Removing Duplicate Values

The Unique Tool

Duplicates are for official documents, not data. Duplicate values can seriously damage your confidence in data and even lead to inaccurate analysis. You may find that you need to remove duplicate rows from your data set based on a value in a column or a combination of values across multiple columns. The Unique tool is an efficient, easy way to identify unique values and remove duplicates.

A sample dataset contains information on server access and use by a network of machines. The data contains three columns of information: date, the name of the machine that accessed the server and the duration of that machine's session on the server. Notice that columns the "Date" and "Machine Name" contain duplicate values. Use the unique tool to identify which dates the server was accessed.

Drag a Unique tool onto the canvas.

Two Outputs

The Unique tool has two output anchors. The top anchor, indicated by the letter "U", outputs the values identified as unique from the dataset. The bottom output anchor of the Unique tool, indicated by the letter "D", outputs the values determined to be duplicates. In the configuration window, select the column "Date" to identify the unique values within that column.

After running the workflow, the U anchor's output contains 5 unique dates. The D anchor outputs the other 16 rows containing duplicate dates.

One Row at a Time

Designer processes datasets from the top down and will classify a value as unique if it has not encountered that value before. Any values that were previously found in the dataset will be classified as duplicates. Adding a Sort tool is a common way of ensuring the unique values are optimal for the analysis to follow.

Next, determine the days on which each machine accessed the server. Rather than drag on another Unique tool, re-configure the current tool by selecting "Date" and "Machine Name".

Multiple Columns

When multiple columns are selected, the values in each selected column are combined, and then compared across the dataset. Different combinations of values are treated as unique values. If values differ in any column, that row is treated as a unique value. All columns in the dataset can be selected.

With both “Machine Name” and “Date” selected, the same machine can appear in the unique output if the date values differ, and the same date can appear if the machine names differ. Only when the machine name and date match another row exactly will a value appear in the duplicate output. After running the workflow, the unique output contains 18 records. Upon further review, a row appears to contain a match to another row on both machine name and date. The Unique tool is case-sensitive, and therefore, data should be consistently cased prior to using a Unique tool. Adding a Data Cleansing tool and re-running the workflow correctly compares the values.