Deduplicate Data

Contents:

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.

Deduplicate Transform

Trifacta® provides a single transform, 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 transform.

```
deduplicate
```

Limitations:

- This transform 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 transform.
- Whitespace and the beginning and ending of values is not ignored.

Before applying the `deduplicate` transform, 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 `derive` transform. The `set` transform replaces the original values.

```
derive type:single value:TRIM(Item)
```

You can paste Wrangle steps into the *Transformer Page*.

Since the `deduplicate` transform is case-sensitive, you can use the `LOWER` function to make the case of each entry in a column to be consistent:

```
derive type:single value:LOWER(Description)
```

For more information, see *Normalize Numeric Values*.

Deduplicate Rows Based on a Primary Key

An easier 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` transform to bring the values into a single column:

   ```
   merge col:RestaurantName,Address,Zip with:'-'
   ```

2. Rename the generated column: `PrimaryKey`.
3. Use the following transform to generate a new column, comparing each value in the `PrimaryKey` column to the previous one:

   ```
   window value: PREV(PrimaryKey, 1) order: PrimaryKey
   ```

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:

   ```
   derive type:single value:((window==PrimaryKey) ? true : false)
   ```

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

   ```
   delete row:(IsDupe==true)
   ```

6. Drop 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 transform similar to the following:

```
derive type:single value: (Column1 == Column2) as:'dupeColVals'
```

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