Google Cloud Storage Import Integration

Learn more about Google Cloud Storage Export Integration.

The Data Connector for Google Cloud Storage enables import of the contents of *.tsv and* .csv files stored in your GCS bucket.

For sample workflows importing data from GCS, view the Treasure Boxes.

Prerequisites

- Basic knowledge of Treasure Data
- An existing Google Service Account

You also need to generate and obtain a JSON key file from Google Developers Console. See Generating a service account credential.

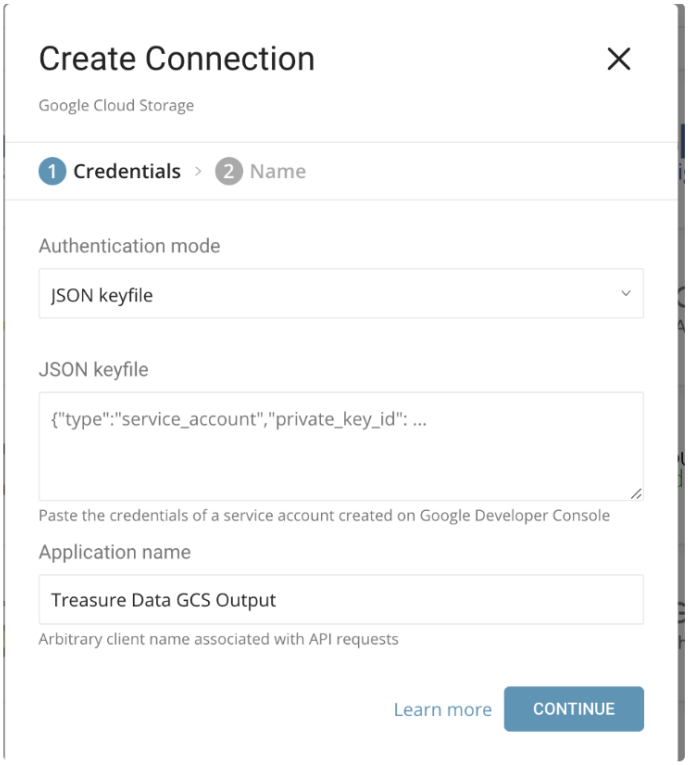
Use TD Console

Create a New Connection

When you configure a data connection, you provide authentication to access the integration. In Treasure Data, you configure the authentication and then specify the source information.

1. Open TD Console.
2. Navigate to Integrations Hub > Catalog
3. Search and select Google Cloud Storage.

4. The following dialog opens.



5. Create a New Google Cloud Storage Connector

6. Set the following parameters:

Parameters	Description
Authentication mode	Select a JSON keyfile. This method uses the JSON keyfile generated from the Google Developers Console.
JSON Keyfile	Copy and paste the contents of the JSON keyfile generated from the Google Developers Console in this field
Application Name	*Treasure Data GCS Output* is the default value. As this is an arbitrary client name associated with API requests, you can leave the default value (Treasure Data GCS Output).

Name Your Connection

1. Type a name for your connection.
2. Select **Done.**

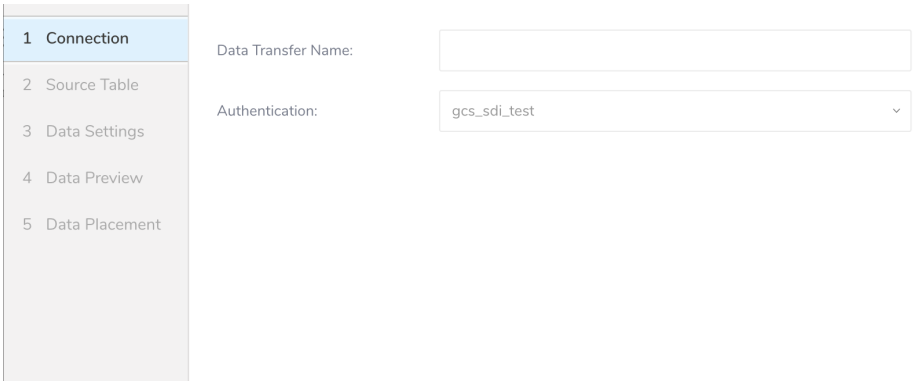
Transfer Your Google Cloud Storage Account Data to Treasure Data

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

1. Search for the connection you created.
2. Select **New Source**.

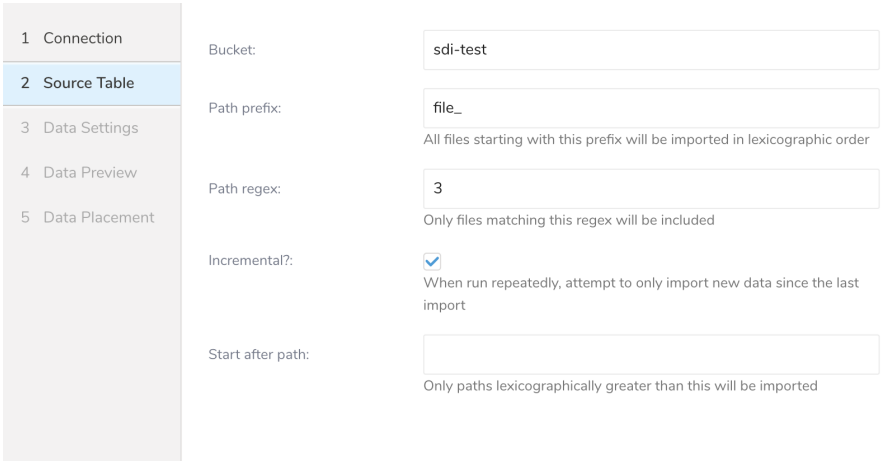
Connection

1. Type a name for your **Source** in the Data Transfer field**.**
2. Click **Next**.



Source Table

1. Select **Next**.
2. The Source Table dialog opens. Edit the following parameters



Parameters	Description
Bucket	Google Cloud Storage bucket name (Ex. *your_bucket_name*)
Path Prefix	Prefix of target keys. (Ex. *logs/data_*)
Path Regex	regexp to match file paths. If a file path doesn't match with this pattern, the file is skipped. (Ex. *.csv$ #* in this case, a file is skipped if its path doesn't match with this pattern)
Start after path	Inserts last_path parameter so that the first execution skips files before the path. (Ex. logs/data_20170101.csv)
Incremental	Enables incremental loading. If incremental loading is enabled, config diff for the next execution will include last_path parameter so that next execution skips files before the path. Otherwise, last_path will not be included.

Example: CloudFront

Amazon CloudFront is a web service that speeds up the distribution of your static and dynamic web content. You can configure CloudFront to create log files that contain detailed information about every user request that CloudFront receives. If you enable logging, you can save CloudFront logfiles, shown as follows:

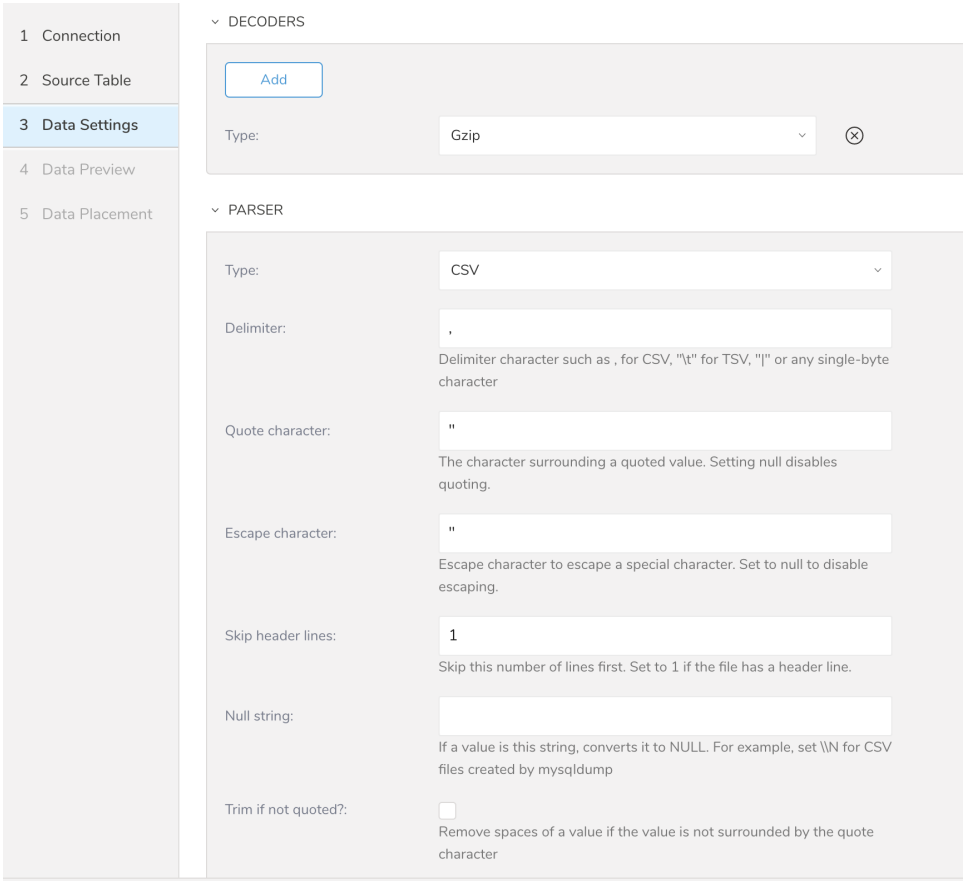
```
[your_bucket] - [logging] - [E231A697YXWD39.2017-04-23-15.a103fd5a.gz]
[your_bucket] - [logging] - [E231A697YXWD39.2017-04-23-15.b2aede4a.gz]
[your_bucket] - [logging] - [E231A697YXWD39.2017-04-23-16.594fa8e6.gz]
[your_bucket] - [logging] - [E231A697YXWD39.2017-04-23-16.d12f42f9.gz]
```

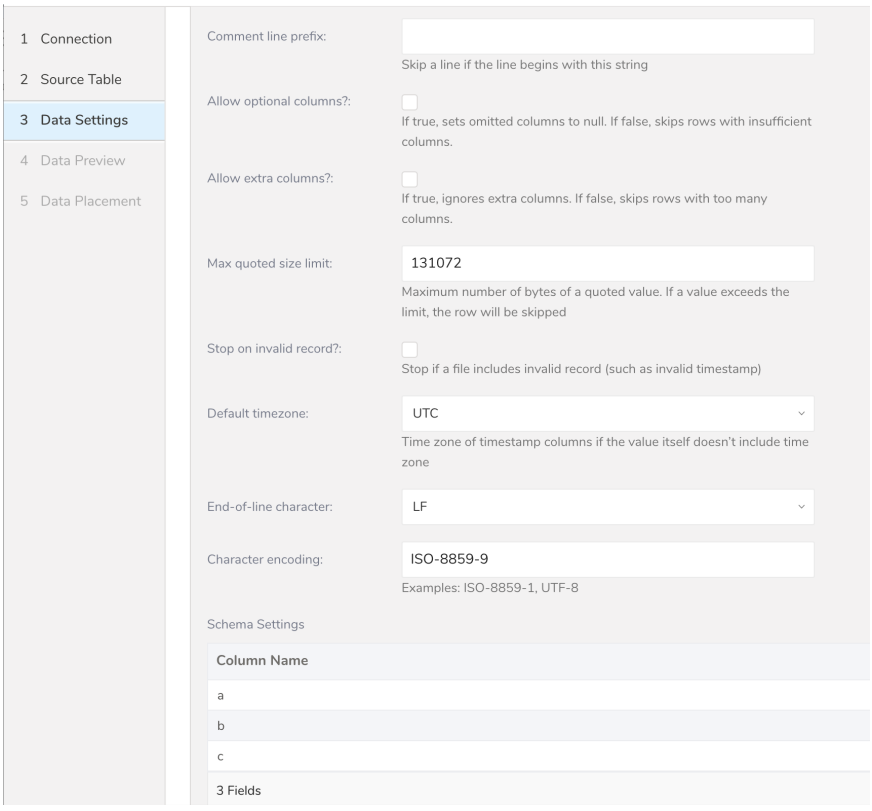
In this case, the Source Table setting should be as shown:

- **Bucket**: your_bucket
- **Path Prefix**: logging/
- **Path Regex**: *.gz$* (Not Required)
- **Start after path**: logging/E231A697YXWD39.2017-04-23-15.b2aede4a.gz (Assuming that you want to import the logfiles from 2017-04-23-16.)
- **Incremental**: true (if you want to schedule this job.)

Data Settings

1. Select **Next**.
   The Data Settings page opens.
2. Optionally, edit the data settings or skip this page of the dialog.


1 Connection
2 Source Table
3 Data Settings
4 Data Preview
5 Data Placement
DECODERS
Add
Type:
Gzip
PARSER
Type:
CSV
Delimiter:
,
Delimiter character such as , for CSV, "\t" for TSV, "|" or any single-byte character
Quote character:
"
The character surrounding a quoted value. Setting null disables quoting.
Escape character:
"
Escape character to escape a special character. Set to null to disable escaping.
Skip header lines:
1
Skip this number of lines first. Set to 1 if the file has a header line.
Null string:
If a value is this string, converts it to NULL. For example, set \\N for CSV files created by mysqldump
Trim if not quoted?:
Remove spaces of a value if the value is not surrounded by the quote character


1 Connection
2 Source Table
3 Data Settings
4 Data Preview
5 Data Placement
Comment line prefix:
Skip a line if the line begins with this string
Allow optional columns?:
If true, sets omitted columns to null. If false, skips rows with insufficient columns.
Allow extra columns?:
If true, ignores extra columns. If false, skips rows with too many columns.
Max quoted size limit:
131072
Maximum number of bytes of a quoted value. If a value exceeds the limit, the row will be skipped
Stop on invalid record?:
Stop if a file includes invalid record (such as invalid timestamp)
Default timezone:
UTC
Time zone of timestamp columns if the value itself doesn't include time zone
End-of-line character:
LF
Character encoding:
ISO-8859-9
Examples: ISO-8859-1, UTF-8
Schema Settings
Column Name
a
b
c
3 Fields

PARSER

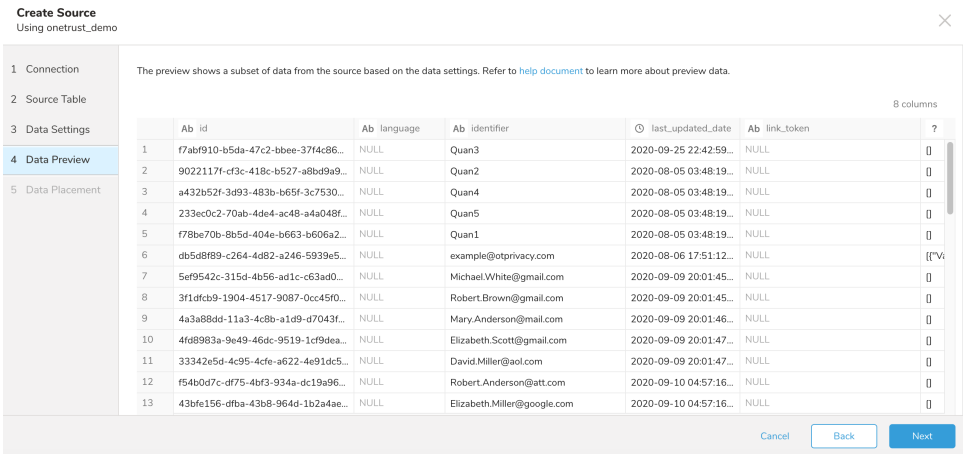
Parameters	Description
Type	Parses a value as a specified type. And then, it stores after converting to Treasure Data schema. • **boolean** • **long** • **timestamp**: will be imported as String type at Treasure Data (Ex. 2017-04-01 00:00:00.000) • **double** • **string** • **json**
Default timezone	Changes time zone of timestamp columns if the value itself doesn't include time zone.
Total file count limit	Maximum number of files to read. (optional)
Schema Settings	You can name the columns and set the data type.

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.

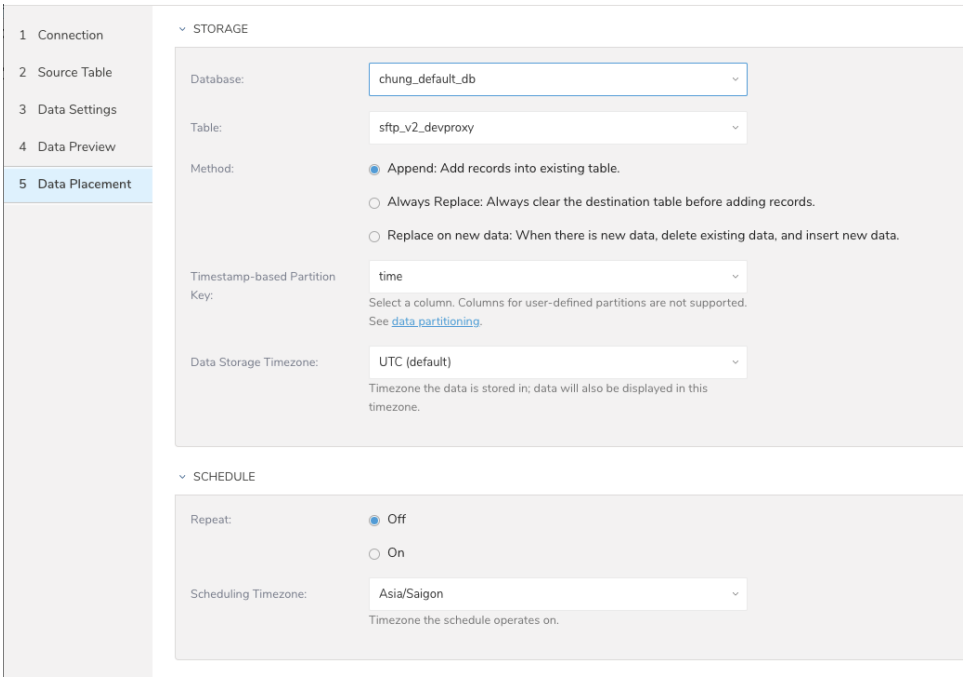


4. Select **Next**.

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**.