

ViewpointTM Deduplication Methodology

Application Version: 7.0
August 19, 2026

© 2021 Conduent, Inc. All rights reserved. Conduent and Conduent Agile Star are trademarks of Conduent, Inc. and/or its subsidiaries in the United States and/or other countries.

Other company trademarks are also acknowledged.

Document Version: 7.0 (August 2021).

Document Conventions

Convention	Explanation
Bold	For file names, commands, fields, menus, options, and window names.
<code>Courier New</code>	Commands as you should type them.
<code>Lucida Console</code>	Example output generated by the system.
<i>Italics</i>	For configuration variables, including variable portions of file names and URLs. Also indicates a document name.
Note / Blue Callout	The Blue Callout text indicates information that is of special interest or importance, an idea that could be useful or additional information about a product or a feature.
	The Caution icon along with the text indicates actions that can lead to problems in system operation or configuration settings if the instructions are not followed properly.

Document Revision History

This section tracks the initial creation of the document after each major version thereafter.

Ver:	Date	Description	Reviewed / Approved By
1.0	Feb 28, 2020	Initial Version	Team
1.1	August 10, 2021	Rebranded adhering to the latest Conduent brand central documentation standards/guidelines.	Technical Writer

Table of Contents

- 1. Deduplication Methodology.....1-4
 - 1.1. Email Deduplication.....1-4
 - 1.1.1. Deduplication Method.....1-4
 - 1.2. E-file Deduplication (non-email).....1-4

1. Deduplication Methodology

Deduplication is the process of identifying logically equal files to cull duplicates from a corpus of documents (a Viewpoint project). The Viewpoint deduplication process utilizes *hashes* for some of this functionality. A hash is a (large) number that represents the binary content of a file. A hash value does not take into account metadata about a given file. Viewpoint supports the use of MD5 or SHA-1 hashes (configurable per project). A file must be hashed to be considered for deduplication. Never hash files using two different algorithms (MD5 or SHA-1) in the same Viewpoint Project – select one algorithm and apply it to the entire Project.

Using the Viewpoint Dedupe Set feature, it is possible to not only identify and suppress duplicates across multiple custodians but also within an individual custodian's data. This allows complete flexibility to manage multiple deduplication methods during the review.

1.1. Email Deduplication

For deduplication, Viewpoint generates a hash for the text file of an email. The text contains the following information:

- To, From, CC and Bcc
- Subject
- Sent date (Time not included in the text file)
- Attachments (hash values of attachments are appended to the end of the text file)
- Body

1.1.1. Deduplication Method

The hash of an email is compared against the hashes of the other emails in the Dedupe Set. The number of attachments and the hash of those attachments are also checked. If all are the same, the family is marked as a duplicate.

1.2. E-file Deduplication (non-email)

An E-file's hash is compared against the hashes of the other E-files in the Dedupe set. If the hash matches another, it is marked as a duplicate.

Note: Standalone E-files are never compared to email attachments in the deduplication process.