Identity Resolution with Informatica Data Quality 9.0.1
Abstract

You can use Informatica Data Quality to find records that share identity information. An identity is a set of data values that collectively identify a unique entity, such as an individual, household, or organization. This article describes methodologies and best practices for identity analysis in Informatica Data Quality.

Supported Versions

- Data Quality 9.0.1

Table of Contents

Overview ................................................................. 2
How Identity Analysis Works ........................................... 3
Preparing for Identity Matching ........................................ 3
  - Step 1. Profile Source Data ........................................ 3
  - Step 2. Define a Matching Process ................................. 4
Creating an Identity Match Mapping ............................... 4
  - Configuring the Match Transformation .......................... 4
  - Configuring Identity Match Types ............................... 4
  - Configuring Identity Match Strategies .......................... 6
  - Configuring Identity Match Output .............................. 8
Viewing Identity Match Results ................................. 8
Customizing Matching Operations ............................... 9
Conclusion .............................................................. 9

Overview

Use Informatica Developer to build mappings that discover duplicate or matching information in records. You can match records within a data set or across two data sets.

You can define the following types of matching in the Developer tool:

- Identity matching. Finds identities shared by multiple records. Some identity matches indicate record duplication. Other identity matches indicate a connection between records, such as a shared family identity or a shared employer identity.

- Field matching. Finds similar or duplicate data in two or more records.

In both identity matching and field matching, you configure a Match transformation to compare records and generate a numeric score that indicates the degree of similarity. You can compare records within a data set or compare the records in two data sets.

Match Strategies

Compare records by configuring predefined match strategies in the Match transformation.

Match strategies compare data in each record against data in every other record and compute the degree of similarity as a numeric match score. The score is a decimal value between 0 and 1, where a value of 1 indicates a perfect match.

An identity matching strategy can analyze multiple fields in each record.
Each field matching strategy analyzes one field in each record. To use field matching to analyze multiple fields in a record, configure the Match transformation with multiple field match strategies.

When two records contain multiple fields that are perfect matches or close matches, the records are likely to be related.

Thresholds and Clusters

You can set a threshold for match scores to determine the level of similarity that indicates a match.

The Match transformation organizes records with match scores above the threshold into clusters. A cluster is a set of records that contain similar or duplicate data. For a record to enter a cluster, it must link to at least one record in the cluster with a score that exceeds the threshold. The Match transformation assigns records in each cluster a common $clusterID$.

If a record does not generate any match scores above the threshold, the Match transformation assigns the record to its own cluster.

How Identity Analysis Works

To perform identity analysis, configure identity match strategies, populations, and identity key properties.

Identity match strategies identify entities such as individuals, organizations, families, households, and addresses. Some identity match strategies find duplicate records, while others find connected records, such as individuals that belong to the same family. Each identity match strategy reads one or more data columns. For more information about identity match strategies, see “Identity Match Strategies” on page 7.

Identity populations are reference data files that contain key-building algorithms customized for locales and languages. These files do not install with Informatica Data Quality. You must download population data files and install them using the Data Quality Content Installer. Contact your Informatica Administrator user for information about the populations available on your system.

Identity key properties determine how the Match transformation groups records. Grouping is the process of organizing similar records into groups to enable faster match analysis. The Match transformation can create identity group keys from input columns that contain person names, organization names, or addresses.

Preparing for Identity Matching

Successful identity matching depends on successful data analysis earlier in the project.

The high-level prerequisites for identity matching are:

1. Profile source data.
2. Design a matching process.

After you complete these steps, create an identity match mapping to implement the matching process you design.

Step 1. Profile Source Data

Create a profile to analyze the content and structure of source data and to confirm that the data is suitable for identity analysis.

Use Informatica Analyst or Informatica Developer to create and run the profile. When you analyze the profile results, identify the columns you want to use during identity analysis. Verify that these columns contain valid data.
Step 2. Define a Matching Process

Define the objectives for your identity matching process before creating an identity match mapping.

Use the profile analysis to answer the following questions:

- What type of identity are you looking for?
- Which comparison strategies can you use to find this type of identity?
- Which columns contain data that you can analyze with these comparison strategies?
- Do the columns you want to analyze contain usable data?

Creating an Identity Match Mapping

After you profile the source data and define a match process, create an identity match mapping.

1. Create a mapping.
2. Add a data source and Match transformation to the mapping. If your data records do not contain unique sequence IDs, add a Key Generator transformation to create IDs.
3. Set up the Match transformation by connecting the ports and configuring the transformation properties.
4. Run the Data Viewer and view the identity match results.

Note: Unlike field matching, identity matching does not require a group key. If you connect the Key Generator transformation Group Key output port to a Match transformation configured for identity matching, you may change the results of identity matching operations.

Configuring the Match Transformation

Set up the Match transformation for identity matching by connecting source data ports and configuring identity match options.

Connect source data ports to the Match transformation before configuring match strategies. When you configure a match strategy, you must select the source data columns that the strategy analyzes.

After you connect the source data to the Match transformation, configure identity match options in the following Match transformation views:

- Match Type
- Strategies
- Match Output

Configuring Identity Match Types

Use the Match Type view to select an identity match type and configure its properties.

1. In the Match Type view, select the type of identity matching to perform. Select Identity Match (Single Source) or Identity Match (Dual Source) depending on the source data.
2. Configure the properties for the match type you choose.
The following table describes identity match type properties:

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Determines the reference data file that contains key-building algorithms customized for locales and languages.</td>
</tr>
<tr>
<td>Key Level</td>
<td>Determines the number of keys generated by identity matching algorithms.</td>
</tr>
<tr>
<td>Key Type</td>
<td>Describes the category of data contained in the key field. Identity matching can generate keys for person names, organizations, and addresses. Select the key type that best describes the column you choose for the Key Field property.</td>
</tr>
<tr>
<td>Search Level</td>
<td>Indicates the balance of search quality and search speed. The search speed is inversely related to the number of matches returned. For example, the Exhaustive option takes the longest time and returns fewer matches.</td>
</tr>
<tr>
<td>Key Field</td>
<td>Specifies the column that the Match transformation uses to generate group keys. Verify that the column you select contains the kind of data you specified in the Key Type option.</td>
</tr>
<tr>
<td>Index Folder</td>
<td>Specifies the directory where the Match transformation writes the identity key index. The Data Integration service must be able to access this directory.</td>
</tr>
<tr>
<td>Cache Directory</td>
<td>Specifies the directory that contains the identity matching cache. The Data Integration service must be able to access this directory.</td>
</tr>
<tr>
<td>Master Data Set</td>
<td>Specifies the data source that contains the master data. Used in dual source matching only.</td>
</tr>
</tbody>
</table>

**Key Level Settings**

Key level settings determine the balance of disk usage and matching accuracy that the matching process uses during identity matching.

The following table provides the disk space usage, matching accuracy, and usage scenario for each key level setting.

<table>
<thead>
<tr>
<th>Key Level</th>
<th>Disk Space Usage</th>
<th>Matching Accuracy</th>
<th>Usage Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited</td>
<td>Low</td>
<td>Finds likely matches. Does not find all probable matches.</td>
<td>Use on systems with limited disk space.</td>
</tr>
<tr>
<td>Standard</td>
<td>High</td>
<td>Overcomes most variations in word order, missing words, and extra words.</td>
<td>Use for most matching operations.</td>
</tr>
<tr>
<td>Extended</td>
<td>Very high</td>
<td>Finds most possible matches, regardless of word order variation and concatenation.</td>
<td>Use for high-risk or mission-critical matching operations.</td>
</tr>
</tbody>
</table>
Search Level Settings

Search level settings determine the balance of speed and accuracy that the matching process uses during identity matching. The following table provides the search speed, matching criteria, and a description for each search level setting.

<table>
<thead>
<tr>
<th>Search Level Setting</th>
<th>Search Speed</th>
<th>Matching Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrow</td>
<td>Fastest</td>
<td>Nearly exact</td>
<td>Performs the fastest and most exact matches. For example, you can use this search level for person name matching to find exact matches and name abbreviation matches.</td>
</tr>
<tr>
<td>Typical</td>
<td>Fast</td>
<td>Strict</td>
<td>Performs fast searches with strict matching criteria. For example, you can use this search level for person name matching to find data with name abbreviation matches and potential errors such as incorrect initials.</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>Average</td>
<td>Loose</td>
<td>Performs average speed searches with loose matching criteria. For example, use this search level to find matches that may represent substantial spelling errors.</td>
</tr>
<tr>
<td>Extreme</td>
<td>Slow</td>
<td>Very loose</td>
<td>Performs slow searches with very loose matching criteria. For example, use this search level to find matches with a wide variety of spelling errors.</td>
</tr>
</tbody>
</table>

Configuring Identity Match Strategies

Use the Strategies view to configure identity match strategies.

1. In the Strategies view, select Define match strategies.
2. Click New to add one or more identity match strategies to the transformation. The Match Strategy window opens.
3. Choose an identity matching strategy and click OK. The strategy appears in the Strategies view.
4. Click the row that contains the strategy. Clickable selection arrows appear in the row.
5. Click the selection arrow on the Match Fields column. The Match Fields window opens.
6. Double-click the Available column to assign source data columns to identity inputs. To increase match accuracy, populate as many identity inputs as possible.

Important: Each identity strategy has required inputs, indicated by an asterisk next to the input name. You must assign source data columns to these identity inputs.
7. Click **OK**.
8. If you add multiple identity strategies, optionally configure the **Weight** column. You can use the **Weight** setting to change the priority of a strategy in relation to other strategies.

**Note**: You can also define match strategies by choosing a match mapplet in the **Strategies** view. A match mapplet is a type of mapplet that you can embed in a Match transformation. For more information about match mapplets, see the Informatica Developer User Guide.

### Identity Match Strategies

Choose a strategy that satisfies your identity analysis objectives.

The following table describes each identity match strategy and lists the required inputs for each strategy:

<table>
<thead>
<tr>
<th>Identity Match Strategy</th>
<th>Match Operation</th>
<th>Required Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>Identifies records that share an address.</td>
<td>Address</td>
</tr>
<tr>
<td>Contact</td>
<td>Identifies records that share a contact at a single organization location.</td>
<td>Person_Name Organization_Name Address_Part1</td>
</tr>
<tr>
<td>Corp Entity</td>
<td>Identifies records that share organization data. Optionally, configure this strategy to analyze address and telephone data.</td>
<td>Organization_Name</td>
</tr>
<tr>
<td>Division</td>
<td>Identifies records that share an organization office at a specific address.</td>
<td>Organization_Name Address_Part1</td>
</tr>
<tr>
<td>Family</td>
<td>Identifies individuals that belong to a family by analyzing name, address, and telephone number data.</td>
<td>Person_Name Address_Part1 Telephone Number</td>
</tr>
<tr>
<td>Fields</td>
<td>Identifies records that share data for custom fields that you select.</td>
<td>N/A</td>
</tr>
<tr>
<td>Household</td>
<td>Identifies individuals that belong to a household by analyzing name and address data.</td>
<td>Person_Name Address_Part1</td>
</tr>
<tr>
<td>Individual</td>
<td>Identifies duplicate individuals by analyzing name, date of birth, and ID data. ID examples include Social Security numbers, account numbers, and vehicle identification numbers.</td>
<td>Person_Name ID Date</td>
</tr>
<tr>
<td>Organization</td>
<td>Identifies records that share organization data.</td>
<td>Organization_Name</td>
</tr>
<tr>
<td>Person Name</td>
<td>Identifies duplicate individuals by analyzing name data.</td>
<td>Person_Name</td>
</tr>
<tr>
<td>Resident</td>
<td>Identifies duplicate individuals at an address. Optionally, configure this strategy to analyze ID information.</td>
<td>Person_Name Address_Part1</td>
</tr>
</tbody>
</table>
Identity Match Strategy | Match Operation | Required Inputs
--- | --- | ---
Wide Contact | Identifies records that share a contact at an organization. | Person_Name, Organization_Name
Wide Household | Identifies individuals that belong the same household. | Address_Part1

Configuring Identity Match Output

Use the **Match Output** view to configure the threshold value. The threshold defines the minimum match score that a record needs to join a cluster.

1. Open the **Match Output** view and configure the value of the **Threshold** property.
2. Click File > Save to save the transformation.

Viewing Identity Match Results

To view match results, open a match mapping and run the **Data Viewer** on the Match transformation.

Read the results and verify that the mapping ran successfully. The mapping results identify records that match according to the match threshold values, weights, and strategies that you selected. If you change these selections, the match results may also change.

The following table describes the output ports in a Match transformation:

<table>
<thead>
<tr>
<th>Port</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClusterId</td>
<td>The ID of the cluster to which the record belongs.</td>
</tr>
<tr>
<td>Group Key</td>
<td>The group key that the Match transformation uses to process the record.</td>
</tr>
<tr>
<td>ClusterSize</td>
<td>The number of records in the cluster to which a record belongs. Records that are not matched with other records have a cluster size of 1.</td>
</tr>
<tr>
<td>RowId</td>
<td>A unique row ID for the record. The Match transformation creates this ID by combining the Sequence ID and the Source ID.</td>
</tr>
<tr>
<td>DriverId</td>
<td>The row ID of the driver record in a cluster. The Match transformation designates a record in each cluster as a driver record to enable post-matching operations.</td>
</tr>
<tr>
<td>DriverScore</td>
<td>The match score between a record and the driver record in its cluster.</td>
</tr>
<tr>
<td>LinkId</td>
<td>The row ID of the record that matched with the current record and linked it to the cluster.</td>
</tr>
<tr>
<td>LinkScore</td>
<td>The match score between a record and its linking record in a cluster. The Match transformation uses the LinkScore value to create clusters.</td>
</tr>
</tbody>
</table>
Customizing Matching Operations

You can design customized matching processes by creating *match mapplets*. A match mapplet is a type of mapplet that you can embed in Match transformations.

You create match mapplets by saving the configuration of a Match transformation as a match mapplet. When you create a match mapplet, you convert Match transformation settings into Comparison and Weighted Average transformations that configure matching strategies and apply weights to match scores.

After you create a match mapplet, you can add transformations to customize matching processes. For example, you can add an Expression transformation to evaluate the link scores of two strategies and choose the highest score.

Unlike Match transformations, match mapplets are passive. This means that you can use match mapplets as rules within the Analyst tool, which enables you to match records as part of data profiling processes. You can also configure multiple Match transformations to read the same match mapplet.

For more information about match mapplets, see the *Informatica Developer User Guide*.

Conclusion

Identity analysis provides information about the identities in data sets. Perform identity analysis to determine the identities shared by multiple records.

Author

Charlie Craig
Senior Technical Writer