Separating Data into Columns and Rows

Introduction
The data you work with isn't always in a readily usable format for analysis. For example, you may find that you have a series of data values that have been chained together, or concatenated, into a single cell with a delimiter. Separating these values into a more efficient and usable format, like separate columns or rows, can be accomplished in a couple of different ways in Designer. The Text to Columns tool in the Parse tool palette parses data values into specified tables.

Delimiters
Delimiters, such as a space, comma, pipe, or another symbol, can be used to distinguish data values and indicate where to split data into separate columns or rows. Explore some of the characters that are commonly used to delimit data.

Click the character or characters in Column1 that is used to separate the data into the additional columns in the table.

Designer
An input dataset contains information on loans issued by various international financial institutions. Each row of data in the column "Loan Data" contains values, such as the start and end dates of the loan, the country code and country name to which the loan was issued, and the loaned amount, separated by delimiters: commas and pipes. In its current state, the values in the column [Loan Data] are unusable for analysis. Parse the data values associated with each loan into five separate columns using the Text to Columns tool.

Drag the Text to Columns Tool and connect it to the Input Data tool.

Configuration Window
Configure the Text to Columns tool by -first- selecting the column to split and specifying the delimiters used to separate values.

Use the Drop Down to select the column to split: Loan Data

Delimiters
Specify the delimiter or delimiters used to separate data by manually entering them into the text box. A comma is used as the default delimiter. Since the data in the column "Loan Data" does contain comma-delimited values, this delimiter can remain in the text box.

In addition to commas, data in the column "Loan Data" is also delimited by pipes. Add a pipe (Shift + Backslash) to the text box of delimiters. Then, click Submit.

When multiple delimiters are specified in the Text to Columns tool, they are not recognized as a combined character. In this case, a new column or row will be created when either a comma or a pipe is found in the column to split.

Splitting to Columns
By default, the Text to Columns tool is configured to split data into separate columns. When splitting data into columns, it helps to know the number of delimited values a cell contains. The column [Loan Data] includes five values: a start date, an end date, two-letter country code, country name and loaned amount.

Type five (5) in the text entry to specify the number of output columns. Then, click Submit.

In the event that the number of specified output columns is not great enough to fit all delimited values, decide how to handle the extra characters. In the Drop Down, choose to leave the un-parsed data in the last column, drop the data with or without a warning, or receive an error message. By default, extra characters will be left in the last created column. Overestimating the number of columns needed to output data creates extra columns of null values.

Output columns that are produced from the Text to Columns tool are given a name that combines the Output Root name, which is the name of the original column that is being split, and the number of the column. The Output Root Name can be changed by manually entering a root name in the text box.

After running the workflow, five new columns have been added to the dataset. The columns Loan Data 1 through 5 contain the delimited values. However, it appears that some rows in the last column contain extra characters: part of the country's name as well as the loaned amount.

Look more closely at the data in the column [Loan Data]. Some countries, include commas in their names in this dataset. Since commas are specified as a delimiter, the presence of commas in data values causes inconsistent results. Luckily, it appears that the commas that should not be used as a delimiter are surrounded by quotes. This formatting can be leveraged in the Text to Columns tool's configuration.

Advanced Options
The Text to Columns tool includes Advanced Options to Ignore Delimiters when they are offset by certain punctuation such as quotes, single quotes, parentheses and brackets. Columns that are empty can be skipped as well.

Select the box to ignore delimiters in quotes.

After re-running the workflow, the data is parsed as expected and five new columns of data appear in the Results window.

Split to Rows

The Text to Columns tool can also separate delimited data values into rows, which can be useful when you do not know the number of delimited values a cell contains, or when you anticipate any aggregation or pivoting of your data downstream in the workflow.

Before entering the Text to Columns tool, it's recommended that each row of data is assigned some kind of unique identifier. This will help with understanding the origin of each data value once it is split by a delimiter. In this case, a column in the Input Data called "Loan ID" assigns a unique number to each row of data.

In the Text to Columns tool's configuration window, select the option to split the values in the column [Loan Data] into rows.

After re-running the workflow, the Text to Columns tool has assigned each value its own row. You can also see how the Loan ID column helps organize the data; each delimited value is associated with its original identifier, "grouping" the values that originated in the same row before parsing.