Configure for Hortonworks

Contents:

This section provides additional configuration requirements for integrating the Trifacta® platform with the Hortonworks Data Platform.

- This section applies only to the versions of HDP that Trifacta supports. For more information, see *Supported Deployment Scenarios for Hortonworks*.

> **NOTE:** Except as noted, the following configuration items apply to the latest supported version of Hortonworks Data Platform.

Pre-requisites

Before you begin, it is assumed that you have completed the following tasks:

1. Successfully installed a supported version of Hortonworks Data Platform into your enterprise infrastructure.
2. Installed the Trifacta software in your environment. For more information, see *Install Software*.
3. Reviewed the mechanics of platform configuration. See *Required Platform Configuration*.
4. Configured access to the Trifacta database. See *Configure the Databases*.
5. Performed the basic Hadoop integration configuration. See *Configure for Hadoop*.
6. You have access to platform configuration either via the Trifacta node or through the Admin Settings page.

Hortonworks Cluster Configuration

The following changes need to be applied to Hortonworks cluster configuration files or to configuration areas inside Ambari.

> **Tip:** Ambari is the recommended method for configuring your Hortonworks cluster.

Configure for Ranger

Configure Ranger to use Kerberos

If you have deployed Ranger in a Kerberized environment, you must verify and complete the following changes in Ambari.

Steps:

1. If you have enabled Ranger, navigate to **Hive > Configs > Settings**.
   a. Choose Authorization: **Ranger**.

    b. Hiveserver2 Authentication: **Kerberos**.
2. If you have enabled Ranger and Hive, navigate to **Hive > Configs > Advanced > General**.
    a. hive.security.authorization.manager: **org.apache.ranger.authorization.hive.authorizer.RangerHiveAuthorizerFactory**
3. Navigate to **Hive > Configs > Advanced > Advanced hive-site**.
    a. hive.security.authentication.manager: **org.apache.hadoop.hive.ql.security.SessionStateUserAuthenticator**
    b. hive.conf.restricted.list: **hive.security.authenticator.manager,hive.security.authorization.manager,*hive.users.in*.admin.role,hive.security.authorization.enabled**
4. Navigate to **Hive > Configs > Advanced > Custom hive-site**. Changes in this area update `hive-site.xml`.
    a. hadoop.proxyuser.trifacta.groups: `[hadoop.group` (default=`trifactausers)]`
    b. hadoop.proxyuser.trifacta.hosts: *****
    c. hive2.jdbc.url:**<your_jdbc_url>**
    d. hive.metastore.sasl.enabled: **true**
5. Save your configuration changes.

Configure for Spark Profiling

Disable intermediate caching for Hive profiling jobs

For Hortonworks 3.0 and later, the intermediate dataset files that are generated as part of Spark profiling of your job can cause the job to hang when the source is a Hive table. As a precaution, if you are profiling jobs from Hive sources, you should disable the following property on Hortonworks 3.0 and later.

Steps:

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

   . For more information, see *Platform Configuration Methods*.
2. Locate the `spark.props` setting.
3. Insert the following setting:

```
"transformer.dataframe.cache.reused": "false"
```

4. Save your changes and restart the platform.

Additional configuration for Spark profiling on S3

If you are using S3 as your datastore and have enabled Spark profiling, you must apply the following configuration, which adds the `hadoop-aws` JAR and the `aws-java-sdk` JAR to the extra class path for Spark.

Steps:

1. In Ambari, navigate to **Spark2 > Configs**.
2. Add a new parameter to **Custom Spark2-defaults**.
3. Set the parameter as follows, which is specified for HDP 2.5.3.0, build 37:

```
spark.driver.extraClassPath=/usr/hdp/2.5.3.0-37/hadoop/hadoop-aws-2.7.3.2.5.3.0-37.jar:/usr/hdp/2.5.3.0-37/hadoop/lib/aws-java-sdk-s3-1.10.6.jar
```

4. Restart Spark from Ambari.
5. Restart the Trifacta platform.

Additional configuration for Spark profiling

If you are using Spark for profiling, you must add environment properties to your cluster configuration. See *Configure for Spark*.

Set up directory permissions

On all Hortonworks cluster nodes, verify that the YARN user has access to the YARN working directories:

```
chown yarn:hadoop /mnt/hadoop/yarn
```

If you are upgrading from a previous version of Hortonworks, you may need to clear the YARN user cache for the `[hadoop.user` (default=`trifacta)]` user:

```
rm -rf /mnt/hadoop/yarn/local/usercache/trifacta
```

Configure Trifacta platform

The following changes need to be applied to the Trifacta node.

Except as noted, these changes are applied to the following file in the Trifacta deployment:

`/opt/trifacta/conf/trifacta-conf.json`

Configure WebHDFS port

1. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.
2. **WebHDFS:** Verify that the port number for WebHDFS is correct:

   ```
   "webhdfs.port": <webhdfs_port_num>,
   ```

3. Save your changes.

Configure Resource Manager port

Hortonworks uses a custom port number for Resource Manager. You must update the setting for the port number used by Resource Manager. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.

> **NOTE:** By default, Hortonworks uses 8050 for Resource Manager. Please verify that you have the correct port number.

```
"yarn.resourcemanager.port": 8032,
```

Save your changes.

Configure location of Hadoop bundle JAR

1. Set the value for the Hadoop bundle JAR to the appropriate distribution. The following is for Hortonworks 2.6:

```
"hadoopBundleJar": "hadoop-deps/hdp-2.6/build/libs/hdp-2.6-bundle.jar"
```

2. Save your changes.

Configure Hive Locations

If you are enabling an integration with Hive on the Hadoop cluster, there are some distribution-specific parameters that must be set. For more information, see *Configure for Hive*.

Restart

To apply your changes, restart the platform. See *Start and Stop the Platform*.

After restart, you should verify operations. For more information, see *Verify Operations*.