

Prepare Hadoop for Integration with the Platform

Contents:

- *Create Trifacta user account on Hadoop cluster*
- *HDFS directories*
- *Kerberos authentication*
- *Acquire cluster configuration files*

Before you deploy the Trifacta® software, you should complete the following configuration steps within your Hadoop environment.

NOTE: The Trifacta platform requires access to a set of Hadoop components. See *System Requirements*.

Create Trifacta user account on Hadoop cluster

The Trifacta platform interacts with Hadoop through a single system user account. A user for the platform must be added to the cluster.

NOTE: In a cluster without Kerberos or SSO user management, the `[hadoop.user (default=trifacta)]` user must be created on each node of the cluster.

If LDAP is enabled, the `[hadoop.user]` user should be created in the same realm as the cluster.

If Kerberos is enabled, the `[hadoop.user]` user must exist on every node where jobs run.

For POSIX-compliant Hadoop environments, the user IDs of the Trifacta user accessing the cluster and the Hadoop user must match exactly.

UserID:

If possible, please create the user ID as: `trifacta`

This user should belong to the group: `trifactausers`

User requirements:

- Access to HDFS
- Permission to run YARN jobs on the cluster.

Verify that the following HDFS paths have been created and that their permissions enable access to the Trifacta user account:

NOTE: Depending on your Hadoop distribution, you may need to modify the following commands to use the Hadoop client installed on the Trifacta node.

Below, change the values for `trifacta` to match the `[hadoop.user]` user for your environment:

```
hdfs dfs -mkdir /trifacta
hdfs dfs -chown trifacta /trifacta
hdfs dfs -mkdir -p /user/trifacta
hdfs dfs -chown trifacta /user/trifacta
```

HDFS directories

The following directories must be available to the [hadoop.user] on HDFS. Below, you can review the minimum permissions set for basic and impersonated authentication for each default directory. Secure impersonation is described later.

NOTE: Except for the `dictionaries` directory, which is used to hold smaller reference files, each of these directories should be configured to permit storage of a user's largest datasets.

Directory	Minimum required permissions	Secure impersonation permissions
/trifacta/uploads	700	770 Set this to 730 to prevent users from browsing this directory.
/trifacta/queryResults	700	770
/trifacta/dictionaries	700	770
/trifacta/tempfiles	770	770

You can use the following commands to configure permissions on these directories. Following permissions scheme reflects the secure impersonation permissions in the above table:

```
$ hdfs dfs -mkdir -p /trifacta/uploads
$ hdfs dfs -mkdir -p /trifacta/queryResults
$ hdfs dfs -mkdir -p /trifacta/dictionaries
$ hdfs dfs -mkdir -p /trifacta/tempfiles
$ hdfs dfs -chown -R trifacta:trifacta /trifacta
$ hdfs dfs -chmod -R 770 /trifacta
$ hdfs dfs -chmod -R 730 /trifacta/uploads
```

If these standard locations cannot be used, you can configure the HDFS paths. You can apply this change through the *Admin Settings Page* (recommended) or

```
trifacta-conf.json
```

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

```
"hdfs.pathsConfig.fileUpload": "/trifacta/uploads",
"hdfs.pathsConfig.batchResults": "/trifacta/queryResults",
"hdfs.pathsConfig.dictionaries": "/trifacta/dictionaries",
```

Kerberos authentication

The Trifacta platform supports Kerberos authentication on Hadoop.

NOTE: If Kerberos is enabled for the Hadoop cluster, the keytab file must be made accessible to the Trifacta platform. See *Configure for Kerberos Integration* in the Configuration Guide.

Acquire cluster configuration files

The Hadoop cluster configuration files must be made available to the Trifacta platform. You can either copy the files over from the cluster or create a local symlink to them.

For more information, see *Configure for Hadoop* in the Configuration Guide.