Track Data Changes

Contents:

Create Backup

After you have created the flow and the datasets within the flow and before applying recipe steps to change the data, create a duplicate of the flow. This becomes a snapshot of your original dataset. Since the imported datasets are not affected, the storage overhead for creating backups is relatively low. See *Flow View Page*.

Track Source Filepath and Filename

When you first load your dataset in the Transformer page, you can add the following to capture the full path to the original file that is the source of the data:

Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`$filepath`
Parameter: New column name	`sourceRowNumber`

With a few extra steps, you can extract the filename from the above output. For more information, see *Source Metadata References*.

Track Source Row Information

You can mark the original row numbers of your source data. In the first step in your recipe after initial parsing, add the following:

Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`$sourcerownumber`
Parameter: New column name	`sourceRowNumber`

This step generates a new column that contains the source row number from the source dataset.

> **NOTE:** Source row information can become invalid if you perform multi-dataset operations such as lookups, unions, and joins. For more precise tracking of source information, you should consider creating multi-column keys, including the source row number information. For more information, see *Generate Primary Keys*.

See *Source Metadata References*.

Track Steps Affecting a Column

To see all of the steps in your current recipe that reference a specific column, select **Show related steps...** from the column menu.

All steps are highlighted in the Recipe panel.

> **NOTE:** If another column is dependent on the selected column, all steps pertaining to that column are highlighted as well.

For more information, see *Column Menus*.

Track Column Value Changes

Trifacta® Self-Managed Enterprise Edition enables you to easily move between steps in your transform recipe so that you can check the state of your dataset at any point during the transformation. In some cases, you may want to be able to track the changes made to an individual column side-by-side with the original column. This section provides a generalized approach for tracking column changes in this manner.

> **NOTE:** Use this workflow only if it is important to monitor which values have changed in a column. For most use cases, the Transformer page provides sufficient visibility over your sample data to manage column values.

Steps:

In the following sequence, the original column is called `String`. For numeric columns, you can perform more detailed analysis between original and modified column values.

1. After you have completed your general setup steps of your transform, create a copy of the original column:

Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`String`
Parameter: New column name	`String_orig`

2. You now have a copy of the original column before any manipulations were applied to it.
3. Add any transforms to your recipe, including any that change the values of `String`. In the example below, the following transform has been applied:

Transformation Name	`Edit with formula`
Parameter: Columns	`String`

Parameter: Formula	`TRIM(String)`

4. At the point in your recipe where you would like to test the column for changes, insert the following:

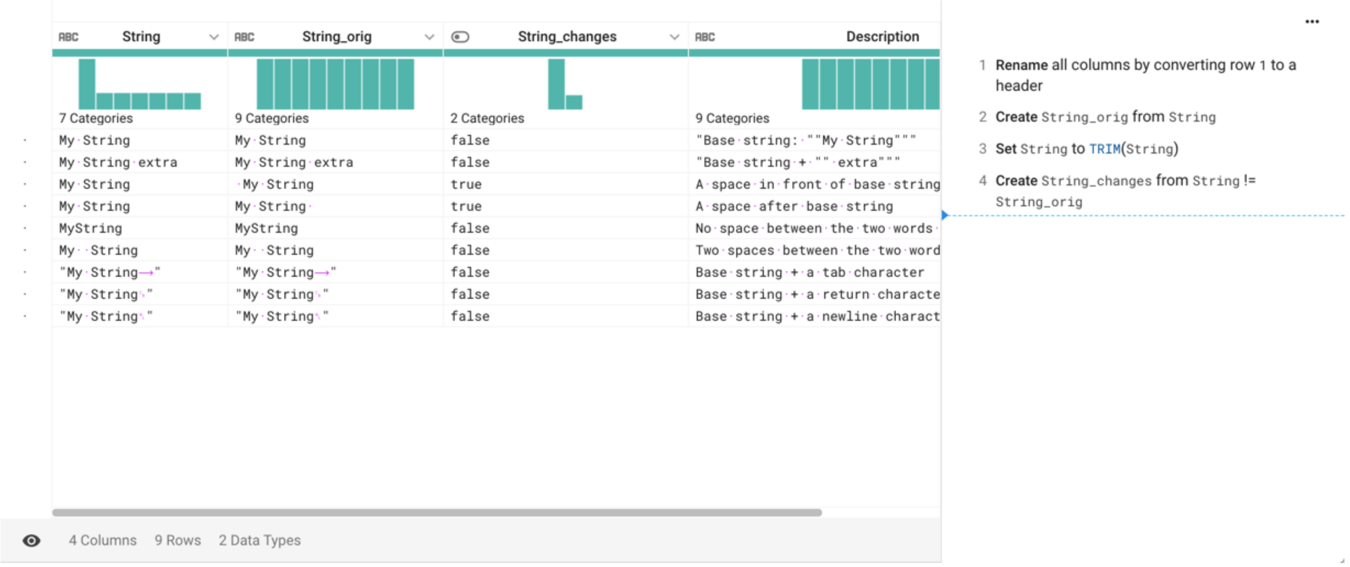
Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`String <> String_orig`
Parameter: New column name	`String_changes`

5. The `String_changes` column now contains `true` values where the values in `String` have been changed from their original values (`String_orig`).
6.

   To see just the values that are different, sort in descending order.

   > **Tip:** You can reposition this test anywhere in your recipe after you have created the `String_orig` column.

7. Before you run your recipe, you may want to remove the tracking columns that you generated (`String_orig` and `String_changes` in our example).



ABC String	ABC String_orig	String_changes	ABC Description
7 Categories	9 Categories	2 Categories	9 Categories
My·String	My·String	false	"Base·string:·""My·String"""
My·String·extra	My·String·extra	false	"Base·string·+·""·extra"""
My·String	·My·String	true	A·space·in·front·of·base·string
My·String	My·String·	true	A·space·after·base·string
MyString	MyString	false	No·space·between·the·two·words·
My··String	My··String	false	Two·spaces·between·the·two·word
"My·String→"	"My·String→"	false	Base·string·+·a·tab·character
"My·String↵"	"My·String↵"	false	Base·string·+·a·return·characte
"My·String↵"	"My·String↵"	false	Base·string·+·a·newline·charact

1 **Rename** all columns by converting row 1 to a header

2 **Create** `String_orig` from `String`

3 **Set** `String` to `TRIM(String)`

4 **Create** `String_changes` from `String` != `String_orig`

Figure: Example tracking column changes

Track Row Changes

Steps:

1. Create a copy of the flow. In its name, identify that it is your original. See *Flow View Page*.
2. In the other flow, create your recipes as normal.
3. When done, you can add the following steps:
   a. Union the two datasets together.
   b. Sort them by a key column.

    c. Add the `deduplicate` transform.

    > **NOTE:** This method may not work if your recipe includes joins or added or removed columns.

4. If the rows are exact duplicates, they are removed. The remaining rows contain data that has been changed.