Import Data Page

Through the Import Data page, you can upload datasets or select datasets from sources that are stored on connected datastores. From the Library page, click **Import Data**.

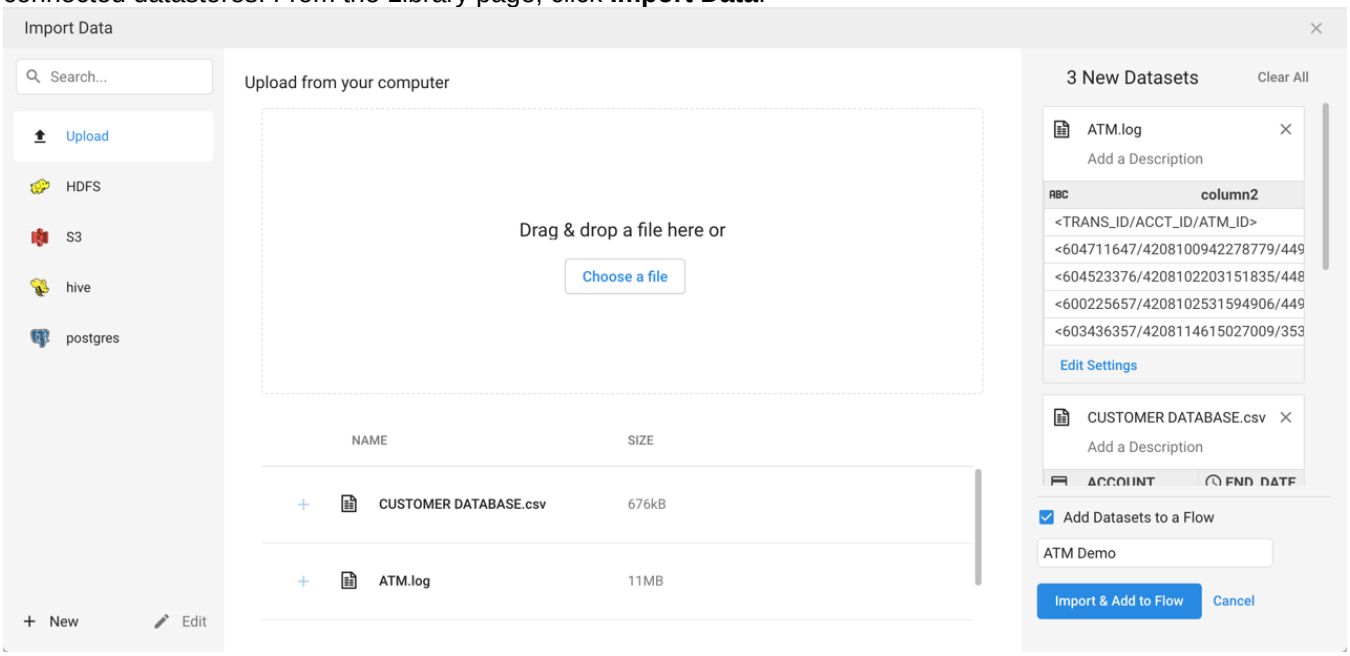

Figure: Import Data page

To import new data:

> **NOTE:** For file-based sources, Trifacta® Self-Managed Enterprise Edition expects that each row of data in the import file is terminated with a consistent newline character, including the last one in the file.
>
> - For single files lacking this final newline character, the final record may be dropped.
> - For multi-file imports lacking a newline in the final record of a file, this final record may be merged with the first one in the next file and then dropped in the Trifacta Photon running environment.

> **NOTE:** To be able to import datasets from the base storage layer, your user account must include the `dataAdmin` role.

> **NOTE:** An imported dataset requires about 15 rows to properly infer column data types and the row, if any, to use for column headers.

1. Connect to the source of your data:

   > **NOTE:** Compressed files are recognized and can be imported based on their file extensions.

   a. **Upload:** Trifacta® Self-Managed Enterprise Edition can also load files from your local file system.

> **Tip:** You can drag and drop files from your desktop to to upload them.

b.

HDFS: If connected to a Hadoop cluster, you can select file(s) or folders to import. See *HDFS Browser*.

S3: If connected to an S3 instance, you can browse your S3 buckets to select source files.

> **Tip:** For HDFS and S3, you can select folders, which selects each file within the directory as a separate dataset.

See *S3 Browser*.

Redshift: If connected to an S3 datawarehouse, you can import source from the connected database. See *Redshift Browser*.

Hive: If connected to a Hive instance, you can load datasets from individual tables within the set of Hive databases. See *Hive Browser*.

Alation: If connected to Alation, you can search for and import Hive tables as imported datasets. For more information, see *Using Alation*.

Waterline: If connected to Waterline, you can search for and import datasets through the data catalog. For more information, *Using Waterline*.

Databases: If connected to a relational datastore, you can load tables or views from your database. See *Database Browser*.

WASB: If enabled, you can import data into your Azure deployment from WASB. For more information, see *WASB Browser*.

ADL: If enabled, you can import data into your Azure deployment from ADLS. The ADLS browser is very similar to the one for HDFS. See *HDFS Browser*.

c.

d. For more information on the supported input formats, see *Supported File Formats*.

2.

New/Edit: Click to create or edit a connection.

Search: Enter a search term to locate a specific connection.

> **NOTE:** This feature may be disabled in your environment. For more information, contact your Trifacta administrator.

See *Create Connection Window*.

3. Add datasets:
   a. When you have found your source directory or file:

i. You can hover over the name of a file to preview its contents.

> **NOTE:** Preview may not be available for some sources, such as Parquet.

ii. Click the Plus icon next to the directory or filename to add it as a dataset.

> **Tip:** You can import multiple datasets at the same time. See below.

b. **Excel files:** Click the Plus icon next to the parent workbook to add all of the worksheets as a single dataset, or you can add individual sheets as individual datasets. See *Import Excel Data*.

c.

If custom SQL query is enabled, you can click **Create Dataset with SQL** to enter a customized SQL statement to pre-filter the relational or Hive table within the database to include only the rows and columns of interest.

> **Through this interface, it is possible to enter SQL statements that can delete data, change table schemas, or otherwise corrupt the targeted database. Please use this feature with caution.**

For more information, see *Create Dataset with SQL*.

d.

If parameterization has been enabled, you can apply parameters to the source paths of your datasets to capture a wider set of sources. Click **Create Dataset with Parameters**.

See *Create Dataset with Parameters*.

This feature must be enabled. For more information, see *Overview of Parameterization*.

4. When a dataset has been selected, the following fields appear on the right side of the screen. Modify as needed:
   a. **Dataset Name:** This name appears in the interface.
   b. **Dataset Description:** You may add an optional description that provides additional detail about the dataset. This information is visible in some areas of the interface.

> **Tip:** Click the Eye icon to inspect the contents of the dataset prior to importing.

5. You can select a single dataset or multiple datasets for import.
6. You can modify settings used during import for individual files. In the card for an individual dataset, click **Edit Settings**.

> **NOTE:** In some cases, there may be discrepancies between row counts in the previewed data versus the data grid after the dataset has been imported, due to rounding in row counts performed in the preview.

   a. **Per-file encoding:** By default, Trifacta Self-Managed Enterprise Edition attempts to interpret the encoding used in the file. In some cases, the data preview panel may contain garbled data, due to a mismatch in encodings. In the Data Preview dialog, you can select a different encoding for the file. When the correct encoding is selected, the preview displays the data as expected.For more information on supported encodings, see *Configure Global File Encoding Type*.

b. **Detect structure:** By default, Trifacta Self-Managed Enterprise Edition attempts to interpret the structure of your data during import. This structuring attempts to apply an initial tabular structure to the dataset.
    i. Unless you have specific problems with the initial structure, you should leave the Detect structure setting enabled. Recipes created from these imported datasets automatically include the structuring as the first, hidden steps. These steps are not available for editing, although you can remove them through the Recipe panel. See *Recipe Panel*.
    ii. When detecting structure is disabled, imported datasets whose schema has not been detected are labeled, **unstructured datasets**. When recipes are created for these unstructured datasets, the structuring steps are added into the recipe and can be edited as needed.
    iii. For more information, see *Initial Parsing Steps*.
c. **Remove special characters from column names:** When selected, characters that are not alphanumeric or underscores are stripped, and space characters are converted to underscores.

    > **Tip:** This feature matches the column renaming behavior in Release 5.0 and earlier.

    For more information, see *Sanitize Column Names*.

d. **Column data type inference:** You can choose whether or not to apply Trifacta type inference to your individual dataset.

    i. In the preview panel, you can see the data type that is to be applied after the dataset is imported. This data type may change depending on whether column data type inference is enabled or disabled for the dataset.
    ii. To enable Trifacta type inference, select the Column Data Type Inference checkbox.

    > **Tip:** To see the effects of Trifacta type inference, you can toggle the checkbox and review data type listed at the top of individual columns. To override an individual column's data type, click the data type name and select a new value.

    iii. You can configure the default use of type inference at the individual connection level. For more information, see *Create Connection Window*.

    For schematized sources that do not require connections, such as uploaded Avro files, the default setting is determined by the global setting for initial type inference. For more information, see *Configure Type Inference*.

7. If you have selected a single dataset for import:
    a. To immediately wrangle it, click **Import & Wrangle**. The dataset is imported. A recipe is created for it, added to a flow, and loaded in the Transformer page for wrangling. See *Transformer Page*.
    b. To import the dataset, click **Import**. The imported dataset is created. You can add it to a flow and create a recipe for it later. See *Library Page*.
8. If you have selected multiple datasets for import:
    a. To import the selected datasets, click **Import Datasets**. The imported datasets are created. You can begin working with these imported datasets now or at a later time.
    b. To import the selected datasets and add them to a flow:
        i. Click the Add Dataset to a Flow checkbox.
        ii. Click the textbox to see the available flows, or start typing a new name.
        iii. Click **Import & Add to Flow**.
        iv. The datasets are imported, and the associated recipes are created. These datasets and recipes are added to the selected flow.
        v. For any dataset that has been added to a flow, you can review and perform actions on it. See *Flow View Page*.

9. If you are not wrangling the datasets immediately, the datasets you just imported are listed at the top of the Library page. See *Library Page*.

Import Multiple Datasets

You can import multiple datasets from multiple sources at the same time. In the Import Data page, continue selecting sources, and additional dataset cards are added to the right panel.

> **NOTE:** If you are importing from multiple files at the same time, the files are not necessarily read in a regular or predictable order.

> **NOTE:** When you import a dataset with parameters from multiple files, only the first matching file is displayed in the right panel.

In the right panel, you can see a preview of each dataset and make changes as needed.

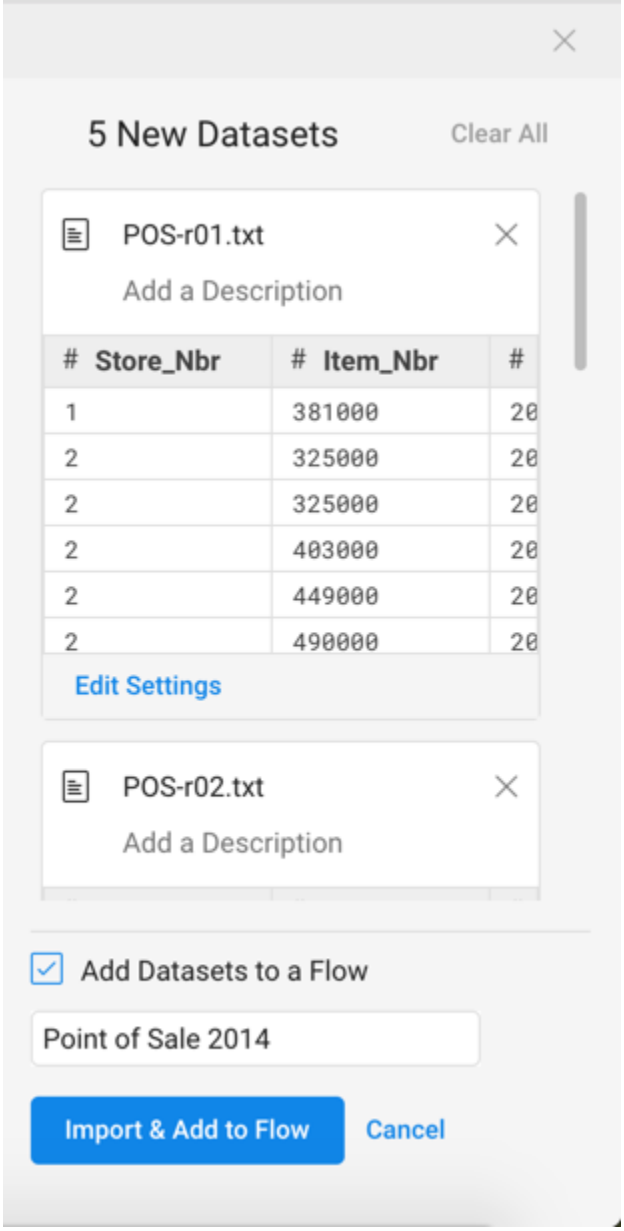

Figure: Import Multiple Datasets

- To remove a dataset from import, click the X in the dataset card.
- To add the datasets to a flow, click the checkbox. Then, select an existing flow or enter the name of a new flow to contain your datasets.
- To import the datasets, click **Import** or **Import & Add to Flow**.