Optimize Job Processing

Contents:

This page contains a set of tips for how to improve the overall performance of job execution.

Filter data early

If you know that you are deleting some rows and columns from your dataset, add these transformation steps early in your recipe. This reduction simplifies working with the content through the application and, at execution, speeds the processing of the remaining valid data. Since you may be executing your job multiple times before it is finalized, it should also speed your development process.

- To delete columns:
  - Select **Delete** from the column drop-down for individual columns. See *Column Menus*.
  - Use the Delete Columns transformation to remove multiple discrete columns or ranges of columns.
- To delete rows: The following example removes all rows that lack a value for the `id` column:

Transformation Name	`Filter rows`
Parameter: Condition	`Is missing`
Parameter: Column	`id`
Parameter: Action	`delete matching rows`

- To keep rows: The following example keeps all rows that lack a value in the `id` column:

Transformation Name	`Filter rows`
Parameter: Condition	`Is missing`
Parameter: Column	`id`
Parameter: Action	`keep matching rows`

- See *Filter Data*.

Perform joins early

After you have filtered out unneeded rows and columns, join operations should be performed in your recipe. These steps bring together your data into a single consistent dataset. By doing them early in the process, you reduce the chance of having changes to your join keys impacting the results of your join operations. See *Join Panel*.

Perform unions late

Union operations should generally be performed later in the recipe so that you have a small chance of changes to the union operation, including dataset refreshes, affecting the recipe and the output.

> **NOTE:** If your dataset requires a significant amount of data cleaning, you should perform your unions early in your recipe, so that all cleaning steps can be applied once across the dataset.

See *Union Page*.

Run jobs on the default running environment

When configuring a job, Trifacta Self-Managed Enterprise Edition analyzes the size of your dataset to determine the best of the available running environments on which to execute the job. This option is presented as the default option in the dialog. Unless you have specific reasons for doing otherwise, you should accept the default suggestion.