# CA DataMinder

# De-duplication Filter for Symantec Enterprise Vault Configuration Guide

## Release 14.6

# CA Technologies Product References

This document references the following CA Technologies products:

- CA DataMinder™

# Contact CA Technologies

**Contact CA Support**

For your convenience, CA Technologies provides one site where you can access the information that you need for your Home Office, Small Business, and Enterprise CA Technologies products. At http://ca.com/support, you can access the following resources:

- Online and telephone contact information for technical assistance and customer services
- Information about user communities and forums
- Product and documentation downloads
- CA Support policies and guidelines
- Other helpful resources appropriate for your product

**Providing Feedback About Product Documentation**

If you have comments or questions about CA Technologies product documentation, you can send a message to techpubs@ca.com.

To provide feedback about CA Technologies product documentation, complete our short customer survey which is available on the CA Support website at http://ca.com/docs.

# Contents

# Chapter 1: About SEV De-duplication Filter Configuration

This section contains the following topics:



## Overview

The De-Duplication Filter for Symantec Enterprise Vault supports Symantec Enterprise Vault 2007. This guide describes the steps required to set up the De-Duplication Filter. Briefly, you need to:

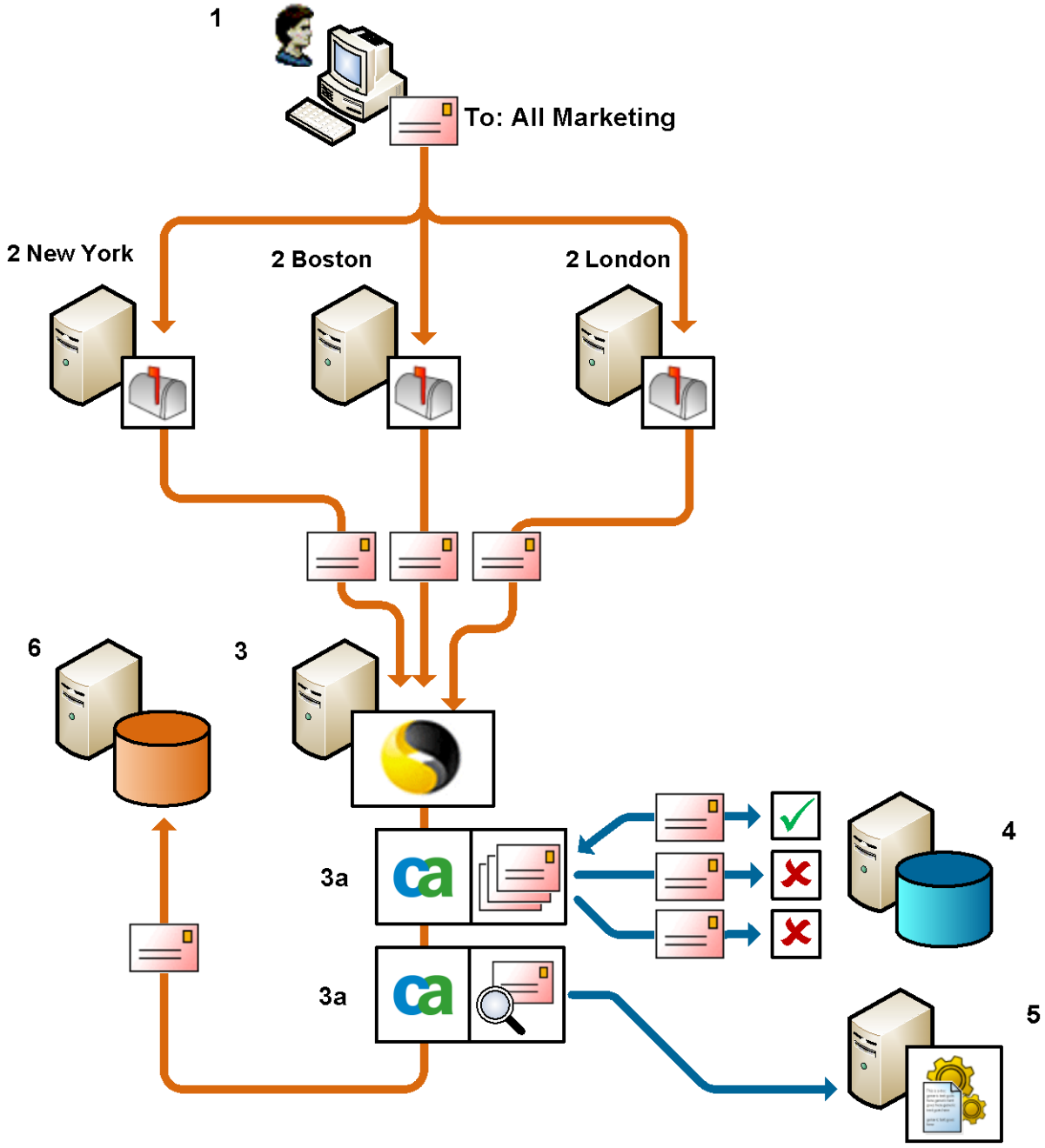
1. Install the De-Duplication filter. This involves the following steps:
    a. Install the De-Duplication database.
    b. Install the De-Duplication filter.
    c. Register the De-Duplication filter.
2. Configure the De-Duplication filter. This involves the following steps:
    a. Configure the De-Duplication filter.
    b. Configure the unique ID property list.

**More information:**

Install the De-Duplication Filter (see page 12)
Configure the De-Duplication Filter (see page 15)
Configure the Unique ID Property List (see page 16)

# De-Duplication Filter Architecture Diagram

The diagram below shows how theDe-Duplication Filter and optionally the EV server agent work with the SEV host server to: process emails from journal mailboxes; remove duplicate emails; optionally integrate with policy engines to apply policy and assign smart tags; and then archive them.



De-Duplication Filter for Symantec Enterprise Vault architecture

1. **Single user**

   An email is sent from a user in the New York office to the 'All Marketing' distribution list.

2. **Journal mailboxes**

   There are members of the distribution list in the New York, Boston and Toronto offices, each of which runs its own email server with its own journal. A copy of the email now exists on each journal server.

3. **SEV host server**

   The SEV server hosts two custom filters: the De-Duplication Filter (**3a**) and, optionally, the EV server agent (**3b**).

4. **De-Duplication database**

   The De-Duplication Filter processes these emails via the De-Duplication database, which stores a unique ID for each email. The ID is identical for each duplicate email, enabling the De-Duplication database to filter out emails already processed by the filter. The New York copy is processed first, and the other two copies are identified as duplicates.

5. **Policy engine**

   After De-Duplication, the single remaining e-mail is passed by the EV server agent and sent to a CA DataMinder policy engine for processing.

6. **Email Archive**

   SEV then sends a single copy of the email to the archive.

# Terminology

**De-Duplication Filter**

The De-Duplication Filter (wgnsevdd.dll) passes e-mails to the De-Duplication database, which stores a unique ID for each email. Using this unique ID, the De-Duplication database can inform the De-Duplication Filter if an e-mail has already been processed. Using the example (De-Duplication filter architecture diagram), the New York copy of the email is processed first, and the other two copies are then identified as duplicates.

**De-Duplication database**

The De-Duplication database enables you to locate and remove duplicate emails. To do this, the database stores a unique ID for each email, which is identical for all duplicate e-mails. This enables the De-Duplication database to filter out e-mails that have already been processed.

In the example (De-Duplication filter architecture diagram), an email is sent from a user in the New York office to the All Marketing distribution list. Without the De-Duplication filter, three identical e-mails would be stored in the email archive, using up valuable storage space. Instead, only a single copy of the email is archived.

**EV Server Agent**

CA DataMinder can integrate with the Symantec Enterprise Vault archive solution through a CA DataMinder custom filter, wgnsev.dll. This is a custom filter for an Exchange Journaling task that works with Enterprise Vault. In this guide, the term 'EV server agent' refers to this custom filter.

**Policy Engine Integration**

You can optionally link the EV server agent to policy engines in order to apply policy and add smart tags to each email. You can then also use the policy engine to generate CA DataMinder email events if required.

# Chapter 2: Installing the De-Duplication Filter

The De-Duplication Filter for Symantec Enterprise Vault (SEV) prevents multiple copies of the same email being archived. For example, if an e-mail is sent to multiple recipients, it is sometimes the case that recipient accounts are hosted across multiple Exchange servers. If these servers host their own journals, there will then be multiple copies of the original email, one in each journal on the recipient mailbox servers, plus another one in the journal on the sender's mailbox server.

Integration with the De-Duplication Filter for Symantec Enterprise Vault (SEV) is provided through a custom filter, wgnsevdd.dll. In this guide, the term 'De-Duplication Filter' refers to this custom filter.

This section contains the following topics:



## Install the De-Duplication Database

**To install the de-duplication database**

1. From CA DataMinder distribution media, copy the contents of the \UA_DDUP folder to any location on the de-duplication database host machine.
2. In the copied \UA_DDUP folder on the de-duplication database host machine, run the following command from a Windows command prompt to create the de-duplication database, tables and stored procedure:

   ```
   ua_dbms <user> <password>
   ```

   where <user> and <password> parameters are the SQL Server de-duplication database user credentials. This is the admin user running the EV server agent.
3. Using the SQL Enterprise Manager, assign the DB_Owner privilege to the SEV admin user for the UA_DDUP database. This privilege provides the SEV admin user with sufficient access rights to the de-duplication database.

**Note:** The SEV admin user is the user created when you installed the Enterprise Vault.

# Install the De-Duplication Filter

This section describes how to install the De-Duplication Filter using the Windows De-Duplication Filter installation wizard.

**To launch the De-Duplication Filter installation wizard**

1. Run sevddup.msi.

   Find this in the \Win32\Support\Sevddup folder on your distribution media.

2. In the Customer Information screen, enter your user name and organization. This information is required for licensing purposes.

3. In the Custom Setup screen, choose the CA De-Duplication Filter feature.

4. In the De-Duplication Database Location screen, specify the name or IP address of the De-Duplication Database Server.

   This is the De-Duplication database created in the previous section.

5. The installation wizard now has all the information it needs. Click Install to start the file transfer.

**Note:** By default, the De-Duplication Filter is set up in test mode.

# Register the De-Duplication Filter

The De-Duplication Filter is a custom filter for an Exchange Journaling task that works with Enterprise Vault. After installing the De-Duplication Filter, you must manually register this filter with Enterprise Vault.

**To register the De-Duplication Filter**

1. Edit the registry on the Agent machine.

   On the Enterprise Vault server, modify values in the Journaling registry key:

   `HKEY_LOCAL_MACHINE\SOFTWARE\KVS\Enterprise Vault\External Filtering\Journaling`

   If the \Journaling registry key does not already exist, you must create it.

2. Add a new string value with these details to the registry key:

   **Name**

   Set this to be 1.

   The number determines the filter processing order, with registry value 1 being processed first.

   The De-Duplication Filter **must** always be the first filter that Journaled emails go through in order to avoid processing duplicate emails any other filter.

   If you already have another filter in \journaling, you must rename that filter's registry value's name to 2.

   **Type**

   REG_SZ

   **Data**

   Set this to be:

   WgnSEVDD.WgnSEVDDFilter

3. Configure the EV archive agent, the hub, and the policy engines.
4. Restart all Exchange Journaling tasks on the Enterprise Vault server.

   Restarting these tasks starts CA DataMinder integration with Enterprise Vault.

# Chapter 3: Configuring the De-Duplication Filter

This section contains the following topics:



## Configure the De-Duplication Filter

To configure the De-Duplication filter, you need to edit registry values located in the following registry key:

```
HKEY_LOCAL_MACHINE\Software\ComputerAssociates\CA DataMinder
    \CurrentVersion\SEV_Deduplication
```

Within this registry key, edit these registry values:

**EnableFilter**

Type: REG_DWORD

Data: Defaults to 0. Set this to 1 to enable the De-Duplication filter.

**TestMode**

Type: REG_DWORD

Data: Defaults to 1. This value enables you to test that the De-Duplication filter is working correctly before you go live. It mimics how the filter will operate when an email is processed. For example, if this value does not exist or is set to 1 then, when a duplicate e-mail is detected, the filter logs the action that would be taken had TestMode not been enabled. However, no emails are actually deleted.

This value must exist and be set to 0 to disable test mode and enable normal operation of the De-Duplication filter, where duplicate emails are actually deleted.

**EliminateArchiveDuplicates**

Type: REG_DWORD

Data: Defaults to 0. Set this value to 1, to delete duplicate emails from the SEV archive. This value does not control the De-Duplication of e-mails from the CA DataMinder database, which always takes place, unless TestMode is enabled.

**LogLevel**

Type: REG_DWORD

Data: Defaults to 2. This determines the level of logging for the De-Duplication filter. For example, you can configure the De-Duplication filter to only log errors or important system messages. Log entries are written to WgnSevDD_<date>.log file, where <date> is the date and time when the log file was created. The location of the log file is specified in the LogFilePath registry value.

Supported log levels are:

- Errors only
- Errors and warnings
- Errors, warnings, plus any extra information

**LogFileMaxBytes**

Type: REG_SZ

Data: Defaults to 1,000,000 (1MB). Specifies the maximum size for each log file. When the current log file reaches its maximum size, the De-Duplication filter creates a new log file.

**LogMaxNumFiles**

Type: REG_DWORD

Data: Defaults to 10. Specifies the maximum number of log files. When the maximum number of log files exists and the maximum size of the most recent is reached, the oldest log file is deleted to enable a new one to be created.

**LogDuplicatesAsWarnings**

Type: REG_DWORD

Data: Defaults to 0. Usually, duplicate emails generate an Informational level log entry to indicate that the e-mail will not be archived as it is a duplicate. If this value is set to 1, then the same logs are generated but at a Warning level.

# Configure the Unique ID Property List

When the De-Duplication Filter processes an email, it applies a SHA-256 cryptographic hash function to a set of MAPI properties from the email. From this, it generates a string that represents a digital fingerprint of that e-mail. In De-Duplication terms, this fingerprint is the unique ID, which is stored in the De-Duplication database.

Each time an email is processed, the De-Duplication Filter compares the unique ID with those already stored in the De-Duplication database. If these differ, the email is stored in the database; if they are the same, the email is deleted as a duplicate and the De-Duplication Filter is informed.

Each MAPI property used to generate the unique ID must be defined in the following registry key:

```
HKEY_LOCAL_MACHINE\Software\ComputerAssociates\CA DataMinder
  \CurrentVersion\SEV_Deduplication\UniqueIDPropListExch
```

Within this registry key, you can define the message properties that will uniquely identify the email.

**HashOrder**

Type: REG_SZ

Data: This value specifies a comma-separated list of MAPI properties used to generate a unique ID for each email processed. For normal use, we recommend you use the default configuration values of:

```
InternetMessageID, ClientSubmitTime
```

## Define a MAPI Property

MAPI properties exist in groups and either have an ID or a name specific to that group. This section defines the three types of MAPI property used most frequently.

**COMMON**

These are the built-in MAPI properties. These properties must be referenced in the following format:

```
<COMMON>, <ID=0xvalue>
```

For example, to specify the MAPI property Internet Message ID, use the following:

```
COMMON, ID=0x1035001E
```

**PS_PUBLIC_STRINGS {00020329-0000-0000-C000-000000000046}**

This property must be referenced in the following format:

```
<PS_PUBLIC_STRINGS>,<ID=0xvalue or NAME=value>
```

You must provide the MAPI group, followed by either the property ID or the property name. For example:

```
PS_PUBLIC_STRINGS, NAME=CustomProperty1
```

**PS_MIME_HEADERS {00020386-0000-0000-C000-000000000046}**

This property must be referenced in the following format:

```
<PS_MIME_HEADERS>, <ID=0xvalue or NAME=value>
```

You must provide the MAPI group, followed by either the property ID or the property name. For example:

```
PS_MIME_HEADERS, NAME=X-Archive-ID
```

The example above will use the value set by the Internet-transmittable MIME tag X-Archive-ID.

**Note:** For any other MAPI groups, the GUID must be specified along with either the property ID or the name.

# Create Custom Registry Values

You now need to add registry values for each property to be used to generate the unique ID. You can optionally create a new registry value if you want to specify an alternative folder for the log files.

**Properties used to generate the unique ID**

You must add registry values for each property to be used to generate the unique ID. To use the properties for Internet Message ID and Client Submit Time, create two registry values in this registry key and configure them to specify those properties:

```
HKEY_LOCAL_MACHINE\Software\ComputerAssociates\CA DataMinder
  \CurrentVersion\SEV_Deduplication\UniqueIDPropListExch
```

For example:

**InternetMessageID**

**Type:** REG_SZ

**Data:** COMMON, ID=0X1035001E

**ClientSubmitTime**

**Type:** REG_SZ

**Data:** COMMON, ID=0X00390040

You can rename these registry values if required.

**Alternative log file location**

If you want to specify an alternative folder for log files, create the following value in this registry key:

```
HKEY_LOCAL_MACHINE\Software\ComputerAssociates\CA DataMinder
  \CurrentVersion\SEV_Deduplication
```

**LogFilePath**

**Type:** REG_SZ

**Data:** Defaults to CA's \data\log subfolder of the Windows All Users profile. Specify the folder where you want to write log files.

# Performance Counters

The De-Duplication filter provides the following performance counters to monitor the filter operation.

**Messages Processed**

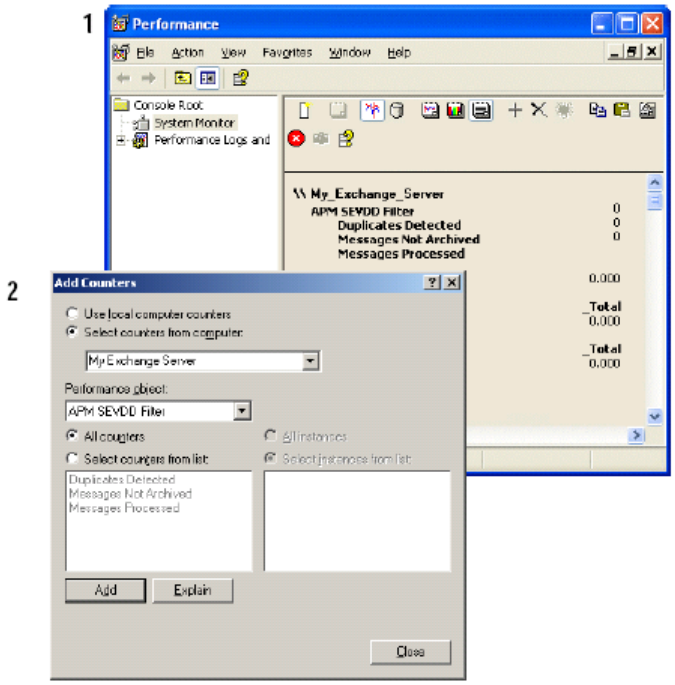
This counter is incremented for each email passed to the De-Duplication filter. It will be reset to zero on restart of the Symantec Enterprise Vault archive server process in which the filter is loaded.

**Duplicates Detected**

This counter is incremented for each email recognized as a duplicate.

**Messages not Archived**

This counter is incremented for every duplicate email deleted from the archive.




**Performance object counters for Universal Adapter**

1. Performance applet, accessible from Administrative Tools. This displays counters for the De-Duplication filter inputs and outputs.
2. Add Counters dialog. For each performance object, specify which counters and, where relevant, instances you want to view.

# Index