Import PDF Data

> **NOTE:** This feature is in Beta release.

Trifacta® can directly import Adobe® Acrobat® PDF files containing one or more tables. The tables of a PDF can be imported as:

- Individual datasets
- A single dataset
- A dataset with parameters

> **NOTE:** When importing as a parameterized dataset, all selected tables are imported into a single dataset.

PDF files can be uploaded from your local system. If Trifacta is connected to a backend file storage system, you can also import PDF files stored in readable directories.

Limitations

- PDF ingest is limited to 100 MB per file.
- Filepath and source row number information is not available from original PDF files. These references return values from the CSV files that have been converted on the backend. For more information, see *Source Metadata References*.
- You cannot import password-protected PDF files.
- Compressed PDF files are not supported.
- Conversion of large PDF files require non-linear increases in memory requirements on the Trifacta node.

- If loading your PDF-based dataset in the Transformer page results in a blank screen, please take a new sample. The file requires conversion again with each generated sampling.
- Latest state of the PDF file may not be reflected in the Transformer page due to caching. When you run a job, the platform always collects the latest version of the data and converts it to CSV for execution.

Enable

This feature is disabled by default. To enable, please complete the following:

Steps:

1. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.
2. Locate the following parameter and set it to `true`:

   ```
   "feature.enablePDFSupport": false,
   ```

3. Add references to the PDF format to the following parameters:

   ```
   "webapp.convertableExtensions": "xls,XLS,xlsx,XLSX,pdf,PDF",
   "webapp.client.allowedFileExtensions": "<other_options>,pdf,PDF",
   ```

4. Save your changes and restart the platform.

Table Import

The PDF file format is a publishing format designed around visual layout of information, some of which may include tabular data. Table data in PDF files must be detected and converted into CSV data for proper ingestion in the platform. This ingest process occurs on the backend datastore.

To facilitate ingestion, the following requirements must be met for tables in your source PDF files:

- Non-tabular data in the file is ignored.
- Tables must be enclosed in a border. Each cell in the table must be bordered.
- Tabular data in the PDF cannot be scanned data, which is stored as an image. Data must be written into the file.
- When a table spans multiple pages, it is ingested as two separate CSV files, which can be combined later.
- If a file contains multiple tables, each table is converted as a separate dataset.

> **Tip:** After import, separate datasets can be unioned together or integrated using as a dataset with parameters.

Import Steps

1. In the menu bar, click **Library**.
2. In the Library page, click **Import Data**. Select the connection to use. See *Import Data Page*.

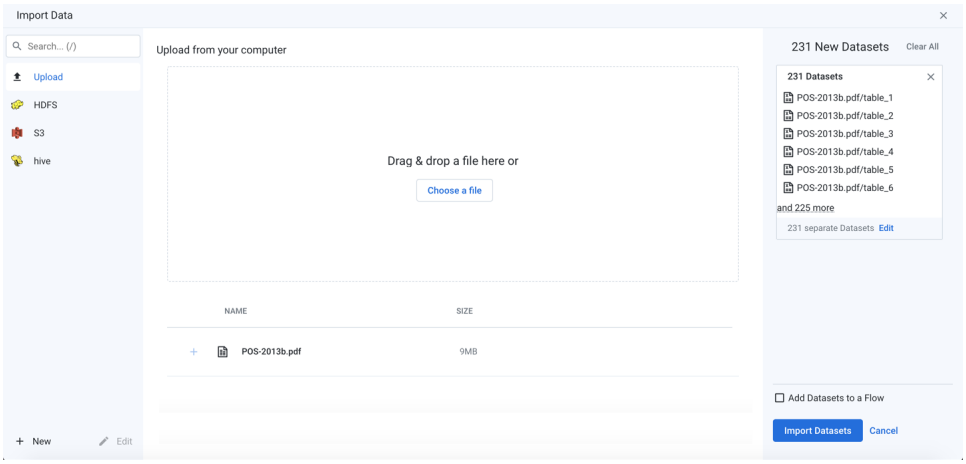

Figure: Import PDF file containing multiple pages

3. After you select the file, it is uploaded and converted to into individual CSV files for each page in the PDF file and then stored by the platform. Depending on the size of the file, this process may take a while.

4. By default, all pages in the PDF are imported as individual datasets. To change how the data is imported, click **Edit** in the right panel.

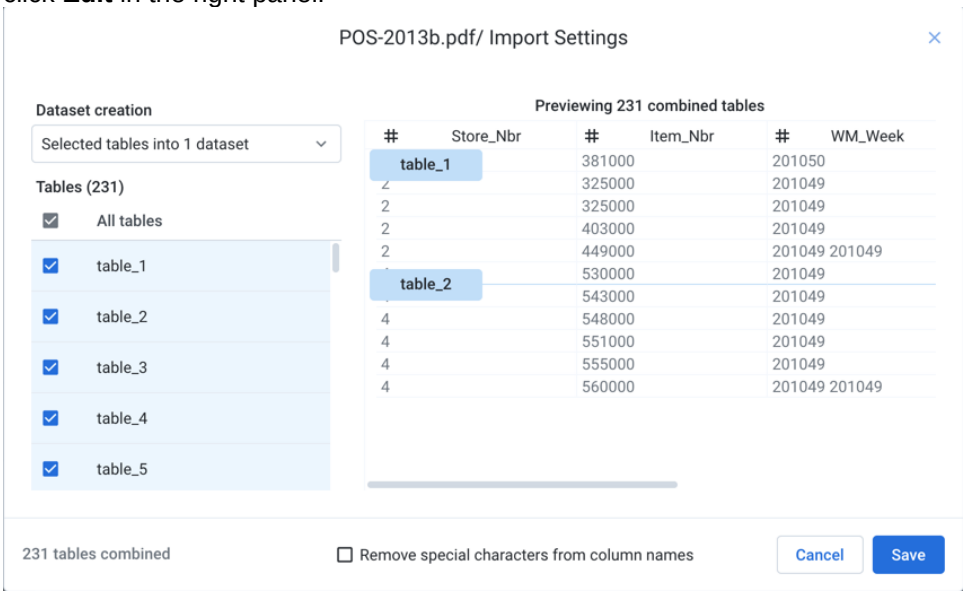

Figure: Import settings for PDF datasets

5. Dataset creation:
   a. **1 dataset per table:** (Default) Each selected table in the PDF is imported as a separate dataset. Specify the base name of the datasets that you are creating. If you are creating a single dataset, the name of the PDF file is used.
   b. **Selected tables into 1 dataset:** All selected tables in the PDF are combined and imported as a single dataset.

      > **NOTE:** The schemas of each dataset must match. Columns must be listed in the same order in each dataset. The column headers are taken from the first selected dataset.

   c.

      All and future tables into 1 dataset: If the PDF is updated periodically with new tables that you would like to add in the future, select this option. After initial selection of the tables to include, all PDF pages that are added to the PDF file in the future are automatically added as part of the imported dataset.

      > **NOTE:** This option is available only if you are connected to a backend file storage system.

      > **NOTE:** When an imported dataset based on this option is first loaded into the Transformer page, the data grid displays an initial sample taken from rows in the first sheet only. When you take another sample from the Samples panel, data is collected from other sheets. For more information, see *Samples Panel*.

6. Selected tables:
   a. You can select the tables to import. A table can be a single page, or a single table among multiple on a page.

      > **NOTE:** If you are importing a folder of PDF files, data preview and initial sampling are executed against the first file found in the folder.

    b. To preview the data of an individual table, mouse over a dataset and click **Jump to**.
7. Remove special characters from column names: Select this option to remove any special characters from the inferred column headers during import.
8. To save changes, click **Save**.
9. After your datasets have been added, you can edit the name and description information for each in the right navigation panel.
10. Optionally, you can assign the new dataset(s) to an existing flow or create a new one to contain them.

See *Import Data Page*.