Overview of Cluster Clean

Contents:

Cluster clean enables users of Trifacta® to standardize values in a column by clustering similar values together. Using one of the supported matching algorithms, Trifacta can cluster together similar column values. You can review the clusters of values to determine if they should be mapped to the same value. If so, you can apply the mapping of these values within the application.

- For more information on how to apply cluster clean, see *Standardize Page*.
- For more information on other methods of standardization, see *Overview of Standardization*.

Artifacts:

When a cluster clean step is added to your recipe, the number of individual changes can be many megabytes of data. Instead of storing these objects within the recipe definition, they are stored as a set of artifacts in the artifact storage database and referenced from the recipe.

- These artifacts exist outside the scope of the recipe file.
- These artifacts must be stored in a Trifacta database for the step to be editable and exportable.

  > **NOTE:** If the artifact storage service is disabled, this feature is unusable.

- When a flow is exported, an `artifact.data` file is included as part of the export. This file must be imported with the flow definition, or the cluster clean step in the imported flow is broken. For more information, see *Export Flow*.

Example - Multiple methods of clustering

Source:

The following dataset includes some values that could be standardized:

RowId	Values
Row01	Apple
Row02	pear
Row03	apple
Row04	pair
Row05	Åpple
Row06	pare

When you standardize using a spelling-based algorithm, the following values are clustered:

Source Value	New Value

Apple	
apple	
Åpple	
	Unclustered values
pear	
pair	
pare	

After you select the cluster of values at top, you can enter `apple`, in the right context panel to replace that cluster of values with a single string.

In the above, the unclustered values are dissimilar in spelling, but in English, they sound the same (homonyms). When you select the Pronunciation-based algorithm, these values are clustered:

Source Value	**New Value**
pear	
pair	
pare	
	Unclustered values
Apple	apple
apple	apple
Åpple	apple

When you select the top values clustered by pronunciation, you can enter `pear` in the right context panel.

Results:

The six source values have been reduced to two final values through two different methods of clustering. See below for more information on the clustering algorithms.

Source Value	**New Value**
pear	pear
pair	pear
pare	pear
Apple	apple
apple	apple
Åpple	apple

You can apply cluster-based standardization through the Standardize Page.

Clustering Algorithms

The following algorithms for clustering values are supported.

Similar strings

For comparing similar strings, the following methods can be applied:

Fingerprint

The fingerprint method compares values in the column by applying the following steps to the data before comparing and clustering:

> **NOTE:** These steps are applied to an internal representation of the data. Your dataset and recipe are not changed by this comparison. Changes are only applied if you choose to modify the values and add the mapping.

1. Remove accents from characters, so that only ASCII characters remain.
2. Change all characters to lowercase.
3. Remove whitespace.
4. Split the string on punctuation, any remaining whitespace, and control characters. Remaining characters are assembled into groups called **tokens**.
5. Sort the tokens and remove any duplicates.
6. Join the tokens back together.
7. Compare all tokenized values in the column for purposes of clustering.

Fingerprint Ngram

This method follows the same steps as those listed above, except that tokens are broken up based on a specific (N) number of characters. By default, Trifacta uses 2-character tokens.

> **Tip:** This method can provide higher fidelity matching, although there may be performance impacts on columns with a high number of unique values.

Pronunciation

Values are clustered based on a language-independent pronunciation.

This method uses the double metaphone algorithm for string comparison. For more information, see *Compare Strings*.

Job Execution

When a job is executed, clustering that has been applied through the data grid is applied to the full dataset. Implications:

- If you have auto-standardized values, the most common value that is applied during job execution is the value that appeared most frequently in the sample that was displayed when the cluster clean step was defined. The most common value is not redetermined based on the entire dataset.
- Values that were not part of the displayed sample may not be factored in the standardization process during job execution.