Informatica® Cloud Data Quality
December 2022

Deduplicate assets

Informatica Cloud Data Quality Deduplicate assets
December 2022
December 2022

Publication Date: 2022-12-13

Table of Contents

Preface

Refer to *Deduplicate* for information on how to analyze the degrees of similarity between records and how to create a single, preferred record from similar records. A deduplicate asset examines the identity information in each record and creates assignments between the records based on their similarity to each other. The asset then uses criteria that you define to create a preferred record from each set of similar records. To perform the analysis and to consolidate similar records into a single record, add the asset to a Deduplicate transformation in a mapping in Data Integration.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Documentation

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

If you have questions, comments, or ideas about the product documentation, contact the Informatica Documentation team at infa_documentation@informatica.com.

Informatica Intelligent Cloud Services web site

You can access the Informatica Intelligent Cloud Services web site at http://www.informatica.com/cloud.

Informatica Intelligent Cloud Services Communities

Use the Informatica Intelligent Cloud Services Community to discuss and resolve technical issues. You can also find technical tips, documentation updates, and answers to frequently asked questions.

Access the Informatica Intelligent Cloud Services Community at:

https://network.informatica.com/community/informatica-network/products/cloud-integration

Developers can learn more and share tips at the Cloud Developer community:

https://network.informatica.com/community/informatica-network/products/cloud-integration/cloud-developers

Informatica Intelligent Cloud Services Marketplace

Visit the Informatica Marketplace to try and buy Data Integration Connectors, templates, and mapplets:

https://marketplace.informatica.com/

Informatica Knowledge Base

Use the Informatica Knowledge Base to find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit https://search.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Intelligent Cloud Services Trust Center

The Informatica Intelligent Cloud Services Trust Center provides information about Informatica security policies and real-time system availability.

You can access the trust center at https://www.informatica.com/trust-center.html.

Subscribe to the Informatica Intelligent Cloud Services Trust Center to receive upgrade, maintenance, and incident notifications. The Informatica Intelligent Cloud Services Status page displays the production status of all the Informatica cloud products. All maintenance updates are posted to this page, and during an outage, it will have the most current information. To ensure you are notified of updates and outages, you can subscribe to receive updates for a single component or all Informatica Intelligent Cloud Services components. Subscribing to all components is the best way to be certain you never miss an update.

To subscribe, go to https://status.informatica.com/ and click **SUBSCRIBE TO UPDATES**. You can then choose to receive notifications sent as emails, SMS text messages, webhooks, RSS feeds, or any combination of the four.

Informatica Global Customer Support

You can contact a Customer Support Center by telephone or online.

For online support, click **Submit Support Request** in Informatica Intelligent Cloud Services. You can also use Online Support to log a case. Online Support requires a login. You can request a login at https://network.informatica.com/welcome.

The telephone numbers for Informatica Global Customer Support are available from the Informatica web site at https://www.informatica.com/services-and-training/support-services/contact-us.html.

CHAPTER 1

Introduction to deduplicate assets

Use a deduplicate asset to analyze the levels of duplication in a data set and optionally to consolidate sets of duplicate records into a single, preferred record. In a data quality context, an *identity* is a group of data values in a record that identify a person or an organization.

You create and test a deduplicate asset in Data Quality. You add the asset to a Deduplication transformation in a mapping in Data Integration. When the mapping runs, Data Integration performs the deduplication and consolidation processes that the asset defines.

The deduplicate asset specifies the type of identity that the transformation looks for at run time. The type of identity that you select determines the input fields that the transformation will analyze.

In the deduplication process, Data Integration generates a set of numerical values that represent the degrees of similarity between the records in the data set. When two or more records match each other with a score that exceeds a threshold value, Data Integration adds them to a set of potential duplicates. You can specify the threshold value when you configure the asset.

In the consolidation process, the Deduplication transformation evaluates the sets of matching records that the deduplication process creates. The transformation selects or constructs a preferred or optimal version of the records in each set according to the criteria that you specify.

Use a deduplicate asset in the following data projects:

- Customer Relationship Management. For example, a store designs a mail campaign and must check the customer database for duplicate customer records.
- Regulatory compliance initiatives. For example, a business operates under government or industry regulations that insist all data systems are free of duplicate records.
- Financial risk management. For example, a bank may want to search for relationships between account holders.
- Any project that must identify or eliminate records that store duplicate identity information.

Duplicate analysis methodology

A deduplicate asset is a set of instructions for a Deduplicate transformation. When you configure a deduplicate asset, you select the type of identity that the Deduplicate transformation will search for at run time, and you define the search criteria that the transformation will apply to the input data.

The deduplicate asset provides a list of identity types that you must choose from. Each identity type requires a different set of input fields at run time. Each input field represents a different type of information about the identity.

When you configure the deduplicate asset, consider the following rules and guidelines:

- To determine the similarity between the identities in a set of records, the Deduplicate transformation creates an index from the values on the input fields that are relevant to the identity. The index contains raw values from the input fields and a range of alternative versions of the raw values.

  The transformation performs duplicate analysis on the index and calculates a numerical score for each pair of records in the input data set. A pair of identical records will return a score of 1.00.
- You identify the field on which the deduplication process will build the identity index. You also select the options that determine the levels of speed, performance, and depth of the duplicate analysis.
- The Deduplicate transformation compares field data from every input record with data from every other record in the input data set. The transformation begins with the first record in the data set, or the record with the lowest sequence ID value if one is available. If the transformation finds any record that matches the input record, it adds the records to a discrete set of matching records. The discrete set is called a cluster.

  The transformation proceeds to the next record in the data set and repeats the process. If the second record matches any record in the data set, it creates a cluster for the records. Alternatively, if the second record matches a record that is part of a cluster, the transformation adds the record to the current cluster.

  The process continues for all records in the input data set. In this manner, each record in the input data set is compared with every other record in the data set.
- The Deduplicate transformation uses a threshold value to identify records that match each other. The threshold is a percentage value that you can specify in the deduplicate asset. The default value is 0.9, or 90 percent. If two records match each other with a score that meets or exceeds the threshold, the transformation identifies the records as duplicates.
- The final contents of the clusters depends on the type of deduplication output that you specify. You can configure the transformation to retain every pair of matching records in the clusters that it creates. Or, you can configure the transformation to retain only the records that are the closest matches with each other. Set the type of deduplication outputs to return as an advanced option on the **Deduplication** tab.

Deduplicate asset structure

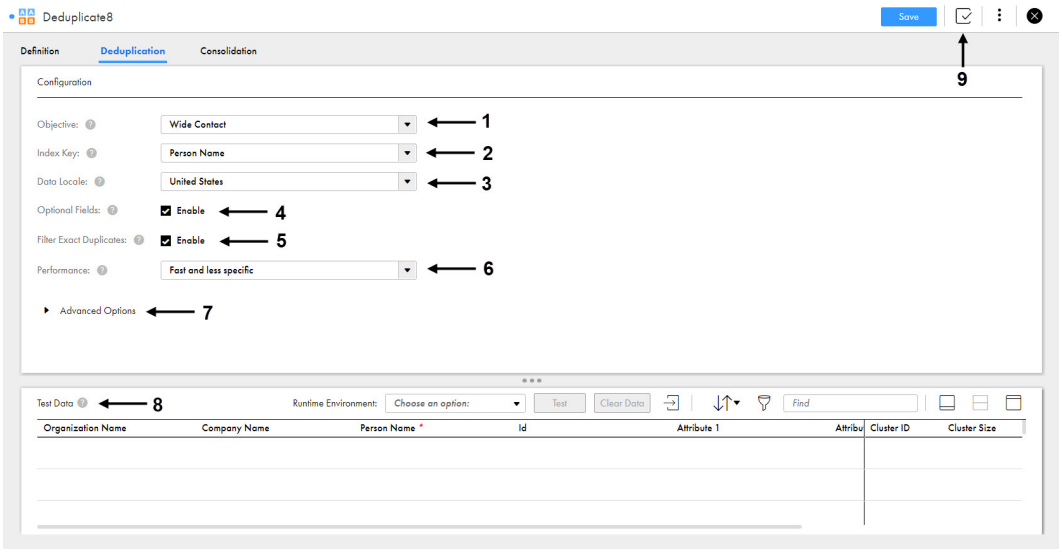
A Deduplicate asset contains options on a **Definition** tab, a **Deduplication** tab, and a **Configuration** tab.

Use the Definition tab options to enter a name for the asset, optionally enter a description for the asset, and select the folder in which to store the asset. Use the Deduplication tab options to configure the type of identity analysis that a mapping will perform. Optionally, use the Consolidation options to configure the strategy that the mapping will apply to the discrete sets of duplicate records that arise from the identity analysis.

Deduplication tab options

Use the Deduplication tab options to configure the type of duplicate analysis that a mapping will perform.

The following image shows the **Deduplication** tab:



The Deduplication tab includes the following options:

1. Objective.
   Identifies the type of identity information that the Deduplicate transformation will analyze when it compares the input records to each other.

   Note: The objective that you select determines the input fields that asset displays in other options.
2. Index Key.
   Identifies the type of information that the Deduplicate transformation will use to create an index of the input records. Select the most relevant type of identity information as the index key. The source data that the mapping reads must include a field that contains the information.
3. Data Locale.
   Identifies the county or region for which the Deduplicate transformation loads identity population reference data. You can select one of the following options: Arabic, Australia, Brazil, Canada, France, Germany, Japan, the United Kingdom, the United States, and International.
4. Optional fields.
   Enables the Deduplicate transformation to use additional input fields to create an index of input records at run time.
5. Filter Exact Duplicates.

   Determines whether the transformation applies the comparison algorithm in a match strategy to pairs of identical records in the input data. When you select the option, the Deduplicate transformation passes records that are exact duplicates of each other directly to the consolidation stage or to the downstream objects in the mapping without additional analysis.

   The Deduplicate transformation output contains the same record data when you select or clear the option.
6. Performance.

Indicates the relative speed and depth of the identity analysis. The default value is fast and less specific, which delivers reasonable analytical depth and faster mapping execution. Other performance options enable more exhaustive identity analysis with correspondingly longer mapping run times.

To view the performance criteria that determine the depth of the identity analysis, expand the Advanced Options. To customize the performance criteria, select Custom as the performance option.

7. Advanced Options.

   Displays the performance criteria that the Deduplicate transformation applies to the input data at run time.

   For more information about the performance criteria, see "Advanced options on the Deduplication tab" on page 9.

8. Test data panel.

   Shows the fields for which the asset expects data, based on the objective that you select. Each field appears as a column name. To test the similarity between data records, enter two or more rows of data. Populate at a minimum each mandatory field and one or more of any required fields that the asset specifies for the current objective.

   The panel also includes a Runtime Environment option and options to search, sort, and filter the test data. Use the Runtime Environment option to specify a Secure Agent.

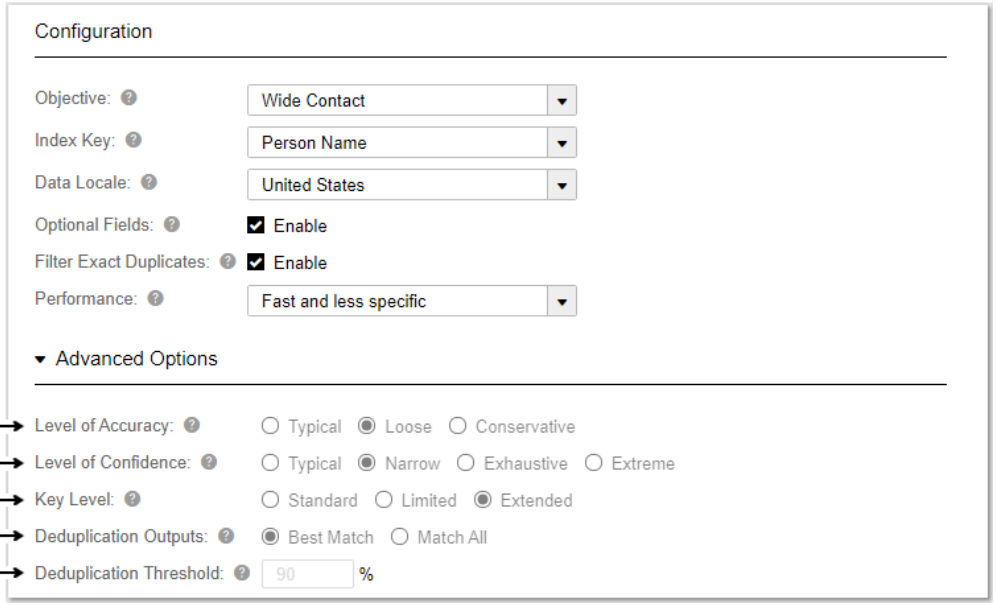
9. Validation.

   Verifies that a deduplicate asset is ready for use in a Deduplication transformation.

Advanced options on the Deduplication tab

The advanced options display the performance criteria that the Deduplicate transformation defines for duplicate analysis at run time. The Performance field value determines the criteria. Expand the advanced options to view the criteria for the Performance field value. To update the criteria that apply at run time, select the Custom performance value.

The following image shows the advanced options:



You can review or update the following options:

1. Level of Accuracy.

   Determines the degree of similarity that must exist between two identities before the deduplication process considers them to be good matches with each other.

   Conservative accuracy requires a very high degree of similarity. Typical accuracy requires a reasonable degree of similarity. Loose accuracy allows for a wider degree of latitude when evaluating the similarity between identities.

2. Level of Confidence.

   Determines the extent of the differences between data values that the deduplication process will tolerate. The level of confidence represents the level of error tolerance that the deduplication process supports.

   Typical confidence configures the process to find common errors and variations in data values. The typical confidence level provides a practical balance between quality and mapping execution time.

   Narrow confidence configures the process to find a narrow range of errors. Narrow confidence prioritizes mapping speed and may miss some duplicate identities.

   Exhaustive confidence configures the process to find a broader range of errors than the typical analysis. Mappings with exhaustive confidence may take longer to run.

   Extreme confidence configures the process to use every possibility to find a candidate match. Select the extreme level when you have a critical need to find every possible duplicate in the input data. Extreme confidence extends the time that the mapping takes to run.

3. Key Level.

   Determines the complexity of the index keys that the deduplication process creates.

   Standard-level keys address most variations in word order, missing words, and extra words. They also maximize the likelihood of finding candidate matches in cases of severe spelling errors in multi-word names.

   Extended-level keys improve match reliability by finding matches regardless of word order or concatenation. Extended keys increase disk space requirements and result in larger sets of matching candidates.

   Limited Keys are a subset of standard keys. Limited keys reduce the use of disk space but may also reduce the reliability of identity search operations.

4. Deduplication Outputs.

   Determines the composition of the clusters of matching records that the transformation creates at run time. Select Best Match to create a cluster that contains only the best match for each record in the cluster. Select Match All to create a cluster that contains all records in the input data that match each record in the cluster.

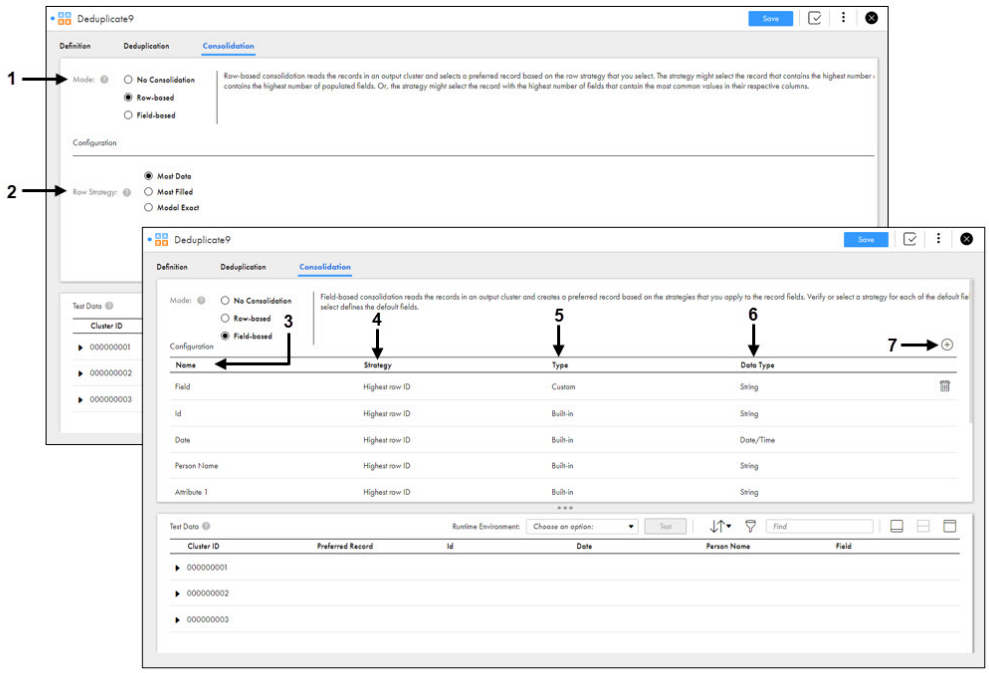
5. Deduplication Threshold.

   Specifies the minimum score between two records that identify the records as plausible duplicates of each other.

Consolidation tab options

Use the Consolidation tab options to configure the type of consolidation that a mapping will perform.

The following image shows the Consolidation tab options:



The Consolidation tab includes the following options:

1. Consolidation mode.

   Identifies the type of consolidation that the Deduplicate transformation will perform when the mapping runs. The type that you select determines how the transformation selects the preferred record in each set of duplicate records.

   Choose the row-based option to select a preferred record based on the quantity of data in the identity fields. Choose the field-based to build a preferred record from the data values across one or more records. You can also choose not to consolidate the duplicate record sets.

2. Row strategy.

   Determines how the transformation will select the preferred record when you choose the row-based consolidation mode.

   Choose Most Data to specify the record with the greatest number of characters as the preferred record. Choose Most Filled to specify the row with the highest number of populated fields. Choose Modal Exact to select the record with the highest number of fields that contain the most common values in their respective columns.

3. Field name column.

   Lists the fields in the input records that the Deduplicate transformation will read. The field name column is visible when you select the field-based consolidation mode. You can specify a consolidation strategy for each field when you select field-based consolidation.

4. Strategy.

Determines how the transformation selects the value in each field for the preferred record when you choose the field-based consolidation mode.

You can select one of the following strategies:

- Highest row ID. Use the value from the record with the highest row ID or sequence ID. Highest row ID is the default strategy.
- Average. Use the average value across the records.
- Longest. Use the longest value in the field across the records.
- Maximum. Use the highest number in the field across the records. Or, choose the last value in alphabetical order.
- Minimum. Use the lowest number in the field across the records. Or, choose the first value in alphabetical order.
- Most frequent. Use the most frequently-occurring value in the field across the records, including blank, empty, or zero-length string fields.

  Note: The consolidation operation will not add a null value to the preferred record.
- Most frequent non-blank. Use the most frequently-occurring value in the field across the records, excluding null, blank, empty, or zero-length string fields.
- Shortest. Use the shortest value in the field across the records.

5. Type

   Indicates whether the asset created the field during the deduplication operation or whether you added the field to the asset in the Consolidation pane.

6. Data Type

   Identifies the data type of the field. The default data type on all fields is String. You can modify the data type in field-based consolidation to suit your data requirements.

   You can select one of the following data types for a field:

   - Date/Time
   - Float
   - Integer
   - String

   Note: If you modify the data type of a field, do not change the mode to No Consolidation or Row-based consolidation without first saving the asset. The asset discards any update that you make to a data type in Field-based consolidation mode if you change to another mode.

7. Add field button.

   Adds one or more inputs to the consolidation operation when you select the field-based mode.

   Add fields in the following cases:

   - The identity analysis that you define on the **Deduplication** tab does not cover all of the fields that the transformation will analyze.
   - You want to specify a non-default strategy for the additional fields.

   The field-based strategies will apply to every input field that you map to the deduplicate asset in the transformation. If you do not specify a strategy for a field in field-based mode, the transformation applies the default strategy.

Deduplication objectives

The **Objective** option on the **Deduplication** tab defines the type of identity that the Deduplicate transformation analyzes when you run a mapping with the transformation. The objective also indicates the types of information that the transformation expects to read for the identity. When you configure the transformation, you map the input fields that contain the identity information to the most appropriate fields on the deduplicate asset.

Each objective supports one or more index key fields. You select an index key on the **Deduplication** tab in the asset. Some objectives define similar identity types, for example Family and Household. In each case, the deduplication process uses unique comparison logic.

Each objective additionally supports a set of input fields that can contain additional data about the identity. The set of potential input fields on an objective contains at least one mandatory field and may also identify one or more required fields. You must map each mandatory field on the asset to a transformation input field. You must map at least one required field on the asset to a transformation input field. Optionally, map any additional asset field to a transformation field that contains the appropriate information. To optimize the analysis of the identity data, ensure that the Deduplication transformation reads as many of the fields as possible.

For more information about mandatory and required fields on each objective, see Appendix A, “Field types and identity objectives” on page 21.

The following table descries the objectives that you can select and identifies the fields that you can select as the index key field:

Identity objective	Description	Index Keys
Address	Identifies records that share an address.	Address Part 1 Date Geocode
Author ISBN	Identifies records that share information about an author who published work with an ISBN number.	Email Person Name
Contact	Identifies records that share a contact at a single organization and location.	Address Part 1 Company Name Date Email Geocode Person Name Organization Name
CC Issuer	Identifies records that share information about a credit card issuer.	Organization Name
CC Owner	Identifies records that share information about a credit card holder.	Address Part 1 Date Email Geocode Person Name
Corp Entity	Identifies records that share corporate identification data.	Address Part 1 Organization Name

Identity objective	Description	Index Keys
Division	Identifies records that share an office location within an organization.	Address Part 1 Company Name Geocode Organization Name
Family	Identifies individuals that belong to the same family.	Address Part 1 Email Geocode Person Name
Fields	Identifies records that share identity data across multiple fields that you select.	Address Part 1 Company Name Date Email Geocode Organization Name Person Name [Generic field]
Generic	Identifies records that share identity data on a field that you select.	[Generic field]
Geocode	Identifies records that share geocode data.	Geocode
Household	Identifies individuals that belong to the same household.	Address Part 1 Email Geocode Person Name
Individual	Identifies duplicate individuals.	Date Email Person Name
Organization	Identifies records that share organization data.	Address Part 1 Company Name Date Geocode Organization Name
Person Name	Identifies records that share information about individuals.	Address Part 1 Date Email Geocode Person Name
Publisher ISBN	Identifies records that share information about a publishing company. The information includes ISBN data for published works.	Address Part 1 Company Name Geocode Organization Name

Identity objective	Description	Index Keys
Resident	Identifies duplicate individuals at the same address.	Address Part 1 Date Email Geocode Person Name
VIN Manufacturer	Identifies records that share information about a vehicle manufacturer.	Address Part 1 Company Name Geocode Organization Name
VIN Owner	Identifies records that share information about a vehicle owner.	Address Part 1 Company Name Date Email Geocode Organization Name Person Name
Wide Contact	Identifies records that share a contact at an organization.	Company Name Email Person Name Organization Name
Wide Household	Identifies individuals that belong the same household.	Address Part 1 Email Geocode Person Name

Identity reference data

The duplicate analysis process use reference data files called population files to evaluate the identity information in the input data. When you run a mapping with a Deduplicate transformation, the transformation compares the input data to the population file for the country in which the input data originates.

Data Quality downloads the population files to the Secure Agent host machine when you install the Secure Agent. You do not need to take any action to download the files.

You can review the location of the population files on the Administrator service.

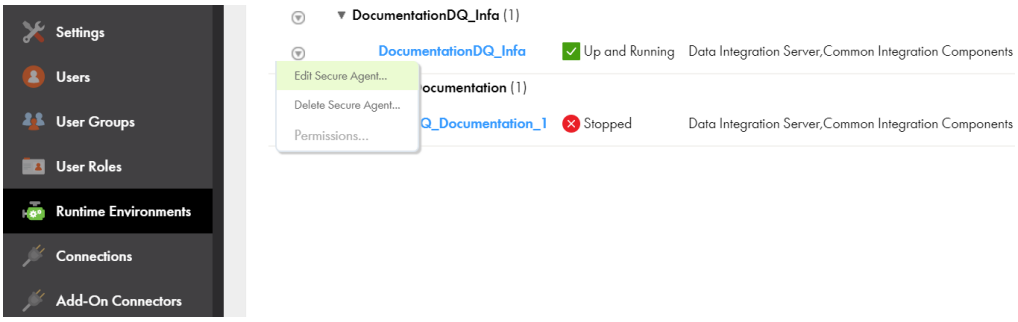
Reviewing the Administrator properties for identity population data

You can review and configure the directory to which the Secure Agent downloads identity population files in the Administrator service. You can also review and configure the directories that the deduplication process uses for index and cache data.

1. From the **My Services** page, select the **Administrator** service.

2. Choose the **Runtime Environments** option.
3. Select the **Secure Agent** that you will use to run mappings with the Deduplicate transformation.
4. Hover over the **Actions** icon for the Secure Agent, and select the **Edit Secure Agent** option.

   The following image shows the option:



   The **Secure Agent** page appears.
5. Under **System Configuration Details**, select **Data Integration Server** in the **Service** field and select **IDQ** in the **Type** field.

   The **System Configuration Details** pane returns a list of properties based on the type that you specified.
6. Review the following properties:
   - IdentityReferenceDataLocation. Identifies the location of the population files.
   - IdentityReferenceIndexLocation. Identifies the location of the index files that the deduplication process creates at run time.
   - IdentityCacheLocation. Identifies the location of the temporary files that the deduplication process creates at run time.

   By default, the properties identify directories in the Informatica installation.
7. Optionally, update the property values to suit your system and the mappings that you will run.

Configuring the deduplication process

To define a deduplication process, configure the options on the **Deduplication** tab.

1. Select the **Deduplication** tab on the asset.
2. Select an objective. The objective represents the type of identity that the Deduplication transformation will look for during duplicate analysis.
3. Select an index key.

   The index key represents the field on which the transformation will build identity data index. The objective that you select determines the set of keys from which you can select the index key.

   Tip: The objective also indicates the set of inputs that the transformation expects to read at run time. You can preview the fields in Test Data panel.
4. Select the data locale in which the data set originates.

   The duplicate analysis process reads identity reference data for the locale that you select.
5. Select or clear the option to define optional fields for the objective.

Select the option if your source data contains one or more columns of relevant data that the objective does not specify. For example, your source data might contain a discrete field for corporate suffixes.

6. Select or clear the option to filter exact duplicates.

   When you select the option, the transformation passes records that are duplicates of each other directly to the consolidation stage or to the downstream objects in the mapping.

   You might select the option when the input data contains many identical rows.

   Note: The output from the analysis contains the same records whether you select or clear the option. The Deduplicate transformation might assign different scores to the output records when you select and clear the option.

7. Select a level of performance for the duplicate analysis.

   The performance level describes the relationship between the speed and the granularity of the analysis. Faster analysis is less granular and might miss some duplicate records.

8. Optionally, review or update the criteria that apply for a performance option.

   To review the criteria, select the option and expand the **Advanced Options**.

   To customize the criteria, select the **Custom** option as the performance level. For example, you might decide to update the threshold score value.

9. Save the asset.

After you configure the deduplication process, you can optionally configure a consolidation process for the duplicate records that deduplication identifies.
For more information about the deduplication options, see "Deduplication tab options" on page 8.

Configuring the consolidation process

To define a consolidation process, configure the options on the **Consolidation** tab.

1. Select the **Consolidation** tab on the asset.
2. Select a consolidation mode. The mode defines the method that the Deduplicate transformation uses to consolidate sets of duplicate records into single, preferred records.

   You can select one of the following modes:

   - No consolidation. The transformation does not create preferred records from the sets of matching records that the duplicate analysis defines.
   - Row-based consolidation. The transformation selects a preferred record from each set of matching records based on a strategy that you select.
   - Field-based consolidation. The transformation creates a preferred record based on a strategy that you specify for each field in the records. Field-based consolidation can combine values from more than one record to create a preferred record that is not present in the source data.

3. Configure the row-based or field-based strategies for the fields in the input records.
4. If you selected the field-based consolidation mode, optionally add one or more fields to the consolidation model.
5. Save the transformation.

For more information about the consolidation options, see "Consolidation tab options" on page 11

CHAPTER 2

Validating and testing deduplicate assets

Validate a deduplicate asset in Data Quality before you add it to a Deduplicate transformation in a mapping.

Test the asset in Data Quality to verify that the asset logic generates the results that you expect.

To test an asset, you must have access to an active Secure Agent. Use the Runtime Environment option to specify the Secure Agent.

Validate a deduplicate asset

Validate a deduplicate asset to verify that the asset is ready for use in a Deduplicate transformation.

1. Open the deduplicate asset.
2. On the **Deduplication** tab or **Consolidation** tab, click **Validation**.

If the validation process reports any error, fix the error before using the asset.

Testing a deduplicate asset

Test a deduplicate asset to verify that the data flows through the asset in the ways that you expect.

1. Open the deduplicate asset.
2. Select the **Deduplication** tab.
3. Select a runtime environment in which to perform the test.
4. Enter data values in the test panel, or import the data to test. To import the data, click the **Import** option in the test panel.

   Consider the following guidelines before you add or import data to the test panel:

   - If you add a string composed of over 255 characters, the asset will use the first 255 characters.
   - Configure the asset as completely as you can before you enter or import the test data. The asset may discard your test data if you make further changes to the configuration.

- Verify that the test data structure matches the column structure in the test panel. Provide test data for each mandatory field in the test panel. If the test panel includes any required fields, provide test data for at least one required field.
- If the objective accepts a date value, you can use the calendar and clock options to add a date and time to an input row. The calendar and clock options are synchronized. You can update the date or time after you add either value.

For more information see "Rules and guidelines for importing test data" on page 20.

5. Optionally, save the asset to preserve the current test data and configuration.
6. Click **Test**.

   You can sort, search, and filter the test results in the following ways:

   - Click the Up and Down arrows to select a field on which to sort the test data. To reverse the sort order, select the field a second time. open a menu of the sorting categories. The categories reflect the configuration of the asset. The test results refresh to show the values for the category that you choose.
   - Click the Filter icon to add a filter option to the test panel. Select a field on which to filter the data, and add a data filter for the field. The test results refresh to show any row that contains the filter value in the field that you specify.
   - You can also enter a value in the Find field to search the test results.

   The sort, search, and filter options work together. For example, if you apply a filter to the test data and you enter a value in the Find field, the asset displays any row that meets both the filter and search criteria.
7. Verify that the deduplication process analyzes the test data in the manner that you expect.
8. Optionally, select the **Consolidation** tab.

   Use the options on the tab to verify that the consolidation process generates preferred records in the manner that you expect.

   The test panel on the **Consolidation** tab contains the results of any test that you ran on the **Deduplication** tab.
9. Click **Test**, and review the results of the test. You can test the consolidation options in row-based mode and field-based mode.

Understanding the test results

When you run a test on the **Deduplication** or **Consolidation** tab, the test results include a number of predefined fields. The input data and the test results on the Deduplication tab form the basis of the test that you can perform on the Consolidation tab.

The test results on the Deduplication tab include the following predefined fields:

Cluster ID

Contains the identifier of the cluster to which the input record belongs.

In the deduplication process, a cluster is a set of records whose data values match each other to a degree that exceeds the duplicate threshold. Records in the same set are likely to identify the same identity. A set may contain a single record, as every unique record is a perfect match with itself.

Cluster Size

Contains the number of records in the set to which the current record belongs. When a set contains a unique record, the cluster size is 1.

The test results on the Consolidation tab include the following predefined fields:

Cluster ID

Contains the identifier of the cluster to which the input record belongs. The Cluster ID fields on the Deduplication and Consolidation tabs contain identical information for a given test.

Preferred Record

Contains the values in the preferred record that the test creates for the current input.

The Deduplication and Consolidation tabs also display the mandatory and required fields that apply for the objective and index key that you select. In addition, the Consolidation tab displays all of the fields that the objective can use and any custom field that you add.

Rules and guidelines for importing test data

You can import data to the test panel in the deduplicate asset and save the test data in the asset configuration.

Consider the following rules and guidelines when you add data to the test panel:

- The import option supports CSV and Microsoft Excel files.
- You can import up to 200 consecutive rows of data from a delimited file. You can specify the row at which the import starts.

  Note: Before you import, check the file for column headings. If the first row in the import file contains column headings, start the import at line 2 or lower.
- You can import or enter an input string of up to 255 characters.
- If you import a CSV file that contains multiple columns or uses a text qualifier, verify that the file uses a delimiter or a text qualifier that the Secure Agent recognizes. By default, the *Comma* option is the delimiter for the column data. By default, the *No quotes* option is the text qualifier for the data. You can update the delimiter and text qualifier characters when you select the data to import. The Delimiter and Text Qualifier options are not required when you import a Microsoft Excel file.
- The Secure Agent saves the data that you import to the asset when you save the asset. If you change an option in the asset configuration, you may lose any unsaved test data.

The following rules and guidelines apply to the deduplicate asset:

- The test panel structure can change based on the objective and the index key that you select. The structure of the test data that you import must match the structure in the test panel.
  The test panel can include two types of input field:

  - Mandatory: You must populate all mandatory fields with test data.

  - Required: You must populate at least one required field with test data.

  To read a list of the mandatory and required fields on each objective, see Appendix A, “Field types and identity objectives” on page 21.

APPENDIX A

Field types and identity objectives

Each objective that you select on the Deduplication tab contributes data from one or more input fields to the process of identity analysis.

An objective can contribute data from one or more of the following field types:

Mandatory field

A field that you must map to an input on the Deduplicate transformation.

Required fields

A set of fields from which you must map at least one field to an to an input on the transformation.

Optional fields

One or more fields that you can optionally map to transformation inputs.

When you select an objective on the **Deduplication** tab, the test pane displays the mandatory and required fields on the objective. The test pane displays an asterisk (*) beside any mandatory field. The test pane displays a plus symbol (+) beside any required field.

An objective can specify any combination of one or more mandatory, required, or optional fields. The combination of fields can depend on the index key that you select for the objective. The test panel on the Deduplication tab displays the current mandatory and required fields for the objective and index key that you select. The index key is always a mandatory field for an objective.

Tip: If you browse the objectives, the index keys for the objectives can change. Before you save and close a deduplicate asset, verify that the Deduplication tab displays the index key that you intend for the objective that you select.

For more information about the fields that you can choose in each objective, see "Mandatory, required, and optional fields" on page 21.

Mandatory, required, and optional fields

Each objective on the **Deduplication** tab specifies a set of fields that the deduplication process can analyze at run time. You must provide input data for every mandatory field that the objective specifies, and you must provide data for at least one required field. You can also provide data for the other fields that the objective specifies. The tables in this appendix list the sets of mandatory, required, and optional fields that appear by default for each objective.

Note: The mandatory and required fields can vary according to the index key that you select for the objective. The current index key on any objective is always a current mandatory field. For example, the deduplicate asset specifies *Organization Name* as a single index key option for the *CC Issuer* objective, which means that *Organization Name* is mandatory in all cases for *CC Issuer*.

The following tables list the default sets of mandatory, required, and optional fields on each objective:

Address

Field	Field Type
Address Part1	Mandatory
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Date	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional

Author ISBN

Field	Field Type
Person Name	Mandatory
ISBN10	Required, unless you specify ISBN13
ISBN13	Required, unless you specify ISBN10
Email	Optional

CC Issuer

Note: Because Organization Name is the single index key option on the CC Issuer objective, Organization Name becomes a mandatory key in addition to CreditCard.

Field	Field Type
CreditCard	Mandatory
Organization Name	Required

CC Owner

Field	Field Type
Person Name	Mandatory
CreditCard	Mandatory
Telephone Number	Optional
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Date	Optional
Geocode	Optional
Email	Optional

Contact

Field	Field type
Person Name	Mandatory
Address Part1	Mandatory
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Date	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

Corp Entity

Note: Because Organization Name is the single required field on the Corp Entity objective, Organization Name becomes a mandatory field.

Field	Field type
Organization Name	Required
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional

Division

Field	Field type
Address Part1	Mandatory
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional

Family

Field	Field type
Person Name	Mandatory
Address Part1	Mandatory
Telephone Number	Mandatory
Address Part2	Optional

Field	Field type
Postal Area	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

Fields

Note: In the core field list for the Fields objective, all fields are optional. The deduplicate asset identifies the current index key as a manadatory field.

Field	Field type
Person Name	Optional
Organization Name	Optional
Company Name	Optional
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Date	Optional
CreditCard	Optional
VIN	Optional
ISBN10	Optional
ISBN13	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional

Field	Field type
Generic Field	Optional
Email	Optional

Generic

Field	Field Type
Generic Field	Mandatory

Geocode

Field	Field Type
Geocode	Mandatory

Household

Field	Field type
Person Name	Mandatory
Address Part1	Mandatory
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

Individual

Field	Field type
Person Name	Mandatory
ID	Required, unless you specify Date

Field	Field type
Date	Required, unless you specify ID
Attribute1	Optional
Attribute2	Optional
Code	Optional
Email	Optional

Organization

Field	Field type
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Date	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional

Person Name

Field	Field type
Person Name	Mandatory
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional

Field	Field type
ID	Optional
Date	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

Publisher ISBN

Field	Field Type
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
ISBN10	Required, unless you specify ISBN13
ISBN13	Required, unless you specify ISBN10
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Geocode	Optional

Resident

Field	Field type
Person Name	Mandatory
Address Part1	Mandatory
Address Part2	Optional
Postal Area	Optional
Telephone Number	Optional
ID	Optional
Date	Optional

Field	Field type
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

VIN Manufacturer

Field	Field Type
VIN	Mandatory
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
Address Part1	Optional
Address Part2	Optional
Postal Area	Optional
Geocode	Optional

VIN Owner

Field	Field Type
VIN	Mandatory
Address Part1	Mandatory
Organization Name	Optional
Company Name	Optional
Person Name	Optional
Address Part2	Optional
Postal Area	Optional
Attribute1	Optional
Attribute2	Optional
Date	Optional

Field	Field Type
Geocode	Optional
Code	Optional
Email	Optional

Wide Contact

Field	Field type
Person Name	Mandatory
Organization Name	Required, unless you specify Company Name
Company Name	Required, unless you specify Organization Name
ID	Optional
Attribute1	Optional
Attribute2	Optional
Code	Optional
Email	Optional

Wide Household

Field	Field type
Address Part1	Mandatory
Telephone Number	Mandatory
Person Name	Optional
Address Part2	Optional
Postal Area	Optional
Attribute1	Optional
Attribute2	Optional
Geocode	Optional
Code	Optional
Email	Optional

INDEX