Append Datasets

If you are wrangling datasets that represent transactional or serialized data, you can append together slices of data to build a larger dataset for richer analysis. For example, you are cleansing log messages on a weekly basis. You can create separate datasets for each day's log messages and then bring them altogether into a single dataset for processing through a single recipe. This method works best for datasets that have identical or very similar structures.

Below, you can see two datasets of contact information. These simplified datasets track customer contact records.

Dataset01:

Name	Email	Last Contact
Jack Jones	jack@example.com	06/15/2015
Tina Toms	tinat@example.com	08/02/2015
Larry Lyons	larry.lyons@example.com	03/22/2015

Dataset02:

Name	Last Contact Date	Email
Amy Abrams	07/24/2015	amy.abrams@example.com
Tina Toms	05/12/2015	tinat@example.com
Samantha Smith	04/22/2015	samantha@example.com

Notes:

- There is one overlapping record for Tina Toms.
- There is a mismatch in one column name ("Last Contact" vs. "Last Contact Date").
- The columns are in a different order.

Steps:

1. Load your first dataset (`Dataset01`) .
2. In the recipe panel, add a step. In the Transformation textbox, enter `union`.
3. In the Union page, you bring together two or more datasets based on a shared set of fields.
    a. A **union** operation appends datasets together. For more information, see *Union Page*.
4. To add another dataset, click **Add datasets**. Navigate to select the file to add to the union (`Dataset02`).
5. Initially, fields are mapped based on the column names. However, in this example, the `Last_Contact_Date` field from `Dataset02` is not included. You can:
    a. Click the + icon next to the `Last_Contact_Date` field in the left panel. The field is added as a separate field. However, it is not matched with the other contact date field from the original dataset.
    b. From the Match columns drop-down menu, select **By Position**. In this case, you can see that there are only three fields, but the order is mismatched.

> **Tip:** When possible, you should try to rename or align columns in your datasets prior to building a union transform step. Otherwise, you might have to edit the columns after the union has been completed.
>
> To rename a column, click **Rename** from the column drop-down in the Transformer page. You can use the same drop-down to move a column.

6. In this case, you can cancel the union and reposition the `Email` column after the `Last Contact` column in `Dataset01`.
7. Then, open the Union page again and add `Dataset02`. Select **By Position** from the Match columns drop-down menu. Your columns are matched.
8. Click **Add to Recipe**.

`Dataset02` records have now been added to `Dataset01`, which now contains all of the records from both datasets. Note that the record for Tina Toms appears twice in the appended dataset.

- If the appended dataset is a record of all contacts, you should leave the duplicate record in place.
- If the appended dataset is a record of the most recent contact with each customer, you should remove the duplicate record. For more information, see *Deduplicate Data*.

> **NOTE:** Be sure to verify that the data type for each column is accurate.