Using ADLS

Contents:

- *Uses of ADLS*
- *Before You Begin Using ADLS*
  - *Secure Access*
- *Storing Data in ADLS*
- *Reading from Sources in ADLS*
- *Creating Datasets*
- *Writing Job Results*
  - *Creating a new dataset from results*

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

- ADLS is a scalable file storage system for use across all of the nodes (servers) of an HDI cluster. Many interactions with ADLS are similar with desktop interactions with files and folders. However, what looks like a "file" or "folder" in ADLS may be spread across multiple nodes in the cluster. For more information, see *https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-overview*.

Uses of ADLS

The Trifacta platform can use ADLS for the following reading and writing tasks:

1. **Creating Datasets from ADLS Files:** You can read in from a data source stored in ADLS. A source may be a single ADLS file or a folder of identically structured files. See *Reading from Sources in ADLS* below.
2. **Reading Datasets:** When creating a dataset, you can pull your data from another dataset defined in ADLS. See *Creating Datasets* below.
3. **Writing Job Results:** After a job has been executed, you can write the results back to ADLS. See *Writing Job Results* below.

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

- **Read/Write Access:** Your HDI administrator must configure read/write permissions to locations in ADLS. Please see the ADLS documentation.

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

- Your HDI administrator should provide a place or mechanism for raw data to be uploaded to your HDI datastore.
- Your HDI administrator should provide a writeable home output directory for you, which you can review. See *Storage Config Page*.

Secure Access

Depending on the security features you've enabled, the technical methods by which Trifacta users access ADLS may vary. For more information, see *Enable ADLS Access*.

Storing Data in ADLS

Your HDI administrator should provide raw data or locations and access for storing raw data within ADLS. All Trifacta users should have a clear understanding of the folder structure within ADLS where each individual can read from and write their job results.

- Users should know where shared data is located and where personal data can be saved without interfering with or confusing other users.

> **NOTE:** The Trifacta platform does not modify source data in ADLS. Sources stored in ADLS are read without modification from their source locations, and sources that are uploaded to the platform are stored in `/trifacta/uploads`.

Reading from Sources in ADLS

You can create a dataset from one or more files stored in ADLS.

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:

> **NOTE:** Avoid including spaces in the paths to your ADLS sources. Spaces in the path value can cause errors during execution on Databricks.

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

- This option selects all files in all sub-folders. If your sub-folders contain separate datasets, you should be more specific in your folder selection.
- All files used in a single 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 ADLS, the following file types are ignored:
  - `*_SUCCESS` and `*_FAILED` files, which may be present if the folder has been populated by HDI.
  - If you have stored files in ADLS that begin with an underscore (_), these files cannot be read during batch transformation and are ignored. Please rename these files through ADLS so that they do not begin with an underscore.

Creating Datasets

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

- ADLS 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 ADLS above.

In the Import Data page, click the ADLS tab. See *Import Data Page*.

Writing Job Results

When your job results are generated, they can be stored back in ADLS for you at the location defined for your user account.

- The ADLS location is available through the Publishing dialog in the Output Destinations tab of the Job Details page. See *Publishing Dialog*.
- Each set of job results must be stored in a separate folder within your ADLS output home directory.
- For more information on your output home directory, see *Storage Config Page*.

If your deployment is using ADLS, 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 interface for managing uploads from ADLS.

Users can specify a default output home directory and, during job execution, an output directory for the current job.

Access to results:

Depending on how the platform is integrated with ADLS, other users may or may not be able to access your job results.

- If user mode is enabled, results are written to ADLS through the ADLS account configured for your use. Depending on the permissions of your ADLS account, you may be the only person who can access these results.
- If user mode is not enabled, then each Trifacta user writes results to ADLS using a shared account. Depending on the permissions of that account, your results may be visible to all platform users.

Creating a new dataset from results

As part of writing job 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 job results, the file or files are written to the designated output location for your user account. Depending on how your HDI permissions are configured, this location may not be accessible to other users.