TRIFACTA

Getting Started

Version: 6.0
Doc Build Date: 04/03/2019

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

Getting Started with Trifacta Wrangler Pro

Contents:

Welcome to Trifacta® Wrangler Pro!

1. Administrators should complete the first section to set up the product for use.
2. After set up is complete, individual users should complete the second section to get started using the product.

Step 1 - Start Process

You can begin using the product in either of the following ways, which are described in the following sections:

Register for Free Trial	Sign up for a free trial of the product, which provides limited access to the full product. See below.
Create Workspace	If you have licensed the full Trifacta Wrangler Pro product, you begin by submitting a request to *Trifacta Support*. See below.

Register for Free Trial

To begin the process, an administrator should complete the registration form available here: *https://www.trifacta.com/gated-form/free-trial-redshift/*.

Limitations:

- 100 Trifacta Compute Units
- 10 users

After you submit the registration form, an email is sent to your provided email address to confirm registration.

> **NOTE:** This process can take up to 24 hours to complete.

Key fields:

Field	Description
Email	This email address will receive a registration email, which contains a link that you must follow to complete registration.
Current AWS services utilized	Please add a comma-separated list of the AWS services that are currently used by your organization. Example: `AWS, S3, EC2, Redshift, VPC`
Primary AWS region	The region you select should be the same as your S3 and Redshift storage locations, if possible. **NOTE:** If you are integrating with Redshift, the region for your Redshift resources must be in the same location as your default S3 bucket, which is specified later.

Create Workspace

When you are ready to create your workspace, please contact *Trifacta Support* to create the workspace.

Key considerations:

- Number of workspace members
- Data volumes
- Primary AWS region

After the workspace has been created, an email is sent to your registered address with next steps.

Step 2 - Configure AWS for Trifacta Wrangler Pro

While your workspace is being created, you must make some decisions and, if applicable, perform some configuration within AWS to provide access to your datasets.

Verify Permissions

Members of your workspace must have access to the following:

- S3 bucket to be used
- EMR cluster to be used
- Resources required for AWS access mode (see below)

Please verify that all members that you wish to invite to the workspace have access to the above resources.

AWS Access Modes

You must determine how workspace members are able to access your resources. Trifacta Wrangler Pro supports the following access modes.

Method	Description

workspace	AWS access is managed through a single account. This configuration is set up by the workspace administrator for all users. **Tip:** This configuration is easiest to manage. After the administrator configures credentials, all invited members can immediately access the product. However, they all have the same permissions, which may be problematic for security reasons.
per-user	Each user must enter configuration settings in the Storage Config page after login. **NOTE:** If using an IAM role for per-user access, you must insert a trust relationship between AWS and the platform within the IAM role. For more information, see *Insert Trust Relationship in AWS IAM Role*. **Tip:** This method is more secure. However, each user must enter his or her own AWS credentials to access the product, which requires extra steps in the login process. These steps are described for non-admin users later.

Each of the above modes can be managed through one of the following credential methods:

- IAM role that provides access to their designated storage

  Tip: This method is recommended.

- Valid key-secret pair

After you make your decisions, your AWS administrator should provide any necessary credentials for access to these resources.

Encryption

Trifacta Wrangler Pro supports the following types of encryption.

- None
- SSE-S3
- SSE-KMS

NOTE: The method of encryption must be provided to the product to communicate with your S3 resources. If per-user authentication is in use individual users must configure the appropriate setting in their accounts.

Create S3 Policies

Access S3 bucket location

Trifacta Wrangler Pro must be able to access the location of the S3 bucket. Please apply the following policy definition to any IAM role used to access your S3 instance:

```
{
         "Sid":"statement1",
         "Effect":"Allow",
         "Action":[
            "s3:GetBucketLocation"
         ],
         "Resource":[
            "arn:aws:s3:::*"
         ]
}
```

For more information, see *https://docs.aws.amazon.com/AmazonS3/latest/dev/using-with-s3-actions.html*.

Access Trifacta Assets

To access S3 assets that are managed by Trifacta, you must apply the following policy definition to any IAM role that is used to access Trifacta Wrangler Pro:

```
{
            "Sid": "VisualEditor1",
            "Effect": "Allow",
            "Action": "s3:*",
            "Resource": [
                "arn:aws:s3:::trifacta-public-datasets/*",
                "arn:aws:s3:::trifacta-public-datasets"
            ]
        }
```

For more information on creating policies, *https://console.aws.amazon.com/iam/home#/policies*.

Whitelist the IP address of the Trifacta Service

If you are enabling any relational source, including Redshift, you must whitelist the IP address of the Trifacta service in the relevant security groups. The IP address of the Trifacta service is the following:

```
35.199.3.51/32
```

For Redshift:

For Redshift, there are two ways to whitelist the IP depending on if you are using EC2-VPC or EC2-Classic (not common).

- **EC2-VPC (Security group):** Add the IP address to the inbound rule for the security group associated with the cluster. For more information, see *https://docs.aws.amazon.com/redshift/latest/gsg/rs-gsg-authorize-cluster-access.html#rs-gsg-how-to-authorize*.
- **EC2-Classic:** Add the IP address to the inbound rule for the security group associated with the EC2 instance. For more information, see *https://docs.aws.amazon.com/redshift/latest/gsg/rs-gsg-authorize-cluster-access.html#rs-gsg-how-to-authorize*.

For details on this process with RDS in general, see *https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.RDSSecurityGroups.html*

For more information, please contact *Trifacta Support*.

Step 3 - Admin Login

Steps:

1. Login to your workspace.
2. You may receive an error message similar to the following:

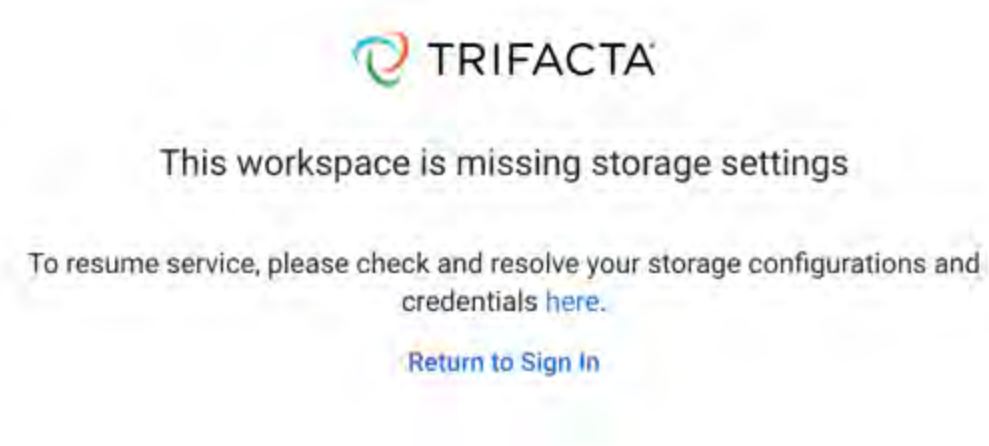

Figure: Missing Storage Settings

3. Click the Here link.
4. Specify the settings below.

Step 4 - AWS Config

In the AWS Credentials and Storage Settings page, workspace administrators can define the credentials mode for the workspace and apply settings for the selected mode, including selecting the credential provider.

Modes:

Mode	Description
Workspace	In Workspace mode, the workspace administrator applies a single set of AWS credentials for all users in the workspace. These credentials are used by each member of the workspace to authenticate with AWS and to gain access to S3 buckets. **Tip:** This mode requires more up-front setup but is easy to manage. However, all members of the workspace have the same set of access controls.
Per User	In Per User mode, individual members of the workspace must apply their AWS credentials to their accounts. **Tip:** This mode is easy to set up but turns responsibility for access controls over to the individual members. If members encounter problems gaining access to S3 assets, the workspace administrator may not be able to troubleshoot them.

Credential Provider:

Whether the credentials are provided by the workspace admin or by individual users, the following providers can be used to manage authentication with AWS.

Credential Provider	Description
IAM Role	Trifacta Wrangler Pro can use any IAM role that has been defined for workspace members to access S3 data sources. **Tip:** This credential provider method is recommended.
AWS Key and Secret	You can apply key and secret combinations to gate access to S3 data sources. These combinations can be applied in workspace mode or in per user mode by individual members.

Workspace Mode

In Workspace mode, you must select the credential provider and then specify the relevant settings.

IAM Role Settings

Pre-requisites:

- The IAM role must include a trust relationship for the Trifacta platform. For more information, see *Insert Trust Relationship in AWS IAM Role*.
- If you want workspace members to be able to use the on-boarding tour, they must have access to the Trifacta assets required for the tour.

Apply the following settings to define the IAM role and related settings.

Setting	Description
Account ID	This value is pre-populated when the workspace is created. **NOTE:** Do not modify.
External ID	This value is pre-populated when the workspace is created. **NOTE:** Do not modify.
IAM Role Arn	Insert the Arn value for the IAM role for workspace members to use.

AWS Key and Secret Settings

For key-secret authentication to AWS, please specify the following settings.

NOTE: The AWS key and secret must provide read/write access to the default S3 bucket at least.

The account must have the ListAllMyBuckets ACL among its permissions.

Setting	Description

AWS Access Key	The AWS access key to use.
AWS Secret Key	The AWS secret associated with the access key.

Per-User Mode

For per-user mode, the workspace administrator only must specify the encryption settings. See below.

Common Settings

S3 Buckets

For key-secret authentication to AWS, please specify the following settings.

Setting	Description
Default S3 bucket	Specify the name of the default S3 bucket.
Extra S3 buckets	You can specify any additional S3 buckets in a comma-separated list of names.

Server-side Encryption

Trifacta Wrangler Pro supports the use of server-side encryption when writing results.

> **NOTE:** When encryption is enabled, all buckets to which you are writing must share the same encryption policy.

Setting	Description
Encryption Type	Supported encryption types: • None • SSE-S3 • SSE-KMS
KMS Key ID	If SSE-KMS has been selected, you can paste the KMS Key ID value in this field.

Step 5 - Access Documentation

At this point, you can access online documentation for the product.

> **NOTE:** Content referenced in the PDF guide is not accessible through the PDF. You must login to the online documentation to access the referenced pages.

Steps:

1. From the left navigation bar, select **Help menu > Product Docs**.
2. You are automatically logged in.
3. PDF content is located in the following pages:
    a. *Getting Started with Trifacta Wrangler Pro*
    b. *AWS Config Page*
    c. *Create Redshift Connections*

Initial Configuration

Before you invite members to the workspace, you should review and modify the basic configuration for the workspace. See *Workspace Admin Page*.

Step 6 - Verify Operations

> **NOTE:** Workspace administrators should complete the following steps to verify that the product is operational end-to-end.

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.

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 **U pload**.
    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**.
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. To generate results in other formats or output locations, click **Add Publishing Destination.** Configure the output formats and locations.
    b. To test dataset profiling, click the Profile Results checkbox. Note that profiling runs as a separate job and may take considerably longer.
    c. See *Run Job Page*.
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. 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.

Step 7 - Invite Members

1. You can invite other people to join your workspace.
    a. When members initially join your workspace, they are assigned a non-admin role. Through the Workspace Members page, you can assign roles.
    b. For more information, see *Workspace Members Page*.
2. If you have enabled per-user authentication, credentials must be provided for each workspace member account:
    a. Administrators can apply per-user authentication for individual accounts. See *Workspace Members Page*.
    b. If individual members need to apply the credentials, the process is the same as for administrators.
        i. Please share Step 4 - AWS Config with them.
        ii. This page is also located online: *AWS Config Page*.

Getting Started for Workspace Members

This section walks through the process of getting started as a new member of a Trifacta Wrangler Pro workspace.

Steps:

1. You should have received an email like the following:

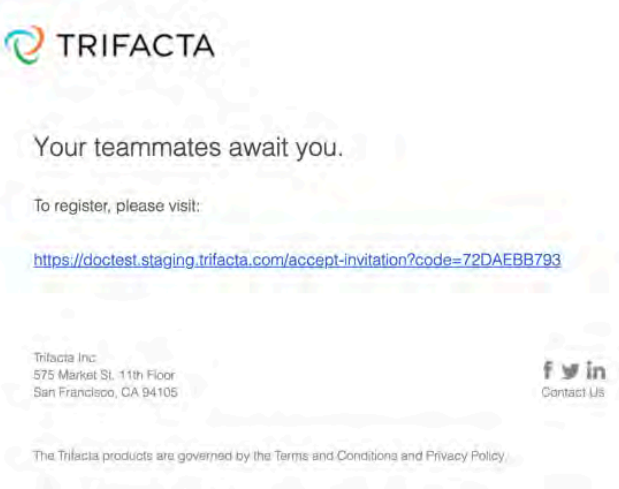

Figure: Welcome email

2. Click the link. If you see a Missing Storage Settings error message, then you must provide your individual user storage credentials and default bucket. To do so, click the Here link.
3. In your User Profile page, you may be required to enter your S3 credentials. For more information, see *Configure Your Access to S3*.
4. After the credentials have been entered, you can begin using the product.
5. **Access documentation:** To access the full customer documentation, from the left nav bar, select **Help menu > Product Docs**.

The following resources can assist workspace members in getting started with wrangling.

- If product walkthroughs have been enabled, each new member can step through an onboarding tour of the product after first login.
- For an overview of the product, see *Product Overview*.
- Check out the Trifacta Community: *https://community.trifacta.com*
    - Try the free Wrangler certification course. See *https://community.trifacta.com/s/certification*.
- For a basic summary of each step of the wrangling process, see *Workflow Basics*.

Create Redshift Connections

Contents:

This section provides information on how to enable Redshift connectivity and create one or more connections to Redshift sources.

- Amazon Redshift is a hosted data warehouse available through Amazon Web Services. It is frequently used for hosting of datasets used by downstream analytic tools such as Tableau and Qlik. For more information, see *https://aws.amazon.com/redshift/*.

Pre-requisites

Before you begin, please verify that your Trifacta® environment meets the following requirements:

> **NOTE:** If you are connecting to any relational source of data, such as Redshift or Oracle, you must add the Trifacta Service to your whitelist for those resources. For more information, see *Getting Started with Trifacta Wrangler Pro*.

1.
2. **Same region:** The Redshift cluster must be in the same region as the default S3 bucket.

Limitations

1. When publishing to Redshift through the Publishing dialog, output must be in Avro or JSON format. This limitation does not apply to direct writing to Redshift.
2. You can publish any specific job once to Redshift through the export window. See *Publishing Dialog*.
3. The Redshift cluster with which you are integrating must be hosted in a public subnet.

Create Connection

You can create Redshift connections through the following methods.

> **Tip:** SSL connections are recommended. Details are below.

Create through application

Any user can create a Redshift connection through the application.

Steps:

1. Login to the application.
2. In the menu, click **Settings menu > Connections**.
3. In the Create Connection page, click the Redshift connection card.
4. Specify the properties for your Redshift database connection. The following parameters are specific to Redshift connections:

Property	Description
IAM Role ARN for Redshift-S3 Connectivity	(Optional) You can specify an IAM role ARN that enables role-based connectivity between Redshift and the S3 bucket that is used as intermediate storage during Redshift bulk COPY/UNLOAD operations. Example: `arn:aws:iam::1234567890:role/MyRedshiftRole`

For more information, see *Create Connection Window*.

5. Click **Save**.

Enable SSL connections

To enable SSL connections to Redshift, you must enable them first on your Redshift cluster. For more information, see *https://docs.aws.amazon.com/redshift/latest/mgmt/connecting-ssl-support.html*.

In your connection to Redshift, please add the following string to your Connect String Options:

```
;ssl=true
```

Save your changes.

Testing

Import a dataset from Redshift. Add it to a flow, and specify a publishing action. Run a job.

> **NOTE:** When publishing to Redshift through the Publishing dialog, output must be in Avro or JSON format. This limitation does not apply to direct writing to Redshift.