hash is calculated as for EML items, but with one extra field of MsgType. For SharePoint items that do not have a document type of EML, the metadata hash is calculated as for files.

Email attachment properties

Field	Document type	Availability	Comments
Content (of pre-8.0 items only; otherwise, the Enterprise Vault-generated FPCN is used for hash generation)	• Exchange emails • Domino emails • EML mail files • Files	All Enterprise Vault versions	Discovery Accelerator considers only the content of the HTML Body element, and not the full HTML. For hash-generation purposes, Discovery Accelerator removes the metadata tags from the HTML.
Attachment File Name	• Exchange emails • Domino emails • EML mail files • Files	All Enterprise Vault versions	Discovery Accelerator uses the file names of all attachments when it generates the hash.
FPCN	• Exchange emails • Domino emails • EML mail files • Files	Enterprise Vault 8.0 and later	Enterprise Vault provides the fingerprint (hash) for email attachments and binary content files.
Modified Date	• Files	All Enterprise Vault versions	

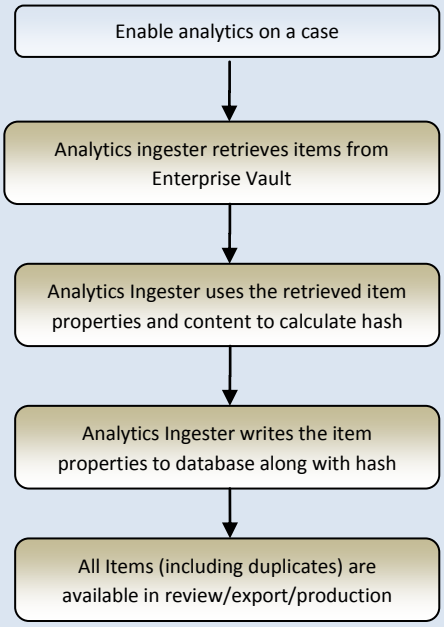
The hash of an email item is calculated based on the properties of the top-level message as well as the properties of all its attachments.

LIMITATIONS WITH PRE-8.0 ARCHIVED ITEMS

Pre-8.0 versions of Enterprise Vault do not generate FPCN properties for the items that they archive. When these properties are not available, Discovery Accelerator uses the HTML content of items to generate the hash values. When the HTML content is not available (as for embedded objects and JPEG or BMP attachments), Discovery Accelerator uses only the item metadata properties to generate the hash values.

HOW DISCOVERY ACCELERATOR PERFORMS HASH CALCULATION WHEN ANALYTICS IS ENABLED

Discovery Accelerator performs hash calculation as part of the ingestion process, as shown below.



The following table identifies the components that Discovery Accelerator uses to generate the hash values for pre-8.0 and post-8.0 archived items, especially when there are attachments.

Enterprise Vault version	Description
2007 and earlier	• HTML content of attachment after pre-process, if available. (The HTML content is not available for JPEG, BMP, and some other non-convertible contents.) • File name of attachment for mail items. • File name of FSA item if HTML content is not available. • Modified date of item.
8.0 and later	• FPCN of attachment that Enterprise Vault has provided. • File name of attachment for mail items. • File name of FSA item if HTML content is not available. • Modified date of item.

EXAMPLES TO SHOW THE DIFFERENCE BETWEEN METADATA ANALYSIS AND CONTENT AND METADATA ANALYSIS

EXAMPLE 1

Consider the following example where Steve exchanges emails with Jane and Jim using Outlook 2000. The mailboxes of all three users are hosted on the same Microsoft Exchange server, and they are journaled to the same journal mailbox.

1. Steve sends the following email (Email 1) to Jane and Jim:

FROM:	Steve
TO:	Jane, Jim
DATE MODIFIED:	Thu 2/3/2011 4:44 PM
SUBJECT:	Financial forecast for next quarter
CONTENT:	Hi, this is an estimate of our... (more than 120 characters of content)

2. Jane sends an immediate reply (Email 2), adding a short comment at the end of Steve's mail.

FROM:	Jane
TO:	Steve, Jim
DATE MODIFIED:	Thu 2/3/2011 4:49 PM
SUBJECT:	Re: Financial forecast for next quarter
CONTENT:	Hi, this is an estimate of our…. (more than 120 characters of content) … "Steve, you missed our planned expense on R&D"

3. A moment later, Jane finds another missing piece and replies to the original mail from Steve with another comment at the end (Email 3).

FROM:	Jane
TO:	Steve, Jim
DATE MODIFIED:	Thu 2/3/2011 4:51 PM
SUBJECT:	Re: Financial forecast for next quarter
CONTENT:	Hi, this is an estimate of our…. (more than 120 characters of content) … "Steve, you also need to account for deferred purchases by customers"

4. When Enterprise Vault journaling and mailbox archiving are enabled, four copies of each email are archived (making a total of 12 archived emails).

Email	Copies archived from
Email 1	• Steve's Sent Items folder in his mailbox • Jane's inbox • Jim's inbox • Journal mailbox (assume everyone is journaled to the same mailbox)
Email 2	• Steve's inbox • Jane's Sent Items folder in her mailbox • Jim's inbox • Journal mailbox (assume everyone is journaled to the same mailbox)
Email 3	• Steve's inbox • Jane's Sent Items folder in her mailbox • Jim's inbox • Journal mailbox (assume everyone is journaled to the same mailbox)

5. A Discovery Accelerator search captures all 12 emails from Enterprise Vault.

EFFECT OF METADATA-ONLY DEDUPLICATION ON THE 12 CAPTURED EMAILS

When metadata-only deduplication is used, Discovery Accelerator groups the emails into two sets, as follows:

Contents of Set 1	Contents of Set 2	
Four emails, all identified as "Similar":	Eight emails, all identified as "Similar":	
• Steve's copy of Email 1 • Jane's copy of Email 1 • Jim's copy of Email 1 • Journal copy of Email 1	• Steve's copy of Email 2 • Jane's copy of Email 2 • Jim's copy of Email 2 • Journal copy of Email 2	• Steve's copy of Email 3 • Jane's copy of Email 3 • Jim's copy of Email 3 • Journal copy of Email 3

This deduplication method considers Emails 2 and 3 to be identical because both their metadata and the first 120 characters of their content are identical.

The following table compares the metadata values for Emails 2 and 3.

Field	Value for Email 2	Value for Email 3	Comments
Author Display Name	Jane	Jane	
CC Display Name			Enterprise Vault truncates this list to 256 characters. The hash-generation algorithm sorts this field into alphabetical order before calculating the fingerprint.
Content	First 120 characters of the content	First 120 characters of the content	Enterprise Vault truncates the content to 120 characters, 118 of which are used for fingerprint calculation. In this example, the first 118 characters of the content of both emails match, as the new text is appended to the end.

Field	Value for Email 2	Value for Email 3	Comments
ConversationID			Absent, as all three users have used Outlook 2000.
Modified Date			The modified date is not considered for hash calculation when processing email items.
Number of attachments	0	0	
Number of CC	0	0	The number of recipients in the CC field.
Number of To	2	2	The number of recipients in the To field.
Subject	Re: Financial forecast for next quarter	Re: Financial forecast for next quarter	For files and SharePoint documents, the subject is the file name.
To Display Name	Jim, Steve	Jim, Steve	Enterprise Vault truncates this list to 256 characters. The hash-generation algorithm sorts this field into alphabetical order before calculating the fingerprint.
Message Type	Exchange	Exchange	In version 10.0 and later of the Accelerators, the message type is also included for SharePoint items with a document type of EML.

EFFECT OF CONTENT AND METADATA DEDUPLICATION ON THE 12 CAPTURED EMAILS

When content and metadata deduplication is used, Discovery Accelerator groups the emails into three sets.

Contents of Set 1	Contents of Set 2	Contents of Set 3
Four emails, all identified as "Duplicate": • Steve's copy of Email 1 • Jane's copy of Email 1 • Jim's copy of Email 1 • Journal copy of Email 1	Four emails, all identified as "Duplicate": • Steve's copy of Email 2 • Jane's copy of Email 2 • Jim's copy of Email 2 • Journal copy of Email 2	Four emails, all identified as "Duplicate": • Steve's copy of Email 3 • Jane's copy of Email 3 • Jim's copy of Email 3 • Journal copy of Email 3

EXAMPLE 2

In this example, File System Archiving (FSA) archives two files from two different file shares. Steve and Jane are colleagues whose public file shares are configured to be archived by Enterprise Vault FSA into different FSA archives.

1. Steve creates a Microsoft Word file called **Financial_Forecast_Announcement_Q1.doc**, saves it on his public file share, and mails its location to Jane.

FILE NAME:	Financial Forecast Announcement Q1.doc
LOCATION:	\\FileServer\PublicShare\Steve
MODIFIED DATE:	03-Feb-2011 9:30pm
AUTHOR:	Steve
CONTENT:	We are happy to announce to our shareholders that our… (more than 120 characters of content)

2. Jane asks Steve to copy the file to her public share.

3. Next morning, Steve opens the document and then saves it to Jane's public file share through Microsoft Word. There is now a second copy of the document, as follows:

FILE NAME:	Financial Forecast Announcement Q1.doc
LOCATION:	\\FileServer\PublicShare\Jane
MODIFIED DATE:	04-Feb-2011 8:40am
AUTHOR:	Steve
CONTENT:	We are happy to announce to our shareholders that our… (more than 120 characters of content)

4. Enterprise Vault archives both items in different archives.

5. A Discovery Accelerator search captures both items from Enterprise Vault.

EFFECT OF METADATA-ONLY DEDUPLICATION ON THE TWO ITEMS

Discovery Accelerator identifies the items as different (not "Similar") because they have different Modified Dates. As the following table shows, however, all the other attributes of the items are the same.

Field	Value for file in Steve's folder	Value for file in Jane's folder	Comments
Author Display Name	Steve	Steve	
CC Display Name			Not considered, as the two items are files.
Content	First 120 characters of the content	First 120 characters of the content	Enterprise Vault truncates the content to 120 characters, 118 of which are used for fingerprint calculation. In this example, the first 118 characters of the content of both files match.

Field	Value for file in Steve's folder	Value for file in Jane's folder	Comments
ConversationID			Not considered, as the two items are files.
Modified Date	03-Feb-2011 9:30pm	04-Feb-2011 8:40am	Modified Date is considered for hash calculation, as these are FSA items.
Number of attachments	0	0	
Number of CC	0	0	Not considered, as the two items are files.
Number of To	0	0	Not considered, as the two items are files.
Subject	Financial Forecast Announcement Q1.doc	Financial Forecast Announcement Q1.doc	For files and SharePoint documents, the subject is the file name.
To Display Name			Not considered, as the two items are files.
Message Type	FSA File	FSA File	In version 10.0 and later of the Accelerators, the message type is also included for SharePoint items with a document type of EML.

EFFECT OF CONTENT AND METADATA DEDUPLICATION ON THE TWO ITEMS

When content and metadata deduplication is used, Discovery Accelerator considers the two items to be duplicates. This type of deduplication ignores the Modified Date when it calculates the hash.