Standardize Page

Through the Standardize page, you can review similar column values and standardize them to values that you specify.

For example, master data on customers and products may use different names for the same product. For the Web team, the product may be called, "ACME Cookies Chocolate Chip," while the data from the Accounting team refers to this product as, "Cookies - Choc Chip." Through the Standardize page, you can normalize these values to a single consistent value for easier consumption downstream.

- Standardization can be applied to a single column at a time.
- For more information on how Cloud Dataprep by TRIFACTA® INC. standardizes values, see *Overview of Standardization*.

Limitations:

- Standardizations applied through this page are stored in a connected database. These standardizations cannot be migrated between instances or workspaces.

To open the Standardize page:

- From a specific column, click the column drop-down and then select **Standardize...**.
- In the Search panel, enter `standardize column`. Then, select the column whose values you wish to standardize and click **Next**. See *Search Panel*.

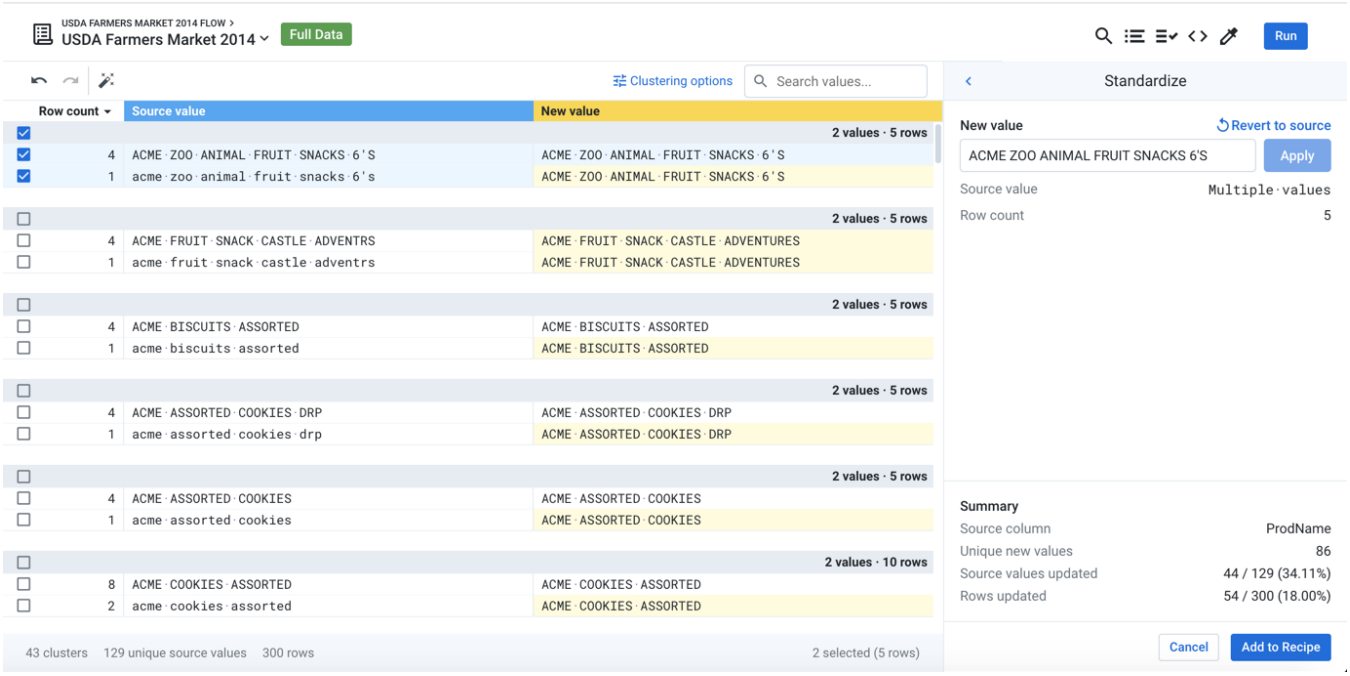

Figure: Standardize page

In the above image, Cloud Dataprep by TRIFACTA® INC. groups the various references to product names into its interpretation of meaningful clusters. This clustering is based on pattern-matching between values in the column.

> **NOTE:** The Standardize page displays column values from the currently selected sample only. If the sample does not span the entire dataset, column values that are not captured in the display are not affected by standardization changes. You may need to take additional samples to capture column values outside the current sample.

In the left side of the screen, you can review the clusters of values that have been detected in the column. In the above image, you can see that the platform has identified a number of clusters based on simple differences in capitalization.

- For each cluster, you can review the number of unique values and the total number of rows where the values appear in the column.
- At the bottom of the left pane, you can review the total number of unique values in the source column and the total number of rows in the displayed sample.

Toolbar

Figure: Standardize toolbar

- **Undo:** Undo the last action in the Standardize page

  > **NOTE:** This action does not undo recipe steps that have already been added.

- **Redo:** Redo the last undo action.
- **Auto Standardize:** The Wand tool automatically standardizes values based in a cluster to the most common value, as long as the most common value occurs in 25% of the clustered rows.

  > **NOTE:** For auto-standardization, the most common value is determined based on the cluster of values that are displayed in the current sample.

  > **Tip:** The Wand tool is recommended for beginning the standardization process. If values within a cluster have been modified, the cluster is not affected by the Wand tool. You can also apply the Wand tool on selected values in a cluster.

- **Clustering options:** By default, values are clustered based on similar spellings. To change the algorithm by which values are clustered, click **Clustering options**.
  - **None:** Do not cluster values. Individual values must be matched.
  - **Similar strings:** Cluster values based on similarities between the text of each value.
  - **Pronunciation:** Cluster values based on phonetic pronunciation of the values.
  - For more information on these options and their variations, see *Overview of Standardization*.
- **Search values:** To locate specific values in the column, enter a search string in the Search textbox.
- To reverse the sort order within your clusters, click Row Count.
  - To sort cluster values alphabetically, click the Source Value header.

Steps to Standardize

To standardize values:

1. If needed, change the clustering algorithm to apply to the values.
2. From the left panel, select the set of values that you wish to standardize.

   > **Tip:** Unclustered values are listed at the bottom of the panel. You should review these values when you are selecting clustered values.

a. To select multiple values, press `COMMAND/CTRL` + click or select multiple values in the left column.

> **Tip:** To select all values in a cluster, click the cluster header.

b. To select a range of sequentially listed values, use `SHIFT` + click, which works across clusters of values.

3. After you have selected all values to standardize, you specify the new value to apply to these selected values in the right panel. This new value applies to all instances of the selected value or values. Changes are previewed next to the source values.

> **Tip:** When you have values selected in a cluster, you can use one of the source values as your standardized value. Hover over the value in the left panel, and then click the icon that appears.

a. Below the value you enter, you can review the number of rows in the sample that are affected by this change.
b. At the bottom of the right panel, you can review the total effects of standardization on the dataset after this change is applied.
c. If the new value is empty, then the values are kept as-is. No change is applied.
d. To apply the standardized value to the affected clustered values, click **Apply**.
e. At any time, you can revert the changes to the cluster values. Click **Revert to source**.

4. Repeat the previous steps as needed.

> **Tip:** You can perform multiple replacements in a single recipe step. So, you can configure all of your standardization steps before adding the single step to your recipe. For debugging purposes, you may want to separate some or all standardization into separate steps.

5. To add the standardizations, click **Add to Recipe**.

> **NOTE:** You cannot copy and paste standardization steps in the Recipe panel.