Sampling Basics

A **sample** is a selection of rows from your dataset, which can be used as the basis for building the transformation steps in your recipe. The Cloud Dataprep application automatically creates initial samples of your data whenever you create a new recipe for a dataset and enables you to create additional samples at any time using a variety of sampling techniques.

Initial Sample

When you create a new recipe and load it in the Transformer page, the Cloud Dataprep application displays the initial sample of the dataset. The **initial sample** consists of the first X rows of the datasets, where X is determined by the following factors:

- The number of columns in the dataset
- The amount of data in each cell
- The maximum permitted size of each sample

Take a Sample

These first rows are displayed for you to begin your work in the Transformer page. However, you may begin to run into limitations with this sample. For example, suppose your dataset is organized by date, with earliest dates listed first. There may be significant changes in the data later in the time period that do not appear in the initial sample. You may decide that you need to take a different sample that captures some of these changes.

Steps:

1. In the Transformer page, click the Eyedropper icon at the top of the page.
2. The Samples panel is displayed.

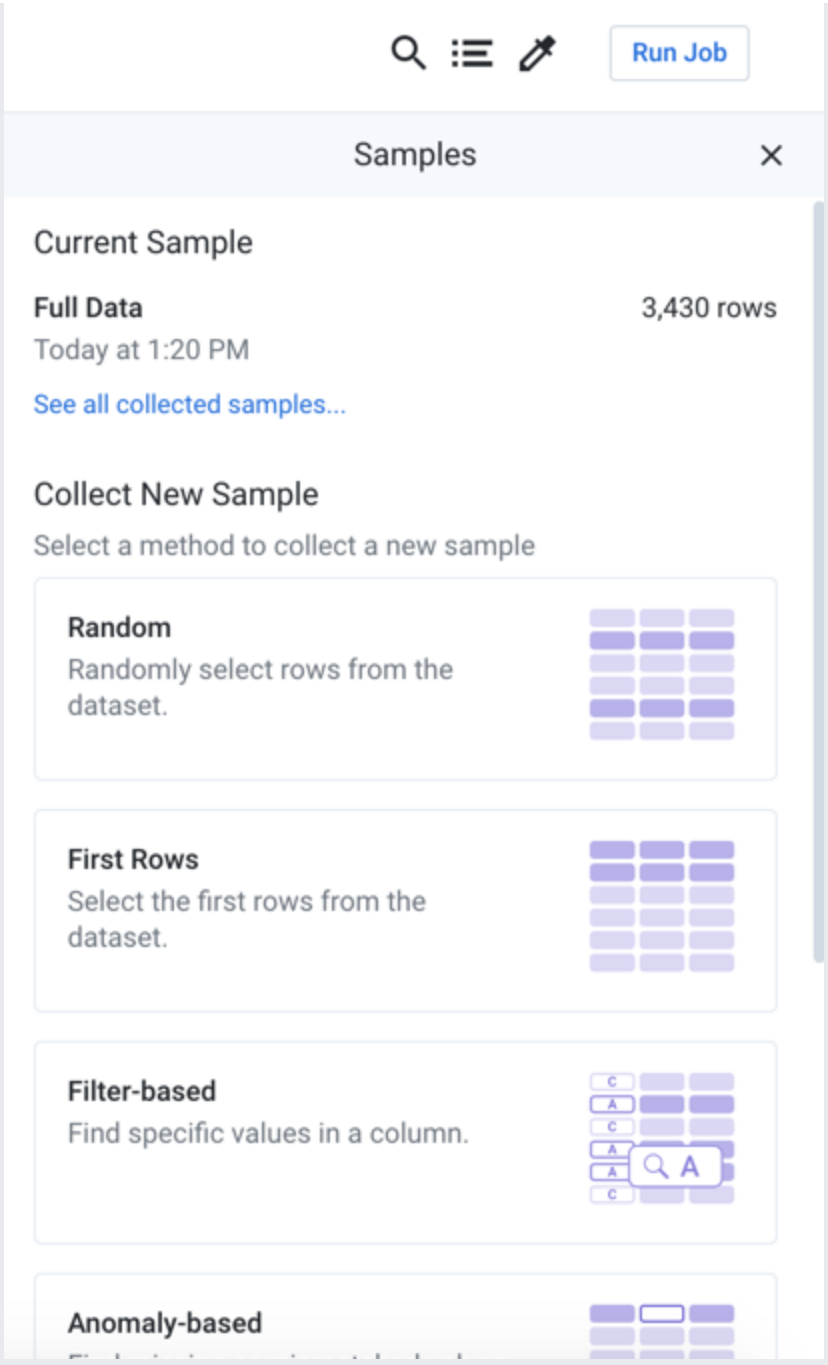

Figure: Samples panel

3. At the top of the panel, you can review the Current Sample.

   Tip: If the current sample indicates Full Data, then the entire dataset is displayed in the data grid. Unless you wish to use a specific sampling technique to filter down your data, sampling may not be useful across the entire dataset.

4. Below the current sample, you can see the available sample types. To take a new random sample:

a. Click the Random card.

b. Depending on your product edition, you may be able to select Quick Scan or Full Scan.

i. Quick Scan creates your sample by making some assumptions about the data when it scans.

ii. Full Scan creates your sample by scanning across all rows of the dataset. This option can take awhile across a large dataset.

c. Click **Collect**.

5. The sampling job is queued for execution. When it completes, click **Load Sample**.

6. The data grid is refreshed to display the rows gathered in the new random sample.

For more information, see Samples Panel (/dataprep/docs/html/Samples-Panel_57344905).

Sampling and Memory

NOTE: After you generate a sample, all steps in a recipe that occur after the step selected when you generated the sample are executed in browser memory on the sample data and then displayed in the data grid.

The above statement is best explained by example:

Action	Sampling
1. Create a new recipe and open it in Transformer page.	The initial sample is generated and displayed.
2. Add 3 steps to your recipe.	The 3 new steps are applied to the initial sample in the browser's memory.
3. Generate a new random sample.	The random sample is generated. When you load the sample, it is displayed in the data grid.
4. Add 25 steps to your recipe.	The 25 new steps are applied to the random sample in the browser's memory.
5. Select one of the first 3 steps of your recipe.	The initial sample is loaded and displayed.

6. Insert a new step below the current one.	Now, the first 4 steps are displayed using the initial sample.

Implications:

- As you add steps to your recipe without resampling, your recipe and sample consume more memory in your browser.
- When you perform complex multi-dataset operations, such as joins or unions, your recipe/sample combination consumes a lot more memory.
- If you continue adding steps:
  - Performance in the browser can be impacted. Basic operations such as selection of data or new recipe steps can become slow to respond.
  - The browser can crash.

Sampling Considerations

Tip: When resources permit, it's a good habit to take a new sample after a few multi-dataset operations or operations that otherwise change the number of rows in your dataset have been added to your recipe.

Other considerations:

- **Generating samples takes time.** This is particularly true for Full Scan samples.
- **Sampling can cost money.** In some cloud-based environments, generating a sample costs compute resources, which can add to your computing bill.
- **You may need multiple samples.** For long or complex recipes, you may need to take multiple samples.
- **Reference datasets should begin with a sample.** When you create a recipe for a reference dataset, you should start by generating a new sample for it.

Invalid samples

Samples can become invalid. If you recipe steps change the number of rows or otherwise reshape your dataset using transformations such as pivot or join in the steps leading up to where you took the current sample, your existing sample may no longer be valid.

When the application determines that a sample is invalid:

- The sample can no longer be used. It is now listed under the Unavailable tab in the Samples panel.
- The application automatically reverts to the last known good sample.

  NOTE: Depending on when the last known good sample was generated, this reversion could suddenly force a large number of steps to be processed in the browser's memory.
- You should consider generating a new sample immediately.

For more information, see Overview of Sampling (/dataprep/docs/html/Overview-of-Sampling_90112099).For more information on best practices, see https://community.trifacta.com/s/article/Best-Practices-Managing-Samples-in-Complex-Flows (https://community.trifacta.com/s/article/Best-Practices-Managing-Samples-in-Complex-Flows).

Last updated 2020-07-13 UTC.