Using S3

Contents:

This section describes how you interact through the Trifacta® platform with your S3 environment.

- Simple Storage Service (S3) is an online data storage service provided by Amazon, which provides low-latency access through web services. For more information, see *https://aws.amazon.com/s3/*.

Uses of S3

The Trifacta platform can use S3 for the following tasks:

1. **Enabled S3 Integration:** The Trifacta platform has been configured to integrate with your S3 instance.For more information, see *Enable S3 Access*.
2. **Creating Datasets from S3 Files:** You can read in source data stored in S3. An imported dataset may be a single S3 file or a folder of identically structured files. See *Reading from Sources in S3* below.
3. **Reading Datasets:** When creating a dataset, you can pull your data from a source in S3. See *Creating Datasets* below.
4. **Writing Results:** After a job has been executed, you can write the results back to S3. See *Writing Results* below.

In the Trifacta application, S3 is accessed through the S3 browser. See *S3 Browser*.

> **NOTE:** When the Trifacta platform 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 Using S3

- **Access:** If you are using system-wide permissions, your administrator must configure access parameters for S3 locations. If you are using per-user permissions, this requirement does not apply. See *Enable S3 Access*.

> **Avoid using `/trifacta/uploads` for reading and writing data. This directory is used by the Trifacta application.**

- Your administrator should provide a writeable home output directory for you. This directory location is available through your user profile. See *Storage Config Page*.

Secure Access

Your administrator can grant access on a per-user basis or for the entire Trifacta platform.

The Trifacta platform utilizes an S3 key and secret to access your S3 instance. These keys must enable read /write access to the appropriate directories in the S3 instance.

> **NOTE:** If you disable or revoke your S3 access key, you must update the S3 keys for each user or for the entire system.

For more information, see *Enable S3 Access*.

Storing Data in S3

Your administrator should provide raw data or locations and access for storing raw data within S3. All Trifacta users should have a clear understanding of the folder structure within S3 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.
- The Trifacta application stores the results of each job in a separate folder in S3.

> **NOTE:** The Trifacta platform does not modify source data in S3. Source data stored in S3 is read without modification from source locations, and source data uploaded to the Trifacta platform is stored in `/trifacta/uploads`.

Reading from Sources in S3

You can create an imported dataset from one or more files stored in S3.

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

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*.

Folder selection:

When you select a folder in S3 to create your dataset, you select all files in the folder to be included.

Notes:

- 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.

When a folder is selected from S3, the following file types are ignored:

- `*_SUCCESS` and `*_FAILED` files, which may be present if the folder has been populated by the running environment.

NOTE: If you have a folder and file with the same name in S3, search only retrieves the file. You can still navigate to locate the folder.

Creating Datasets

When creating a dataset, you can choose to read data in from a source stored from S3 or local file.

- S3 sources are not moved or changed.
- Local file sources are uploaded to `/trifacta/uploads` where they remain and are not changed.

Data may be individual files or all of the files in a folder.

- For more information, see Reading from Sources in S3 above.
- In the Import Data page, click the S3 tab. See *Import Data Page*.

Writing Results

When you run a job, you can specify the S3 bucket and file path where the generated results are written. By default, the output is generated in your default bucket and default output home directory.

- Each set of results must be stored in a separate folder within your S3 output home directory.
- For more information on your output home directory, see *Storage Config Page*.

If Trifacta installation is using S3, do not use the `trifacta/uploads` directory. This directory is used for storing uploads and metadata, which may be used by multiple users. Manipulating files outside of the Trifacta application can destroy other users' data. Please use the tools provided through the Trifacta application interface for managing uploads from S3.

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.

Purging Files

Other than temporary files, the Trifacta platform does not remove any files that were generated or used by the platform, including:

- Uploaded datasets
- Generated samples
- Generated results

If you are concerned about data accumulation, you should create a bucket policy to periodically backup or purge directories in use. For more information, please see the S3 documentation.