Enable HttpFS

Contents:

- *Pre-requisites*
- *Configuration*
- *Enable SSL*

This section describes how to enable the Trifacta® platform to use the HttpFS service for communicating with Hadoop HDFS. HttpFS is commonly used in the following scenarios:

1. **High Availability.** WebHDFS does not support High Availability failover. You must use HttpFS instead.
2. **HDFS user is not available for secure impersonation.** If you have enabled secure impersonation in an environment where the HDFS superuser is restricted from use, you can enable HttpFS and use the HttpFS superuser for secure impersonation.

Pre-requisites

Before you begin, please verify that you have done the following in your environment:

- Enabled HDFS in your Hadoop cluster.
- Installed hadoop-httpfs into your Hadoop cluster.
- HttpFS has been enabled on a known port on the cluster.

  NOTE: If you are enabling HttpFS for use with High Availability, you should avoid enabling the HttpFS service on the primary namenode of the cluster. For more information, see *Enable Integration with Cluster High Availability*.

  NOTE: By default, HttpFS is available on port `14000`. Please verify the port number in use for your cluster.

- Started HttpFS service on the cluster.

Configuration

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

Steps:

1. The configuration settings for HttpFS are within the HDFS configuration area:

   ```
   "hdfs.webhdfs.host": "",
   "hdfs.webhdfs.port": 14000,
   "hdfs.webhdfs.httpfs": true,
   ```

2. Set `hdfs.webhdfs.httpfs` to `true`.
3. Specify the host and port for the HttpFS service. You can use one of the following methods:

a. Specify `hdfs.webhdfs.host` and `hdfs.webhdfs.port` values to point to the node hosting HttpFS.
b. Leave the `hdfs.webhdfs.host` value empty, in which case the platform falls back to using the namenode host as the WebHDFS host. Modify that value if required.

> **NOTE:** By default, the platform expects this service to be available on port `14000`. Please apply the value that matches your cluster environment.

4. Save your changes and restart the platform.

Enable SSL

Optionally, you can enable secure (SSL) communications between the platform and HttpFS.

> **NOTE:** The most secure method requires the creation and deployment of an SSL certificate for the HDFS instance. These steps provide instructions for how to do so.
>
> If this certificate is not available, you can still enable communication over SSL over WebHDFS or HttpFS. Please skip steps 1 and 2 and complete the secure configuration without certificate export.

Steps:

1. Deploy a PEM file certificate that can be read by the `[os.user` (default=`trifacta)]` user account on the Trifacta node.

   > **NOTE:** The following security configuration requires export of and access to an SSL certificate in PEM file format for the HDFS instance. Creation and deployment of this certificate exceeds the scope of this document. Please see the documentation provided with your Hadoop distribution.

   Certificates are commonly stored in Java keystores. They can be exported to PEM file format using the following command:

   ```
   keytool -exportcert -rfc  -alias <node_alias>   -storepass <pwd> -keystore cacerts -file <filename.pem>
   ```

   where:
   <pwd> is the keystore password.
   `<filename.pem>` is the output filename for the certificate.
   `<node_alias>` is the alias for the certificate in the keystore.
2. Place this generated certificate on the Trifacta node in a place where it is readable by the `[os.user` (default=`trifacta)]` user. The following location is suitable:

   ```
   /opt/trifacta
   ```

3. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`

   . For more information, see *Platform Configuration Methods*.
4. Locate the following setting and enable it:

Setting	Description

`"hdfs.webhdfs.ssl.enabled": true,`	Set to `true` to enable SSL communications with WebHDFS or (if enabled) HttpFS.

5. There is no need to update the port number. Port `14000` applies to HTTP and HTTPS.
6. **Security Level:** The level of security is determined by the following configuration options:

   a. Secure without certificate export:

Setting	Description
`"hdfs.webhdfs.ssl. certificateValidationRequired": false,`	Set to `false` to disable use of trusted certificate validation.
`"hdfs.webhdfs.ssl.certificatePath: "",`	Leave this value empty.

   b. Secure with certificate:

Setting	Description
`"hdfs.webhdfs.ssl. certificateValidationRequired": false,`	Set to `true` to require SSL use of trusted certificate validation.
`"hdfs.webhdfs.ssl.certificatePath: "",`	Configure the path on the Trifacta node to the location where you stored the certificate.

7. Save your changes and restart the platform.