



Free Training Transcript: The Input Step

Welcome to this video on the Input Step in Tableau Prep. You can download the data sets and packaged flow file underneath the video to follow along in your own copy of Tableau Prep. We're working with data for bestselling books.

Adding Data to the Flow

Every flow starts with data, and a flow must start with at least one input step, though new data (new input steps) can continue to be added to the flow at any stage. To add data in an existing flow, we can click the plus in the Connections pane. But let's start from the beginning. To begin a flow and add data, we'll click Connect to Data from the start screen.

We see the list of supported data sources. If we chose to connect to a server, we'll need the connection information specific to that data source. For now, we'll choose Excel. Navigate to the file on your machine, and click Open.

The Connections pane will display the data connection at the top, and below, a list of the tables (or sheet tabs) in that data connection. We have the option to turn on the Data Interpreter, which can identify subtables and clean up things like header rows and blank columns. The Data Interpreter functions just as it does in Tableau Desktop.

To create an input step, we need to drag a table out from the connections pane. This creates a step in the flow, and the input pane below provides options for configuration.

For now, let's leave this as a single table.

To the right we see the fields in the data and can pick exactly which fields to bring into Tableau Prep. Sample values show the type of data in each field, so we can see how things are formatted. We can also set filters here, limiting what data is brought into the rest of the flow. For example, let's only bring in bestsellers ranked 1-5. If we click "Add Filter", note that it opens as a calculation window. We can enter our parameters, such as Rank < 6, and see that filter reflected. To remove it, simply click to open the drop down menu and remove.

Wildcard Union

But let's say we want to bring in more than a single table as part of the data input. If the

structure of the tables is consistent, we can perform a wildcard union. A wildcard union during the input step is different than a Union step in the flow in that we cannot merge fields or modify the schema. (For more information on unions, check out the video on the Union Step.)

Data Sampling

If we connect to a large file, Tableau Prep may sample by default. This takes a subset of the data and brings it into the flow for us to manipulate and prepare. When we run the flow, it is run against all the data, not just the sample. The sample keeps the process of building the flow more performant than trying to profile all the data and apply changes to the much larger data volume as we're working.

We can configure the sample under the Data Sample tab in the input pane. Data over 1 million rows will likely be sampled, but if there is a large number of columns, the sample could contain fewer rows or be triggered earlier. The default sample amount is based on the size of the data, not an explicit number of rows. This means if you have 300 fields, you'll get fewer rows in your sample than if you had 5 fields.

Alternatively, we can choose to sample a specified number of rows (with an upper limit of a million), or we can choose to use all the data (though this will have a performance impact when building the flow).

We can also select a sampling method. Quick Select contains whichever rows the data source provides. It may run into sampling issues such as not covering the full range of values in the data (for example, only showing one year's worth of data when the original data spans a decade). By contrast, we can select a Random sample. This will take longer to generate and may have performance impacts if any operations require the sample to be regenerated, though it may result in a better (though not perfect) representation of the entire data set.

To get the most representative sample possible, we should apply any relevant filters and deselect any irrelevant fields in the input step, rather than the flow. By narrowing the scope of fields or rows in the input, which is to say, before sampling, we'll maximize the sample's usefulness. If we don't remove fields or filter until after the input step, the sample will not be as robust as it could be.

Additional Inputs

Input steps can be added along any point in the flow. Each data source comes in as its

own step, then can be unioned or joined with other steps to combine the data. (For more information on combining data, check out the videos on The Union Step and The Join Step.)

Conclusion

Thank you for watching this video on bringing data in Tableau Prep with an input step. We invite you to continue with the free training videos to learn more about using Tableau.

