Enriching Data

Contents:

Dataprep by Trifacta® provides multiple tools for bringing data from other sources into your dataset.

Union

A **union** operation concatenates multiple datasets together. An example is below.

> **Tip:** The following example unions two datasets based on the position of the columns. Unions may also be performed based on the column names.

Dataset 1:

CName1	CName2	CName3
C1.1	C2.1	C3.1
C1.2	C2.2	C3.2
C1.3	C2.3	C3.3

Dataset 2:

CName1	CName2	CName4
C4.1	C5.1	C6.1
C4.2	C5.2	C6.2
C4.3	C5.3	C6.3

When a union is performed based on the position of the columns in each dataset, all of the rows of Dataset 1 are included, followed by all of the rows of Dataset 2. You can choose which columns to include from each of the source datasets.

Output:

In the above, note that the name of the third column in each dataset is different (`CName3` and `CName4`).

CName1	CName2	CName3	CName4
C1.1	C2.1	C3.1	
C1.2	C2.2	C3.2	
C1.3	C2.3	C3.3	
C4.1	C5.1		C6.1
C4.2	C5.2		C6.2

Copyright © 2021 Trifacta Inc.

C4.3	C5.3		C6.3

When to use:

> **Tip:** You should perform union operations as early as possible in your recipes.

- If your datasets include event or log information, you can use the union operation to create a longer sequence of those transactions. For example, you might union together all of your log data for a week from daily log files.

To union your dataset to another, enter `Union datasets` the Transformation textbox in the recipe panel. See *Recipe Panel*.

See *Union Page*.

Join

A join operation brings together two datasets based on a column that appears in both datasets and contains the same unique values used to identify records. Based on the values in this column, called the **primary key**, records in the second dataset are joined to records in the first dataset. As part of the join definition, you may select the fields from both datasets to include, filtering out any duplicated or unnecessary fields in the combined dataset.

The way in which the two datasets are joined is defined by the type of join:

- inner join - include only the records in which key (**primary key**) values in the first dataset appear as key (**foreign key**) values in the second dataset.
- left join - include only the records that contain a primary key value that appears in the first (left) dataset.
  - If a primary key value from the first dataset does not appear as a foreign key in the second dataset, any columns brought in from the second dataset contain missing values.
  - Foreign key values that appear in the second dataset and not the first one do not generate rows in the output dataset.
- right join - include only the records that contain a foreign key value that appears in the second (right) dataset. The other conditions above apply in reverse.
- outer join - include all records from both datasets. If a key value is missing from either dataset, the column values included from that dataset are missing.
- For more information, see *Join Types*.

When to use:

> **Tip:** Generally, you should perform join operations as late as possible in your recipes.

- A join is useful for pulling in **selected** fields from a second dataset based on matches of key values. These operations can be expensive to execute but can generate a much wider range of output datasets.

To join your dataset with another, enter `join datasets` in the Search panel. See *Join Window*.

Lookup

A **lookup** operation is used to pull in reference fields from another dataset based on the values contained in a selected column of the first dataset. These second datasets are typically static or changing infrequently.

> **NOTE:** A lookup is similar to a left join. However, with a lookup, all fields from the reference dataset are brought into the generated dataset and all fields from the original dataset are included automatically. When you create a join, you may specify the fields to include in the output dataset.

Copyright © 2021 Trifacta Inc.

For example, you might create a dataset like the following:

State-2letters	State-full
AL	Alabama
AK	Alaska
AZ	Arizona
...	...
WI	Wisconsin
WY	Wyoming

If you have a dataset containing the two-letter abbreviations, you can perform a lookup into the above dataset to retrieve the corresponding full names, which are inserted as an adjacent column called `State-full` in the original dataset.

> **NOTE:** If a value in the column from the first dataset does not appear in the second dataset, there is no corresponding value in the generated `State-full` column.

When to use:

- Lookups are useful for referencing shared datasets whose meaning must be consistent across multiple datasets. You can use lookups to pull in customer or product master data (Customer name, address, etc.) based on CustomerId or ProductId values.

To perform a lookup from a column in your dataset, open the column drop-down and select **Lookup...**. See *Lookup Wizard*.

Aggregation

A single-dataset operation, an aggregation is used to perform summary calculations on columns in your dataset, optionally grouping your data by the values in one or more columns.

For example, your dataset contains point-of-sale transactions from all of the stores in your organization. You can use an aggregation to summarize total sales by performing a sum operation on your `Total_Sale` column. If you group this calculation by month and by `StoreId`, you can acquire monthly sales per month per store.

When to use:

- An aggregation is useful for performing exploratory calculations on your entire dataset or segments of your dataset.
- You can perform aggregations and run jobs to generate the results. After you have these summary reports, you can return to the Transformer page and remove the aggregation to continue wrangling your data.

For more information on in-column aggregations, see *Create Aggregations*.

For more information on building aggregated pivot tables, see *Pivot Data*.

Copyright © 2021 Trifacta Inc.