TRIFACTA

Install Guide

Version: 6.0.2
Doc Build Date: 07/11/2019

For third-party license information, please select **About Trifacta** from the User menu.

Install Overview

Contents:

Basic Install Workflow

1. Review the pre-installation checklist and other system requirements. See *Install Preparation*.
2. Review the requirements for your specific installation scenario in the following sections.
3. Install the software. See *Install Software*.
4. Install the databases. See *Install Databases*.
5. Configure your installation.
6. Verify operations.

Notation

In this guide, JSON settings are provided in dot notation. For example, `webapp.selfRegistration` refers to a JSON block `selfRegistration` under `webapp`:

```
{
...
  "webapp": {
    "selfRegistration": true,
    ...
  }
...
}
```

Install for High Availability

Contents:

The Trifacta® platform can be installed across multiple nodes for high availability failover. This section describes the general process for installing the platform across multiple, highly available nodes.

- The Trifacta platform can also integrate with a highly available Hadoop cluster. For more information, see *Enable Integration with Cluster High Availability*.

Limitations

The following limitations apply to this feature:

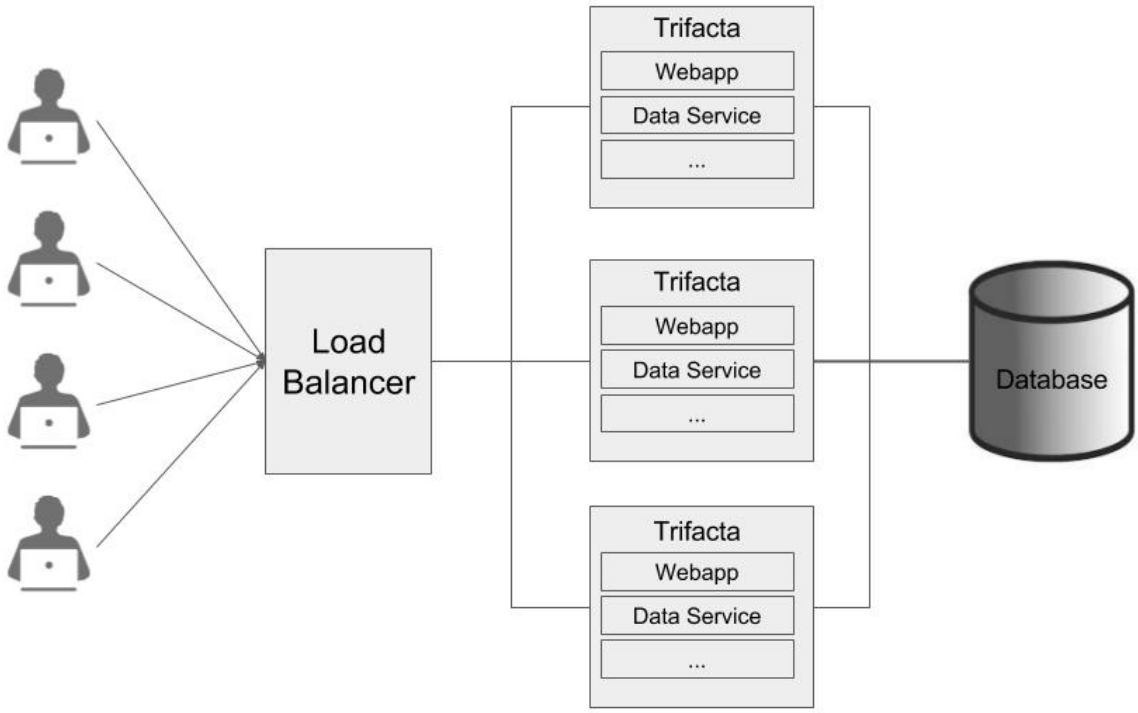
- This form of high availability is not supported for Marketplace installations.
- Job canceling does not work.
- When HA is enabled, the restart feature in the Admin Settings page does not work. You must restart using the command line.
- The platform must be installed on `/opt/trifacta` on every failover node.
- This feature does not apply to the following components:
  - Hadoop cluster (See previous link.)
  - webhdfs/httpfs
  - Sentry
  - Navigator
  - Atlas
  - any other application/infrastructure with which the Trifacta platform can integrate

For more information, see *Configure for High Availability*.

Overview

The Trifacta platform supports an Active-Active HA deployment model, which works well at scale. The architecture features a single load balancer sitting in front of multiple nodes running the Trifacta platform. Each node:

- communicates with the same database
- shares the `/opt/trifacta/conf` and `/opt/trifacta/logs` directories through NFS.



- **Database:** PostGreSQL supports HA. The HA-enabled database runs outside of the cluster of platform nodes and appears to each node as a single database. No application code changes are required.
- **Load balancer:** HAProxy is used for its capabilities on health checking the other HA nodes. This load balance periodically checks the health of the other nodes in the setup.
    - If the health for a given node fails, then the load balancer stops routing traffic to that node while continuing to poll its health.
    - If the node recovers, the load balancer resumes sending traffic to it.
    - Node health is described below.
- **Synchronized configuration:** All nodes share the `/opt/trifacta/conf` mount point, which allows the same configuration files to be visible and accessible on each node.

Job interruption

In case of a failover event, any in-progress job should be marked as failed.

Failover events/scenarios around jobs:

#	Job	Event	Resulting job state
1	In progress	The batch job runner is fine, but executor running the job fails.	Failed
2	In progress	The batch job runner or the node dies.	In Progress
3	Queued	The batch job runner or the node dies.	In Progress[1]
4	Pending	The batch job runner or the node dies.	In Progress[1 2]

[1] It may not be "In Progress". However, the job has not truly failed.

2 A nuance around #3. There is a feature flag that can be enabled and is enabled by default, which causes pending jobs to be marked as failed on (re)start of batch job runner. However, because this feature indiscriminately marks *all* pending jobs as failed, it cannot be safely enabled in an environment that has multiple running batch job runners.

Installation Topography

The Trifacta platform supports a single load balancer placed in front of multiple nodes, each of which runs the same version of Trifacta Wrangler Enterprise. Content between nodes is shared using an NFS resource mount.

- **master node:** This node is the default one used for hosting and serving the Trifacta platform. Example node information:

```
NFS Server Hostname: server.local
NFS Server IP Address: 192.168.1.101
```

- **client node(s):** These nodes are failover nodes in case the master node is unavailable. Example node information:

```
NFS Client Hostname: client.local
NFS Client IP Address: 192.168.1.102
```

- **load balancer:** This documentation references set up for HAProxy as an example. If you are using a different load balancer, please consult the documentation that came with your product.

Shared resources:

Each node shares the following resources:

- Trifacta databases
- Directories shared via NFS mount:

```
/opt/trifacta/logs
/opt/trifacta/conf
```

Order of Installation

Steps:

1. All nodes must meet the system requirements. See *System Requirements*.
2. All nodes must have the appropriate ports opened. See *System Ports*.
3. Install the databases.

   NOTE: The databases must be installed in a location that is accessible to all nodes.

   NOTE: When installing databases for high availability access, you should deploy standard access and replication techniques that are consistent with the policies of your enterprise.

   See *Install Databases*.
4. Complete the installation process for the server node.

   NOTE: After install, do not start the Trifacta node.

   See *Install Software*.
5. Repeat the above process for each of the client nodes.
6. The software is installed on all nodes. No node is running the software.

Configuration

Additional configuration is required.

NOTE: Starting and stopping the platform in high availability mode requires additional steps.

For more information, see *Configure for High Availability*.

Install On-Premises

Contents:

To install Trifacta® Wrangler Enterprise inside your enterprise infrastructure, please review and complete the following sections in the order listed below.

Scenario Description

- Installation of Trifacta Wrangler Enterprise on a server on-premises
- Installation of Trifacta databases on a server on-premises
- Integration with a supported Hadoop cluster on premises.
- Base storage layer of HDFS

Preparation

1. **Review Planning Guide:** Please review and verify *Install Preparation* and sub-topics.
2. **Acquire Assets:** Acquire the installation package for your operating system and your license key. For more information, contact *Trifacta Support*.
    1. If you are completing the installation without Internet access, you must also acquire the offline versions of the system dependencies. See *Install Dependencies without Internet Access*.
3. **Deploy Hadoop cluster:** In this scenario, the Trifacta platform does not create a Hadoop cluster. See below.

    > **NOTE:** Installation and maintenance of a working Hadoop cluster is the responsibility of the Trifacta Wrangler Enterprise customer. Guidance is provided below on the requirements for integrating the platform with the cluster.

4. **Deploy Trifacta node:**Trifacta Wrangler Enterprise must be installed on an edge node of the cluster. Details are below.

Limitations: For more information on limitations of this scenario, see *Product Limitations* in the *Install Preparation* area.

Deploy the Cluster

In your enterprise infrastructure, you must deploy a cluster using a supported version of Hadoop to manage the expected data volumes of your Trifacta jobs. For more information on suggested sizing, see *Sizing Guidelines* in the *Install Preparation* area.

When you configure the platform to integrate with the cluster, you must acquire information about the cluster configuration. For more information on the set of information to collect, see *Pre-Install Checklist* in the *Install Preparation* area.

> **NOTE:** By default, smaller jobs are executed on the Trifacta Photon running environment . Larger jobs are executed using Spark on the integrated Hadoop cluster. Spark must be installed on the cluster. For more information, see *System Requirements* in the *Install Preparation* area.

The Trifacta platform supports integration with the following cluster types. For more information on the supported versions, please see the listed sections below.

- See *Supported Deployment Scenarios for Cloudera*.
- See *Supported Deployment Scenarios for Hortonworks*.

Prepare the cluster

Before installing software, please complete the following steps if you are integrating with a Hadoop cluster.

Before you begin, please verify or complete the following:

1. On the Hadoop cluster:
    1. Create a user `[hadoop.user` (default=`trifacta)]` and a group for it `[hadoop.group` (default=`trifactausers)]`.
    2. Create the following directories:
        1. `/trifacta`
        2. `/user/trifacta`
    3. Change the ownership of `/trifacta` and `/user/trifacta` to `trifacta:trifacta` or the corresponding values for the Hadoop user in your environment.

        > **NOTE:** You must verify that the `[hadoop.user]` user has complete ownership and full access to Read, Write and Execute on these directories recursively.

2. Verify that WebHDFS is configured and running on the cluster.

3. Software installation is completed on a dedicated node in the cluster. The user installing the Trifacta software must have sudo access.

4. If you are installing on a server with an older instance of Postgres, you should remove the older version or change the default ports.

For more information, see *Prepare Hadoop for Integration with the Platform*.

Additional users may be required. For more information, see *Required Users and Groups* in the *Install Preparation* area.

Deploy the Trifacta node

An edge node of the cluster is required to host the Trifacta platform software. For more information on the requirements of this node, see *System Requirements*.

Install Workflow

Please complete these steps listed in order:

1. **Install software:** Install the Trifacta platform software on the cluster edge node. See *Install Software*.
2. **Install databases:** The platform requires several databases for storage.

   > **NOTE:** The default configuration assumes that you are installing the databases on a PostgreSQL server on the same edge node as the software using the default ports. If you are changing the default configuration, additional configuration is required as part of this installation process.

   For more information, see *Install Databases*.
3. **Start the platform:** For more information, see *Start and Stop the Platform*.
4. **Login to the application:** After software and databases are installed, you can login to the application to complete configuration:
   1. See *Login*.
   2. As soon as you login, you should change the password on the admin account. In the left menu bar, select **Settings > Admin Settings**. Scroll down to Manage Users. For more information, see *Change Admin Password*.

> **Tip:** At this point, you can access the online documentation through the application. In the left menu bar, select **Help menu > Product Docs**. All of the following content, plus updates, is available online. See Documentation below.

Configure for Hadoop

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 - non-edge node

If the Trifacta platform is being installed on a non-edge node, you must copy over the Hadoop Client Configuration files from the cluster.

> **NOTE:** When these files change, you must update the local copies. For this reason, it is best to install on an edge node.

1. Download the Hadoop Client Configuration files from the Hadoop cluster. The required files are the following:
   1. `core-site.xml`
   2. `hdfs-site.xml`
   3. `mapred-site.xml`
   4. `yarn-site.xml`
   5. `hive-site.xml` (if you are using Hive)
2. These configuration files must be moved to the Trifacta deployment. By default, these files are in `/etc/hadoop/conf`:

```
sudo cp <location>/*.xml /opt/trifacta/conf/hadoop-site/
sudo chown trifacta:trifacta /opt/trifacta/conf/hadoop-site/*.xml
```

For more information, see *Configure for Hadoop*.

Apply cluster configuration files - edge node

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.

For more information, see *Configure for Hadoop*.

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
   ```

   For more information, see *Configure for Hadoop*.

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.

NOTE: Beginning in Release 4.0, the Spark Job Service and running environment are enabled by default. If you are upgrading from an earlier release, you may be required to enable the service through the following configuration changes.

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-5.4/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/libs
/spark-job-server-bundle.jar:%(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.
    1. **For Cloudera 5.x:** This property can be left blank.
    2. **For Hortonworks 2.x:** This property configuration is covered later in this section.
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 5.10:

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

```
"spark-job-service.hiveDependenciesLocation":"%(topOfTree)s/hadoop-deps
/cdh-5.10/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*.

Configure for Trifacta platform

Set base storage layer

The platform requires that one backend datastore be configured as the base storage layer. This base storage layer is used for storing uploaded data and writing results and profiles.

> **NOTE:** By default, the base storage layer for Trifacta Wrangler Enterprise is set to HDFS. You can change it now, if needed. After this base storage layer is defined, it cannot be changed again.

See *Set Base Storage Layer*.

Verify Operations

> **NOTE:** You can try to verify operations using the Trifacta Photon running environment at this time. While you can also try to run a job on the Hadoop cluster, additional configuration may be required to complete the integration. These steps are listed under Next Steps below.

Prepare Your Sample Dataset

To complete this test, you should locate or create a simple dataset. Your dataset should be created in the format that you wish to test.

Characteristics:

- Two or more columns.
- If there are specific data types that you would like to test, please be sure to include them in the dataset.
- A minimum of 25 rows is required for best results of type inference.
- Ideally, your dataset is a single file or sheet.

Store Your Dataset

If you are testing an integration, you should store your dataset in the datastore with which the product is integrated.

> **Tip:** Uploading datasets is always available as a means of importing datasets.

- You may need to create a connection between the platform and the datastore.
- Read and write permissions must be enabled for the connecting user to the datastore.
- For more information, see *Connections Page*.

Verification Steps

Steps:

1. Login to the application. See *Login*.
2. In the application menu bar, click **Library**.
3. Click **Import Data**. See *Import Data Page*.
    1. Select the connection where the dataset is stored. For datasets stored on your local desktop, click **Upload**.
    2. Select the dataset.
    3. In the right panel, click the Add Dataset to a Flow checkbox. Enter a name for the new flow.
    4. Click **Import and Add to Flow**.
    5. **Troubleshooting:** At this point, you have read access to your datastore from the platform. If not, please check the logs, permissions, and your Trifacta® configuration.
4. In the left menu bar, click the Flows icon. Flows page, open the flow you just created. See *Flows Page*.
5. In the Flows page, click the dataset you just imported. Click **Add new Recipe**.
6. Select the recipe. Click **Edit Recipe**.
7. The initial sample of the dataset is opened in the Transformer page, where you can edit your recipe to transform the dataset.
    1. In the Transformer page, some steps are automatically added to the recipe for you. So, you can run the job immediately.
    2. You can add additional steps if desired. See *Transformer Page*.
8. Click **Run Job**.
    1. If options are presented, select the defaults.
    2. To generate results in other formats or output locations, click **Add Publishing Destination.** Configure the output formats and locations.
    3. To test dataset profiling, click the Profile Results checkbox. Note that profiling runs as a separate job and may take considerably longer.
    4. See *Run Job Page*.
    5. **Troubleshooting:** Later, you can re-run this job on a different running environment. Some formats are not available across all running environments.
9. When the job completes, you should see a success message under the Jobs tab in the Flow View page.
    1. **Troubleshooting:** Either the Transform job or the Profiling job may break. To localize the problem, try re-running a job by deselecting the broken job type or running the job on a different running environment (if available). You can also download the log files to try to identify the problem. See *Job Details Page*.
10. Click **View Results** from the context menu for the job listing. In the Job Details page, you can see a visual profile of the generated results. See *Job Details Page*.
11. In the Output Destinations tab, click a link to download the results to your local desktop.
12. Load these results into a local application to verify that the content looks ok.

> **Checkpoint:** You have verified importing from the selected datastore and transforming a dataset. If your job was successfully executed, you have verified that the product is connected to the job running environment and can write results to the defined output location. Optionally, you may have tested profiling of job results. If all of the above tasks completed, the product is operational end-to-end.

Documentation

> **Tip:** You should access online documentation through the product. Online content may receive updates that are not present in PDF content.

You can access complete product documentation online and in PDF format. From within the Trifacta application, select **Help menu > Product Docs**.

Next Steps

After you have accessed the documentation, the following topics are relevant to on-premises deployments. Please review them in order.

> **NOTE:** These materials are located in the *Configuration Guide*.

Topic	Description
Required Platform Configuration	This section covers the following topics, some of which should already be completed: • *Set Base Storage Layer* - The base storage layer must be set once and never changed. • *Create Encryption Key File* - If you plan to integrate the platform with any relational sources, including Hive or Redshift, you must create an encryption key file and store it on the Trifacta node • *Running Environment Options* - Depending on your scenario, you may need to perform additional configuration for your available running environment(s) for executing jobs. • *Profiling Options* - In some environments, tweaks to the settings for visual profiling may be required. You can disable visual profiling if needed. • *Configure for Spark* - If you are enabling the Spark running environment, please review and verify the configuration for integrating the platform with the Hadoop cluster instance of Spark.

Configure for Hadoop	• *Configuration by Hadoop Distribution*:
	  • *Configure for Cloudera*
	  • *Configure for Hortonworks*
	• *Configure Hadoop Authentication*

Enable Integration with Compressed Clusters	If the Hadoop cluster uses compression, additional configuration is required.
Enable Integration with Cluster High Availability	If you are integrating with high availability on the Hadoop cluster, please complete these steps. • If you are integrating with high availability on the Hadoop cluster, HttpFS must be enabled in the platform. HttpFS is required in other, less-common cases. See *Enable HttpFS*.
Configure for Hive	Integration with the Hadoop cluster's instance of Hive.
Configure for KMS	Integration with the Hadoop cluster's key management system (KMS) for encrypted transport. Instructions are provided for distribution-specific versions of Hadoop.
Configure Security	A list of topics on applying additional security measures to the Trifacta platform and how integrates with Hadoop.
Configure SSO for AD-LDAP	Please complete these steps if you are integrating with your enterprise's AD/LDAP Single Sign-On (SSO) system.

Install for AWS

Contents:

This install process applies to installing Trifacta® Wrangler Enterprise on an AWS infrastructure that you manage.

AWS Marketplace deployments:

> **NOTE:** Content in this section does not apply to deployments from the AWS Marketplace, which provide fewer deployment and configuration options. For more information, see the AWS Marketplace.

Scenario Description

> **NOTE:** All hardware in use for supporting the platform is maintained within the enterprise infrastructure on AWS.

- Installation of Trifacta Wrangler Enterprise on an EC2 server in AWS
- Installation of Trifacta databases on AWS
- Integration with a supported EMR cluster.
- Base storage layer and backend datastore of S3

> **NOTE:** When the above installation and configuration steps have been completed, the platform is operational. Additional configuration may be required, which is referenced at the end of this section.

For more information on deployment scenarios, see *Supported Deployment Scenarios for AWS*.

Product Limitations

The following limitations apply to installations of Trifacta Wrangler Enterprise on AWS:

- No support for Hive integration
- No support for secure impersonation or Kerberos
- No support for high availability and failover
- Job cancellation is not supported on EMR.
- When publishing single files to S3, you cannot apply an `append` publishing action.

Pre-requisites

Desktop Requirements

- All desktop users of the platform should have a supported version of Google Chrome installed on their desktops.
    - For more information. see *Desktop Requirements*.
    - If a supported browser is not available within your enterprise, desktop users can install the Trifacta enterprise application as a separate application. For more information, see *Install for Wrangler Enterprise Application*.
- All desktop users must be able to connect to the EC2 instance through the enterprise infrastructure.

AWS Pre-requisites

Depending on which of the following AWS components you are deploying, additional pre-requisites and limitations may apply. Please review these sections as well.

- *Configure for EMR*
- *Enable S3 Access*
- *Create Redshift Connections*

Prep

Before you begin, please verify that you have completed the following:

1. **Review Planning Guide:** Please review and verify *Install Preparation* and sub-topics.
    1. **Limitations:** For more information on limitations of this scenario, see *Product Limitations* in the *Install Preparation* area.
2. **Read:** Please read this entire document before you create the EMR cluster or install the Trifacta platform.
3. **Acquire Assets:** Acquire the installation package for your operating system and your license key. For more information, contact *Trifacta Support*.
    1. If you are completing the installation without Internet access, you must also acquire the offline versions of the system dependencies. See *Install Dependencies without Internet Access*.
4. **VPC:** Enable and deploy a working AWS VPC.
5. **S3:** Enable and deploy an AWS S3 bucket to use as the base storage layer for the platform. In the bucket, the platform stores metadata in the following location:

    ```
    <S3_bucket_name>/trifacta
    ```

    See *https://s3.console.aws.amazon.com/s3/home*.
6. **IAM Policies:** Create IAM policies for access to the S3 bucket. Required permissions are the following:
    - The system account or individual user accounts must have full permissions for the S3 bucket:

```
Delete*, Get*, List*, Put*, Replicate*, Restore*
```

- These policies must apply to the bucket and its contents. Example:

```
"arn:aws:s3:::my-trifacta-bucket-name"
"arn:aws:s3:::my-trifacta-bucket-name/*"
```

- See *https://console.aws.amazon.com/iam/home#/policies*

7. **EC2 instance:** Deploy an AWS EC2 with SELinux where the Trifacta software can be installed.
   1. The required set of ports must be enabled for listening. See *System Ports*.
   2. This node should be dedicated for Trifacta use.

   > **NOTE:** The EC2 node must meet the system requirements. For more information, see *System Requirements*.

8. **EC2 instance role:** Create an EC2 instance role for your S3 bucket policy. See *https://console.aws.amazon.com/iam/home#/roles*.
9. **EMR cluster:** An existing EMR cluster is required.
   1. **Cluster sizing:** Before you begin, you should allocate sufficient resources for sizing the cluster. For guidance, please contact your Trifacta representative.
   2. See Deploy the Cluster below.
10. **Databases:**
    1. The platform utilizes a set of databases that must be accessed from the Trifacta node. Databases are installed as part of the workflow described later.
    2. For more information on the supported databases and versions, see *System Requirements*.
    3. For more information on database installation requirements, see *Install Databases*.
    4. If installing databases on Amazon RDS an admin account to RDS is required. For more information, see *Install Databases on Amazon RDS*.

AWS Information

Before you begin installation, please acquire the following information from AWS:

- **EMR:**
  - AWS region for the EMR cluster, if it exists.
  - ID for EMR cluster, if it exists
    - If you are creating an EMR cluster as part of this process, please retain the ID.
    - The EMR cluster must allow access from the Trifacta node. This configuration is described later.
- **Subnet:** Subnet within your virtual private cloud (VPC) where you want to launch the Trifacta platform.
  - This subnet should be in the same VPC as the EMR cluster.
  - Subnet can be private or public.
  - If it is private and it cannot access the Internet, additional configuration is required. See below.
- **S3:**
  - Name of the S3 bucket that the platform can use
  - Path to resources on the S3 bucket
- **EC2:**
  - Instance type for the Trifacta node

Internet access

From AWS, the Trifacta platform requires Internet access for the following services:

NOTE: Depending on your AWS deployment, some of these services may not be required.

- AWS S3
- Key Management System [KMS] (if sse-kms server side encryption is enabled)
- Secure Token Service [STS] (if temporary credential provider is used)
- EMR (if integration with EMR cluster is enabled)

NOTE: If the Trifacta platform is hosted in a VPC where Internet access is restricted, access to S3, KMS and STS services must be provided by creating a VPC endpoint. If the platform is accessing an EMR cluster, a proxy server can be configured to provide access to the AWS ElasticMapReduce regional endpoint.

Deploy the Cluster

In your AWS infrastructure, you must deploy a supported version of EMR across a recommended number of nodes to support the expected data volumes of your Trifacta jobs.

- For more information on suggested sizing, see *Sizing Guidelines* in the *Install Preparation* area.

For more information on the supported EMR distributions, see *Supported Deployment Scenarios for AWS*.

When you configure the platform to integrate with the cluster, you must acquire some information about the cluster resources. For more information on the set of information to collect, see *Pre-Install Checklist* in the *Install Preparation* area.

Deploy the EC2 Node

An EC2 node of the cluster must be deployed to host the Trifacta platform software. For more information on the requirements of this node, see *System Requirements*.

When you configure the platform to integrate with the cluster, you must acquire some information about the cluster resources. For more information on the set of information to collect, see *Pre-Install Checklist* in the *Install Preparation* area.

Here are some guidelines for deploying the EC2 cluster from the EC2 cluster:

1. **Instance size:** Select the instance size.
2. **Network:** Configure the VPC, subnet, firewall and other configuration settings necessary to communicate with the instance.
3. **Auto-assigned Public IP:** You must create a public IP to access the Trifacta platform.
4. **EC2 role:** Select the EC2 role that you created.
5. **Local storage:** Select a local EBS volume. The default volume includes 100GB storage.

   NOTE: The local storage environment contains the Trifacta databases, the product installation, and its log files. No source data is ever stored within the product.

6. **Security group:** Use a security group that exposes access to port 3005, which is the default port for the platform.
7. **Create an AWS key-pair for access:** This key is used to provide SSH access to the platform, which may be required for some admin tasks.
8. Save your changes.

Install Workflow

> **NOTE:** These steps are covered in greater detail later in this section.

After you have completed, the above, please complete these steps listed in order:

1. **Install software:** Install the Trifacta platform software on the EC2 node you created. See *Install Software*.
2. **Install databases:** The platform requires several databases for storing metadata.

   > **NOTE:** The software assumes that you are installing the databases on a PostgreSQL server on the same node as the software. If you are not or are changing database names or ports, additional configuration is required as part of this installation process.

   For more information, see *Install Databases*.
3. **Start the platform:** For more information, see *Start and Stop the Platform*.
4. **Login to the application:** After software and databases are installed, you can login to the application to complete configuration:
   1. See *Login*.
   2. As soon as you login, you should change the password on the admin account. In the left menu bar, select **Settings > Admin Settings**. Scroll down to Manage Users. For more information, see *Change Admin Password*.

> **Tip:** At this point, you can access the online documentation through the application. In the left menu bar, select **Help menu > Product Docs**. All of the following content, plus updates, is available online. See Documentation below.

Configure for EMR

> **NOTE:** If you are creating a new EMR cluster as part of this installation process, please skip this section. That workflow is covered later in the document. For more information, see *Configure for EMR*.

Please complete the following configuration to enable access to your pre-existing EMR cluster from the Trifacta platform.

IAM and Security Group updates

You must make changes to your IAM and Security Group changes to enable the Trifacta instance to communicate with your existing EMR cluster and your EMR cluster to read/write to the Trifacta data bucket. Below are the requirements and suggested implementation details. Please adapt these suggestions to fit your environment as long as the requirements are satisfied.

For additional documentation around these changes:

- *https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-iam-roles.html*
- *https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-additional-sec-groups.html*

Requirement	Example

Trifacta EC2 instance role must be permitted to use your EMR cluster.	<pre>{
	"Version": "2008-10-17",
	"Statement": [
	{</pre>

```
            "Action": [

"elasticmapreduce:
DescribeStep",

"elasticmapreduce:
ListBootstrapActions",

"elasticmapreduce:
ListClusters",

"elasticmapreduce:
DescribeCluster",

"elasticmapreduce:
AddJobFlowSteps",

"elasticmapreduce:
DescribeJobFlows",

"elasticmapreduce:
ListInstanceGroups"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

EMR EC2 instance role must be permitted to use the Trifacta data bucket.

```
{
    "Version": "2008-10-17",
    "Statement": [
        {
            "Action": [

"elasticmapreduce:Describe*",

"elasticmapreduce:List*",
                "s3:
ListAllMyBuckets",
                "ec2:
Describe*"
            ],
            "Resource": "*",
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:
PutObject",
                "s3:
```

```
ListBucket",
                "s3:
GetObject",
                "s3:
DeleteObject"
            ],
            "Resource": [
                "arn:aws:
s3:::YOUR-TRIFACTA-BUCKET",
                "arn:aws:
s3:::YOUR-TRIFACTA-BUCKET/*"
            ],
            "Effect": "Allow"
        }
    ]
}
```

Your EMR Service Role should permit access to the Trifacta bucket.

> **NOTE:** This example is not a complete policy. You should update your existing policy with these statements.

```
        {
            "Action": [
                "s3:
HeadBucket",
                "s3:
ListAllMyBuckets"
            ],
            "Resource": "*",
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:
PutObject",
                "s3:
GetObject",
                "s3:
ListBucket",
                "s3:
DeleteObject"
            ],
            "Resource": [
                "arn:aws:
s3:::YOUR-TRIFACTA-BUCKET",
                "arn:aws:
s3:::YOUR-TRIFACTA-BUCKET/*"
            ],
            "Effect": "Allow"
        },
```

Your EMR cluster master node must permit the Trifacta EC2 instance to access it.

	• The Trifacta EC2 instance must be able to communicate with your EMR master node on TCP ports 18080 and 8088. • You should create a security group and then associate it with your EMR master node using the "additional security groups" functionality. • For future ease of use, you should specify the security group associated with your Trifacta EC2 instance as the source.

Additional configuration must be applied within the platform. These steps are described later.

Additional Configuration for AWS Installs

Apply license key to EC2 node

Steps:

1. Acquire the `license.json` license key file that was provided to you by your Trifacta representative.
2. Transfer the license key file to the EC2 node that is hosting the Trifacta platform. Navigate to the directory where you stored it.
3. Make the Trifacta user the owner of the file:

```
sudo chown trifacta:trifacta license.json
```

4. Make sure that the Trifacta user has read permissions on the file:

```
sudo chmod 644 license.json
```

5. Copy the license key file to the proper location:

```
cp license.json /opt/trifacta/license/
```

Launch the platform

For more information on how to launch the platform, see *Start and Stop the Platform*.

When the instance is spinning up for the first time, performance may be slow. When the instance is up, navigate to the following:

```
http://<public_hostname>:3005
```

When the login screen appears, enter the default admin credentials provided to you.

> **NOTE:** As soon as you login as an admin for the first time, you should immediately change the password. From the left nav bar, select **Settings > Settings > User Profile**. Change the password and click **Save** to restart the platform.

Configure for EMR clusters

The following steps apply to configure the platform to integrate with the EMR cluster:

1. From the application menu, select the Settings menu. Then, click **Settings > Admin Settings**.
2. In the Admin Settings page, you can configure many aspects of the platform, including user management tasks, and perform restarts to apply the changes.
    1. In the Search bar, enter the following:

    ```
    aws.s3.bucket.name
    ```

    2. Set the value of this setting to be S3 bucket name.
3. Check the following setting. Verify that it is set to `2.3.0`:

    ```
    "spark.version": "2.3.0",
    ```

4. The following setting must be specified.

    ```
    "aws.mode":"system",
    ```

    You can set the above value to either of the following:

aws.mode value	Description
`system`	Set the mode to `system` to enable use of EC2 instance-based authentication for access.
`user`	Set the mode to `user` to utilize user-based credentials. This mode requires additional configuration.

    Details on the above configuration are described later.
5. Set the following parameter to `true`, which instructs the Trifacta application to run jobs on the integrated EMR cluster:

    ```
    "webapp.runinEMR" = true,
    ```

6. In the Admin Settings page, locate the External Service Settings section.
7. In the Admin Settings page, locate the External Service Settings section.

    1. **AWS EMR Cluster ID:** Paste the value for the EMR Cluster ID for the cluster to which the platform is connecting.
    2. **AWS Region:** Enter the region where your EMR cluster is located.
    3. **Resource Bucket:** Enter the name of the S3 bucket to use.
    4. **Resource Path:** you should use something like `EMRLOGS`.
8. Click **Save** underneath the External Service Settings section.

Set base storage layer

The platform requires that one backend datastore be configured as the base storage layer. This base storage layer is used for storing uploaded data and writing results and profiles.

NOTE: By default, the base storage layer for Trifacta Wrangler Enterprise is set to HDFS. You must change this value for S3. After this base storage layer is defined, it cannot be changed again.

See *Set Base Storage Layer*.

Verify Operations

NOTE: You can try to verify operations using the Trifacta Photon running environment at this time. While you can also try to run a job in the Spark running environment, additional configuration may be required to complete the integration. These steps are listed under Next Steps below.

Prepare Your Sample Dataset

To complete this test, you should locate or create a simple dataset. Your dataset should be created in the format that you wish to test.

Characteristics:

- Two or more columns.
- If there are specific data types that you would like to test, please be sure to include them in the dataset.
- A minimum of 25 rows is required for best results of type inference.
- Ideally, your dataset is a single file or sheet.

Store Your Dataset

If you are testing an integration, you should store your dataset in the datastore with which the product is integrated.

Tip: Uploading datasets is always available as a means of importing datasets.

- You may need to create a connection between the platform and the datastore.
- Read and write permissions must be enabled for the connecting user to the datastore.
- For more information, see *Connections Page*.

Verification Steps

Steps:

1. Login to the application. See *Login*.
2. In the application menu bar, click **Library**.
3. Click **Import Data**. See *Import Data Page*.
    1. Select the connection where the dataset is stored. For datasets stored on your local desktop, click **Upload**.
    2. Select the dataset.
    3. In the right panel, click the Add Dataset to a Flow checkbox. Enter a name for the new flow.
    4. Click **Import and Add to Flow**.
    5. **Troubleshooting:** At this point, you have read access to your datastore from the platform. If not, please check the logs, permissions, and your Trifacta® configuration.
4. In the left menu bar, click the Flows icon. Flows page, open the flow you just created. See *Flows Page*.

5. In the Flows page, click the dataset you just imported. Click **Add new Recipe**.
6. Select the recipe. Click **Edit Recipe**.
7. The initial sample of the dataset is opened in the Transformer page, where you can edit your recipe to transform the dataset.
    1. In the Transformer page, some steps are automatically added to the recipe for you. So, you can run the job immediately.
    2. You can add additional steps if desired. See *Transformer Page*.
8. Click **Run Job**.
    1. If options are presented, select the defaults.
    2. To generate results in other formats or output locations, click **Add Publishing Destination.** Configure the output formats and locations.
    3. To test dataset profiling, click the Profile Results checkbox. Note that profiling runs as a separate job and may take considerably longer.
    4. See *Run Job Page*.
    5. **Troubleshooting:** Later, you can re-run this job on a different running environment. Some formats are not available across all running environments.
9. When the job completes, you should see a success message under the Jobs tab in the Flow View page.
    1. **Troubleshooting:** Either the Transform job or the Profiling job may break. To localize the problem, try re-running a job by deselecting the broken job type or running the job on a different running environment (if available). You can also download the log files to try to identify the problem. See *Job Details Page*.
10. Click **View Results** from the context menu for the job listing. In the Job Details page, you can see a visual profile of the generated results. See *Job Details Page*.
11. In the Output Destinations tab, click a link to download the results to your local desktop.
12. Load these results into a local application to verify that the content looks ok.

> **Checkpoint:** You have verified importing from the selected datastore and transforming a dataset. If your job was successfully executed, you have verified that the product is connected to the job running environment and can write results to the defined output location. Optionally, you may have tested profiling of job results. If all of the above tasks completed, the product is operational end-to-end.

Documentation

> **Tip:** You should access online documentation through the product. Online content may receive updates that are not present in PDF content.

You can access complete product documentation online and in PDF format. From within the Trifacta application, select **Help menu > Product Docs**.

Next Steps

After you have accessed the documentation, the following topics are relevant to AWS enterprise infrastructure deployments.

> **NOTE:** These materials are located in the *Configuration Guide*.

Please review them in order.

Topic	Description
Required Platform Configuration	This section covers the following topics, some of which should already be completed: • *Set Base Storage Layer* - The base storage layer must be set once and never changed. Set this value to `s3`.

	• *Create Encryption Key File* - If you plan to integrate the platform with any relational sources, including Redshift, you must create an encryption key file and store it on the Trifacta node • *Running Environment Options* - Depending on your scenario, you may need to perform additional configuration for your available running environment(s) for executing jobs. • *Profiling Options* - In some environments, tweaks to the settings for visual profiling may be required. You can disable visual profiling if needed. • *Configure for Spark* - If you are enabling the Spark running environment, please review and verify the configuration for integrating the platform with the Spark running environment.
Configure for EMR	Set up for a new EMR cluster. Some content may apply to existing EMR clusters.
Enable Integration with Compressed Clusters	If the Hadoop cluster uses compression, additional configuration is required.
Enable Integration with Cluster High Availability	If you are integrating with high availability on the Hadoop cluster, please complete these steps. • If you are integrating with high availability on the Hadoop cluster, HttpFS must be enabled in the platform. HttpFS is required in other, less-common cases. See *Enable HttpFS*.
Enable Relational Connections	Enable integration with relational databases, including Redshift. • For more information on creating a connection to Redshift, see *Create Redshift Connections*.
Configure for KMS	Integration with the Hadoop cluster's key management system (KMS) for encrypted transport. Instructions are provided for distribution-specific versions of Hadoop.
Configure Security	A list of topics on applying additional security measures to the Trifacta platform and how integrates with Hadoop.
Configure SSO for AD-LDAP	Please complete these steps if you are integrating with your enterprise's AD/LDAP Single Sign-On (SSO) system.

Upgrade

For more information on upgrading your Trifacta Wrangler Enterprise on AWS, please contact *Trifacta Customer Success Services*.

Install for Azure

Contents:

This install process applies to installing Trifacta® Wrangler Enterprise on an Azure infrastructure that you manage.

Azure Marketplace deployments:

> **NOTE:** Content in this section does not apply to deployments from the Azure Marketplace, which provide fewer deployment and configuration options. For more information, see the Azure Marketplace.

Scenario Description

> **NOTE:** All hardware in use for supporting the platform is maintained within the enterprise infrastructure on Azure.

- Installation of Trifacta Wrangler Enterprise on a node in Microsoft Azure
- Installation of Trifacta databases on the same node
- Integration with a supported cluster for running jobs.
- Base storage layer and backend datastore of ADLS or WASB
- High availability or failover of the Trifacta node is not supported in Azure.
- High availability of cluster components is automatically managed by the HDI cluster.
    - Auto-management does not apply to non-Hadoop clusters, such as Azure Databricks.

For more information on deployment scenarios, see *Supported Deployment Scenarios for Azure*.

Product Limitations

- The application user credentials are used to access to the HDI cluster. Details are provided below.
- ADLS/Storage Blob access is only for the HDInsight cluster's primary storage. Additional storage accounts are not supported.
- HDFS must be set as the base storage layer of the Trifacta platform. Details are provided later.
    - S3 integration and AWS-based integrations such as Redshift are not supported.
- Use of HttpFS is not supported.
- Security features such as Kerberos and secure impersonation are not supported.

For more information on the limitations of this deployment scenario, see *Product Limitations*.

Pre-requisities

Desktop Requirements

- All desktop users of the platform should have a supported version of Google Chrome installed on their desktops.
    - For more information. see *Desktop Requirements*.
    - If a supported browser is not available within your enterprise, desktop users can install the Trifacta enterprise application as a separate application. For more information, see *Install for Wrangler Enterprise Application*.
- All desktop users must be able to connect to the EC2 instance through the enterprise infrastructure.

Azure Pre-requisites

Depending on which of the following Azure components you are deploying, additional pre-requisites and limitations may apply. Please review these sections as well.

- Cluster:
  - *Configure for HDInsight*
  - *Configure for Azure Databricks*
- Storage:
  - *Enable ADLS Access*
  - *Enable WASB Access*
- *Configure SSO for Azure AD*

Preparation

Before you begin, please verify that you have completed the following:

1. **Review Planning Guide:** Please review and verify *Install Preparation* and sub-topics.
2. **Read:** Please read this entire document before you create the EMR cluster or install the Trifacta platform.
3. **Acquire Assets:** Acquire the installation package for your operating system and your license key. For more information, contact *Trifacta Support*.
    1. If you are completing the installation without Internet access, you must also acquire the offline versions of the system dependencies. See *Install Dependencies without Internet Access*.
4. **Cluster sizing:** Before you begin, you should allocate sufficient resources for the cluster. For guidance, please contact your Trifacta representative.
5. **Node:** Review the system requirements for the node hosting the Trifacta platform. See *System Requirements*.
    1. The required set of ports must be enabled for listening. See *System Ports*.
    2. This node should be dedicated for Trifacta use.
6. **Databases:**
    1. The platform utilizes a set of databases that must be accessed from the Trifacta node. Databases are installed as part of the workflow described later.
    2. For more information on the supported databases and versions, see *System Requirements*.
    3. For more information on database installation requirements, see *Install Databases*.

Limitations: For more information on limitations of this scenario, see *Product Limitations* in the *Install Preparation* area.

Deploy the Cluster

Deploy and provision a cluster of one of the supported types. The Trifacta platform supports integrations with multiple cluster types.

> **NOTE:** Before you deploy, you should review cluster sizing options. For guidance, please contact your Trifacta representative.

Primary storage of the cluster may be set to an existing Azure Data Lake Store or Blob Storage.

- Any additional storage associated with the cluster is not available through the Trifacta application.

For more information, see *Supported Deployment Scenarios for Azure*.

Deploy the Trifacta node

In your Azure infrastructure, you must deploy a suitable VM for the installation of the Trifacta platform.

The operating system requirements for the VM for installing the platform vary depending on the type of job execution cluster with which you are running.

Cluster Type	Supported O/S for VM	Notes
HDInsight	Ubuntu only	

		Trifacta platform must be installed on an edge node of the HDInsight cluster.
Azure Databricks	CentOS and Ubuntu	

- When you configure the platform to integrate with the cluster, you must acquire some information about the cluster resources. For more information on the set of information to collect, see *Pre-Install Checklist* in the *Install Preparation* area.
- For more information, see *System Requirements* in the *Install Preparation* area.
- A set of ports must be opened on the VM for the platform. For more information, see *System Ports* in the *Install Preparation* area.

For more information on the supported EMR distributions, see *Supported Deployment Scenarios for Azure*.

Prepare the cluster

1. Create the following directories, which are specified by parameter in the platform.

Default HDFS path	Platform configuration property
/user/trifacta	
/trifacta	
/trifacta/dictionaries	hdfs.pathsConfig.dictionaries
/trifacta/libraries	hdfs.pathsConfig.libraries
/trifacta/queryResults	hdfs.pathsConfig.batchResults
/trifacta/tempfiles	hdfs.pathsConfig.tempFiles
/trifacta/uploads	hdfs.pathsConfig.fileUpload
/trifacta/.datasourceCache	hdfs.pathsConfig.globalDatasourceCache

2. Change the ownership of the above directories to `trifacta:trifacta` or the corresponding values for the S3 user in your environment.

Additional users may be required. For more information, see *Required Users and Groups* in the *Install Preparation* area.

Install Workflow

Please complete these steps listed in order:

1. **Install Software:** Install the Trifacta platform software on the node you created.

   > **NOTE:** You must follow the instructions provided for Ubuntu installation.

   See *Install Software*.
2. **Install Databases:** The platform requires several databases for storing metadata.

   > **NOTE:** The software assumes that you are installing the databases on a PostgreSQL server on the same node as the software. If you are not or are changing database names or ports, additional configuration is required as part of this installation process.

   For more information, see *Install Databases*.

3. **Login to the application:** After software and databases are installed, you can login to the application to complete configuration:
    1. See *Login*.
    2. As soon as you login, you should change the password on the admin account. In the left menu bar, s elect **Settings > Admin Settings**. Scroll down to Manage Users. For more information, see *Change Admin Password*.

> **Tip:** At this point, you can access the online documentation through the application. In the left menu bar, select **Help menu > Product Docs**. All of the following content, plus updates, is available online. See Documentation below.

Configuration Workflow

After you have completed the above topics, you can complete the configuration for your deployment below.

> **NOTE:** The following configuration topics are not part of this installation guide. See links below.

1. **Configure for Azure:** Configure the platform to work with Azure.
2. **Integrate with cluster:** If the application is up and running, you can configure to the backend cluster for running jobs. Choose one of the following:
    1. HDInsight
    2. Azure Databricks
3. **Integrate with backend storage:**
    1. **Set base storage layer:** The base storage layer must be set at the time of install and cannot be changed. See *Set Base Storage Layer*.
    2. ADLS
    3. WASB
4. **Verify operations:** At this point, you should be able to run a job. See *Verify Operations*.
5. **Create additional connections:** Through connections, you can access other sources of data and, optionally, publish job results.

Documentation

You can access complete product documentation online and in PDF format. From within the product, select **Help menu > Product Docs**.

After you have accessed the documentation, the following topics are relevant to Azure deployments. Please review them in order.

Topic	Description
Supported Deployment Scenarios for Azure	Matrix of supported Azure components.
Configure for Azure	Top-level configuration topic on integrating the platform with Azure. **Tip:** You should review this page.
Configure for HDInsight	Review this section if you are integrating the Trifacta platform with a pre-existing HDI cluster.
Configure for Azure Databricks	Review this section if you are integrating with a pre-existing Azure Databricks cluster.
Enable ADLS Access	Configuration to enable access to ADLS.
Enable WASB Access	Configuration to enable access to WASB.

Verify Operations	You should be able to verify platform operations by running a simple job at this time.
Relational Connections	To enable, see *Enable Relational Connections*. Azure-specific relational connections: • *Create SQL DW Connections* • *Create SQL DB Connections*
Configure SSO for Azure AD	How to integrate the Trifacta platform with Azure Active Directory for Single Sign-On.

Install from AWS Marketplace

Contents:

This guide steps through the requirements and process for installing Trifacta® Data Preparation for Amazon Redshift and S3 through the AWS Marketplace.

Product Limitations

- Connectivity to sources other than S3 and Redshift is not supported.
- Jobs must be executed in the Trifacta Photon running environment. No other running environment integrations are supported.
- Anomaly and stratified sampling are not supported in this deployment.
- When publishing single files to S3, you cannot apply an `append` publishing action.
- Trifacta Data Preparation for Amazon Redshift and S3 must be deployed into an existing Virtual Private Cloud (VPC).
- The EC2 instance, S3 buckets, and any connected Redshift databases must be located in the same Amazon region. Cross-region integrations are not supported at this time.

  NOTE: HDFS integration is not supported for Amazon Marketplace installations.

- The S3 bucket automatically created by the Marketplace CloudFormation template is not automatically deleted when you delete the stack in CloudFormation. You must empty the bucket and delete it, which can be done through the AWS Console.

Internet access

From AWS, the Trifacta platform requires Internet access for the following services:

NOTE: Depending on your AWS deployment, some of these services may not be required.

- AWS S3
- Key Management System [KMS] (if sse-kms server side encryption is enabled)
- Secure Token Service [STS] (if temporary credential provider is used)
- EMR (if integration with EMR cluster is enabled)

NOTE: If the Trifacta platform is hosted in a VPC where Internet access is restricted, access to S3, KMS and STS services must be provided by creating a VPC endpoint. If the platform is accessing an EMR cluster, a proxy server can be configured to provide access to the AWS ElasticMapReduce regional endpoint.

SELinux

By default, Trifacta Data Preparation for Amazon Redshift and S3 is installed on a server with SELinux enabled. Security-enhanced Linux (SELinux) provides a set of security features for, among other things, managing access controls.

Tip: The following may be applied to other deployments of the Trifacta platform on servers where SELinux has been enabled.

In some cases, SELinux can interfere with normal operations of platform software. If you are experiencing connectivity problems related to SELinux, you can do either one of the following:

1. Disable SELinux on the server. For more information, please see the CentOS documentation.
2. Apply the following commands on the server, as root:
    1. Open ports on the server for listening.
        1. By default, the Trifacta application listens on port 3005. The following opens that port when SELinux is enabled:

```
semanage port -a -t http_port_t -p tcp 3005
```

        2. Repeat the above step for any other ports that you wish to open on the server.
    2. Permit nginx, the proxy on the Trifacta node, to open websockets:

```
setsebool -P httpd_can_network_connect 1
```

Install

Desktop Requirements

- All desktop users of the platform must have the latest version of Google Chrome installed on their desktops.
- All desktop users must be able to connect to the EC2 instance over the port on which Trifacta Data Preparation for Amazon Redshift and S3 is listening. Default is 3005.

NOTE: Trifacta Data Preparation for Amazon Redshift and S3 enforces a maximum limit of 30 users.

Pre-requisites

Before you install the platform, please verify that the following steps have been completed.

1. **EULA.** Before you begin, please review the End-User License Agreement. See *End-User License Agreement*.
2. **SSH Key-pair.** Please verify that there is an SSH key pair available to assign to the Trifacta node.

Install Steps - CloudFormation template

This install process creates the following:

- Trifacta node on an EC2 instance
- An S3 bucket to store data
- IAM roles and policies to access the S3 bucket from the Trifacta node.

Steps:

1. In the Marketplace listing, click **Deploy into an existing VPC**.
2. **Select Template:** The template path is automatically populated for you.
3. **Specify Details:**
    1. **Stack Name:** Display name of the application

        NOTE: Each instance of the Trifacta platform should have a separate name.

    2. **Instance Type:** Only one instance type is enabled.
    3. **Key Pair:** Select the SSH pair to use for Trifacta Instance access.
    4. **Allowed HTTP Source:** Please specify the IP address or range of address from which HTTP /HTTPS connections to the application are permitted.
    5. **Allowed SSH Source:** Please specify the IP address or range of address from which SSH connections to the EC2 instance are permitted.
4. **Options:** None of these is required for installation. Specify your options as needed for your environment.
5. **Review:** Review your installation and configured options.
    1. Select the checkbox at the end of the page.
    2. To launch the configured instance, click **Create**.
6. In the Stacks list, select the name of your application. Click the Outputs tab and collect the following information. Instructions in how to use this information is provided later.
    1. No outputs appear until the stack has been created successfully.

Parameter	Description	Use
TrifactaUrl value	URL and port number to which to connect to the Trifacta application	Users must connect to this IP address and port number to access.
TrifactaBucket	The address of the default S3 bucket	This value must be applied through the application.
TrifactaInstanceId	The identifier for the instance of the platform	This value is the default password for the admin account. **NOTE:** This password must be changed immediately.

7. When the instance is spinning up for the first time, performance may be slow. When the instance is up, please navigate to the `TrifactaUrl` location:

   ```
   http://<public_hostname>:3005
   ```

8. When the login screen appears, enter the following:
   1. Username: `admin@trifacta.local`
   2. Password: (the `TrifactaInstanceId` value)

   > **NOTE:** As soon as you login as an admin for the first time, you should change the password.

9. From the application menu, select the Settings menu. Then, click **Settings > Admin Settings**.
10. In the Admin Settings page, you can configure many aspects of the platform, including user management tasks, and perform restarts to apply the changes.
    1. In the Search bar, enter the following:

       ```
       aws.s3.bucket.name
       ```

    2. Set the value of this setting to be the `TrifactaBucket` value that you collected from the Outputs tab.
11. Click **Save**.
12. When the platform restarts, you can begin using the product.

SSH Access

If you need to SSH to the Trifacta node, you can use the following command:

```
ssh -i <path_to_key_file> <userId>@<tri_node_DNS_or_IP>
```

Parameter	Description
`<path_to_key_file>`	Path to the key file stored on your local computer.
`<userId>`	The user ID is always `centos`.
`<tri_node_DNS_or_IP>`	DNS or IP address of the Trifacta node

Upgrade

For more information, see *Upgrade for AWS Marketplace*.

Documentation

You can access complete product documentation online and in PDF format. From within the product, select **Help menu > Product Docs**.

Install from AWS Marketplace with EMR

Contents:

This documentation applies to installation from a supported Marketplace. Please use the installation instructions provided with your deployment.

If you are installing or upgrading a Marketplace deployment, please use the available PDF content. You must use the install and configuration PDF available through the Marketplace listing.

Scenario Description

CloudFormation templates enable you to install Trifacta® Wrangler Enterprise with a minimal amount of effort.

- After install, customizations can be applied by tweaking the resources that were created by the CloudFormation process.
- If you have additional requirements or a complex environment, please contact *Trifacta Support* for assistance with your solution.

Install

The CloudFormation template creates a complete working instance of Trifacta Wrangler Enterprise, including the following:

- VPC and all required networking infrastructure
- EC2 instance with all supporting policies/roles
- S3 bucket
- EMR cluster
  - Configurable autoscaling instance groups
  - All supporting policies/roles

Pre-requisites

If you are integrating the Trifacta platform with an EMR cluster, you must acquire a Trifacta license first. Additional configuration is required. For more information, please contact aws-marketplace@trifacta.com.

Before you begin:

1. **Read:** Please read this entire document before you begin.
2. **EULA.** Before you begin, please review the End-User License Agreement. See *https://docs.trifacta.com/display/PUB/End-User+License+Agreement+-+Trifacta+Wrangler+Enterprise*.
3. **Trifacta license file:** If you have not done so already, please acquire a Trifacta license file from your Trifacta representative.

Internet access

From AWS, the Trifacta platform requires Internet access for the following services:

> **NOTE:** Depending on your AWS deployment, some of these services may not be required.

- AWS S3
- Key Management System [KMS] (if sse-kms server side encryption is enabled)
- Secure Token Service [STS] (if temporary credential provider is used)
- EMR (if integration with EMR cluster is enabled)

> **NOTE:** If the Trifacta platform is hosted in a VPC where Internet access is restricted, access to S3, KMS and STS services must be provided by creating a VPC endpoint. If the platform is accessing an EMR cluster, a proxy server can be configured to provide access to the AWS ElasticMapReduce regional endpoint.

SELinux

By default, Trifacta Wrangler Enterprise is installed on a server with SELinux enabled. Security-enhanced Linux (SELinux) provides a set of security features for, among other things, managing access controls.

> **Tip:** The following may be applied to other deployments of the Trifacta platform on servers where SELinux has been enabled.

In some cases, SELinux can interfere with normal operations of platform software. If you are experiencing connectivity problems related to SELinux, you can do either one of the following:

1. Disable SELinux on the server. For more information, please see the CentOS documentation.
2. Apply the following commands on the server, as root:
    1. Open ports on the server for listening.
        1. By default, the Trifacta application listens on port 3005. The following opens that port when SELinux is enabled:

```
semanage port -a -t http_port_t -p tcp 3005
```

        2. Repeat the above step for any other ports that you wish to open on the server.
    2. Permit nginx, the proxy on the Trifacta node, to open websockets:

```
setsebool -P httpd_can_network_connect 1
```

Product Limitations

- The EC2 instance, S3 buckets, and any connected Redshift databases must be located in the same Amazon region. Cross-region integrations are not supported at this time.
- No support for Hive integration
- No support for secure impersonation or Kerberos
- No support for high availability and failover
- Job cancellation is not supported on EMR.
- When publishing single files to S3, you cannot apply an `append` publishing action.

Install

Desktop Requirements

- All desktop users of the platform must have the latest version of Google Chrome installed on their desktops.
- All desktop users must be able to connect to the EC2 instance through the enterprise infrastructure.

Steps:

1. In the Marketplace listing, click **Deploy into a new VPC**.
2. **Choose a Template:** The template path is automatically populated for you.
3. **Specify Details:**
    1. **Stack Name:** Display name of the stack is used in the names of resources created by the stack and as an identifier for the stack.

       > **NOTE:** Each instance of the Trifacta platform must have a separate name.

    2. **Instance Type:** Please select the appropriate instance depending on the number of users and data volumes of your environment. For more information, see the Sizing Guide above.
    3. **Key Pair:** This SSH key pair is used to access the Trifacta Instance and the EMR cluster instances.
    4. **Allowed HTTP Source:** This range of addresses are permitted access to the Trifacta Instance on port 80, 443, and 3005.

        1. Port numbers 80 and 443 do not have any services by default, but you may modify the Trifact a configuration to enable access via these ports.
    5. **Allowed SSH Source:** This range of addresses is permitted access to port 22 on the Trifacta Instance.
    6. **EMR Cluster Node Configuration:** Allows you to customize the configuration of the deployed EMR nodes
        1. Reasonable values are used as defaults.
        2. If you do customize these values, you should upsize. Avoid downsizing these values.
    7. **EMR Cluster Autoscaling Configuration:** Allows you to customize the autoscaling settings used by the EMR cluster.
        1. Reasonable values are used as defaults.
4. **Options:** None of these is required for installation. Specify any options as needed for your environment.
5. **Review:** Review your installation and configured options.
    1. Select the checkbox at the end of the page.
    2. To launch the stack, click **Create**.
6. Please wait while the stack creates all required resources.
7. In the Stacks list, select the name of your application. Click the Outputs tab and collect the following information. Instructions on how to use this information are provided later.

Parameter	Description	Use

Trifacta URL value	URL and port number to which to connect to the Trifacta application	Users must connect to this IP address and port number to access. By default, it is set to `3005`. The access port can be moved to `80` or `443` if desired. Please contact us for more details.
Trifacta Bucket	The address of the default S3 bucket	This value must be applied through the application after it has been deployed.
Trifacta Instance Id	The identifier for the instance of the platform	This value is the default password for the admin account. **NOTE:** You must change this password on the first login to the application.

8. After the Trifacta instance has been created, you must add a license file before starting the Trifacta service. In the following steps, you SSH into the server, create the license file, and paste in the license file content, plus update the ownership and permissions of that file:
    1. SSH into the server as the CentOS user, using the key you specified.
    2. Change to root user:

```
sudo su
```

    3. Add your license:

```
vi /opt/trifacta/license/license.json
```

    4. Into the above file, paste the contents of the `license.json` file that was provided to you by your Trifacta representative.
    5. Verify permissions on the file:

```
chown trifacta:trifacta /opt/trifacta/license/license.json
chmod 644 /opt/trifacta/license/license.json
```

9. Start the Trifacta service:

```
service trifacta start
```

10. It may take some time for the server to finish coming online. Navigate to the Trifacta application.
11. When the login screen appears, enter the following:
    1. Username: `admin@trifacta.local`
    2. Password: (the `TrifactaInstanceId` value)

> **NOTE:** After you login as an admin for the first time, you must change the password.

12. From the application menu, select the Settings menu. Then, click **Settings >Admin Settings**.
13. In the Admin Settings page, you can configure many aspects of the platform, including user management tasks, and perform restarts to apply the changes.

14. Add the S3 bucket that was automatically created to store Trifacta metadata and EMR content. Search for:

```
"aws.s3.bucket.name"
```

    1. Update the value with the Trifacta Bucket value provided when you created the stack in AWS.
15. Verify your Spark version. If the cluster was launched from AWS, this value should be set to `2.3.0`. Search for:

```
"spark.version"
```

    1. Update its value to `2.3.0`, if necessary.
16. Enable the Run in EMR option within the platform. Search for:

```
"webapp.runinEMR"
```

    1. Select the checkbox to enable it.
17. Click **Save** underneath the Platform Settings section.
18. In the Admin Settings page, locate the External Service Settings section.
    1. **AWS EMR Cluster ID:** Paste the value for the EMR Cluster ID for the cluster to which the platform is connecting.
        1. Verify that there are no extra spaces in any copied value.
    2. **AWS Region:** Enter the region where your EMR cluster is located.
    3. **Resource Bucket:** you may use the already created Trifacta Bucket.
        1. Verify that there are no extra spaces in any copied value.
    4. **Resource Path:** you should use something like `EMRLOGS`.
19. Click **Save** underneath the External Service Settings section.
20. When the platform restarts, you can begin using the product.

Note about deleting the CloudFormation stack

If you must delete the CloudFormation stack, please be aware of the following.

1. The S3 bucket that was created for the stack is not removed. If you want to delete it, you must empty it first and then delete it.
2. Any EMR security groups created for the stack cannot be deleted, due to circular references. The stack deletion process informs you of the security groups that it failed to delete. To complete the deletion:
    1. Remove all rules from the security groups.
    2. Delete the security groups manually.
    3. Re-run the stack deletion, which should complete successfully.

Verify

Start and Stop the Platform

Use the following command line commands to start, stop, and restart the platform.

Start:

```
sudo service trifacta start
```

Stop:

```
sudo service trifacta stop
```

Restart:

```
sudo service trifacta restart
```

Verify Operations

After you have installed or made changes to the platform, you should verify end-to-end operations.

NOTE: The Trifacta® platform is not operational until it is connected to a supported backend datastore.

Steps:

1. Login to the application as an administrator. See *Login*.
2. Through the Admin Settings page, run Tricheck, which performs tests on the Trifacta node and any connected cluster. See *Admin Settings Page*.
3. In the application menu bar, click **Library**. Click **Import Dataset**. Select your backend datastore.
4. Navigate your datastore directory structure to locate a small CSV or JSON file.
5. Select the file. In the right panel, click **Create and Transform**.
    1. **Troubleshooting:** If the steps so far work, then you have read access to the datastore from the platform. If not, please check permissions for the Trifacta user and its access to the appropriate directories.
    2. See *Import Data Page*.
6. In the Transformer page, some steps have already been added to your recipe, so you can run the job right away. Click **Run Job.**
    1. See *Transformer Page* .
7. In the Run Job Page:
    1. For Running Environment, some of these options may not be available. Choose according to the running environment you wish to test.
        1. **Photon:** Runs job on the Photon running environment hosted on the Trifacta node. This method of job execution does not utilize any integrated cluster.
        2. **Spark:** Runs the job on Spark on the integrated cluster.

3. **Databricks:** If the platform is integrated with an Azure Databricks cluster, you can test job execution on the cluster.

   > **NOTE:** Use of Azure Databricks is not supported for Marketplace installs.

2. Select CSV and JSON output.
3. Select the Profile Results checkbox.
4. **Troubleshooting:** At this point, you are able to initiate a job for execution on the selected running environment. Later, you can verify operations by running the same job on other available environments .
5. See *Run Job Page*.

8. When the job completes, you should see a success message in the Jobs tab of the Flow View page.
   1. **Troubleshooting:** Either the Transform job or the Profiling job may break. To localize the problem, mouse over the Job listing in the Jobs page. Try re-running a job by deselecting the broken job type or running the job in a different environment. You can also download the log files to try to identify the problem. See *Jobs Page*.

9. Click **View Results** in the Jobs page. In the Profile tab of the Job Details page, you can see a visual profile of the generated results.
   1. See *Job Details Page*.
10. In the Output Destinations tab, click the CSV and JSON links to download the results to your local desktop. See *Import Data Page*.

11. Load these results into a local application to verify that the content looks ok.

Upgrade

For more information, see *Upgrade for AWS Marketplace with EMR*.

Documentation

You can access complete product documentation in online and PDF format. After the platform has been installed, select **Help menu > Product Docs** from the menu in the Trifacta application.

Install from Azure Marketplace

Contents:

- *Documentation*

This documentation applies to installation from a supported Marketplace. Please use the installation instructions provided with your deployment.

If you are installing or upgrading a Marketplace deployment, please use the available PDF content. You must use the install and configuration PDF available through the Marketplace listing.

This guide steps through the requirements and process for installing Trifacta® Wrangler Enterprise from the Azure Marketplace.

Product Limitations

- HDInsight 3.6
- HDInsight Hadoop, Spark and HBase cluster types

Documentation Scope

This document guides you through the process of installing the product and beginning to use it.

- **If you are creating a new cluster as part of this process:** You can begin running jobs. Instructions are provided in this document for testing job execution.
- **If you are integrating the product with a pre-existing cluster:** Additional configuration is required after you complete the installation process. Relevant topics are listed in the Documentation section at the end of this document.

Install

Desktop Requirements

- All desktop users of the platform must have the latest version of Google Chrome installed on their desktops.
- All desktop users must be able to connect to the created Trifacta node instance through the enterprise infrastructure.

Sizing Guide

NOTE: The following guidelines apply only to Trifacta Wrangler Enterprise on the Azure Marketplace.

Use the following guidelines to select your instance size:

Azure virtual machine type	vCPUs	RAM (GB)	Max recommended concurrent users	Avg. input data size of jobs on Trifacta Server (GB)
Standard_A4	8	14	30	1
Standard_A6	4	28	30	2
Standard_A7	8	56	60	5
Standard_A10	8	56	60	5

Standard_A11	16	112	90	11
Standard_D4_v2	8	28	30	2
Standard_D5_v2	16	56	90	5
Standard_D12_v2	4	28	30	2
Standard_D13_v2	8	56	60	5
Standard_D14_v2	16	112	90	11
Standard_D15_v2	20	140	120	14

Pre-requisites

Before you install the platform, please verify that the following steps have been completed.

1. **License:** When you install the software, the installed license is valid for 24 hours. You must acquire a license key file from Trifacta. For more information, please contact Sales.
2. **Supported Cluster Types:**
    1. Hadoop
    2. HBase
    3. Spark
3. **Data Lake Store only:** If you are integrating with Azure Data Lake Store, please review and complete the following section.

Required permissions

- Azure Active Directory registered application permissions are listed below.
- If you are connecting to a SQL DW database, additional permissions are required and are described in the full Install Guide. See *Create SQL DW Connections*.
- If you are integrating with a WASB cluster, you must generate a SAS token with appropriate permissions for each WASB cluster. Instructions are available in the full Install Guide. See *Enable WASB Access*.

Register application and r/w access

You must create an Azure Active Directory registered application with appropriate permissions for the following:

- Read/write access to the Azure Key Vault resources
- Read/write access to the Data Lake Store resource

This can be either the same service principal used for the HDInsight cluster or a new one created specifically for the Trifacta platform. This service principal is used by the Trifacta platform for access to all Azure resources. For more information, see *https://docs.microsoft.com/en-us/azure/azure-resource-manager/resource-group-create-service-principal-portal*.

1. To create a new service principal, see *https://docs.microsoft.com/en-us/azure/azure-resource-manager/resource-group-create-service-principal-portal#create-an-azure-active-directory-application* .
2. Obtain property values:
    1. After you have created a new one, please acquire the following property values prior to install.
    2. For an existing service principal, see *https://docs.microsoft.com/en-us/azure/azure-resource-manager/resource-group-create-service-principal-portal#get-application-id-and-authentication-key* to obtain the property values.

These values are applied during the install process.

Azure Property	Location	Use
Application ID		

	Acquire this value from the Registered app blade of the Azure Portal.	Applied to Trifacta platform configuration: `azure.applicationid`.
Service User Key	Create a key for the Registered app in the Azure Portal.	Applied to Trifacta platform configuration: `azure.secret`.
Directory ID	Copy the Directory ID from the Properties blade of Azure Active Directory.	Applied to Trifacta platform configuration: `azure.directoryId`.

Install Process

Methods

You can install from the Azure Marketplace using one of the following methods:

1. **Create cluster:** Create a brand-new HDI cluster and add Trifacta Wrangler Enterprise as an application.

   > **Tip:** This method is easiest and fastest to deploy.

2. **Add application:** Use an existing HDI cluster and add Trifacta Wrangler Enterprise as an application.

   > **Tip:** This method does not support choosing a non-default size for the Trifacta node. If you need more flexibility, please choose the following option.

3. **Create a custom ARM template:** Use an existing HDI cluster and configure a custom application via Azure Resource Manager (ARM) template.

   > **Tip:** Use the third method only if your environment requires additional configuration flexibility or automated deployment via ARM template.

Depending on your selection, please follow the steps listed in one of the following sections.

> **NOTE:** These steps include required settings or recommendations only for configuring a cluster and the application for use with Trifacta Wrangler Enterprise. Any other settings should be specified for your enterprise requirements.

Install Method - New cluster

Please use the following steps if you are creating a new HDI cluster and adding the Trifacta application to it.

Steps:

1. From the Azure Marketplace listing, click **Get it Now**.
2. In Microsoft Azure Portal, click the New blade.
3. Select **Trifacta Wrangler Enterprise** .
4. Click **Create**. Then, click the Quick Create tab.
    1. Please configure any settings that are not listed below according to your enterprise requirements.
    2. For more information on the Quick Create settings, see *https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-provision-linux-clusters*.
5. Basics tab:
    1. Cluster type:
        1. Hadoop: One of the following:
            1. Hadoop
            2. HBase

3. Spark
2. Version:
    1. HDI 3.6
3. Click **Select**.

2. Click **Next**.

6. Storage tab:
    1. Select your primary storage:

> **NOTE:** Trifacta Wrangler Enterprise does not support Additional storage options. Additional configuration is required after install is complete for the supported storage options.

        1. Azure Storage
        2. Azure Data Lake Store
    2. Select a SQL database for Hive: If you plan to use Hive, you can choose to specify a database. If not, the platform creates one for you.
    3. Click **Next**.
7. Trifacta Wrangler Enterprise tab:
    1. Please review and accept the terms for using the product.
    2. Click **Create**. Click **Ok**.
    3. Click **Next**.
8. Custom tab:

    1. By default, the Trifacta node is defined as: `Standard_D13_V2`.
    2. If you need to change the edge node size, please click the Custom settings tab and make your selections. The following virtual machine types are supported:

```
Standard_A4
Standard_A6
Standard_A7
Standard_A10
Standard_A11
Standard_D4_v2
Standard_D5_v2
Standard_D12_v2
Standard_D13_v2
Standard_D14_v2
Standard_D15_v2
```

    3. Click **Next**.
9. Advanced Settings tab:

    1. No changes are needed here.
10. Summary tab:

    1. Review the specification of the cluster you are creating.
    2. Make modifications as needed.
    3. To create the cluster, click **Create**.
11. After the cluster has been started and deployed:
    1. Login to the application.
    2. Change the admin password.
    3. Perform required additional configuration.
    4. Instructions are provided below.

Install Method - Add application to a cluster

Please use the following steps if you are adding the Trifacta application to a pre-existing HDI cluster.

> **NOTE:** The Trifacta node is set to `Standard_V13_D2` by default. The size cannot be modified.

Steps:

1. In the Microsoft Azure Portal, select the HDInsight cluster to which you are adding the application.
2. In the Portal, select the Applications blade. Click **+ Add**.
3. From the list of available applications, select **Trifacta Wrangler Enterprise** .
4. Please accept the legal terms.
5. Click **Next**.
6. The application is created for you.
7. After the cluster has been started and deployed:
    1. Login to the application.
    2. Change the admin password.
    3. Perform required additional configuration.
    4. Instructions are provided below.

Install Method - Build custom ARM template

Please use the following steps if you are creating a custom application template for later deployment. This method provides more flexible configuration options and can be used for deployments in the future.

> **NOTE:** You must have a pre-existing HDI cluster for which to create the application template.

> **NOTE:** Before you begin, you should review the End-User License Agreement. See *End-User License Agreement*.

Steps:

1. Start here:
    1. *https://github.com/trifacta/azure-deploy/tree/release/5.0*
    2. Click **Deploy from Azure**.
2. From the Microsoft Azure Portal, select the custom deployment link.
3. Resource Group: Create or select one.
4. Cluster Name: Select an existing cluster name.
5. Edge Node Size: Select the instance type. For more information, see the Sizing Guide above.
6. Trifacta version: For the version, select the latest listed version of Trifacta Wrangler Enterprise.
7. Application Name: If desired, modify the application name as needed. This name must be unique per cluster.
8. Subdomain Application URI Suffix: If desired, modify the three-character alphanumeric string used in the DNS name of the application. This suffix must be unique per cluster.
9. Please specify values for the following:
    1. Application ID
    2. Directory ID
    3. Secret
10. Gallery Package Identifier: please leave the default value.
11. Please accept the Microsoft terms of use.
12. To create the template, click **Purchase**.
13. The custom template can be used to create the Trifacta Wrangler Enterprise application. For more information, please see the Azure documentation.
14. After the application has been started and deployed:
    1. Login to the application.

2. Change the admin password.
3. Perform required additional configuration.
4. Instructions are provided below.

Login to Trifacta Wrangler Enterprise

Steps:

1. In the Azure Portal, select the HDI cluster.
2. Select the Applications blade.
3. Select the Trifacta Wrangler Enterprise application.
4. Click the Portal link.
5. You may be required to apply the cluster username and password.

> **NOTE:** You can create a local user of the cluster to avoid enabling application users to use the administrative user's cluster credentials. To create such a user:
>
> 1. Navigate to the cluster's Ambari console.
> 2. In the user menu, select the Manage Ambari page.
> 3. Select **Users**.
> 4. Select **Create Local User**.
> 5. Enter a unique (lowercase) user name.
> 6. Enter a password and confirm that password.
> 7. For User Access, select **Cluster User**.
> 8. Verify that the user is not an Ambari Admin and has Active status.
> 9. Click **Save**.

6. You are connected to the Trifacta application.
7. In the login screen, enter the default username and password:
    1. Username: *admin@trifacta.local*
    2. Password: `admin`
    3. Click **Login**.

> **NOTE:** If this is your first login to the application, please be sure to reset the admin password. Steps are provided below.

Change admin password

Steps:

1. If you haven't done so already, login to the Trifacta application as an administrator.
2. In the menu bar, select **Settings menu > Administrator**.
3. In the User Profile, enter and re-enter a new password.
4. Click **Save**.
5. Logout and login again using the new password.

Post-install configuration

Base parameter settings

If you are integrating with HDI and did not install via a custom ARM template, the following settings must be specified with the Trifacta application .

> **NOTE:** These settings are specified as part of the cluster definition. If you have not done so already, you should acquire the corresponding values for the Trifacta application in the Azure Portal.

Steps:

1. Login to the Trifacta application as an administrator.
2. In the menu bar, select **Settings menu > Admin Settings**.
3. In the Admin Settings page, specify the values for the following parameters:

```
"azure.secret"
"azure.applicationId"
"azure.directoryId"
```

4. Save your changes and restart the platform.
5. When the platform is restarted, continue the following configuration.

Apply license file

When the application is first created, the license is valid for 24 hours. Before the license expires, you must apply the license key file to the Trifacta node. Please complete the following general steps.

Steps:

1. Locate the license key file that was provided to you by Trifacta. Please store this file in a safe location that is not on the Trifacta node.
2. In the Azure Portal, select the HDI cluster.
3. Select the Applications blade.
4. Select the Trifacta Wrangler Enterprise application.
5. From the application properties, acquire the SSH endpoint.
6. Connect via SSH to the Trifacta node. For more information, see *https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-linux-use-ssh-unix*.
7. Drop the license key file in the following directory on the node:

```
/opt/trifacta/license
```

8. Restart the platform.

Review refresh token encryption key

By default, the Trifacta platform includes a static refresh token encryption key for the secure token service. The same default key is used for all instances of the platform.

> **NOTE:** A valid base64 value must be configured for the platform, or the platform fails to start.

If preferred, you can generate your own key value, which is unique to your instance of the platform.

Steps:

1. Login at the command line to the Trifacta node.
2. To generate a new refresh token encryption key, please execute the following command:

```
cd /opt/trifacta/services/secure-token-service/ && java -cp server/build
/install/secure-token-service/secure-token-service.jar:server/build/install
/secure-token-service/lib/* com.trifacta.services.secure_token_service.tools.
RefreshTokenEncryptionKeyGeneratorTool
```

3. The refresh token encryption key is printed to the screen. Please copy this value to the clipboard.
4. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.
5. Paste the value in the following property:

```
"com.trifacta.services.secure_token_service.
refresh_token_encryption_key": "<generated_key>",
```

6. Save your changes.

Configure Spark Execution (Advanced)

The cluster instance of Spark is used to execute larger jobs across the nodes of the cluster. Spark processes run multiple executors per job. Each executor must run within a YARN container. Therefore, resource requests must fit within YARN's container limits.

Like YARN containers, multiple executors can run on a single node. More executors provide additional computational power and decreased runtime.

Spark's dynamic allocation adjusts the number of executors to launch based on the following:

- job size
- job complexity
- available resources

The per-executor resource request sizes can be specified by setting the following properties in the `spark.props` section.

NOTE: In `trifacta-conf.json`, all values in the `spark.props` section must be quoted values.

Parameter	Description
`spark.executor.memory`	Amount of memory to use per executor process (in a specified unit)
`spark.executor.cores`	Number of cores to use on each executor - limit to 5 cores per executor for best performance

A single special process (the application driver) also runs in a container. Its resources are specified in the `spark.props` section:

Parameter	Description
`spark.driver.memory`	Amount of memory to use for the driver process (in a specified unit)
`spark.driver.cores`	Number of cores to use for the driver process

Optimizing "Small" Joins (Advanced)

Broadcast, or map-side, joins materialize one side of the join and send it to all executors to be stored in memory. This technique can significantly accelerate joins by skipping the sort and shuffle phases during a "reduce" operation. However, there is also a cost in communicating the table to all executors. Therefore, only "small" tables should be considered for broadcast join. The definition of "small" is set by the `spark.sql.autoBroadcastJoinThreshold` parameter which can be added to the `spark.props` section of `trifacta-conf.json`. By default, Spark sets this to `10485760` (10MB).

NOTE: You should set this parameter between 20 and 100MB. It should not exceed 200MB.

Example Spark configuration parameters for given YARN configurations:

Property settings in **bold** are provided by the cluster.

Property	Small	Medium	Large	Extra large
YARN NodeManager node (v)CPUs	**4**	**8**	**16**	**40**
YARN NodeManager node memory (GB)	**16**	**32**	**64**	**160**
yarn.nodemanager.resource.memory-mb	**12288**	**24576**	**57344**	**147456**
yarn.nodemanager.resource.cpu-vcores	**3**	**6**	**13**	**32**
yarn.scheduler.maximum-allocation-mb	**12288**	**24576**	**57344**	**147456**
yarn.scheduler.maximum-allocation-vcores	**3**	**6**	**13**	**32**
spark.executor.memory	6GB	6GB	16GB	20GB
spark.executor.cores	2	2	4	5
spark.driver.memory	4GB	4GB	4GB	4GB
spark.driver.cores	1	1	1	1
spark.sql.autoBroadcastJoinThreshold	20971520	20971520	52428800	104857600

Additional Configuration

Integration with a pre-existing cluster

If you did not create an HDI cluster as part of this install process, you must perform additional configuration to integrate the Trifacta platform with your cluster. Please see the links under Documentation below.

Exploring and Wrangling Data in Azure

NOTE: If you are integrating the Trifacta platform with an existing cluster, these steps do not work. Additional configuration is required. See the Documentation section below.

Basic steps:

1. When data is imported to the Trifacta platform, a reference to it is stored by the platform as an **imported dataset**. The source data is not modified.
2. In the application, you modify the recipe associated with a dataset to transform the imported data.
3. When the recipe is ready, you define and run a job, which executes the recipe steps across the entire dataset.
4. The source of the dataset is untouched, and the results are written to the specified location in the specified format.

Steps:

> **NOTE:** Any user with a valid user account can import data from a local file.

1. Login.
2. In the menubar, click **Datasets**. Click **Import Data**.
3. To add a dataset:
    1. Select the connection where your source is located:
        1. WASB (Blob Storage)
        2. ADL (Azure Data Lake Store)
        3. Hive
    2. Navigate to the file or files for your source.
    3. To add the dataset, click the Plus icon next to its name.
4. To begin working with a dataset, you must first add it into a **flow**, which is a container for datasets. Click the Add Dataset to a Flow checkbox and enter the name for a new flow.

    > **Tip:** If you have selected a single file, you can begin wrangling it immediately. Click **Import and Wrangle**. The flow is created for you, and your dataset is added to it.

5. Click **Import & Add to Flow**.
6. After the flow has been created, the flow is displayed. Select the dataset, which is on the left side of the screen.
7. Click **Add New Recipe**. Click **Edit Recipe**.
8. The dataset is opened in the Transformer page, where you can begin building your recipe steps.

Upgrade

For more information, see *Upgrade for Azure Marketplace*.

Documentation

You can access complete product documentation online and in PDF format. From within the product, select **Help menu > Product Docs**.

After you have accessed the documentation, the following topics are relevant to Azure deployments. Please review them in order.

Topic	Description
Supported Deployment Scenarios for Azure	Matrix of supported Azure components.
Configure for Azure	Top-level configuration topic on integrating the platform with Azure. **Tip:** You should review this page.

Configure for HDInsight	Review this section if you are integrating the Trifacta platform with a pre-existing HDI cluster.
Enable ADLS Access	Configuration to enable access to ADLS.
Enable WASB Access	Configuration to enable access to WASB.
Configure SSO for Azure AD	How to integrate the Trifacta platform with Azure Active Directory for Single Sign-On.

Configure Server Access through Proxy

When you attempt to launch the application, you may receive an error message similar to the following:

```
No internet connection
Remote server timed out.
```

In some environments, your desktop machine may need to connect to the Internet through a proxy server. If you are using Wrangler Enterprise desktop application, it needs to know the proxy server to which to connect in order to access the Trifacta® node.

Please complete the following configuration steps to access the Trifacta servers.

Steps:

1. In the No internet connection dialog, click **Configure Proxy Settings**.
2. Please provide the following configuration information for your proxy server:
    1. **Proxy Host:** The URL of the proxy server. Please include the protocol identifier (e.g. `http://` or `https://`).
    2. **Proxy Port:** The port number to use to connect to the proxy server. In a URL, this value appears after a colon (e.g. `http://myproxy.example.com:8080`).
    3. **Username:** (optional) If your proxy requires a username to access, please enter it here.
    4. **Password:** (optional) Password associated with the user name.
3. Click **Save Proxy Settings and Restart**.

When the application restarts, you should be able to connect to the login screen.

> **NOTE:** If you continue to have difficulties connecting to the Internet, please contact your network administrator or Internet provider.

Install Software

To install Trifacta® Wrangler Enterprise, please review and complete the following sections in the order listed below.

Topics:

- *Install Dependencies without Internet Access*

- *Install Enterprise on CentOS and RHEL*
- *Install Enterprise on Ubuntu*
- *License Key*
- *Install for Wrangler Enterprise Application*
- *Start and Stop the Platform*
- *Login*

Install Dependencies without Internet Access

Offline dependencies should be included in the URL location that Trifacta® provided to you. Please use the `\*deps\*` file.

> **NOTE:** If your installation server is connected to the Internet, the required dependencies are automatically downloaded and installed for you. You may skip this section.

Use the steps below to acquire and install dependencies required by the Trifacta platform. If you need further assistance, please contact *Trifacta Support*.

Install dependencies without Internet access for CentOS or RHEL:

1. In a CentOS or RHEL environment, the dependencies repository must be installed into the following directory:

   ```
   /var/local/trifacta
   ```

2. The following commands configure Yum to point to the repository in `/var/local/trifacta`, which yum knows as `local`. Repo permissions are set appropriately. Commands:

   ```
   tar xvzf <DEPENDENCIES_ARCHIVE>.tar.gz
   mv local.repo /etc/yum.repos.d
   mv trifacta /var/local
   chown -R root:root /var/local/trifacta
   chmod -R o-w+r /var/local/trifacta
   ```

3. The following command installs the RPM while disable all repos other than local, which prevents the installer from reaching out to the Internet for package updates:

   > **NOTE:** The disabling of repositories only applies to this command.

   ```
   sudo yum --disablerepo=* --enablerepo=local install <INSTALLER>.rpm
   ```

4. If the above command fails and complains about a missing repo, you can add the missing repo to the `enablerepo` list. For example, if the `centos-base` repo is reported as missing, then the command would be the following:

   ```
   sudo yum --disablerepo=* --enablerepo=local,centos-base install
   <INSTALLER>.rpm
   ```

5. If you do not have a supported version of a Java Developer Kit installed on the Trifacta node, you can use the following command to install OpenJDK, which is included in the offline dependencies:

```
sudo yum --disablerepo=* --enablerepo=local,centos-base install
java-1.8.0-openjdk-1.8.0 java-1.8.0-openjdk-devel
```

Install dependencies without Internet access in Ubuntu:

If you are trying to perform a manual installation of dependencies in Ubuntu, please contact *Trifacta Support*.

Install Enterprise on CentOS and RHEL

Contents:

This guide takes you through the steps for installing Trifacta® Wrangler Enterprise software on CentOS or Red Hat.

For more information on supported operating system versions, see *System Requirements*.

Preparation

Before you begin, please complete the following.

> **NOTE:** Except for database installation and configuration, all install commands should be run as the root user or a user with similar privileges. For database installation, you will be asked to switch the database user account.

Steps:

1. Set the node where Trifacta Wrangler Enterprise is to be installed.
   1. Review the *System Requirements* and verify that all required components have been installed.
   2. Verify that all required system ports are opened on the node. See *System Ports*.
2. Review the *Desktop Requirements*.

> **NOTE:** Trifacta Wrangler Enterprise requires the installation of Google Chrome on each desktop. Additionally, two plugins must be enabled and of sufficient versions to properly use the Trifacta Photon client. For more information,. see *Desktop Requirements*.

3. Review the *System Dependencies*.

> **NOTE:** If you are installing on node without access to the Internet, you must download the offline dependencies before you begin. See *Install Dependencies without Internet Access*.

4. Acquire your *License Key*.
5. Install and verify operations of the datastore, if used.

> **NOTE:** Access to the Spark cluster is required.

6. Verify access to the server where the Trifacta platform is to be installed.
7. **Cluster Configuration:** Additional steps are required to integrate the Trifacta platform with the cluster. See *Prepare Hadoop for Integration with the Platform*.

Installation

1. Install Dependencies

Without Internet access

If you have not done so already, you may download the dependency bundle with your release directly from Trifacta . For more information, see *Install Dependencies without Internet Access*.

With Internet access

Use the following to add the hosted package repository for CentOS/RHEL, which will automatically install the proper packages for your environment.

```
# If the client has curl installed ...
curl https://packagecloud.io/install/repositories/trifacta/dependencies
/script.rpm.sh | sudo bash

# Otherwise, you can also use wget ...
wget -qO- https://packagecloud.io/install/repositories/trifacta
/dependencies/script.rpm.sh | sudo bash
```

2. Install JDK

By default, the Trifacta node uses OpenJDK for accessing Java libraries and components. In some environments, basic setup of the node may include installation of a JDK. Please review your environment to verify that an appropriate JDK version has been installed on the node.

> **NOTE:** Use of Java Development Kits other than OpenJDK is not currently supported. However, the platform may work with the Java Development Kit of your choice, as long as it is compatible with the supported version(s) of Java. See *System Requirements*.

NOTE: OpenJDK is included in the offline dependencies, which can be used to install the platform without Internet access. For more information, see *Install Dependencies without Internet Access*.

The following commands can be used to install OpenJDK. These commands can be modified to install a separate compatible version of the JDK.

```
sudo yum install java-1.8.0-openjdk-1.8.0 java-1.8.0-openjdk-devel
```

NOTE: If `java-1.8.0-openjdk-devel` is not included, the batch job runner service, which is required, fails to start.

JAVA_HOME:

By default, the `JAVA_HOME` environment variable is configured to point to a default install location for the OpenJDK package.

NOTE: If you have installed a JDK other than the OpenJDK version provided with the software, you must set the `JAVA_HOME` environment variable on the Trifacta node to point to the correct install location.

The property value must be updated in the following locations:

1. Edit the following file: `/opt/trifacta/conf/env.sh`
2. Save changes.

3. Install Trifacta package

NOTE: If you are installing without Internet access, you must reference the local repository. The command to execute the installer is slightly different. See *Install Dependencies without Internet Access*.

NOTE: Installing the Trifacta platform in a directory other than the default one is not supported or recommended.

Install the package with yum, using root:

```
sudo yum install <rpm file>
```

4. Verify Install

The product is installed in the following directory:

```
/opt/trifacta
```

JAVA_HOME:

The platform must be made aware of the location of Java.

Steps:

1. Edit the following file: `/opt/trifacta/conf/trifacta-conf.json`
2. Update the following parameter value:

```
"env": {
  "JAVA_HOME": "/usr/lib/jvm/java-1.8.0-openjdk.x86_64"
},
```

3. Save changes.

5. Install License Key

Please install the license key provided to you by Trifacta. See *License Key*.

6. Store install packages

For safekeeping, you should retain all install packages that have been installed with this Trifacta deployment.

7. Install and configure Trifacta databases

The Trifacta platform requires installation of several databases. If you have not done so already, you must install and configure the databases used to store Trifacta metadata. See *Install Databases*.

Configuration

After installation is complete, additional configuration is required.

The Trifacta platform requires additional configuration for a successful integration with the datastore. Please review and complete the necessary configuration steps. For more information, see *Configure*.

Install Enterprise on Ubuntu

Contents:

This guide takes you through the steps for installing Trifacta® Wrangler Enterprise software on Ubuntu.

For more information on supported operating system versions, see *System Requirements*.

Preparation

Before you begin, please complete the following.

> **NOTE:** Except for database installation and configuration, all install commands should be run as the root user or a user with similar privileges. For database installation, you will be asked to switch the database user account.

Steps:

1. Set the node where Trifacta Wrangler Enterprise is to be installed.
    1. Review the *System Requirements* and verify that all required components have been installed.
    2. Verify that all required system ports are opened on the node. See *System Ports*.
2. Review the *Desktop Requirements*.

    > **NOTE:** Trifacta Wrangler Enterprise requires the installation of Google Chrome on each desktop. Additionally, two plugins must be enabled and of sufficient versions to properly use the Trifacta Photon running environment.

3. Review the *System Dependencies*.

    > **NOTE:** If you are installing on node without access to the Internet, you must download the offline dependencies before you begin. See *Install Dependencies without Internet Access*.

4. Acquire your *License Key*.
5. Install and verify operations of the datastore, if used.

    > **NOTE:** Access to the cluster may be required.

6. Verify access to the server where the Trifacta platform is to be installed.
7. **Cluster configuration:** Additional steps are required to integrate the Trifacta platform with the cluster. See *Prepare Hadoop for Integration with the Platform*.

Installation

1. Install Dependencies

Without Internet access

If you have not done so already, you may download the dependency bundle with your release directly from Trifacta . For more information, see *Install Dependencies without Internet Access*.

With Internet access

Use the following to add the hosted package repository for Ubuntu, which will automatically install the proper packages for your environment.

NOTE: Install curl if not present on your system.

Then, execute the following command:

NOTE: Run the following command as the root user. In proxied environments, the script may encounter issues with detecting proxy settings.

```
curl https://packagecloud.io/install/repositories/trifacta/dependencies
/script.deb.sh | sudo bash
```

Special instructions for Ubuntu installs

These steps manually install the correct and supported version of the following:

- nodeJS
- nginX

Due to a known issue resolving package dependencies on Ubuntu, please complete the following steps prior to installation of other dependencies or software.

1. Login to the Trifacta node as an administrator.
2. Execute the following command to install the appropriate versions of nodeJS and nginX.
    1. Ubuntu 14.04:

        ```
        sudo apt-get install nginx=1.12.2-1~trusty nodejs=10.13.0-
        1nodesource1
        ```

    2. Ubuntu 16.04

        ```
        sudo apt-get install nginx=1.12.2-1~xenial nodejs=10.13.0-
        1nodesource1
        ```

3. Continue with the installation process.

2. Install JDK

By default, the Trifacta node uses OpenJDK for accessing Java libraries and components. In some environments, basic setup of the node may include installation of a JDK. Please review your environment to verify that an appropriate JDK version has been installed on the node.

NOTE: Use of Java Development Kits other than OpenJDK is not currently supported. However, the platform may work with the Java Development Kit of your choice, as long as it is compatible with the supported version(s) of Java. See *System Requirements*.

NOTE: OpenJDK is included in the offline dependencies, which can be used to install the platform without Internet access. For more information, see *Install Dependencies without Internet Access*.

The following commands can be used to install OpenJDK. These commands can be modified to install a separate compatible version of the JDK.

```
sudo apt-get install openjdk-8-jre-headless
```

JAVA_HOME:

By default, the `JAVA_HOME` environment variable is configured to point to a default install location for the OpenJDK package.

NOTE: If you have installed a JDK other than the OpenJDK version provided with the software, you must set the `JAVA_HOME` environment variable on the Trifacta node to point to the correct install location.

The property value must be updated in the following locations:

1. Edit the following file: `/opt/trifacta/conf/env.sh`
2. Save changes.

3. Install Trifacta package

NOTE: If you are installing without Internet access, you must reference the local repository. The command to execute the installer is slightly different. See *Install Dependencies without Internet Access*.

NOTE: Installing the Trifacta platform in a directory other than the default one is not supported or recommended.

Install the package with apt, using root:

```
sudo dpkg -i <deb file>
```

The previous line may return an error message, which you may ignore. Continue with the following command:

```
sudo apt-get -f -y install
```

4. Verify Install

The product is installed in the following directory:

```
/opt/trifacta
```

JAVA_HOME:

The platform must be made aware of the location of Java.

Steps:

1. Edit the following file: `/opt/trifacta/conf/trifacta-conf.json`
2. Update the following parameter value:

```
"env": {
  "JAVA_HOME": "/usr/lib/jvm/java-1.8.0-openjdk.x86_64"
},
```

3. Save changes.

5. Install License Key

Please install the license key provided to you by Trifacta. See *License Key*.

6. Store install packages

For safekeeping, you should retain all install packages that have been installed with this Trifacta deployment.

7. Install and configure Trifacta databases

The Trifacta platform requires installation of several databases. If you have not done so already, you must install and configure the databases used to store Trifacta metadata. See *Install Databases*.

Configuration

After installation is complete, additional configuration is required.

> **The Trifacta platform requires additional configuration for a successful integration with the datastore. Please review and complete the necessary configuration steps. For more information, see *Configure*.**

License Key

Contents:

Acquire license key

A valid license key (`license.json`) is provided to each customer prior to installation. Your license key file is a JSON file that contains important information on your license such as the expiration date.

> **NOTE:** If your license key has expired, please contact *Trifacta Support*.

Install your license key

If you are updating your license, you may want to save your previous license key to a new location before overwriting.

> **NOTE:** Do not maintain multiple license key files in this directory.

To apply your new or updated license key, copy the key file to the following location in the Trifacta® deployment:

```
/opt/trifacta/license
```

Update your license key

After you have installed your license key, you can update your license with a new one through the Admin Settings page. See *Admin Settings Page*.

Changing the license key location

By default, the license key file in use must be named: `license.json`.

If needed, you can change the path and filename of the license key. The property is the following:

```
"license.location"
```

See *Admin Settings Page*.

Expired license

> **NOTE:** If your license expires, you cannot use the product until a new and valid license key file has been applied. When administrators attempt to login to the application, they are automatically redirected to a location from which they can upload a new license key file.

Invalid license key file

When you start the Trifacta platform, you may see the following:

Your license key is either missing or has expired. Please contact *Trifacta Support*.

Install for Wrangler Enterprise Application

Contents:

If your environment does not support the use of Chrome, you can install the Wrangler Enterprise desktop application to provide the same access and functionality as the Trifacta® application. This desktop application connects to the enterprise Trifacta instance and provides the same capabilities without requiring a locally installed version of Chrome browser.

Trifacta application is a hybrid desktop application.Your local application instance accesses registered data files located in the datastore to which the Trifacta node is connected.

> **NOTE:** The Wrangler Enterprise desktop application is a 64-bit Microsoft Windows application. It requires a 64-bit version of Windows to execute. The application also supports Single Sign On (SSO), if it is enabled.

Download

To begin, you must download the following Windows MSI file (`TrifactaEnterpriseSetup.msi`) from the location where your software was provided.

If you are planning to automate installation to desktops in your environment, please also download `setTrifactaServer.ps1`.

Setup

Before you begin, you should perform any necessary configuration of the Trifacta node before deploying the instances of the application. See *Configure for Trifacta Enterprise Application*.

Install for Windows

Steps:

1. On your Windows desktop, double-click the MSI file.
2. Follow the on-screen instructions to install the software.

Windows Command Line Installation and Configuration

As an alternative, you can perform installation and initial configuration from the command line. Download the MSI and the PS1 files to a local directory that is accessible.

> **NOTE:** For command line install, you must download from the `setTrifactaServer.ps1` from the download location.

Install software:

```
msiexec /i <path_to_TrifactaEnterpriseSetup.msi> /passive
```

Configure URL of Trifacta node:

```
setTrifactaServer.ps1 -trifactaServer <server_url> -installDir
<local_dir>
```

Parameter	Description
`trifactaServer`	(Required) URL of the server hosting the Trifacta platform. Format: `<http|https>://<host>:<port>`
`installDir`	(Optional) Specifies the installation directory in the local environment. If not specified, installation directory defaults to use the same path as the installer.
common installer parameters	This command supports the following Windows installer parameters: Verbose, Debug, ErrorAction, ErrorVariable, WarningAction, WarningVariable, OutBuffer, PipelineVariable, and OutVariable. For more information, see *about_CommonParameters* here: *http://go.microsoft.com/fwlink/?LinkID=113216*.

After this install is completed, desktop users should be able to use the application normally.

Launch the Application

Steps:

1. When installation is complete, double-click the application icon.
2. For the server, please enter the full URL including port number of the Trifacta instance to which you are connecting.

    1. By default, the server is available over port `3005`. For more information, please contact your IT administrator.
    2. If you connect to the Internet through a proxy, additional configuration is required. See *Configure Server Access through Proxy*.

    > **NOTE:** If you make a mistake in specifying the URL to the server, please uninstall and reinstall the MSI. This step clears the local application cache, and you can enter the appropriate path through the application. See *Uninstall* below.

3. When the proper URL and port number are provided, you may launch the application.
4. If your environment contains multiple server deployments, you can select the one to which to connect:

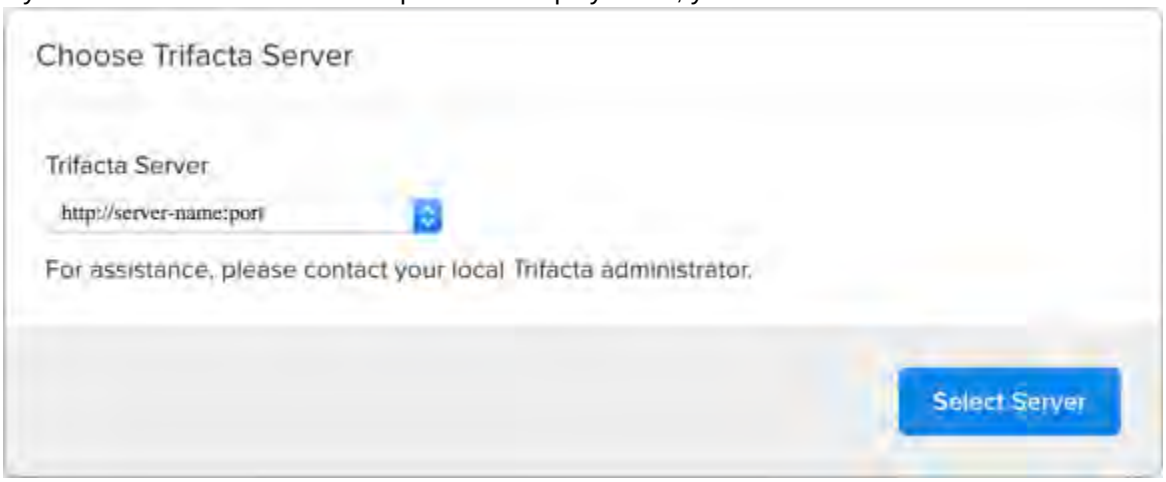


Figure: Choose Server

5. Login with your Trifacta account. See *Login*.

Documentation Note

Unless specifically noted, all features described for Trifacta $prodname or the Trifacta application apply to the Wrangler Enterprise desktop application.

Uninstall

To uninstall from your Windows machine, use the Add or Remove Programs control panel.

Troubleshooting

Cannot connect to server

If you are unable to connect to the server, please do the following:

1. Verify that you are connecting to the appropriate URL.

1. If you are connecting to the incorrect URL, please uninstall the application and re-install using the MSI file. See *Uninstall* above.
2. Verify if you need to connect to the server through a proxy server. If so, additional configuration is required. See *Configure Server Access through Proxy*.
3. Check your firewall settings.

"Does Not Support Your Browser" error

This error message indicates that you are trying to connect to an instance of the server that does not support the Wrangler Enterprise desktop application. Please verify that your connection URL is pointed to a supported instance of the server.

Start and Stop the Platform

Contents:

Tip: The Restart Trifacta button in the Admin Settings page is the preferred method for restarting the platform.

NOTE: The restart button is not available when high availability is enabled for the Trifacta® node.

See *Admin Settings Page*.

Start

NOTE: These operations must be executed under the root user.

Command:

```
service trifacta start
```

Verify operations

Steps:

1. Check logs for errors:

```
/opt/trifacta/logs/*.log
```

    1. You can also access logs through the Trifacta® application for each service. See *System Services and Logs*.
2. Login to the Trifacta application. If available, perform a simple transformation operation. See *Login*.
3. Run a simple job. See *Verify Operations*.

Restart

Command:

```
service trifacta restart
```

When the login page is available, the system has been restarted. See *Login*.

Tip: If you have made any configuration changes, you should verify operations. See *Verify Operations*.

Stop

Command:

```
service trifacta stop
```

Debugging

You can verify operations of WebHDFS. Command:

```
curl -i  "http://<hadoop_node>:<port_number>/webhdfs/v1/?
op=LISTSTATUS&user.name=trifacta"
```

Troubleshooting

Error - SequelizeConnectionRefusedError: connect ECONNREFUSED

If you have attempted to start the platform after an operating system reboot, you may receive the following error message, and the platform start fails to complete:

```
2016-10-04T14:03:17.883Z - error: [ENVIRONMENT] Environment Sanity Test
Failed
2016-10-04T14:03:17.883Z - error: [ENVIRONMENT] Exception Type: Error
2016-10-04T14:03:17.883Z - error: [ENVIRONMENT] Exception Message:
SequelizeConnectionRefusedError: connect ECONNREFUSED
```

Solution:

> **NOTE:** This solution applies to PostgreSQL 9.6 only. Please modify for your installed database version.

This error can occur when the operating system is restarted. Please execute the following commands to check the PostgreSQL configuration and restart the databases.

```
chkconfig postgresql-9.6 on
```

Then, restart the platform as normal.

```
service trifacta restart
```

Login

> **NOTE:** Administrators of the platform should change the default password for the admin account. See *Change Admin Password*.

To login to the Trifacta® application, navigate to the following in your browser:

`http://<host_name>:<port_number>`

where:

- `<host_name>` is the host of the Trifacta application.
- `<port_number>` is the port number to use. Default is `3005`.

If you do not have an account, click **Register**.

- If self-registration is enabled, you you may be able to immediately login after registering.
- If Kerberos or secure impersonation is enabled, an administrator must apply a Hadoop principal value to the account before you can login. Please contact your Trifacta administrator.
- System administrators can enable self-registration. See *Configure User Self-Registration*.

After you login, you are placed in the Flows page, where you can create and manage your datasets and flows. See *Flows Page*.

- If you are using S3 as your base storage layer, you or your Trifacta administrator must provide the AWS access key, secret, and storage bucket identifiers to connect to your storage. To do it yourself, click **Configure Storage Settings**. See *User Profile Page*.
- For a basic walkthrough of the Trifacta application, see *Workflow Basics*.

To logout:

From the Settings menu, select **Logout**.

Install Reference

These appendices provide additional information during installation of Trifacta® Wrangler Enterprise.

Topics:

- *Install SSL Certificate*
- *Change Listening Port*
- *Supported Deployment Scenarios for Cloudera*
- *Supported Deployment Scenarios for Hortonworks*
- *Supported Deployment Scenarios for AWS*
- *Supported Deployment Scenarios for Azure*
- *Uninstall*

Install SSL Certificate

Contents:

You may optionally configure an SSL certificate to secure connections to the web application of the Trifacta® platform.

Pre-requisites

1. A valid SSL certificate for the FQDN where the Trifacta application is hosted
2. Root access to the Trifacta server
3. Trifacta platform is up and running

Configure nginx

There are two separate Nginx services on the server: one service for internal application use, and one service that functions as a proxy between users and the Trifacta application. To install the SSL certificate, all configuration are applied to the proxy process only.

Steps:

1. Log into the Trifacta server as the **centos** user. Switch to the **root** user:

```
sudo su
```

2. Enable the proxy nginx service so that it starts on boot:

```
systemctl enable nginx
```

3. Create a folder for the private key and limit access to it:

```
sudo mkdir /etc/ssl/private/ && sudo chmod 700 /etc/ssl/private
```

4. Copy the following files to the server. If you copy and paste the content, please ensure that you do not miss characters or insert unwanted characters.

    1. The `.key` file should go into the `/etc/ssl/private/` directory.
    2. The .crt file and the CA bundle/intermediate certificate bundle should go into the `/etc/ssl/certs/` directory.

    > **NOTE:** The delivery name and format of these files varies by provider. Please verify with your provider's documentation if this is unclear.

    3. Your certificate and the intermediate/authority certificate must be combined into one file for nginx. Here is an example of how to combine them together:

```
cat example_com.crt bundle.crt >> ssl-bundle.crt
```

5. Update the permissions on these files. Modify the following filenames as necessary:

```
sudo chmod 600 /etc/ssl/certs/ssl-bundle.crt
sudo chmod 600 /etc/ssl/private/your-private-cert.key
```

6. Use the following commands to deploy the example SSL configuration file provided on the server:

> **NOTE:** Below, some values are too long for a single line. Single lines that overflow to additional lines are marked with a `\`. The backslash should not be included if the line is used as input.

```
cp /opt/trifacta/conf/ssl-nginx.conf.sample /etc/nginx/conf.d
/trifacta.conf && \
rm /etc/nginx/conf.d/default.conf
```

7. Edit the following file:

```
/etc/nginx/conf.d/trifacta.conf
```

8. Please modify the following key directives at least:

Directive	Description
server_name	FQDN of the host, which must match the SSL certificate's Common Name
ssl_certificate	Path to the file of the certificate bundle that you created on the server. This value may not require modification.
ssl_certificate_key	Path to the .key file on the server.

Example file:

```
server {
  listen          443;
  ssl             on;
  server_name     EXAMPLE.CUSTOMER.COM;
  # Don't limit the size of client uploads.
  client_max_body_size 0;
  access_log      /var/log/nginx/ssl-access.log;
  error_log       /var/log/nginx/ssl-error.log;
  ssl_certificate      /etc/ssl/certs/ssl-bundle.crt;
  ssl_certificate_key  /etc/ssl/certs/EXAMPLE-NAME.key;
  ssl_protocols        SSLv3 TLSv1 TLSv1.1 TLSv1.2;
  ssl_ciphers RC4:HIGH:!aNULL:!MD5;
  ssl_prefer_server_ciphers on;
  keepalive_timeout    60;
  ssl_session_cache    shared:SSL:10m;
  ssl_session_timeout  10m;
  location / {
    proxy_pass  http://localhost:3005;
    proxy_next_upstream error timeout invalid_header http_500
http_502 http_503 http_504;
    proxy_set_header        Accept-Encoding   "";
    proxy_set_header        Host            $host;
    proxy_set_header        X-Real-IP       $remote_addr;
    proxy_set_header        X-Forwarded-For
$proxy_add_x_forwarded_for;
    proxy_set_header        X-Forwarded-Proto $scheme;
    add_header              Front-End-Https   on;
    proxy_http_version 1.1;
    proxy_set_header Upgrade $http_upgrade;
    proxy_set_header Connection "upgrade";
    proxy_set_header Host $host;
    proxy_redirect     off;
  }
  proxy_connect_timeout       6000;
```

```
  proxy_send_timeout          6000;
  proxy_read_timeout          6000;
  send_timeout                6000;
}
server {
  listen          80;
  return 301 https://$host$request_uri;
}
```

9. Save the file.
10. To apply the new configuration, start or restart the nginx service:

```
service nginx restart
```

Modify listening port for Trifacta platform

If you have changed the listening port as part of the above configuration change, then the `proxy.port` setting in Trifacta platform configuration must be updated. See *Change Listening Port*.

Add secure HTTP headers

If you have enabled SSL on the platform, you can optionally insert the following additional headers to all requests to the Trifacta node:

Header	Protocol	Required Parameters
X-XSS-Protection	HTTP and HTTPS	`proxy.securityHeaders.enabled=true`
X-Frame-Options	HTTP and HTTPS	`proxy.securityHeaders.enabled=true`
Strict-Transport-Security	HTTPS	`proxy.securityHeaders.enabled=true` and `proxy.securityHeaders.httpsHeaders=true`

NOTE: SSL must be enabled to apply these security headers.

Steps:

To add these headers to all requests, please apply the following change:

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 following setting and change its value to `true`:

```
"proxy.securityHeaders.httpsHeaders": false,
```

3. Save your changes and restart the platform.

Enable secure cookies

If you have enabled SSL on the platform, you can optionally enable the use of secure cookies.

> **NOTE:** SSL must be enabled.

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 following setting and change its value to `true`:

```
"webapp.session.cookieSecureFlag": false,
```

3. Save your changes and restart the platform.

Troubleshooting

Problem - SELinux blocks proxy service from communicating with internal app service

If the Trifacta platform is installed on SELinux, the operating system blocks communications between the service that manages the proxy between users and the application and the service that manages internal application communications.

To determine if this problem is present, execute the following command:

```
sudo cat /var/log/audit/audit.log | grep nginx | grep denied
```

The problem is present if an error similar to the following is returned:

```
type=AVC msg=audit(1555533990.045:1826142): avc:  denied  { name_connect
} for  pid=25516 comm="nginx" dest=3005 scontext=system_u:system_r:
httpd_t:s0
```

For more information on this issue, see *https://www.nginx.com/blog/using-nginx-plus-with-selinux*.

Solution:

The solution is to enable the following network connection through the operating system:

```
sudo setsebool -P httpd_can_network_connect 1
```

Restart the platform.

Change Listening Port

If you need to change the listening port for the Trifacta® platform, please complete the following instructions.

Tip: This change most typically applies if you are enabling use of SSL. For more information, see *Install SSL Certificate*.

NOTE: By default, the platform listens on port `3005`. All client browsing devices must be configured to enable use of this port or any port number that you choose to use.

Steps:

1. Login to the Trifacta node as an admin.
2. Edit the following file:

```
/opt/trifacta/conf/nginx.conf
```

3. Edit the following setting:

```
server {
  listen 3005;
  ...
```

4. Save the file.
5. You can apply this change through the *Admin Settings Page* (recommended) or `trifacta-conf.json`. For more information, see *Platform Configuration Methods*.
6. Locate the following setting:

```
"proxy.port": 3005,
```

7. Set this value to the same value you applied in `nginx.conf`.
8. Save your changes and restart the platform.

Supported Deployment Scenarios for Cloudera

Contents:

Supported Cloudera Distributions

NOTE: By default, Cloudera may be installed with Java JDK 1.7 or earlier. If so, you must upgrade each node in the cluster to Java JDK 1.8. For more information, see *https://www.cloudera.com/documentation/enterprise/latest/topics/cdh_ig_jdk_installation.html*.

For this release, the Trifacta® platform supports the following Cloudera versions.

NOTE: Cloudera 6.0 and later requires use of native Hadoop libraries from the cluster. See *Configure for Spark*.

- Cloudera 6.2.x (Release 6.0.1 and later)
- Cloudera 6.1.x (Release 6.0.1 and later)
- Cloudera 6.0.x (recommended)

NOTE: Spark 2.4 is not supported on Cloudera 6.0. Please use Spark 2.2. See *Configure for Spark*.

- Cloudera 5.16.x (Release 6.0.1 and later)
- Cloudera 5.15.x (recommended)
- Cloudera 5.14.x

NOTE: Cloudera 5.13.x and Cloudera 5.12.x are no longer supported. For best results, please upgrade your Hadoop distribution.

Notes:

- **Update Date:** March 21, 2019
- The Trifacta platform supports all variants of patch or point releases (X.Y.* and X.Y.*.* releases) through the Hadoop vendor's backwards compatibility policy.
- For individual versions of Hadoop components (such as HDFS, Spark, and Hive), the Trifacta platform sup ports the component version that is bundled with the vendor's package for the supported Hadoop distribution.
- For more information on how to set up your Hadoop distribution, please consult the documentation provided with your distribution or contact your distribution vendor.

Supported Deployments

NOTE: Unless otherwise noted, all items listed below are supported across all Hadoop distribution versions listed above. Unlisted items are not supported. Please contact *Trifacta Support* or your sales representative for items not listed here.

Deployment System

Item	Description
Physical On Premise Machines	Supported.
VMWare / VXServer	Supported.
Amazon EC2	Supported.

Running Environment

Item	Description
Spark	Supported.
Trifacta Photon	Supported.

Platform Security

Item	Description
HDFS File Permissions	Supported.
HDFS Privileges	Supported through Sentry.
Hive Privileges	Supported through Sentry.
Kerberos-Enabled Hadoop Cluster	Supported. See *Configure for Kerberos Integration*.
Secure User Impersonation	Supported. See *Configure for Secure Impersonation*.

High Availability

Item	Description
Name Node, Resource Manager, HttpFS	Supported. See *Enable Integration with Cluster High Availability*.

Metadata Publishing

Item	Description
Cloudera Navigator	Not supported.
Hive Publishing	Supported. See *Configure for Hive*.
Redshift Publishing	Supported. See *Run Job Page*. See *Publishing Dialog*.

Supported File Formats

See *Supported File Formats*.

Connectivity

Hadoop Connectivity

The Trifacta platform supports connectivity for execution to the following Hadoop environments for this vendor's distributions. Connectivity exceptions are listed below:

Running Environment	HDFS Reader	HDFS Writer	Hive Reader w/ HiveServer2
Spark	Supported.	Supported.	Supported.

Profiling Environment	HDFS Reader	HDFS Writer
Profiling on Spark	Supported.	Supported.

External Connectivity

Storage Platform	HDFS Reader	HDFS Writer
S3	Supported.	Supported.

Storage Platform	Amazon S3 Reader	Amazon S3 Writer
Spark Profiling	Supported.	Supported.

Notes

- none.

Supported Deployment Scenarios for Hortonworks

Contents:

Supported Hortonworks Distributions

For the following release, the Trifacta® platform supports the following Hortonworks versions.

> **NOTE:** Hortonworks 3.0 and later requires use of native Hadoop libraries. See *Configure for Spark*.

- Hortonworks 3.1.x (Release 6.0.1 and later)

  > **NOTE:** Spark 2.4 is not supported on Hortonworks 3.1. Please use Spark 2.3. See *Configure for Spark*.

- Hortonworks 3.0.x

  > **NOTE:** Spark 2.4 is not supported on Hortonworks 3.0. Please use Spark 2.3. See *Configure for Spark*.

- Hortonworks 2.6.x
- Hortonworks 2.5.x

> **NOTE:** Hortonworks 2.4.x is no longer supported. For best results, please upgrade your Hadoop distribution.

Notes:

- **Update Date:** April 1, 2019
- The Trifacta platform supports all variants of patch or point releases (X.Y.* and X.Y.*.* releases) through the Hadoop vendor's backwards compatibility policy.
- For individual versions of Hadoop components (such as HDFS, Spark, and Hive), the Trifacta platform supports the component version that is bundled with the vendor's package for the supported Hadoop distribution.
- For more information on how to set up your Hadoop distribution, please consult the documentation provided with your distribution or contact your distribution vendor.

Supported Deployments

> **NOTE:** After the Trifacta software has been installed, additional configuration is required for integration with the Hortonworks Data Platform. See *Configure for Hortonworks*.

> **NOTE:** Unless otherwise noted, all items listed below are supported across all versions listed above. Unlisted items are not supported. Please contact *Trifacta Support* or your sales representative for items not listed here.

Deployment System

Item	Description
Physical On Premise Machines	Supported.
VMWare / VXServer	Supported.
Amazon EC2	Supported.

Running Environment

Item	Description
Spark	Supported.
Trifacta Photon	Supported.

Platform Security

Item	Description
HDFS File Permissions	Supported.
HDFS Privileges	Supported through Ranger.
Hive Privileges	Supported through Ranger.

Kerberos-Enabled Hadoop Cluster	Supported. See *Configure for Kerberos Integration*.
Secure User Impersonation	Supported. See *Configure for Secure Impersonation*.

High Availability

Item	Description
Name Node, Resource Manager, HttpFS	Supported. See *Enable Integration with Cluster High Availability*.

Metadata Publishing

Item	Description
Hive Publishing	Supported. See *Configure for Hive*.
Redshift Publishing	Supported. See *Run Job Page*. See *Publishing Dialog*.

File Formats

See *Supported File Formats*.

Connectivity

Hadoop Connectivity

The Trifacta platform supports connectivity for execution to the following Hadoop environments for this vendor's distributions.

Running Environment	HDFS Reader	HDFS Writer	Hive Reader w/ HiveServer2
Spark	Supported.	Supported.	Supported.

Profiling Environment	HDFS Reader	HDFS Writer
Profiling on Spark	Supported.	Supported.

External Connectivity

Storage Platform	HDFS Reader	HDFS Writer
S3	Supported.	Supported.

Storage Platform	Amazon S3 Reader	Amazon S3 Writer
Spark Profiling	Supported.	Not supported.

Notes

- None.

Uninstall

To remove Trifacta® Wrangler Enterprise, execute as root user one of the following commands on the Trifacta node.

> **NOTE:** All platform and cluster configuration files are preserved. User metadata is preserved in the Trifact a database.

CentOS/RHEL:

```
sudo rpm -e trifacta
```

Ubuntu:

```
sudo apt-get remove trifacta
```