Storage Deployment Options

Contents:

- *Definitions*
- *HDFS Only*
- *Hybrid Hadoop-based Deployment*
- *Amazon-based Deployment*
- *S3 without Browse or Access*
- *Microsoft Azure with ADLS Access*
  - *ADLS Gen2*
  - *ADLS Gen1*
- *Microsoft Azure with WASB Access*
- *Configuration for Storage Deployments*

The Trifacta® platform can be configured to read and write data from multiple environments at the same time. This page provides information on the supported options.

> **After you have configured the base storage layer and access and browsing capabilities, you cannot switch them for your Trifacta deployment.**

Definitions

Base Storage Layer:

The base storage layer defines where job results are written by default.

> **NOTE:** The base storage layer should be enabled and configured during initial installation. After the base storage layer has been configured, it cannot be switched to another environment.

> **Tip:** The Trifacta platform can enable connectivity to both S3 and HDFS at the same time. Note that `webapp.storageProtocol=s3` should still be specified to write results to S3.

Access and Browse data - S3:

Optionally, you can enable access and the ability to browse your S3 datastore.

JDBC Sources:

Independent of these storage options, you can access database table data through JDBC datastores. See *Relational Access*.

HDFS Only

Base Storage Layer: HDFS

Access and Browse data - S3: Off

Notes:

The default configuration, this deployment should be used for most on-premise Hadoop environments. In this case, the Trifacta platform only has access to HDFS and Hive as sources on a single Hadoop cluster.

Hybrid Hadoop-based Deployment

Base Storage Layer: HDFS

Access and Browse data - S3: On

Notes:

This deployment is recommended for the following:

- On-premises Hadoop clusters that require access to remote S3 data
- Hadoop clusters hosted in the cloud that require access to remote S3 data and want to continue to use HDFS as an output location

In this scenario, the Trifacta platform has access to HDFS and Hive data on the same cluster, as well as access to the remote S3 buckets that have been enabled for the platform.

- HDFS remains the output location for all job results, profiles, and uploads.

Amazon-based Deployment

Base Storage Layer: S3

Access and Browse data - S3: On

Notes:

This deployment is recommended for Hadoop clusters that are completely hosted in AWS and must use S3 as the base storage for all data including job results, profiles, and uploads.

S3 without Browse or Access

> **NOTE:** Before you select your deployment options, you should review additional Amazon information on running Hadoop on S3. For more information, see *https://wiki.apache.org/hadoop/AmazonS3*.

Base Storage Layer: S3

Access and Browse data - S3: Off

Notes:

This configuration is not supported. For more information, please contact *Trifacta Support*.

Microsoft Azure with ADLS Access

ADLS Gen2

Base Storage Layer: ABFSS

Access and Browse data:

- Azure Databricks: Enabled
- ADLS Gen2: Read-write

- WASB: (optional) Read-only

ADLS Gen1

Base Storage Layer: HDFS

Access and Browse data:

- Azure Databricks: Enabled
- ADLS: Read-write
- WASB: (optional) Read-only

Microsoft Azure with WASB Access

Base Storage Layer: WASB

Access and Browse data:

- Azure Databricks: Enabled
- ADLS: (optional) Read-only
- WASB: Read-write

Configuration for Storage Deployments

Base Storage Layer: *Set Base Storage Layer*

Storage Deployments:

- *Configure for Hadoop*
- *S3 Access*
- *WASB Access*
- *ADLS Gen1 Access*