Blending Data with Unions - Transcript

Blending Data with Unions

Your analysis may require that many sources containing the same types of data values be combined into a single data stream. In simple terms, a Union of data vertically expands a data stream by stacking incoming data streams on top of each other based on matching column names, column position, or manual alignment. Not only does this prevent you from having to work with each data source separately and duplicate analytic processes, but it also allows you to develop dynamic workflows that applies a process to a dataset that you, essentially, create! Three input datasets contain data on trees surveyed in New York City. Similar information on the trees that were planted in 2016, 2017 and 2018 are stored in separate files, one for each year. However, for the analysis you want to perform, it makes sense to combine these datasets and analyze the data as one single data stream. Create this new, expanded data stream with the Union tool.

Drag the Union tool from the Favorites tool palette and drop it onto the Canvas.

The Union Tool's Anchors

The Union tool has two anchors, an input and an output anchor. However, the Union tool's input anchor is different from those found on many other tools in Designer. Its input anchor is made up of multiple arrows to indicate that this tool can accept multiple incoming data streams.

The first input that is connected to the Union tool determines the output column names and data types that are used in the output from the Union tool. Because changes to column names and data types have already been applied to the input containing data on trees planted in 2016 with a Select tool, this data stream will serve as the template for all the connected inputs.

Connect the output anchor of the Select tool to the input anchor of the Union tool.

Then, connect the output anchor of the Input Data tool to the input anchor of the Union tool.

Connect the final Input Data tool to the Union tool.

Connection Strings

Connection strings are numbered as the order in which they were connected to the Union tool. To clarify the data being passed through each connection string, rename the connection string to reflect the source of the data.

Double click the connection string #1. In the Configuration window, remove the current name and enter "2016" to rename the connection string. Double click the connection string #2 and change the connection string's name to "2017" in the Configuration window.

Double click the connection string #3 and rename it as "2018" in the Configuration window.

Union Tool Configuration Options

The Union tool can vertically align by one of three ways: column name, column position, and manually. Click a configuration option to learn more.

By Column Name

When combining data by column name, the values in columns with identical names are stacked vertically. Before using this method, it is important to thoroughly investigate the inputs so that values that do represent the same data but are in identically named columns are not incorrectly blended together. Aligning data by column name is the default configuration of the Union tool.

By Column Position

When combining data based on column position, the values in the first column of all the inputs will be stacked on top of each other. The values in the second column will be combined with the values in the second columns of all the inputs, and so on. In this case, it is important that the order of the columns in the inputs match before entering the Union tool.

Manually

You may find that neither column name nor position is reliable to use for automatically aligning data values. Manual configuration allows you to manipulate the data into the proper alignment based on your knowledge of your data.

When Inputs Have Differing Fields

Not all the inputs to the Union tool contain the same number of columns. The 2017input, for example, includes data on the latitude and longitude of each tree and the2018 input contains this information, as well as a spatial object. The 2016 input does not include any of these columns.

When the number of columns in each of the combined inputs differs, you can choose how those differences are handled in terms of workflow processing and outputs. By default, workflow processing will not stop, but you will receive a warning in the Messages window when running the workflow. Alternatively, you may choose to cause an error, stopping the workflow processing, or ignore these differences in schema.

Additionally, you may choose whether to output all the columns that are present in the combined inputs, or only output the columns that all inputs shares.

First, click Run to output all the columns present in the combined inputs.

After combining the data with the Union tool, this single stream of data includes all the information from the three inputs, about 200,000 total rows.

If all the output columns from the Union tool are not present in every input, you may see columns that contain Null values in the Results window. Depending on the analysis you plan to perform, you may consider leaving this data as is, or decide to address the Null values with an additional step later in the workflow.

Outputting the Common Subset

To output only the columns that are present in all the inputs, configure the Union tool to "Output a Common Subset" of the incoming columns. While this configuration may remove columns of data that could be important for downstream analysis, it can also reduce the need for additional data cleansing and preparation later in the workflow.

Outputting in a Specific Order

By default, output data values will be stacked in the order in which they were connected to the Union tool: 2016, 2017 and 2018. However, you may prefer to see the data from each connected input to the Union tool stacked in a particular order on output, such as the most recently planted trees at the top of the dataset.

Select the checkbox to “Set a Specific Output Order”.

Click the connection "2016" and click the down arrow twice.

Click the connection "2018" and click the up arrow once.

After running the workflow, the data is output from the Union tool in a specific order: the values from the trees planted in 2018, then 2017, then 2016. Note that actual data values have not been sorted; the order of rows for each input dataset remains the same.