Deduplicate Data

Contents:

- Validate Duplicate Data
- Remove duplicate rows transformation
- Deduplicate Rows Based on a Primary Key
- Deduplicate Columns

As part of your data cleansing steps, you might need to remove duplicate rows of data from your dataset.

Validate Duplicate Data

In some cases, it might be acceptable to have duplicated data. For example, additional records using the same primary key might be included in a dataset as amendments or detail records.

**NOTE:** Before you remove duplicates from your dataset, you should verify that the data should not contain duplicates at all. If the data structure supports some duplicate elements including key values, you should exercise care in how you identify what constitutes duplicate information.

Remove duplicate rows transformation

Cloud Dataprep by TRIFACTA® INC. provides a single transformation, which can remove identical rows from your dataset:

**Tip:** If you are attempting to identify if there are duplicate rows, check the row count in your dataset before and after you have added this transformation.

| Transformation Name | Remove duplicate rows |

Limitations:

- This transformation is case-sensitive. So, if a column has values Hello and HELLO, the rows containing those values are not considered duplicates and cannot be removed with this transformation.
- Whitespace and the beginning and ending of values is not ignored.

Before applying the Remove deduplicate rows transformation, you should attempt to normalize your data. You can use the following techniques to normalize a few columns of data.

**NOTE:** If you have more than 20 columns of data, you might be better served by trying to identify a primary key method for de-duplicating your dataset. Details are below.

For individual columns, you can use the trim function to remove leading and trailing whitespace:

**NOTE:** To preserve the original column values, use the New formula transformation. The Edit column with formula transformation replaces the original values.

| Transformation Name | New formula |
### Deduplicate Rows Based on a Primary Key

Another method to deduplicate data might be to delete rows based on one or more columns that you identify as a primary key for the dataset. A **primary key** is an identifier that uniquely identifies a row of data within a dataset. It can be a single field (column) or a combination of columns. For example, in a datasets of restaurant locations, the primary key can be a combination of RestaurantName, Address, and Zip.

**NOTE:** Before continuing, you must identify a primary key for your dataset. See *Generate Primary Keys*.

When you have identified your primary key, you should identify the appropriate method for your dataset. Please complete the following steps.

**Steps:**

1. If your primary key spans multiple columns, use the **Merge columns** transformation to bring the values into a single column:

<table>
<thead>
<tr>
<th>Transformation Name</th>
<th>Parameter: Columns</th>
<th>Parameter: Separator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merge columns</td>
<td>RestaurantName, Address, Zip</td>
<td>'-'</td>
</tr>
</tbody>
</table>

2. Rename the generated column: **PrimaryKey**.

3. Use the following transformation to generate a new column, comparing each value in the **PrimaryKey** column to the previous one:

<table>
<thead>
<tr>
<th>Transformation Name</th>
<th>Parameter: Formulas</th>
<th>Parameter: Order by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>PREV(PrimaryKey, 1)</td>
<td>PrimaryKey</td>
</tr>
</tbody>
</table>

4. For each row, the value of the new column is the value in the **PrimaryKey** for the previous row. Now, test if this value is the same as the value in the **PrimaryKey** column for the current row:

<table>
<thead>
<tr>
<th>Transformation Name</th>
<th>Parameter: Formula type</th>
</tr>
</thead>
<tbody>
<tr>
<td>New formula</td>
<td>Single row formula</td>
</tr>
</tbody>
</table>
4. Parameter: Formula
   IF((window==PrimaryKey),true,false)

Parameter: New column name
   IsDupe

5. The new column (IsDupe) contains true for duplicate primary keys. Delete the rows that are duplicates:
   Transformation Name | Delete rows

6. Delete any generated columns that are no longer needed.

Deduplicate Columns

While this form of duplicate data is rarer, you might want to check on the possibility of duplicate data between your columns. To check for duplicate column data, you can use a transformation similar to the following:

<table>
<thead>
<tr>
<th>Transformation Name</th>
<th>New formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter: Formula type</td>
<td>Single row formula</td>
</tr>
<tr>
<td>Parameter: Formula</td>
<td>Column1 == Column2</td>
</tr>
<tr>
<td>Parameter: New column name</td>
<td>'dupeColVals'</td>
</tr>
</tbody>
</table>

In the generated column, values that are true indicate duplicate data. If all values are true, then you can remove one of the columns.