Sample Types

Contents:

This section provides an overview of the types of samples that Trifacta® can generate.

Sample filters:

Several sampling types support the application of filters to the source data. In this case, a **filter** can be defined to limit the scope of rows that are used to generate the sample. For example, suppose you apply a filter like the following:

```
orderId == '100'
```

The rows of data available for generating the sample are reduced to include only the rows where the value of the `orderId` column is `100`.

> **Tip:** Sample filters are very useful for allowing you to generate samples that are much more specific to the steps that are trying to build at the present time in your recipe.

Scan method:

Depending on the type of sample, you may be able to select the method by which the data is scanned:

- **Quick Scan:** Representative sample is scanned and executed in-memory on the Trifacta node. Although the scope of the scanned data is smaller, these samples are much faster to generate.
  - If a Quick Scan sample fails, the Trifacta application may attempt to perform the scan on an available clustered running environment.
- **Full Scan:** Data is sampled from the full set of available data. The sampling job is executed on an available clustered running environment. These sampling jobs can take longer to execute. Depending on your environment, additional costs may be be incurred.

Initial Data Samples

These samples are collected automatically when you first load a new dataset into the Transformer page. These sample contain the first 10 MB of data from the first file or table in the dataset.

> **Tip:** In the Transformer page, these samples are labeled as **Initial Data**.

First Rows Samples

> **NOTE:** The First rows sampling technique requires the Trifacta Photon running environment.

This sample is taken from the first set of rows in the imported dataset based on the current cursor location in the recipe. The first N rows in the dataset are collected based on the recipe steps up to the configured sample size.

- This sample may span multiple datasets and files, depending on how the recipe is constructed.
- The first rows sample is different from the initial sample, which is gathered without reference to any recipe steps.

These samples are fast to generate. These samples may load faster in the application than samples of other types.

> **Tip:** If you have chained together multiple recipes, all steps in all linked recipes must be run to provide visual updates. If you are experiencing performance problems related to this kind of updating, you can select a recipe in the middle of the chain of recipes and switch it off the initial sample to a different sample. When invoked, the recipes from the preceding datasets do not need to be executed, which can improve performance.

Random Samples

Random selection of a subset of rows in the dataset. These samples are comparatively fast to generate. You can apply quick scan or full scan to determine the scope of the sample.

Filter-Based Samples

Find specific values in one or more columns. From the rows that have matching set of values, a random sample is generated.

You must define your filter in the Filter textbox.

Anomaly-Based Samples

Find mismatched or missing data or both in one or more columns.

You specify one or more columns and whether the anomaly is:

1. mismatched
2. missing
3. either of the above

Optionally, you can define an additional filter on any column.

Stratified Samples

Find all unique values within a column and create a sample that contains the unique values, up to the sample size limit. The distribution of the column values in the sample reflects the distribution of the column values in the dataset. Sampled values are sorted by frequency, relative to the specified column.

Optionally, you can apply a filter to this one.

> **Tip:** Collecting samples containing all unique values can be useful if you are performing mapping transformations, such as values to columns. If your mapping contains too many unique values among your key-value pairs, you can try to delete all columns except the one containing key-value pairs in a step, collect the sample, add the mapping step, and then delete the step where all other columns are removed.

Cluster-Based Samples

Cluster sampling collects contiguous rows in the dataset that correspond to a random selection from the unique values in a column. All rows corresponding to the selected unique values appear in the sample, up to the maximum sample size. This sampling is useful for time-series analysis and advanced aggregations.

Optionally, you can apply an advanced filter to the column.