Using Google Cloud Storage

This section describes how you interact through Cloud Dataprep by TRIFACTA® INC. with your Google Cloud Storage environment.

Uses of Google Cloud Storage

Cloud Dataprep by TRIFACTA INC. can use Google Cloud Storage for the following reading and writing tasks:

1. **Upload through application:** When files are imported into Cloud Dataprep by TRIFACTA INC. as datasets, it is uploaded and stored in a location in Google Cloud Storage. For more information, see User Profile Page (/dataprep/docs/html/User-Profile-Page_57344911).
2. **Creating Datasets from Google Cloud Storage Files:** You can read in from source data stored in Google Cloud Storage. A source may be a single Google Cloud Storage file or a folder of identically structured files. See Reading from Sources below.
3. **Reading Datasets:** When creating a dataset, you can pull your data from another dataset defined in Google Cloud Storage.
4. **Writing Results:** After a job has been executed, you can write the results back to Google Cloud Storage.

In Cloud Dataprep by TRIFACTA INC., Google Cloud Storage is accessed through the user interface. See Google Cloud Storage Browser (/dataprep/docs/html/Google-Cloud-Storage-Browser_59736143).

NOTE: When Cloud Dataprep by TRIFACTA INC. executes a job on a dataset, the source data is untouched. Results are written to a new location, so that no data is disturbed by the process.

Before You Begin

Your administrator must configure read/write permissions to locations in Google Cloud Storage. Please see the Google Cloud Storage documentation.

Avoid reading and writing in the following locations:

The Scratch Area location is used by Cloud Dataprep by TRIFACTA INC. for temporary storage.

The Upload location is used for storing data that has been uploaded from local file.

For more information on these locations, see User Profile Page (/dataprep/docs/html/User-Profile-Page_57344911)**.**

Limitations:

- The Requestor Pays feature of Google Cloud Storage is not supported in Cloud Dataprep by TRIFACTA INC..

Storing Data in Google Cloud Storage

Your administrator should provide raw data or locations and access for storing raw data within Google Cloud Storage.

- All Cloud Dataprep users should have a clear understanding of the folder structure within Google Cloud Storage where each individual can read from and write results.
- Users should know where shared data is located and where personal data can be saved without interfering with or confusing other users.

NOTE: Cloud Dataprep by TRIFACTA INC. does not modify source data in Google Cloud Storage. Sources stored in Google Cloud Storage are read without modification from their source locations, and sources that are uploaded to the platform are stored in the designated Upload location for each user. See User Profile Page (/dataprep/docs/html/User-Profile-Page_57344911).

Reading from Sources

You can create a dataset from one or more files stored in Google Cloud Storage.

Wildcards:

You can parameterize your input paths to import source files as part of the same imported dataset. For more information, see Overview of Parameterization

(/dataprep/docs/html/Overview-of-Parameterization_118228665).

Folder selection:

When you select a folder in Google Cloud Storage to create your dataset, you select all files in the folder to be included.

- This option selects all files in all sub-folders and bundles them into a single dataset. If your sub-folders contain separate datasets, you should be more specific in your folder selection.
- All files used in a single imported dataset must be of the same format and have the same structure. For example, you cannot mix and match CSV and JSON files if you are reading from a single directory.

Read file formats:

From Google Cloud Storage, Cloud Dataprep by TRIFACTA INC. can read the following file formats:

- CSV
- JSON
- AVRO
- GZIP
- BZIP2
- TXT
- XLS/XLSX
- LOG
- TSV

Creating Datasets

When creating a dataset, you can choose to read data from a source stored from Google Cloud Storage or from a local file.

- Google Cloud Storage sources are not moved or changed.
- Local file sources are uploaded to the designated Upload location in Google Cloud Storage where they remain and are not changed. This location is specified in your user profile. See User Profile Page (/dataprep/docs/html/User-Profile-Page_57344911).

Data may be individual files or all of the files in a folder. For more information, see Reading from Sources above.

Writing Results

When your results from a job are generated, they can be stored back in Google Cloud Storage. The Google Cloud Storage location is available through the Output Destinations tab in the Job Details page. See Job Details Page (/dataprep/docs/html/Job-Details-Page_57344846).

If your environment is using Google Cloud Storage, do not use the Upload location for storage. This directory is used for storing uploads, which may be used by multiple users. Manipulating files outside of the product can destroy other users' data. Please use the tools provided through the interface for managing uploads from Google Cloud Storage.

Creating a new dataset from results

As part of writing results, you can choose to create a new dataset, so that you can chain together data wrangling tasks.

NOTE: When you create a new dataset as part of your results, the file or files are written to the designated output location for your user account. Depending on how your permissions are configured, this location may not be accessible to other users.

Last updated 2020-06-22 UTC.