Using HDFS

Contents:

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

- HDFS is a scalable file storage system for use across all of the nodes (servers) of a Hadoop cluster. Many interactions with HDFS are similar with desktop interactions with files and folders. However, what looks like a "file" or "folder" in HDFS may be spread across multiple nodes in the cluster. For more information, see *https://en.wikipedia.org/wiki/Apache_Hadoop#HDFS*.

Uses of HDFS

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

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

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

- **Read/Write Access:** Your Hadoop administrator must configure read/write permissions to locations in HDFS. Please see the HDFS documentation provided with your Hadoop distribution.

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

- Your Hadoop administrator should provide a place or mechanism for raw data to be uploaded to your Hadoop datastore.
- Your Hadoop 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 HDFS may vary. For more information, see *Configure Hadoop Authentication*.

Storing Data in HDFS

Your Hadoop administrator should provide raw data or locations and access for storing raw data within HDFS. All Trifacta users should have a clear understanding of the folder structure within HDFS 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 HDFS. Sources stored in HDFS are read without modification from their source locations, and sources that are uploaded to the platform are stored in `/trifacta/uploads`.

Ingest Caching

If JDBC ingest caching has been enabled, users may see a `dataSourceCache` folder in their browser. This folder is used to store per-user caches of JDBC-based data that has been ingested into the platform from its source.

> **NOTE:** The datasourceCache folder should not be used for reading and writing of datasets, metadata, or results.

For more information, see *Configure JDBC Ingestion*.

Reading from Sources in HDFS

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

> **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 HDFS 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 HDFS, the following file types are ignored:

- *_SUCCESS and *_FAILED files, which may be present if the folder has been populated by Hadoop.
- If you have stored files in HDFS that begin with an underscore (_), these files cannot be read during batch transformation and are ignored. Please rename these files through HDFS 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 HDFS or from a local file.

- HDFS 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 HDFS*.

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

Writing Job Results

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

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

If your deployment is using HDFS, 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 HDFS.

NOTE: Users can specify a default output home directory and, during job execution, an output directory for the current job. In an encrypted HDFS environment, these two locations must be in the same encryption zone. Otherwise, writing the job results fails with a `Publish Job Failed` error.

Access to results:

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

- If user impersonation is enabled, results are written to HDFS through the HDFS account configured for your use. Depending on the permissions of your HDFS account, you may be the only person who can access these results.
- If user impersonation is not enabled, then each Trifacta user writes results to HDFS 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 Hadoop 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, please contact your HDFS administrator.