Initial Parsing Steps

Contents:

When a dataset is initially loaded into the Transformer page, one or more steps may be automatically added to the new recipe in order to assist in parsing the data. The added steps are based on the type of data that is being loaded and the ability of the application to recognize the structure of the data.

Automatic Structure Detection

> **NOTE:** By default, these steps do not appear in the recipe panel due to automatic structure detection. If you are having issues with the initial structuring of your dataset, you may choose to re-import the dataset with Detect structure disabled. Then, you can review this section to identify how to manually structure your data. For more information on changing the import settings for a dataset, see *Import Data Page*.

This section provides information on how to apply initial parsing steps to unstructured imported datasets. These steps should be applied through the recipe panel.

> **NOTE:** Imported datasets whose schema has not been detected are labeled, **unstructured datasets**. These datasets are marked in the application. When a recipe for this dataset is first loaded into the Transformer page, the structuring steps are added as the first steps to the associated recipe, where they can be modified as needed.

Overview

When data is first loaded, it is initially contained in a single column, so the initial steps apply to `column1`.

Step 1: Split the rows. In most cases, the first step added to your recipe is a Splitrows transformation, which breaks up the individual rows based on a consistently recognized pattern at the end of each line. Often, this value is a carriage return or a carriage return-new line. These values are written in Wrangle as `\r` and `\r\n`, respectively. See the example below.

Step 2: Split the columns. Next, the application attempts to break up individual rows into columns.

- If the dataset contains no schema, the Split Column transformation used. This transformation attempts to find a single consistent pattern or a sequence of patterns in row data to demarcate the end of individual values (fields).

  > **NOTE:** Avoid creating datasets that are wider than 2500 columns. Performance can degrade significantly on very wide datasets.

- If the dataset contains a schema, that information is used to demarcate the columns in the dataset.

When the above steps have been successfully completed, the data can be displayed in tabular format in the data grid.

Step 3: Add column headers. If the first row of data contains a recognizable set of column names, a Rename Columns with Rows transformation might be applied, which turns the first row of values into the names of the columns.

Example recipe:

1.

Transformation Name	`Split into rows`
Parameter: Column	`column1`
Parameter: Split on	`\r`
Parameter: Ignore matches between	`\"`
Parameter: Quote escape character	`\"`

2.

Transformation Name	`Split column`
Parameter: Column	`column1`
Parameter: Option	`on pattern`
Parameter: Match pattern	`','`
Parameter: Number of matches	`9`
Parameter: Ignore matches between	`\"`

3.

Transformation Name	`Add header`
Parameter: Row number	`1`

After these steps are completed, the data type of each column is inferred from the data in the sample. See *Supported Data Types*.

Excel, CSV

Microsoft Excel files are internally converted to CSV files and then loaded into the Transformer page. CSV files are treated using the general parsing steps. See previous section.

For more information, see *Import Excel Data*.

JSON

If 80% of the records in an imported dataset are valid JSON objects, then the data is parsed as JSON.

Notes:

- For JSON files, it is important to import them in unstructured format.
- **Designer Cloud powered by Trifacta® Enterprise Edition** requires that JSON files be submitted with one valid JSON object per line.
    - Multi-line JSON import is not supported.

- Consistently malformed JSON objects or objects that overlap linebreaks might cause import to fail.

Depending on the shape of your data, you may need to change the following properties. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.

`webapp.maxRecordLength` - determines the maximum length for an individual line.

`webapp.sampleLoadLimit` - determines the maximum size in bytes for a random sample.

> **NOTE:** Be cautious in changing these values. If you set these values too high, you can overload the client and crash the application.

Step 1: Split the rows. JSON data is initially split using Splitrows transformation.

Step 2: Unnest the rows. Then, the data must be broken out from the nested JSON structures into flat rows. Using the Unnest Objects transformation, the application attempts to render the JSON into consistently formatted rows.

> **NOTE:** After initial parsing, you might need to apply the Unnest Objects transformation multiple times on individual columns to completely unnest the data.

Step 3: Delete source. If the data is successfully unnested, the source column is removed with a Delete Columns transformation.

Example recipe:

Transformation Name	`Split into rows`
Parameter: Column	`column1`
Parameter: Split on	`\n`
Parameter: Quote escape character	`\"`

In the following, the values `c1` - `c3` identify the keys used to demarcate top-level nodes in the JSON source. These become individual column headers in the data grid.

Transformation Name	`Unnest Objects into columns`
Parameter: Column	`column1`
Parameter: Paths to elements	`c1,c2,c3`
Parameter: Remove elements from original	`true`

If the above successfully executes, the source column is deleted:

Transformation Name	`Delete columns`
Parameter: Columns	`column1`

Avro

Avro-based sources of data do not require any initial restructuring.

Database Tables

Properly formatted database tables with a provided schema should not require any initial parsing steps.

Known Issues

- Some characters in imported datasets, such as NUL (ASCII character 0) characters, may cause problems with recognizing line breaks. If initial parsing is having trouble with line breaks, you may need to fix the issue in the source data prior to import, since the Splitrows transformation must be the first step in your recipe.

Troubleshooting

Fixing parsing issues from structured source after recipe has been created

If you discover that your dataset has issues related to initial parsing of a structured source after you have started creating your recipe, you can use the following steps to attempt to rectify the problem.

Steps:

1. Open the flow containing your recipe.
2. Select the imported dataset. From the context menu, select **Remove structure...**.
3. For the imported dataset, click **Add new recipe**.
4. Make any changes to the initial parsing steps in this recipe.
5. Select the recipe you were initially modifying. From its context menu, select the new recipe as its source.

The new initial parsing steps are now inserted into recipe flow before the recipe steps in development.