FTP Server Import Integration

Use this data connector to directly import data from your FTP server to Treasure Data.

For sample workflows on importing data from your FTP server, view Treasure Boxes.

Prerequisites

- Basic knowledge of Treasure Data
- Basic knowledge of FTP

Requirements

- Make sure you have a valid protocol. If you intend to *FTP* or *FTPS*, you can use this Data Connector for FTP. If *SFTP*, use the SFTP Integration.
- If you're using a firewall, check your accepted IP range/port. Server administrators sometimes change the default port number for security reasons.
- Be sure that FTP uses *TCP/21* as the default control port but also uses any TCP ports as a data transfer port when you're using passive mode. This port range will depend on your server's settings.
- Check that you're connecting with *passive* mode. *activeP* mode generally doesn't work because it establishes the connection from the FTP server-side.
- If you're using FTPS, there are 2 modes *Explicit* and *Implicit*. Explicit mode is typically used.

About Incremental Data Loading

Incremental loading is the activity of loading only new or updated records from a source into Treasure Data. Incremental loads are useful because they run efficiently when compared to full loads, and particularly for large data sets.

Incremental loading is available for many of the Treasure Data integrations. In some cases, it is a simple checkbox choice and in others, after you select incremental loading you are provided with other fields that must be specified.

Limitations, Supported, Suggestions

- For some integrations, if you choose incremental loading, you might need to make sure that there is an index on the columns to avoid a full table scan.
- Only Timestamp, Datetime, and numerical columns are supported as incremental_columns.
- For the raw query, the incremental_columns is required because it won't be able to detect the Primary keys for a complex query.

About Incremental Loading for Integrations

Treasure Data Incremental loading has 4 patterns (3 types of data connector + 1 workflow `td_load` operator.), then the 3 data connector loading examples are as follows:

- Cloud storage service (e.g. AWS S3, GCS and etc.)
    - Lexicographic order of file name
- Query (e.g. MySQL, BigQuery and etc.)
    - Date time
- Variable period (Google Analytics, etc)

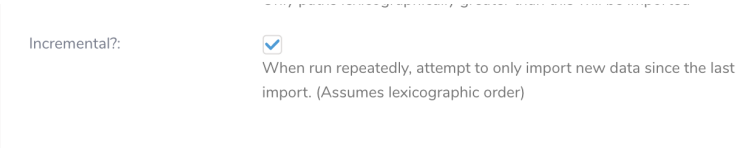
- Use `start_date` for loading

Incremental Loading for Connectors

If incremental loading is selected, data for the connector is loaded incrementally.

This mode is useful when you want to fetch just the object targets that have changed since the previously scheduled run.

For example, in the UI:



Database integrations, such as MySQL, BigQuery, and SQL server, require column or field names to load incremental data. For example:

Learn more About Database-based Integrations.

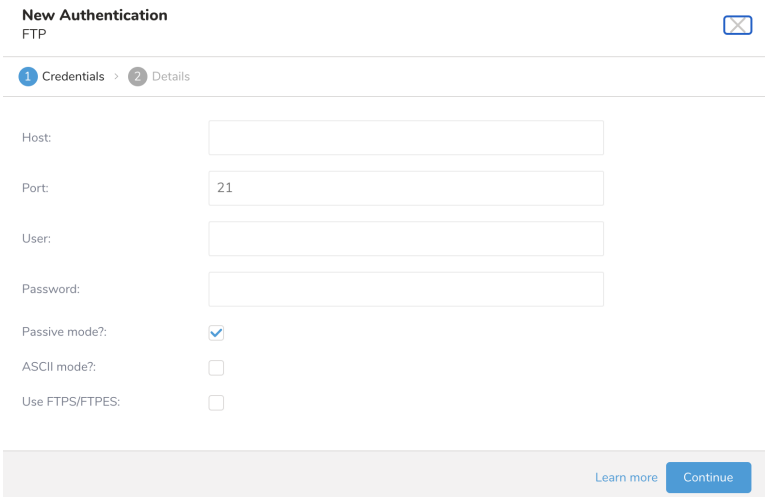
Use the TD Console to Create Your Connection

Create a New Connection

In Treasure Data, you must create and configure the data connection prior to running your query. As part of the data connection, you provide authentication to access the integration.

1. Open **TD Console**.
2. Navigate to **Integrations Hub** > **Catalog**.
3. Search for and select FTP.

4. Select **Create Authentication**.



5. Enter the required credentials for your remote FTP instance. Depending on your selections, the fields you see might vary:

Field	Description
Host	The host information of the remote FTP instance, for example, an IP address.
Port	The connection port on the remote FTP instance the default is 21.
User	The user name used to connect to the remote FTP instance.
Password	The password used to connect to the remote FTP instance.
Passive mode	Use passive mode (default: checked)
ASCII mode	Use ASCII mode instead of binary mode (boolean, default: unchecked)
Use FTPS /FTPES	Use FTPS (SSL encryption). (boolean, default: unchecked)
Verify cert	Verify the certification provided by the server. By default, the connection fails if the server certificate is not signed by one of the CAs in JVM's default trusted CA list.
Verify hostname	Verify server's hostname matches the provided certificate.
Enable FTPES	FTPES is a security extension to FTPS
SSL CA Cert Content	Paste the contents of the certificate file

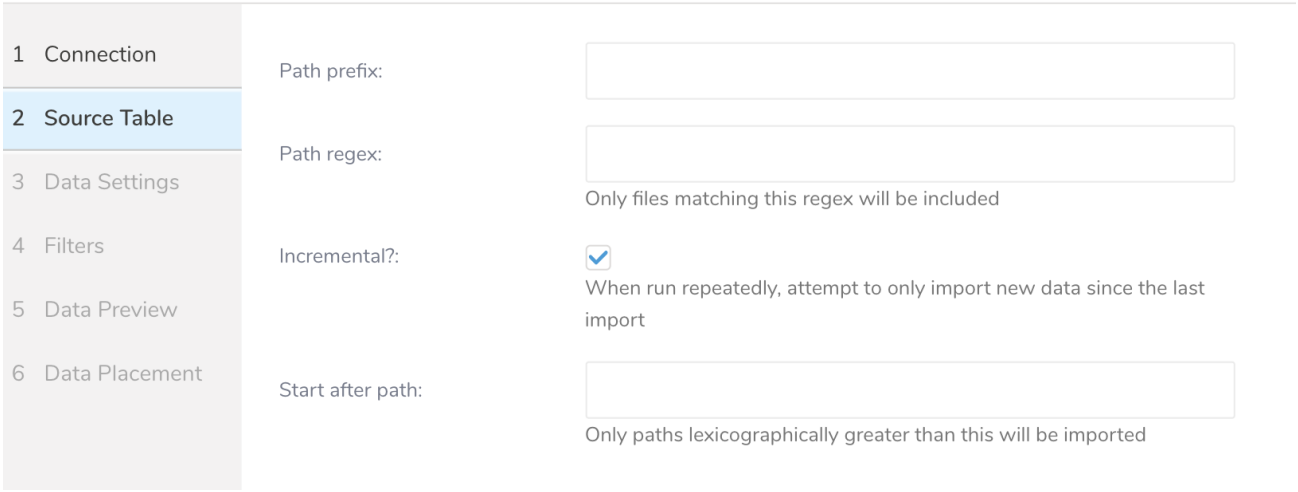
6. Select **Continue.**
7. Enter a name for your connection.
8. Select **Continue.**

Transfer Your Data to Treasure Data

After creating the authenticated connection, you are automatically taken to Authentications.

1. Search for the connection you created.
2. Select **New Source**.
3. Type a name for your **Source** in the Data Transfer field**.**
4. Select **Next**.

   The Source Table dialog opens.

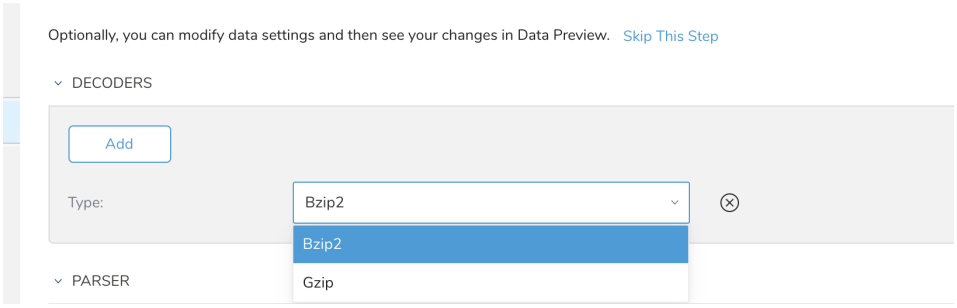


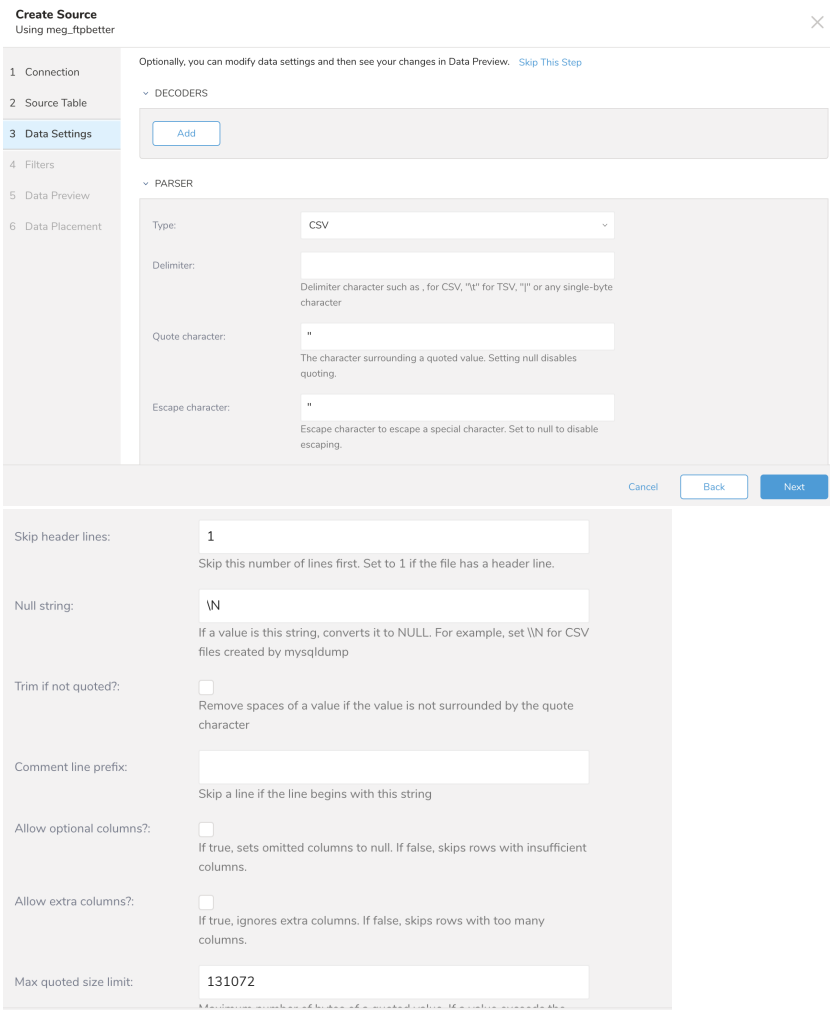
5. Edit the following parameters:

Parameters	Description
Path prefix	The prefix of target files (string, required). For example, resultoutputtest.
Path regex	Type a regular expression to query file paths. If a file path doesn't match the specified pattern, the file is skipped. For example, if you specify the pattern .csv$ #, then a file is skipped if its path doesn't match the pattern.
Incremental	Enables incremental loading (boolean, optional. default: true. If incremental loading is enabled, the config diff for the next execution will include last_path parameter so that the next execution skips files before the path. Otherwise, last_path is not included.
Start after path	Only paths lexicographically greater than this will be imported.

6. Select **Next**.

   The Data Settings page can be modified for your needs or you can skip the page.





7. Optionally, edit the parameters.

8. Select **Next**.

Filters

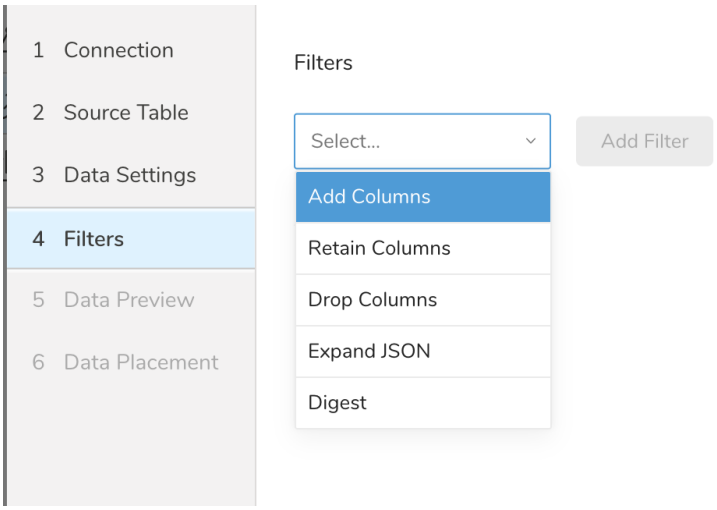
Import Integration Filters enable you to modify your imported data after you have completed Editing Data Settings for your import.

To apply import integration filters:

1. Select **Next** in Data Settings.

   The Filters dialog opens.

2. Select the filter option you want to add.



3. Select **Add Filter**.

The parameter dialog for that filter opens.

4. Edit the parameters.

For information on each filter type, see one of the following:
—Retaining Columns Filter
—Adding Columns Filter
—Dropping Columns Filter
—Expanding JSON Filter
—Digesting Filter

5. Optionally, to add another filter of the same type, select **Add** within the specific column filter dialog.

6. Optionally, to add another filter of a different type, select the filter option from the list and repeat the same steps.

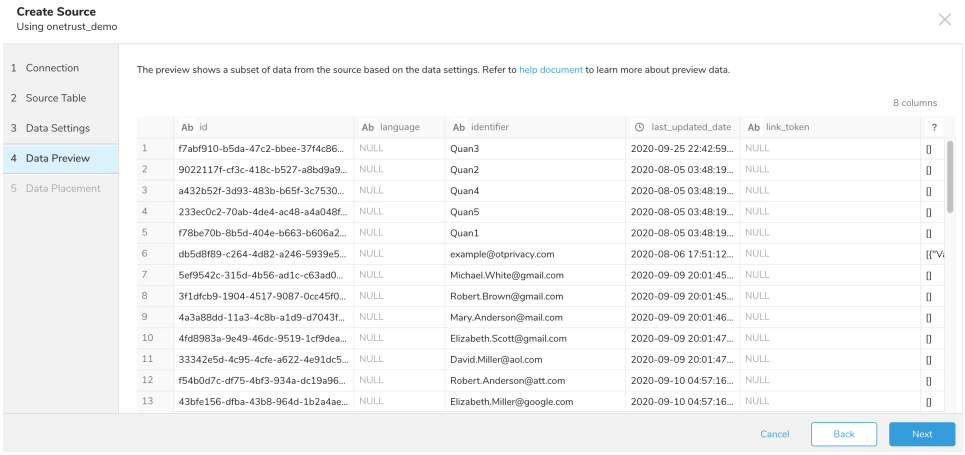
7. After you have added the filters you want, select **Next**.
The Data Preview dialog opens.

Data Preview

You can see a preview of your data before running the import by selecting Generate Preview.

Data shown in the data preview is approximated from your source. It is not the actual data that is imported.

1. Click **Next**.
   Data preview is optional and you can safely skip to the next page of the dialog if you want.
2. To preview your data, select **Generate Preview**. Optionally, click **Next**.
3. Verify that the data looks approximately like you expect it to.

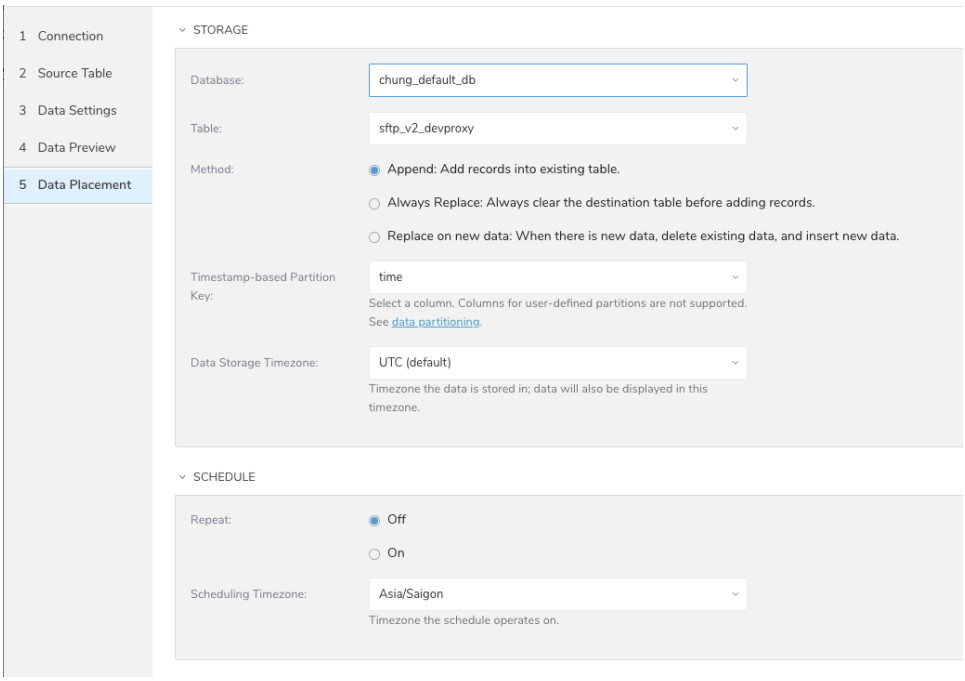


4. Select **Next**.

Data Placement

For data placement, select the target database and table where you want your data placed and indicate how often the import should run.

1. Select **Next.** Under Storage you will create a new or select an existing database and create a new or select an existing table for where you want to place the imported data.



2. Select a **Database** > **Select an existing** or **Create New Database**.
3. Optionally, type a database name.
4. Select a **Table**> **Select an existing** or **Create New Table**.
5. Optionally, type a table name.
6. Choose the method for importing the data.

    - **Append** (default)-Data import results are appended to the table.
      If the table does not exist, it will be created.
    - **Always Replace**-Replaces the entire content of an existing table with the result output of the query. If the table does not exist, a new table is created.
    - **Replace on New Data**-Only replace the entire content of an existing table with the result output when there is new data.
7. Select the **Timestamp-based Partition Key** column.
   If you want to set a different partition key seed than the default key, you can specify the long or timestamp column as the partitioning time. As a default time column, it uses upload_time with the add_time filter.

8. Select the **Timezone** for your data storage.
9. Under **Schedule**, you can choose when and how often you want to run this query.

    - Run once:

        a. Select **Off**.
        b. Select **Scheduling Timezone**.
        c. Select **Create & Run Now**.
    - Repeat the query:

        a. Select **On**.
        b. Select the **Schedule**. The UI provides these four options: *@hourly*, *@daily* and *@monthly* or custom *cron*.
        c. You can also select **Delay Transfer** and add a delay of execution time.
        d. Select **Scheduling Timezone**.
        e. Select **Create & Run Now**.

After your transfer has run, you can see the results of your transfer in **Data Workbench** > **Databases**.

Validate Connection

Review the job log. Warning and errors provide information about the success of your import. For example, you can identify the source file names associated with import errors.

Optionally Configure Export Results in Workflow

Within Treasure Workflow, you can specify the use of this data connector to export data.

Learn more at Using Workflows to Export Data with the TD Toolbelt.

Example Workflow for FTP

```
timezone: UTC

schedule:
  daily>: 02:00:00

sla:
  time: 08:00
  +notice:
    mail>: {data: Treasure Workflow Notification}
    subject: This workflow is taking long time to finish
    to: [meg@example.com]

_export:
  td:
    dest_db: dest_db
    dest_table: dest_table
  ftp:
    ssl: true
    ssl_verify: false

+prepare_table:
  td_ddl>:
  database: ${td.dest_db}
  create_tables: ["${td.dest_table}"]

+load_step:
  td_load>: config/daily_load.yml
  database: ${td.dest_db}
  table: ${td.dest_table}
```