Insert Metadata

Contents:

Metadata is data about your data. For example, you might decide that one or more of the following types of information about your dataset should be tracked:

- Source system(s)

- Source creation date
- Date of import
- Date of wrangling
- Name of person who performed the wrangling

This section provides some methods for how to insert metadata into your dataset.

Insert source row number

You can insert the row number in the source file from which rows in your dataset are sourced, using the `$sourcerownumber` reference.

> **Tip:** Source row number information can be lost when multi-dataset operations, such as unions and joins, are performed on your dataset. These steps should be added very early in your recipe.

In your recipe, insert the following transformation:

Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`$sourcerownumber`
Parameter: New column name	`sourceRowNumber`

For more information, see *Source Metadata References*.

> **Tip:** You can derive the current row number in your dataset. For more information, see *ROWNUMBER Function*.

Insert a single metadata column

The following example describes how to insert a single column of metadata. In this case, the full path to the source is inserted as a new column in the dataset.

Steps:

1. In the Dataset page, locate the imported dataset that is the source for your recipe. Click the Imported filter to show only the imported datasets.
2. For the imported dataset, click **Details**.
3. In the Dataset Details page, select the entire value for the Location, which is the storage location of the source.

   > **Tip:** If the full path of the dataset is too long for screen display, be sure to include the ellipsis (...) at the end of the Location value.

4. Copy the value. Paste the value into a text editor. You should see the full path, like the following:

   ```
   <root_dir>/uploads/1/2580298d-3477-4907-bfa7-f71978eace04/SF Restaurants - businesses.csv
   ```

5. Load the dataset in the Transformer page.
6. Specify the following transformation:

Transformation Name	`New formula`
Parameter: Formula type	`Single row formula`
Parameter: Formula	`'<root_dir>\/uploads\/1\/2580298d-3477-4907-bfa7-f71978eace04\/SF Restaurants - businesses.csv'`
Parameter: New column name	`datasetPath`

Insert multiple columns of metadata

You might need to track more fields of dataset information. While you might be able to perform these kinds of individual inserts, it might be easier to build this information from a separate file.

> **NOTE:** This method uses the FILL function, which should be limited to smaller datasets when applied with a single key. Otherwise, there might be performance impacts when running the job against the full dataset.

> **Tip:** You can perform a similar merging of datasets using the Join tool. See *Join Window*.

For example, you want to track the following fields as metadata:

- source_system
- source_author
- source_date_create

You could create a CSV file that looks like the following:

```
source_system,source_author,source_date_create
Excel,Joe Guy,12/9/15
```

In this case, the column headers are in the first line, and the values for each column are in the second line.

Steps:

1. Use your CSV file as the source for a new dataset within the flow containing the associated dataset.
2. In the data grid, make sure that the first line of data is treated as the header. If not, add a `header` transform to your recipe.
3. Open the other (source) dataset in the Transformer page.
4. In the recipe panel of the Transformer page, add a new step. In the Transformation textbox, enter `union`.
5. Create a union:
    a. Include all columns from both datasets.
    b. Configure the step to perform the union by name, instead of by position.
    c. See *Union Page*.
6. Add this step to your recipe.
7. You should see one row in the union recipe that contains the new data.
8. Sort your data by a key value (e.g. `business_id`).
9. Determine an appropriate grouping parameter. This step is necessary to simplify the filling process when the job runs at scale. Ideally, you should choose a grouping column that contains a relative few number of values in it (e.g. `region`).
10. Fill values in the data rows with metadata column values. For each metadata column, add the following transformation, done here for the `source_system` column of metadata.

Transformation Name	`Window`
Parameter: Formula	`FILL(source_system)`
Parameter: Group by	`region`
Parameter: Order by	`business_id`

11. Repeat the above step for each metadata column you want to insert.
12. Delete the source metadata columns.
13. Rename the `window` columns to use a more appropriate name.
14. Delete the row containing the original metadata values.