Using WASB

Contents:

> **NOTE:** For more information on support for Azure with this release, please contact your Trifacta representative.

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

- WASB is a scalable file storage system for use across all of the nodes (servers) of an HDInsight cluster. As with HDFS, many interactions with WASB are similar with desktop interactions with files and folders. However, what looks like a "file" or "folder" in WASB may be spread across multiple nodes in the cluster. For more information on HDInsight, see *https://azure.microsoft.com/en-us/services/hdinsight/*.

Uses of WASB

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

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

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

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

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

> **NOTE:** If a directory is created on the HDI cluster through WASB, the directory includes a Size=0 blob. The Trifacta platform does not list them and does not support interaction with Size=0 blobs.

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

Secure Access

Client-side encryption is not supported. Through WASBS, HTTPS is supported.

Storing Data in WASB

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

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

Reading from Sources in WASB

You can import a dataset from one or more files stored in WASB.

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 WASB for your imported 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 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. Files of the same format must have identical column structures.
- When a folder is selected from WASB, the following file types are ignored:
  - `*_SUCCESS` and `*_FAILED` files, which may be present if the folder has been populated from the cluster.
  - If you have stored files in WASB that begin with an underscore (_), these files cannot be read during batch transformation and are ignored. Please rename these files through WASB so that they do not begin with an underscore.

Creating Datasets

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

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

Writing Job Results

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

- As part of the job specification, you can create a publishing target in WASB. See *Run Job Page*.
- For ad-hoc publication to WASB, the target location is available through the Job Details page. See *Job Details Page*.
- Each set of job results must be stored in a separate folder within your WASB output home directory.
- For more information on your output home directory, see *Storage Config Page*.

If your deployment is using WASB, 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 WASB.

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 WASB permissions are configured, this location might not be accessible to other users.