TRIFACTA

Install Guide for AWS

Version: 6.0.2
Doc Build Date: 05/24/2019

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 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.
   a. **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*.
   a. 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.
   a. The required set of ports must be enabled for listening. See *System Ports*.
   b. 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.
   a. **Cluster sizing:** Before you begin, you should allocate sufficient resources for sizing the cluster. For guidance, please contact your Trifacta representative.
   b. See Deploy the Cluster below.
10. **Databases:**
    a. 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.
    b. For more information on the supported databases and versions, see *System Requirements*.
    c. For more information on database installation requirements, see *Install Databases*.
    d. 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 Server. 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 Server

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:
    a. See *Login*.
    b. 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.	{
	"Version": "2008-10-17",
	"Statement": [
	{
	"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.
	<pre>{
	"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"
	},</pre>

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.
   a. In the Search bar, enter the following:

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

   b. 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.
    a. **AWS EMR Cluster ID:** Paste the value for the EMR Cluster ID for the cluster to which the platform is connecting.
    b. **AWS Region:** Enter the region where your EMR cluster is located.
    c. **Resource Bucket:** Enter the name of the S3 bucket to use.
    d. **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 Server 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.

> **Tip:** It might be easiest to create your dataset in Microsoft Excel.

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*.
    a. Select the connection where the dataset is stored. For datasets stored on your local desktop, click **Upload**.
    b. Select the dataset.
    c. In the right panel, click the Add Dataset to a Flow checkbox. Enter a name for the new flow.
    d. Click **Import and Add to Flow**.
    e. **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.
    a. In the Transformer page, some steps are automatically added to the recipe for you. So, you can run the job immediately.
    b. You can add additional steps if desired. See *Transformer Page*.
8. Click **Run Job**.
    a. If options are presented, select the default running environment to generate the results.
    b. To generate results in other formats or output locations, click **Add Publishing Destination.** Configure the output formats and locations.
    c. To test dataset profiling, click the Profile Results checkbox. Note that profiling runs as a separate job and may take considerably longer.
    d. See *Run Job Page*.
    e. **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.
    a. **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.
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 the CSV and JSON links 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 Hadoop cluster instance of Spark.
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 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.
   a. Review the *System Requirements* and verify that all required components have been installed.
   b. 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 Photon in-browser engine. 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:** In some cases, access to the Hadoop cluster is required.

6. Verify access to the server where the Trifacta platform is to be installed.
7. **Hadoop:** Additional steps are required to integrate the Trifacta platform with Hadoop. 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 Trifact a. 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:

- *7. Install and configure Trifacta databases*
- *Configuration*

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.
   a. Review the *System Requirements* and verify that all required components have been installed.
   b. 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 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:** In some cases, access to the Hadoop cluster is required.

6. Verify access to the server where the Trifacta platform is to be installed.
7. **Hadoop:** Additional steps are required to integrate the Trifacta platform with Hadoop. 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 Trifact a. 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.
   a. Ubuntu 14.04:

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

   b. 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*
- *Install your license key*
- *Update your license key*
- *Changing the license key location*
- *Expired license*
- *Invalid license key file*

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 server 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 Server 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 Server:

```
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 Trifacta Server, please enter the full URL including port number of the Trifacta instance to which you are connecting.
    a. By default, the server is available over port `3005`. For more information, please contact your IT administrator.
    b. 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 Trifacta 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 Trifacta Server deployments, you can select the one to which to connect:

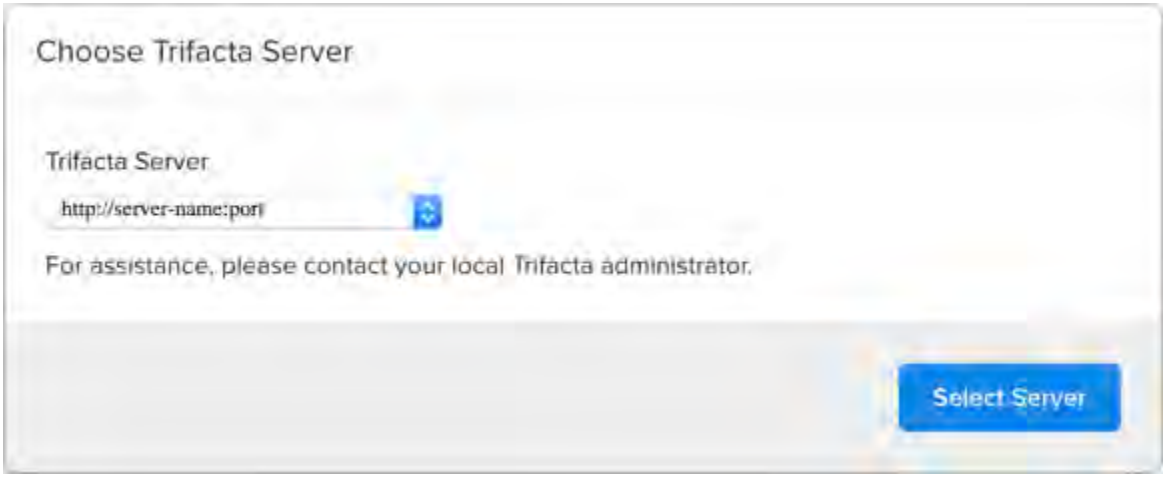

Figure: Choose Server

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

Documentation Note

Unless specifically noted, all features described for Trifacta Wrangler Enterprise 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 Trifacta server, please do the following:

1. Verify that you are connecting to the appropriate URL.
    a. 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 Trifacta server that does not support the Wrangler Enterprise desktop application. Please verify that your connection URL is pointed to a supported instance of the Trifacta 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
```

   a. 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=t
rifacta"
```

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 **Config ure 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.
   a. The `.key` file should go into the `/etc/ssl/private/` directory.
   b. 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.

   c. 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.

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 AWS

Contents:

AWS Deployment Scenarios

The following are the basic AWS deployment scenarios.

Trifacta platform installed on AWS:

Deployment Scenario	Trifacta node installation	Base Storage Layer	Storage - S3	Storage - Redshift	Cluster	Notes
Trifacta Data Preparation for Amazon Redshift and S3 AWS install through AWS Marketplace CloudFormation template	EC2	S3	read/write	read/write	None	Trifacta Data Preparation for Amazon Redshift and S3 does not support integration with any running environment clusters. All job execution occurs on the Trifacta Server. This scenario is suitable for smaller user groups and data volumes.
Trifacta Wrangler Enterprise AWS install through AWS Marketplace CloudFormation template - with integration to EMR cluster	EC2	S3	read/write	read/write	EMR	This deployment scenario integrates by default with an EMR cluster, which is created as part of the process. It does not support integration with a Hadoop cluster.
Trifacta Wrangler Enterprise AWS install through AWS Marketplace - without integration to EMR cluster	EC2	S3	read/write	read/write	EMR	This deployment scenario assumes that the platform is to be integrated at a later time with a pre-existing EMR cluster.
Trifacta Wrangler Enterprise AWS install with S3 read access	EC2	HDFS	read only	Not supported	EMR	When HDFS is the base storage layer, the only accessible AWS resources is read-only access to S3.
Trifacta Wrangler Enterprise AWS install with S3 read/write access	EC2	S3	read/write	read/write	EMR	

Trifacta platform installed on-premises and integrated with AWS resources:

Deployment Scenario	Trifacta node installation	Base Storage Layer	Storage - S3	Storage - Redshift	Cluster	Notes
Trifacta® Wrangler Enterprise on-premises install with S3 read access	On-premises	HDFS	read only	Not supported	Hadoop	When HDFS is the base storage layer, the only accessible AWS resources is read-only access to S3. For more information, see *Install Software*.
Trifacta Wrangler Enterprise on-premises install with S3 read/write access	On-premises	S3	read/write	read/write	Hadoop or EMR	For more information, see *Install Software*.
Microsoft Azure						Integration with AWS-based resources is not supported. See *Install from Azure Marketplace*.

Legend and Notes:

Column	Notes
Deployment Scenario	Description of the AWS-connected deployment
Trifacta node installation	Location where the Trifacta node is installed in this scenario. All AWS installations are installed on EC2 instances.

Base Storage Layer	When the Trifacta platform is first installed, the base storage layer must be set. **NOTE:** In Marketplace deployments, the base storage layer is set for you. After you have begun using the product, you cannot change the base storage layer. **NOTE:** Read/write access to AWS-based resources requires that S3 be set as the base storage layer.
Storage - S3	Trifacta Wrangler Enterprise supports read access to S3 when the base storage layer is set to HDFS. For read/write access to S3, the base storage layer must be set to S3.
Storage - Redshift	For access to Redshift, the base storage layer must be set to S3.
Cluster	List of cluster types that are supported for integration and job execution at scale. • The Trifacta platform can integrate with at most one cluster. It cannot integrate with two different clusters at the same time. • Access to an EMR cluster requires S3 to be the base storage layer. • Smaller jobs can be executed on the Trifacta Server running environment, which is hosted on the Trifacta node itself. • For more information, see *Running Environment Options*.
Notes	Any additional notes

AWS Installations

Trifacta Data Preparation for Amazon Redshift and S3 on AWS Marketplace (AMI)

Through the Amazon Marketplace, you can license and deploy an AMI of Trifacta Data Preparation for Amazon Redshift and S3, which does not require integration with a clustered running environment. All job execution happens within the AMI on the EC2 instance that you deploy. For more information, see the Trifacta Data Preparation for Amazon Redshift and S3 listing for AWS Marketplace.

- For install and configuration instructions, see *Install from AWS Marketplace*.

Trifacta Wrangler Enterprise on AWS Marketplace with EMR

You can deploy an AMI of the Trifacta platform onto an EC2 instance. For more information, see the Trifacta Wrangler Enterprise listing for AWS Marketplace.

You can deploy it in either of the following ways:

1. Auto-create a 3-node EMR cluster. For more information on installation, see *Install from AWS Marketplace with EMR*.
2. Integrate it later with your pre-existing EMR cluster.
    a. For more information on base AWS configuration, see *Configure for AWS*.
    b. For more information on configuring integration with EMR, see *Configure for EMR*.

Trifacta Wrangler Enterprise on EC2 Instance

When the Trifacta platform is installed on AWS, it is deployed on an EC2 instance. Through the EC2 console, there are a few key parameters that must be specified.

NOTE: After you have created the instance, you should retain the instanceId from the console, which must be applied to the configuration in the Trifacta platform.

For more information, see *Install*.

For more information on base AWS configuration, see *Configure for AWS*.

For more information on configuring EC2, see *Configure for EC2 Role-Based Authentication*.

AWS Integrations

The following table describes the different AWS components that can host or integrate with the Trifacta platform. Combinations of one or more of these items constitute one of the deployment scenarios listed in the following section.

AWS Service	Description	Base Storage Layer	Other Required AWS Services
EC2	Amazon Elastic Compute Cloud (EC2) can be used to host the Trifacta node in a scalable cloud-based environment. The following deployments are supported: • Trifacta Wrangler Enterprise with or without access to an EMR cluster • Trifacta Data Preparation for Amazon Redshift and S3 on an AMI	Base storage layer can be S3 or HDFS. If set to HDFS, only read access to S3 is permitted.	
S3	Amazon Simple Storage Service (S3) can be used for reading data sources, writing job results, and hosting the Trifacta databases.	Base storage layer can be S3 or HDFS. If set to HDFS, only read access to S3 is permitted.	
Redshift	Amazon Redshift provides a scalable data warehouse platform, designed for big data analytics applications. The Trifacta platform can be configured to read and write from Amazon Redshift database tables.	Base Storage Layer = S3	S3
AMI	Through the AWS Marketplace, you can license and install an Amazon Machine Image (AMI) instance of Trifacta Data Preparation for Amazon Redshift and S3. This product is intended for smaller user groups that do not need large-scale processing of Hadoop-based clusters.	Base Storage Layer = S3 **NOTE:** HDFS is not supported.	EC2 instance
EMR	Through the AWS Marketplace, you can license and install an AMI specifically configured to work with Amazon Elastic Map Reduce (EMR), a Hadoop-based data processing platform.	Base Storage Layer = S3	EC2 instance, AMI
Amazon RDS	Optionally, the Trifacta databases can be installed on Amazon RDS. For more information, see *Install Databases on Amazon RDS*.	Base Storage Layer = S3	

Uninstall

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

NOTE: All platform and Hadoop configuration files are preserved. User metadata is preserved in the Trifacta database.

CentOS/RHEL:

```
sudo rpm -e trifacta
```

Ubuntu:

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