Install Hadoop Configuration Setup

After you have performed the base installation of the Trifacta® platform, please complete the following steps if you are integrating with a Hadoop cluster.

Apply cluster configuration files via symlink

If the Trifacta platform is being installed on an edge node of the cluster, you can create a symlink from a local directory to the source cluster files so that they are automatically updated as needed.

1. Navigate to the following directory on the Trifacta node:

   ```
   cd /opt/trifacta/conf/hadoop-site
   ```

2. Create a symlink for each of the Hadoop Client Configuration files referenced in the previous steps. Example:

   ```
   ln -s /etc/hadoop/conf/core-site.xml core-site.xml
   ```

3. Repeat the above steps for each of the Hadoop Client Configuration files.

Version update for Hortonworks

If you are using Hortonworks, you must complete the following modification to the site configuration file that is hosted on the Trifacta node.

> **NOTE:** Before you begin, you must acquire the full version and build number of your Hortonworks distribution. On any of the Hadoop nodes, navigate to `/usr/hdp`. The version and build number is a directory in this location, named in the following form: `A.B.C.D-XXXX`.

In the Trifacta deployment, edit the following file:

```
/opt/trifacta/conf/hadoop-site/mapred-site.xml
```

Perform the following global search and replace:

1. Search:

   ```
   ${hdp.version}
   ```

2. Replace with your hard-coded version and build number:

   ```
   A.B.C.D-XXXX
   ```

Save the file.

Restart the Trifacta platform.

Modify Trifacta configuration changes

1. To apply this configuration change, login as an administrator to the Trifacta node. Then, edit

   `trifacta-conf.json`

   . Some of these settings may not be available through the *Admin Settings Page*. For more information, see *Platform Configuration Methods*.

2. **HDFS:** Change the host and port information for HDFS as needed. Please apply the port numbers for your distribution:

   ```
   "hdfs.namenode.host": "<namenode>",
   "hdfs.namenode.port": <hdfs_port_num>
   "hdfs.yarn.resourcemanager": {
   "hdfs.yarn.webappPort": 8088,
   "hdfs.yarn.adminPort": 8033,
   "hdfs.yarn.host": "<resourcemanager_host>",
   "hdfs.yarn.port": <resourcemanager_port>,
   "hdfs.yarn.schedulerPort": 8030
   ```

3. Save your changes and restart the platform.

Configure Spark Job Service

The Spark Job Service must be enabled for both execution and profiling jobs to work in Spark.

Below is a sample configuration and description of each property. You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

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

```
"spark-job-service" : {
  "systemProperties" : {
    "java.net.preferIPv4Stack": "true",
    "SPARK_YARN_MODE": "true"
  },
  "sparkImpersonationOn": false,
  "optimizeLocalization": true,
  "mainClass": "com.trifacta.jobserver.SparkJobServer",
  "jvmOptions": [
"-Xmx128m"
],
  "hiveDependenciesLocation": "%(topOfTree)s/hadoop-deps/cdh-6.2/build/libs",
  "env": {
    "SPARK_JOB_SERVICE_PORT": "4007",
    "SPARK_DIST_CLASSPATH": "",
    "MAPR_TICKETFILE_LOCATION": "<MAPR_TICKETFILE_LOCATION>",
    "MAPR_IMPERSONATION_ENABLED": "0",
    "HADOOP_USER_NAME": "trifacta",
    "HADOOP_CONF_DIR": "%(topOfTree)s/conf/hadoop-site/"
  },
  "enabled": true,
  "enableHiveSupport": true,
  "enableHistoryServer": false,
  "classpath": "%(topOfTree)s/services/spark-job-server/server/build/install/server/lib/*:%(topOfTree)s/conf
/hadoop-site/:%(topOfTree)s/services/spark-job-server/build/bundle/*:%(topOfTree)s/%(hadoopBundleJar)s",
  "autoRestart": false,
},
```

The following properties can be modified based on your needs:

NOTE: Unless explicitly told to do so, do not modify any of the above properties that are not listed below.

Property	Description
`sparkImpersonationOn`	Set this value to `true`, if secure impersonation is enabled on your cluster. See *Configure for Secure Impersonation*.
`jvmOptions`	This array of values can be used to pass parameters to the JVM that manages Spark Job Service.
`hiveDependenciesLocation`	If Spark is integrated with a Hive instance, set this value to the path to the location where Hive dependencies are installed on the Trifacta node. For more information, see *Configure for Hive*.
`env.SPARK_JOB_SERVICE_PORT`	Set this value to the listening port number on the cluster for Spark. Default value is `4007`. For more information, see *System Ports*.
`env.HADOOP_USER_NAME`	The username of the Hadoop principal used by the platform. By default, this value is `trifacta`.
`env.HADOOP_CONF_DIR`	The directory on the Trifacta node where the Hadoop cluster configuration files are stored. Do not modify unless necessary.
`enabled`	Set this value to `true` to enable the Spark Job Service.
`enableHiveSupport`	See below.

After making any changes, save the file and restart the platform. See *Start and Stop the Platform*.

Configure service for Hive

Depending on the environment, please apply the following configuration changes to manage Spark interactions with Hive:

Environment	spark.enableHiveSupport
Hive is not present	`false`
Hive is present but not enabled.	`false`
Hive is present and enabled	`true`

If Hive is present on the cluster and either enabled or disabled: the `hive-site.xml` file must be copied to the correct directory:

```
cp /etc/hive/conf/hive-site.xml /opt/trifacta/conf/hadoop-site/hive-site.xml
```

At this point, the platform only expects that a `hive-site.xml` file has been installed on the Trifacta node . A valid connection is not required. For more information, see *Configure for Hive* .

Configure Spark

After the Spark Job Service has been enabled, please complete the following sections to configure it for the Trifact a platform.

Yarn cluster mode

All jobs submitted to the Spark Job Service are executed in YARN cluster mode. No other cluster mode is supported for the Spark Job Service.

Configure access for secure impersonation

The Spark Job Service can run under secure impersonation. For more information, see *Configure for Secure Impersonation*.

When running under secure impersonation, the Spark Job Service requires access to the following folders. Read, write, and execute access must be provided to the Trifacta user and the impersonated user.

Folder Name	Platform Configuration Property	Default Value	Description
Trifacta Libraries folder	`"hdfs.pathsConfig.libraries"`	`/trifacta/libraries`	Maintains JAR files and other libraries required by Spark. No sensitive information is written to this location.
Trifacta Temp files folder	`"hdfs.pathsConfig.tempFiles"`	`/trifacta/tempfiles`	Holds temporary progress information files for YARN applications. Each file contains a number indicating the progress percentage. No sensitive information is written to this location.
Trifacta Dictionaries folder	`"hdfs.pathsConfig.dictionaries"`	`/trifacta/dictionaries`	Contains definitions of dictionaries created for the platform.

Identify Hadoop libraries on the cluster

The Spark Job Service does not require additional installation on the Trifacta node or on the Hadoop cluster. Instead, it references the spark-assembly JAR that is provided with the Trifacta distribution.

This JAR file does not include the Hadoop client libraries. You must point the Trifacta platform to the appropriate libraries.

Steps:

1. In platform configuration, locate the `spark-job-service` configuration block.
2. Set the following property:

```
"spark-job-service.env.HADOOP_CONF_DIR": "<path_to_Hadoop_conf_dir_on_Hadoop_cluster>",
```

Property	Description
spark-job-service.env.HADOOP_CONF_DIR	Path to the Hadoop configuration directory on the Hadoop cluster.

3. In the same block, the `SPARK_DIST_CLASSPATH` property must be set depending on your Hadoop distribution.
4. Save your changes.

Locate Hive dependencies location

If the Trifacta platform is also connected to a Hive instance, please verify the location of the Hive dependencies on the Trifacta node. The following example is from Cloudera 6.2:

> **NOTE:** This parameter value is distribution-specific. Please update based on your Hadoop distribution.

```
"spark-job-service.hiveDependenciesLocation":"%(topOfTree)s/hadoop-deps/cdh-6.2/build/libs",
```

For more information, see *Configure for Spark*.

Enable High Availability

> **NOTE:** If high availability is enabled on the Hadoop cluster, it must be enabled on the Trifacta platform, even if you are not planning to rely on it. See *Enable Integration with Cluster High Availability*.