TIBCO® Data Science - Team Studio

Installation and Administration

Version 7.1.0 | September 2023

Document Updated: December 2023

Contents

System Requirements

The server on which you install the TIBCO Data Science - Team Studio must meet these minimum requirements.

The following hardware, software, operating systems, third-party tools, and browsers have been tested with this version of TIBCO Data Science - Team Studio.

> **Note:** Some data sources are deprecated and are no longer supported. For a list of supported data sources, see Supported Legacy Database Platforms, Supported Legacy Hadoop Platforms, Supported TIBCO® DV Platforms, and Supported Apache Spark Cluster Platforms.

Hardware Requirements

Item	Requirements
Dedicated server	
Redundant power	
Hard disk space	Disk space storage requirements (approximately 500 GB total, using RAID 1 mirroring) • 240 GB storage file system for TIBCO Data Science - Team Studio installation.
Processor	• 16 core or higher
RAM	• Minimum: 32 GB. • Recommended: 32+ GB (more recommended for improved performance).

Required Third-Party Products

Items	Requirements
Docker	Version 20.10.0 or higher on a Unix/Linux Operating System

The following are the tested operating systems:

Operating System	Supported Versions
Redhat	Red Hat Enterprise Linux 8.6
CentOS	CentOS 7, CentOS Stream 9
Debain	Debian GNU/Linux 12
Ubuntu	Ubuntu 22.04.2 LTS
Oracle Linux	Oracle Linux Server 8.8

> **Note:** When you obtain third-party software or services, it is your responsibility to ensure you understand the license terms associated with such third-party software, or services and comply with such terms.
>
> The included Docker file and script, when used in accordance with the instructions here, downloads and install third-party components to create Docker images. We recommend that you review the script to identify the website(s) from which the components are downloaded to ensure that you understand which license terms apply to these components and what is required to comply with those terms, and to track their security status and determine if and when to update or replace them for security purposes. The `tds-platform-build` script if available in the following directory:
>
> ```
> installer/current/harness
> ```

Tested third-party and TIBCO tools

Tool	Tested version
TIBCO Spotfire Analyst, TIBCO Spotfire Desktop (using the Export to SBDF operator)	12.0.1 or later
TIBCO® Data Virtualization	8.7.0 or later
PMML	1.3

Supported browsers

Browser	Version
Google® Chrome	105.0.5195.127 or later

Supported Legacy Database Platforms

You can use any of the following legacy Database platforms in TIBCO Data Science - Team Studio version 7.1.0.

Platform	Version	Data source	Analytic data source	Sandbox
Amazon RedShift	JDBC 4.2 compatible driver	Yes		
Apache Impala	2.2	Yes		
Azure SQL Data Warehouse		Yes		
Google BigQuery	BigQuery JDBC 4.2	Yes		
Greenplum	5	Yes	Yes	Yes

Platform	Version	Data source	Analytic data source	Sandbox
database				
Hive JDBC	13	Yes		
MS SQL Server	Recommended: 12.0, 13.0 (2014, 2016) or higher Deprecated in 6.5.0: 11.0 (2012)	Yes		
Oracle database	10g, 11g,18c, Exadata	Yes	Yes	
Pivotal HAWQ	Deprecated in 6.5.0: 1.2	Yes	Yes	Yes
PostgreSQL	9.4 or higher. Deprecated in 6.5.0: • 8.4 • 9.3	Yes	Yes	Yes
SAP Hana	1.0 SPS 11 or higher	Yes		
Vertica	7.1 or higher	Yes		

Supported Legacy Hadoop Platforms

You can use the following legacy Hadoop platforms in TIBCO Data Science - Team Studio version 7.1.0.

Platform	Version
Amazon EMR	5.35.x

Supported TIBCO Data Virtualization Platforms

You can use TIBCO® Data Virtualization version 8.7.0 and above in the TIBCO Data Science - Team Studio version 7.1.0.

You can read all the data sources available in the TIBCO® DV and use them in TIBCO Data Science - Team Studio. For a list of supported data sources, see the *Supported Data Sources* section in the TIBCO® Data Virtualization Installation Guide.

However, you can write the tables to the selected data sources. For a list of supported data sources, see the *TDV DDL Feature Support* section in the TIBCO® Data Virtualization Installation Guide.

Supported Spark Cluster Platforms

You can use any of the following Apache Spark cluster platforms in TIBCO Data Science - Team Studio version 7.1.0.

Platform	Version
Amazon EMR	6.x
Cloudera	CDP 7
Apache Spark Standalone	3.3.x

System Administration

As a system administrator for TIBCO Data Science - Team Studio, you might need to know how to work with log files, how to free up memory, and how to configure custom properties for the application from the command line.

You can find articles about these tasks in this documentation.

TIBCO Data Science - Team Studio Licensing

TIBCO Data Science - Team Studio has a licensing model that allows granularity and access control for users of the application. This affects all levels of the business organization, from data scientists to the front-line business user.

This topic explains the roles in the application and the permissions granted with each role. For more information on your license terms, contact your TIBCO Data Science - Team Studio Account Manager.

To view license information from within TIBCO Data Science - Team Studio, in the top right corner, click your username and then click **About**. This displays information about features enabled on your TIBCO Data Science - Team Studio instance. For more information on the licensed roles and user counts, from the sidebar menu, click **Administration > License Information**, or see Administrator Options in TIBCO Data Science - Team Studio.

Installation

Before you install TIBCO Data Science - Team Studio, plan the installation, and then prepare your system for the installation. You can perform a quick installation, or specify options for the installation.

Prerequisites

The following checklists are intended to help you check your system's readiness for TIBCO Data Science - Team Studio.

Hadoop Connection Prerequisites

This checklist is provided to help ensure that all TIBCO Data Science - Team Studio components of a typical Hadoop-based installation are accounted for and completed.

The Hadoop connection configuration requires the HDFS host, HDFS port, Jobtracker host, and Jobtracker port. All Hadoop node hostnames must resolve to the proper computers from the TIBCO Data Science - Team Studio server. TIBCO Data Science - Team Studio needs access to a Hadoop administrator or anyone with access to the Hadoop configuration files (`*-site.xml`) if the inputs provided in the form below are not valid. TIBCO Data Science - Team Studio also might have to make changes to the host file of the TIBCO Data Science - Team Studio server if the Hadoop hostnames do not resolve.

The connection takes approximately two hours to configure and test if the Hadoop cluster is not configured for Kerberos. If it is, the user running TIBCO Data Science - Team Studio on the TIBCO Data Science - Team Studio server must have a key tab to authenticate in Kerberos. TIBCO Data Science - Team Studio requires that key tabs for the NameNode and Jobtracker are on the TIBCO Data Science - Team Studio server. If any of these three elements is missing or invalid, TIBCO Data Science - Team Studio requires that a Hadoop administrator is available to contact during installation. Configuring the initial connection to a cluster configured for Kerberos takes approximately 4 hours.

Hadoop Cluster

Question	Response	For Reference
Which version of Hadoop is installed?		
Is a Hadoop administrator be available during installation?		
Is the NameNode of the resource manager enabled for high availability?		

Question	Response	For Reference
Is the cluster configured for Kerberos?		
Is the cluster running MapReduce (MRv1) or YARN (MRv2)?		
Do the HDFS and JobTracker/resource manager hostnames resolve to the correct computers from the TIBCO Data Science - Team Studio server?		If they do not, configure the host file so that these Hadoop hosts resolve properly.

Hadoop Cluster without High Availability

Question	Response	For Reference
What are the HDFS host and port?		Can be found in `core-site.xml` as `fs.default.name: hdfs://HDFSHOST:HDFSPORT`

Hadoop Cluster with High Availability

Question	Response	For Reference
What is the name of the name service?		Can be found in `hdfs-site.xml` as `dfs.nameservices: hdfs://nameservice1`
What is the value for `dfs.ha.namenodes.<nameservice>`?		Can be found in `hdfs-site.xml` using the name of the name service.
What are the values for `dfs.namenode.rpc-address.<nameservice>.<namenode>`?		Can be found in `hdfs-site.xml` using the name of the name service, and each NameNode specified in the previous row.

Question	Response	For Reference
What is the value for `dfs.client.failover.proxy.provider.<namerservice>`?		Can be found in `hdfs-site.xml` using the name of the name service.

MapReduce (MRv1)

Question	Response	For Reference
What are the Job host and port?		Can be found in `mapred-site.xml` as `mapred.job.tracker:` `hdfs://JOBHOST:JOBPORT`

YARN (MRv2)

Question	Response	For Reference
What is the YARN resource manager's address?		Can be found in `yarn-site.xml` as `yarn.resourcemanager.address`

Kerberos (Ignore if Kerberos is not enabled)

Question	Response	For Reference
Is there a keytab that authenticates the TIBCO Data Science - Team Studio server?		
Are these required keytab files (merged or unmerged) on the TIBCO Data Science - Team Studio server?		

Spark Cluster Connection Prerequisites

The following Spark cluster configuration must be completed for the New Workflow.

The TIBCO® DV and the nodes of the Spark cluster must have a shared file system with a unique and consistent name. Only the system administrator can make these changes.

The shared file system can be HDFS, NFS drive, or Amazon S3 bucket. If you are using HDFS shared drive, you must have access to the shared file system for changing the configuration.

For example, if the name of a NFS shared file system is **tds**, then you can interactively mount as follows:

```
sudo mkdir -p /mnt/tds
sudo mount some.host.net:/path/to/volume /mnt/tds
sudo chown -R someOtherUser /mnt/tds
```

Here, the root is the only Linux user allowed to perform the mount (and unmount). Therefore, you must change the ownership of the mount point **someOtherUser** after the mount is complete. The **someOtherUser** is the Linux username (or user ID) that owns the Spark worker process (or the Hadoop YARN node manager process). If the owner is not changed, then the TIBCO Data Virtualization reader/writer components cannot access the mount point.

Prerequisites Server Configuration

Before installing the TIBCO Data Science - Team Studio, you must configure the server to prepare for the installation.

Verify that your environment meets the prerequisites specified in System Requirements and perform the following task on the Linux server.

Procedure
1. Verify kernel settings as the Chorus user.

   The maximum number of user processes must be 512K. The number of open files must be 512K.

   ```
   $ ulimit -a
   ```

   a. Modify the following lines in the file located at /etc/security/limits.conf and /etc/security/limits.d/90-nproc.conf.

```
*  soft nofile 524288
*  hard nofile 524288
*  soft nproc 524288
*  hard nproc 524288
```

b. If you are using `systemd` based Linux distribution (such as RHEL 8, Ubuntu >= 20.04), then add the following limits under the `Manager` section in the file located at `/etc/systemd/system.conf` and `/etc/systemd/user.conf`.

```
DefaultLimitNOFILE=524288:524288
DefaultLimitNPROC=524288:524288
```

2. Set the following parameters in `/etc/sysctl.conf`.

```
kernel.shmmax = 500000000
kernel.shmall = 4000000000
```

3. Restart the server.

4. Install the **jq** (JSON parser) software package.

5. Install the **Perl** (HTTP client) software package. For more information, refer to the Perl documentation.

Installing TIBCO Data Science - Team Studio

Before you begin

1. Install TIBCO Data Science - Team Studio on a Linux platform. See System Requirements for more information.

2. Docker Engine must be installed in your system. See System Requirements for more information.

3. You must have a Linux user (for example, `tds`), and the user must belong to your Docker Linux group. The system administrator can also create a Linux user. Use the following command to add the user to your Docker Linux group.

```
$ sudo usermod -aG docker $USER
```

For more information, see the Docker documentation.

4. Download the TIBCO Data Science - Team Studio installer package from TIBCO eDelivery to your local machine.

Procedure

1. Switch user to provide ownership for installing TIBCO Data Science - Team Studio (for example, `tds`).

   ```
   $ su – tds
   ```

   > ℹ **Note:** You can switch to any user to install TIBCO Data Science - Team Studio. The user must belong to your Docker Linux group.

2. Create an installer directory and a downloads directory.

   ```
   $ mkdir –p installers downloads
   ```

3. Configure your shell environment for the user.

   a. Update the `.bash_profile`.

   ```
   $ echo "export PATH=\$HOME/installers/current/harness:\$PATH" >> ~/.bash_profile
   ```

   b. Reload the `.bash_profile`.

   ```
   $ source ~/.bash_profile
   ```

4. Copy the TIBCO Data Science - Team Studio downloaded file (for example, *tds-installer-7.1.0.tgz*) to the downloads directory.

5. Extract the TIBCO Data Science - Team Studio installer file to the installers directory.

   ```
   $ tar xvfz downloads/tds-installer-x.x.x.tgz –C installers
   ```

> **ⓘ** **Note:** Here *x.x.x* is the version of the TIBCO Data Science - Team Studio installer package. So, if you are extracting the installer package version *7.0.0*, enter the following command:
>
> ```
> $ tar xvfz downloads/tds-installer-7.0.0.tgz -C installers
> ```
>
> You can download various versions of the TIBCO Data Science - Team Studio installer from TIBCO eDelivery and extract them to the installers directory. Now, you can browse the following directory structure for upgrades and downgrades of TIBCO Data Science - Team Studio.
>
> ```
> /$HOME
> ├── downloads
> │   └── tds-installer-7.0.0.tgz
> │   └── tds-installer-7.0.x.tgz
> │   └── tds-installer-7.x.x.tgz
> └── installers
>     ├── current -> tds-installer-7.0.0
>     ├── tds-installer-7.0.0
>     ├── tds-installer-7.0.x
>     ├── tds-installer-7.x.x
> ```

6. Link the current TIBCO Data Science - Team Studio installer.

```
$ ln -snf tds-installer-x.x.x installers/current
```

> **ⓘ** **Note:** Here *x.x.x* is the version of the TIBCO Data Science - Team Studio installer. So, if you are installing version *7.0.0*, enter the following command:
>
> ```
> $ ln -snf tds-installer-7.0.0 installers/current
> ```

7. To configure the TIBCO Data Science - Team Studio platform, perform the following steps:

   a. Use the following command line to start the configuration.

```
$ tds platform configure
```

b. Specify the target for installing the software. The default is `docker-local`.

c. Specify whether you want to install the R Connector.

d. Specify whether you want to install the Notebooks. If yes, then specify if you want any additional Python packages. For example, `pip:theano`, `conda:matplotlib=3.5.1`, and so on.

e. A private encryption key is generated for encrypting or decrypting the sensitive data. The encryption key is stored in the following path:

resources/tds-chorus/secret/data_encryption_password

> **ⓘ** **Note:** Use the following command line to copy the encryption key.
>
> ```
> cp resources/tds-chorus/secret/data_encryption_
> password /home/tds/encryption.key
> ```
>
> Place the generated encryption key in all the nodes of the Spark cluster in the directory `/etc/tds`. It must be accessible to the user submitting the Spark jobs.

f. A `.car` file for integrating the TIBCO Data Virtualization with TIBCO Data Science - Team Studio is generated. The file is stored in the following path:

build/tds-datavirt

For more information on integration, see Integrating TIBCO Data Virtualization with TIBCO Data Science - Team Studio.

g. Manage your Domain Name System (DNS).

   i. Enter your DNS name.

> **ⓘ** **Note:** Do not use an underscore (_) in the DNS hostname.

   ii. Specify whether you want to enable the Transport Layer Security (TLS). If

yes, enter the following details:

a. Path of the TLS digital certificate.

b. File path of the TLS private key.

All the answers to the questionnaire are saved to the default environment file.

```
./tds/installers/current/default.env
```

> **ⓘ Note:** If you do not want to go through the interactive questionnaire, enter the following command:
>
> ```
> $ tds platform build –env-file/path/to/my-custom.env
> ```

8. Build the Docker container images.

```
$ tds platform build
```

9. Now, deploy the build. This command starts the application.

```
$ tds platform start
```

> **ⓘ Note:** Before performing the next step, wait for 5 minutes for the containers to boot.

10. Use the following command line to create all the base operator.

```
$ tds platform seed
```

11. If you are using NFS as a shared drive, create a directory for the Jupyter Notebook configuration, and then add the following lines to the `tds-notebook` section of the compose file generator in the installer located at `$HOME/installers/current/compose/fragments`.

```
–target: /<directory_path>
```

```
source: /<directory_path>
type: bind
```

12. Restart the TIBCO Data Science - Team Studio application.

13. Log in as `siteadmin` using the default password, and then change your default password.

Installer Command Line Interface

Use the following command line to perform an action on the TIBCO Data Science - Team Studio platform:

Installer help and information

Option	Description
`$ tds help`	Print a help message that contains the available configuration options before you run the installer.
`$ tds platform version`	Get the version of TIBCO Data Science - Team Studio installed.

`$ tds platform switch`	To switch the version of TIBCO Data Science - Team Studio. For example,
	
	`$ tds platform switch 7.1.0`
	`$ tds platform switch latest`

`$ tds platform configure`	To configure TIBCO Data Science - Team Studio.
`$ tds platform`	To build all the container images.

Option	Description
`build`	
`$ tds platform clean`	To remove all the container images and anonymous volumes.
`$ tds platform start`	To start all the container services.
`$ tds platform stop`	To stop and remove all the container services.
`$ tds platform backup`	To create a backup of data and configuration files.
`$ tds platform restore`	To restore the backup of data and configuration files.
`$ tds platform seed`	Creates all the base operators.
`$ tds operators create`	Create all the analytical operators from the manifests at a given path.

Option	Description
`$ tds operators list`	Lists all the analytical operators.
`$ tds operators prune`	Remove or wipe-off all the operators.

Upgrading TIBCO Data Science - Team Studio

To upgrade the TIBCO Data Science - Team Studio to the current version, perform the following steps:

Procedure

1. Back up the previous version. For more information, see Backing up TIBCO® Data Science - Team Studio.

2. Install the current version of the TIBCO® Data Science - Team Studio. For more information, see Installation.

3. Restore the backup that you have taken from the previous version. For more information, see Restoring TIBCO® Data Science - Team Studio.

Integrating TIBCO Data Virtualization with TIBCO Data Science - Team Studio

To enable the integration of TIBCO® Data Virtualization with TIBCO Data Science - Team Studio, the system administrator needs to set up TIBCO® DV to include several TIBCO Data Science - Team Studio-specific data services.

The following steps must be performed for integrating TIBCO Data Virtualization with TIBCO Data Science - Team Studio:

Procedure

1. Installing TIBCO Data Virtualization. For more information about the installation, see TIBCO Data Virtualization.

   > ⓘ **Note:** Do not install TIBCO Data Virtualization and TIBCO Data Science - Team Studio on the same physical machine.

2. Importing the `.car` file to TIBCO Data Virtualization.

3. Running the initial setup.

4. Configuring TIBCO Data Virtualization Data Service.

If you are migrating from a previous version of TIBCO Data Virtualization that is integrated with TIBCO Data Science - Team Studio, then see Migrating TIBCO Data Virtualization Assets.

Importing the `.car` file to TIBCO Data Virtualization

The `.car` files must be imported to the TIBCO® Data Virtualization for integration with TIBCO Data Science - Team Studio. This action has to be performed by the system administrator.

Before you begin

Locate the `.car` file generated while installing the TIBCO Data Science - Team Studio. The file must be accessible to the computer where TIBCO Data Virtualization Studio is installed.

> ⓘ **Note:** The file is located at */home/ec2-user/installers/current/build/tds-datavirt*.

Procedure

1. Open TIBCO Data Virtualization Studio and log in as administrator (with the password created during installation), or log in as a user with administration rights.

2. In the navigation tree, right-click the **Shared** folder, and then select **Import**.
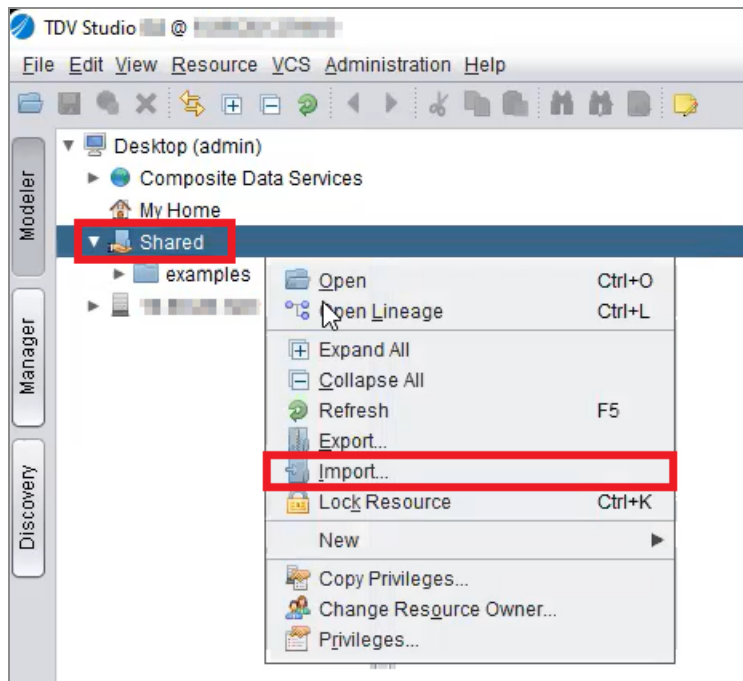
3. The Import into dialog appears. Perform the following steps:

   i. In the **File** column, click **Browse** and navigate to the location of the `.car` file.

   ii. In the **Password** column, enter your TIBCO Data Virtualization password.

   iii. In the **Include Resource Information**, select the **Custom Jars** and **Data Source Connections** checkboxes.

   iv. Click **Import**.

4. If the import is successful, the following status is displayed. Click **OK**.
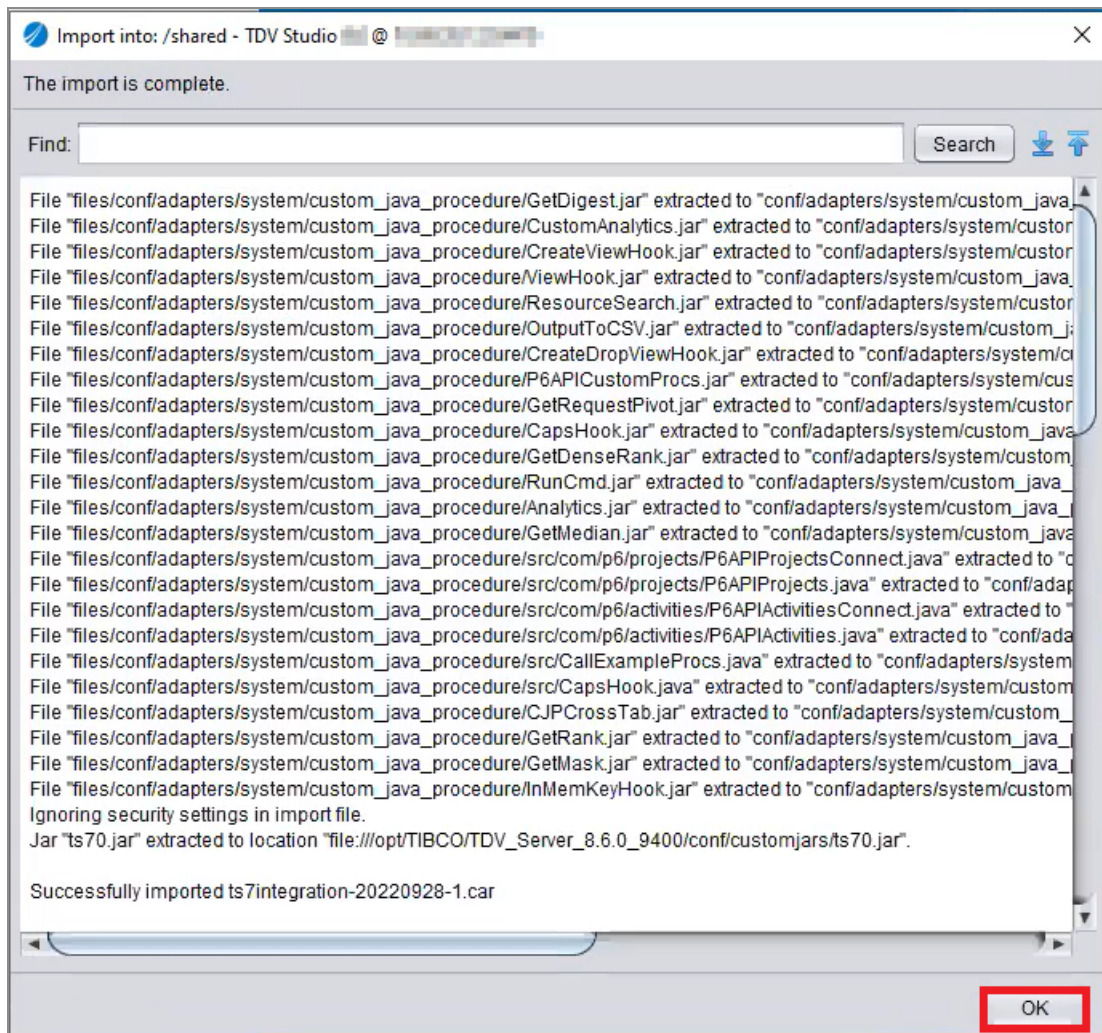
The imported *.car* file can be viewed under the **Shared** folder.
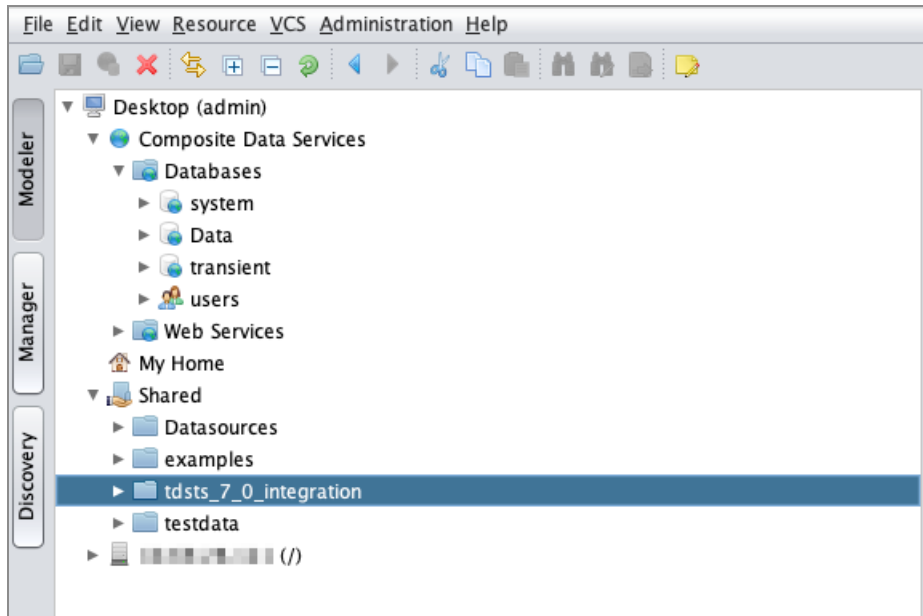
Running the Initial Setup

Once the .car files are imported, perform the initial setup of the imported files as follows:

Before you begin

Make sure that the .car files are imported successfully in TIBCO Data Virtualization Studio.

Procedure

1. Open TIBCO Data Virtualization Studio and log in as administrator (with the password created during installation), or log in as a user with administration rights.

2. In the navigation tree, navigate to **Shared** > **tdsts_7_0_integration** > **Functions** > **p_ init**. Right-click the **p_init** script and select **Open**.
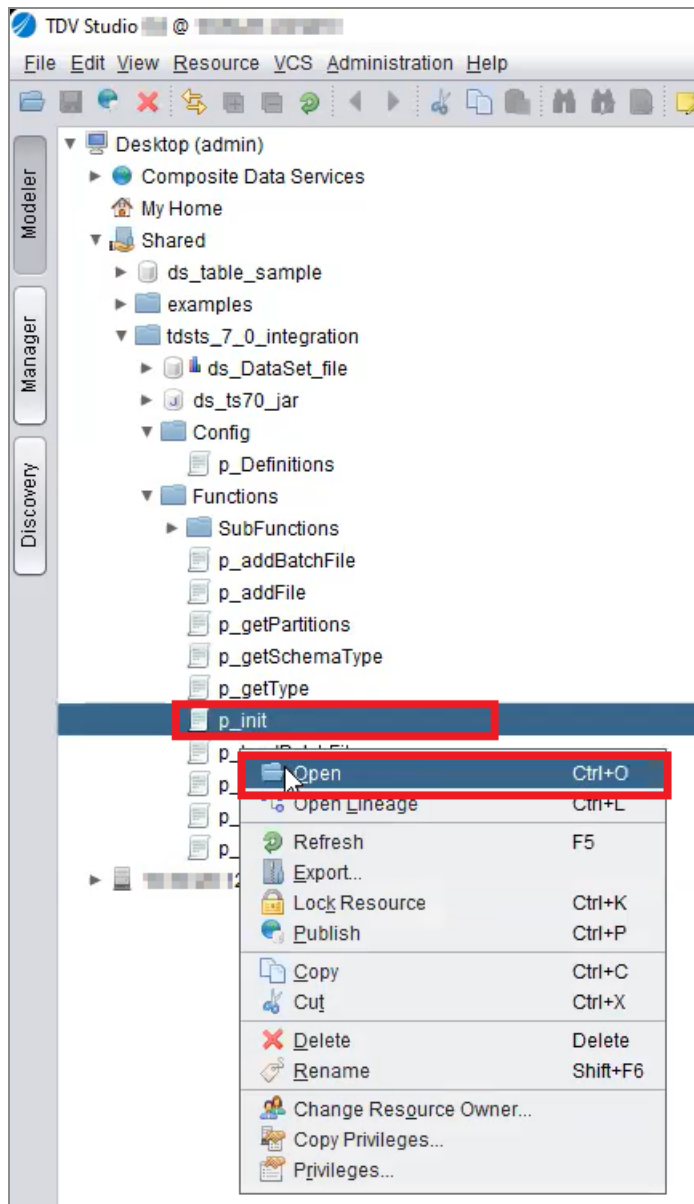
Configuring TIBCO Data Virtualization Data Service

Under Composite Data Services, the data service is created with the name **Data**. This data service is the virtual database that TIBCO Data Science - Team Studio connects to for all the access to TIBCO® Data Virtualization.

Create a schema under the **Data** virtual database. This schema specifies the location where data sets are published from the underlying Data Sources that are accessed by TIBCO Data

Science - Team Studio. In addition, as part of running workflows, TIBCO Data Science - Team Studio creates views in TIBCO® DV or tables in underlying data sources.

For each data source, the following two schemas are created:

1. **Datasets_<ds>**: Specifies the schema where data sets from the underlying data sources are published. Here **<ds>** identifies the data source (for example, `redshift` for Amazon Redshift). This schema is the input to TIBCO Data Science - Team Studio.

2. **Compute_<ds>**: Specifies the schema where TIBCO Data Science - Team Studio creates views or tables. This schema is for output from TIBCO Data Science - Team Studio for the associated data source. When the output type is TABLE, this creates a table in the data source connected to the schema.

To create a schema under the **Data** virtual database, perform the following steps:

Procedure

1. Open TIBCO Data Virtualization Studio and log in as administrator (with the password created during installation), or log in as a user with Administration rights.

2. In the navigation tree, navigate to **Composite Data Services** > **Databases** > **Data**. Right-click the **Data** and select **Open**.
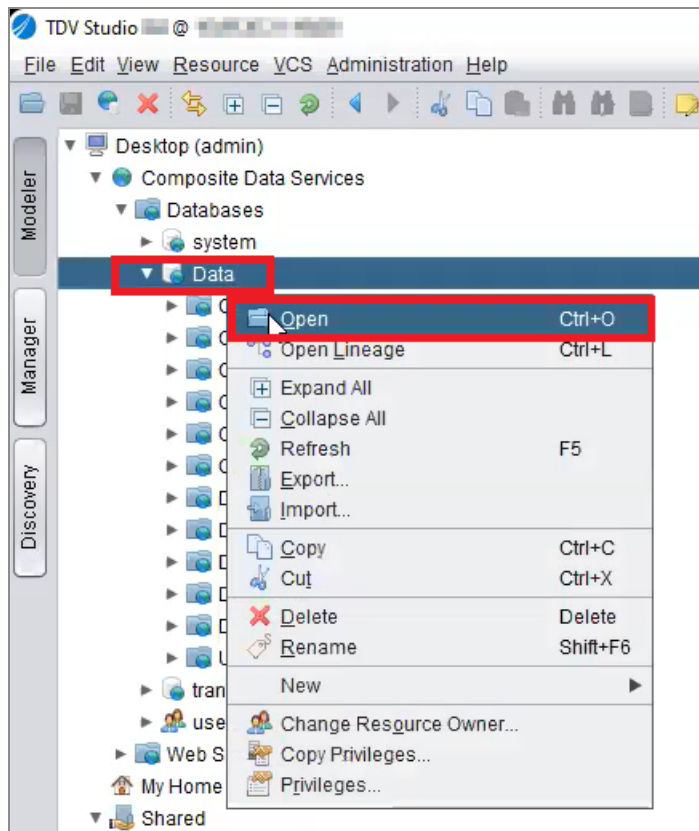
3. The Data page appears. Select the **Composite DDL** tab from the bottom of the page.

4. In the Composite Container DDL Path page, specify the Published Container Path and Physical Container Path.

   - **Published Container Path**: Specifies the schema of the published container.

   - **Physical Container Path**: Specifies the path of the physical container of the data source.

   - **Add New Container Mapping**: To add new rows of published container path and physical container path.
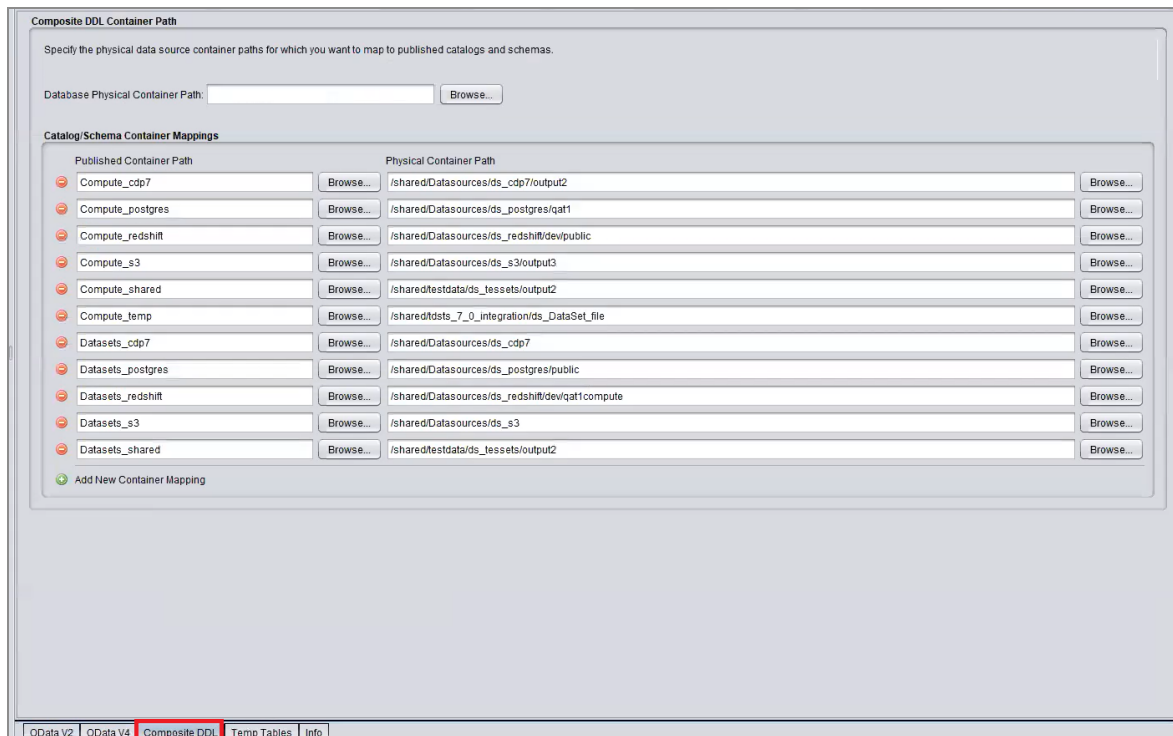
TIBCO Data Virtualization Data Sources

All the data sources that are available in TIBCO Data Virtualization can be viewed by navigating to **Shared** > **Datasources**. You can also add a data source. For more information on Data Sources, refer to the TIBCO DV documentation.

To add the available data source to the schema under **Data** virtual database, perform the following steps:

Procedure

1. Navigate to **Shared** > **Datasources** > **data_source**. For example, data_source is **ds_ postgress**.
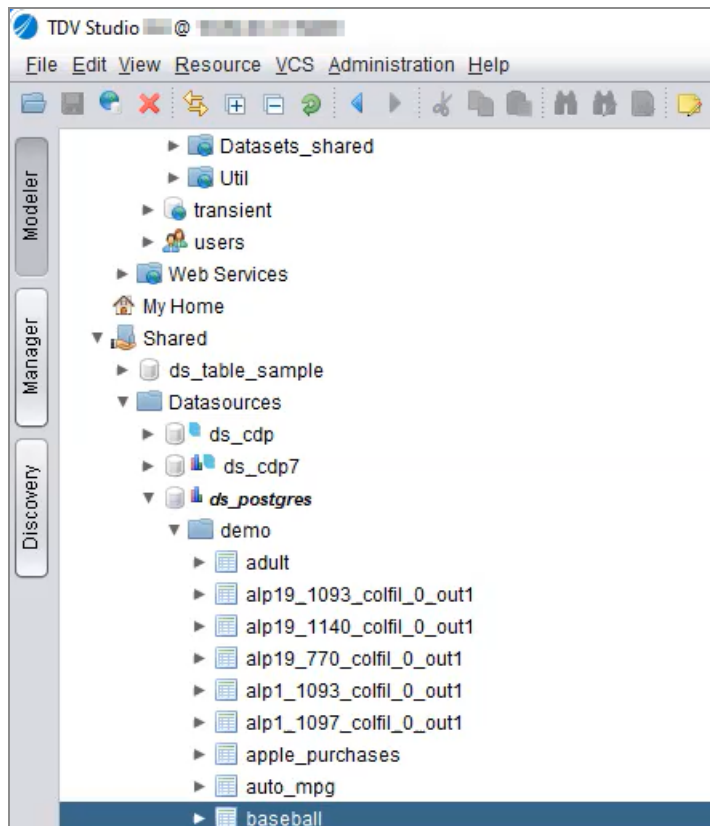
2. Select the data set from the selected data source. You can also select multiple data sets. Right-click on the selected data sets and select **Publish**.

3. A Publish dialog appears. Select the required schema, and then click **OK**.
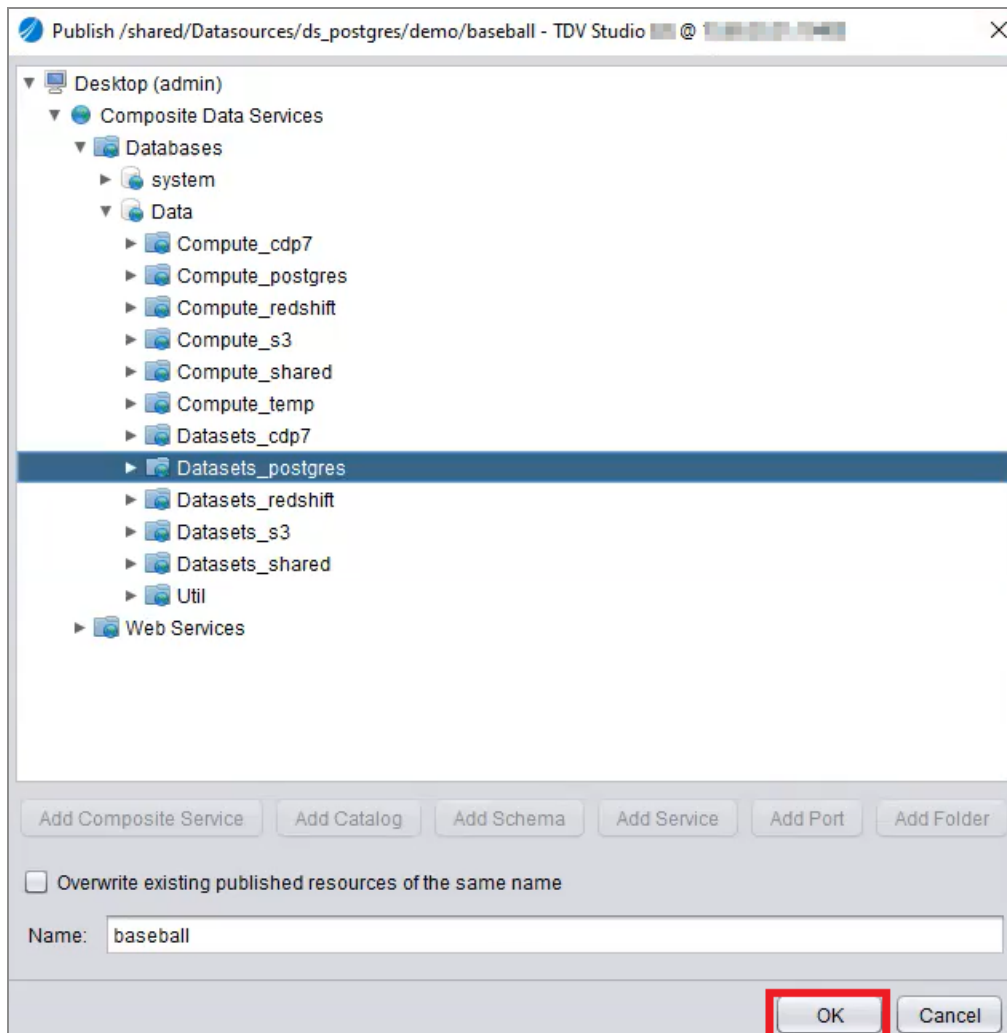
Migrating TIBCO Data Virtualization Assets

If the user has already integrated the TIBCO® Data Virtualization with TIBCO Data Science - Team Studio, then for migrating the TIBCO Data Virtualization assets that are connected with the previous integration of TIBCO Data Science - Team Studio, perform the following steps:

Procedure

1. Perform **Full Server Backup** on your TIBCO Data Virtualization machine. A .car file is generated. For more information on backup, see TIBCO Data Virtualization.

2. Migrate to the latest TIBCO Data Virtualization. For more information about the

installation, see TIBCO Data Virtualization.

3. Import the `.car` files that were generated in step 1 to TIBCO Data Virtualization. For importing the files, see Importing the `.car` file to TIBCO Data Virtualization. Your existing `.car` file is updated.

4. Import the `.car` file to TIBCO Data Virtualization that you have generated while installing the TIBCO Data Science - Team Studio. For importing the files, see Importing the `.car` file to TIBCO Data Virtualization.

5. Run the initial setup.

Accessing Data in TIBCO Data Virtualization from TIBCO Data Science - Team Studio

There are three methods of accessing data in TIBCO Data Virtualization from TIBCO Data Science - Team Studio.

1. TIBCO Data Virtualization managed Data Access

   This is the default data access method and supports a wide range of different types of data sources. The data from different sources, such as the Database, Remote File Systems, and Shared File Systems passes through TIBCO Data Virtualization before they are read from the Apache Spark cluster. Due to the lazy evaluation nature of Apache Spark, only the data that is necessary for processing is retrieved and sent to the Apache Spark cluster. This is made possible by using the push-down features in TIBCO Data Virtualization. After computation, the results are returned to the TIBCO Data Virtualization before writing back to the respective data sources.

   Examples of supported Database data sources are JDBC-compliant databases such as PostgreSQL, Oracle, and Redshift. Examples of the supported Remote File Systems and Shared File System data sources are HDFS, NFS drive, and Amazon S3. This is the slowest method of accessing the data. To use this method, you must remove the *tds.datavirt.sharedDataVolumes* parameter while configuring the Spark Cluster data source.

TIBCO Data Virtualization managed Data Access
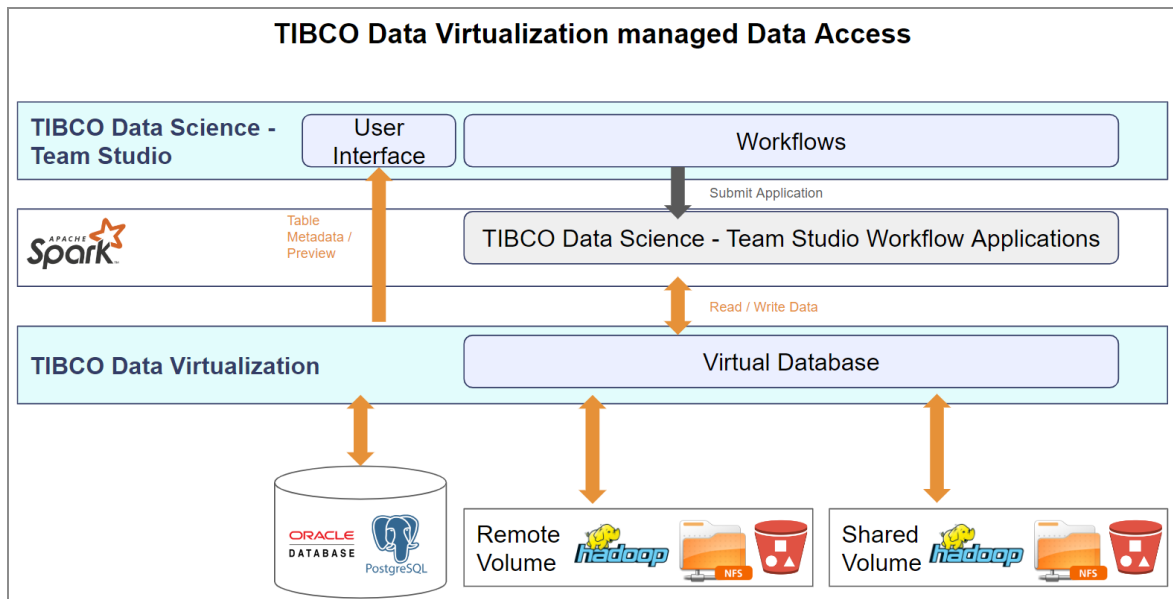
2. Shared Volume Optimized Data Access

This method optimizes access for files in the shared volume. The data from the Shared File System are read directly from the Apache Spark cluster and after computation, they are written back to the same Shared File System. The TIBCO Data Virtualization stores the metadata while writing the data to the Shared File System. Examples of the supported data sources are HDFS, NFS drive, and Amazon S3. This is the fastest method of accessing data because the file read and write is directly from the Apache Spark cluster. In other words, the data is not moved across the cluster. For information on configuring the Shared Volume for data accessing, see Shared Volume Data Access Configuration.
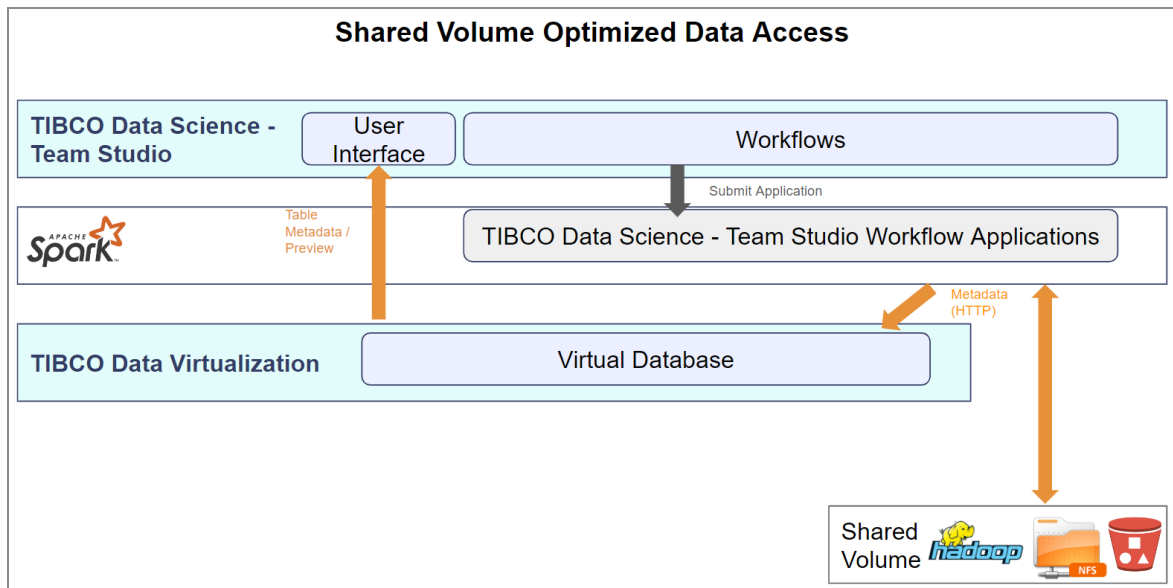
3. Shared Volume Optimized Data Access for Remote Sources

This method optimizes the writing of data to remote sources. In this method, the reading of data from remote sources is through the TIBCO Data Virtualization managed Data Access. After computation, the metadata is stored in TIBCO Data Virtualization and the results are temporarily written to the Shared Auxiliary Volume before writing to the Target Volume (Database or Remote File System). The Shared File System acts as the shared auxiliary volume. The results are moved to the remote sources or target volume using TIBCO Data Virtualization. For information on configuring the Shared Volume for data accessing, see Shared Volume Data Access Configuration.

> **ⓘ** **Note:** This is the recommended method for accessing data.
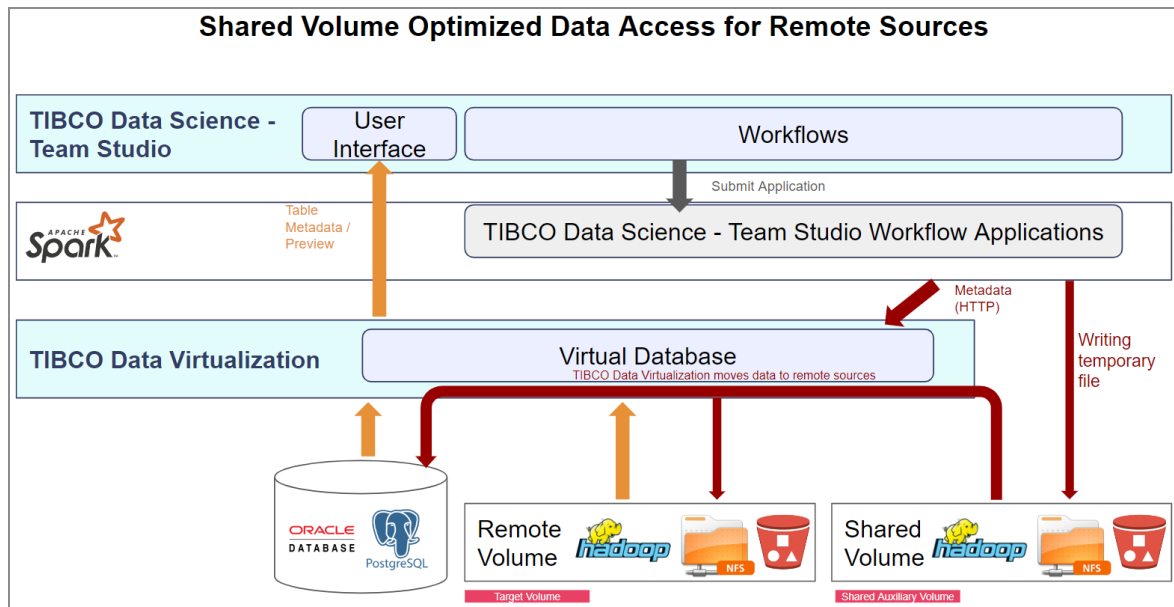
Shared Volume Data Access Configuration

You can configure the Shared Volume between the TIBCO Data Virtualization and the Apache Spark cluster for faster execution of results. It can be configured using the *tds.datavirt.sharedDataVolumes* parameter while configuring the Spark Cluster data source and the URL of the File Adapter in TIBCO Data Virtualization.

The following scenarios are applicable:

1. If the *tds.datavirt.sharedDataVolumes* parameter is not configured while configuring the Spark Cluster data source, it is known as TIBCO Data Virtualization managed data access.

2. If the *tds.datavirt.sharedDataVolumes* parameter is configured with the same value as the File Adapter in TIBCO Data Virtualization or if the file path is under the Shared Volume, it is known as Shared Volume Optimized data access.

3. If the *tds.datavirt.sharedDataVolumes* parameter is configured and the File path is not under the Shared Volume in TIBCO Data Virtualization, it is known as Shared Volume Optimized data access for Remote sources.

Recommendations

- To benefit from optimized file access, do not use an S3 data source as a shared data

volume. You can use any other data sources.

- The input and output schemas can be different.

- Assign the shared data volume to the *@default_schema* workflow variable. Make sure that all the operators use this schema as the output schema unless you need explicit writing to the remote sources.

Integrating TIBCO Data Science - Team Studio with TIBCO ModelOps

To enable the integration of TIBCO Data Science - Team Studio with TIBCO® ModelOps, the system administrator needs to configure TIBCO Data Science - Team Studio.

Before you begin

You must have installed TIBCO Data Science - Team Studio. See Installing TIBCO Data Science - Team Studio for more information.

Perform the following steps for integrating TIBCO Data Science - Team Studio with TIBCO® ModelOps:

> ℹ **Note:** Complete this integration for using the **Export Model to ModelOps** operator.

1. Use the following command line to configure the TIBCO® ModelOps.

   ```
   $ tds configs update modelops
   ```

2. Enter the TIBCO® ModelOps API URL.

3. Enter the TIBCO® ModelOps username.

4. Enter the TIBCO® ModelOps password.

TIBCO Data Science - Team Studio Default Ports

The following default ports are assigned to the TIBCO Data Science - Team Studio:

Ports	Function
31200 to 31211	Notebook Ports **Note:** These ports are used for PySpark initialization. Each connection requires three ports. Hence, by default, you can make four connections. If you want more, then add three ports for each new connection.
9090, 9100, 9101, 2580, 2581	Alpine Ports
80	If you are using HTTP insecure.
443	If you are using HTTP over TLS.

Python Packages in Notebooks Container of TIBCO Data Science - Team Studio

The following Python packages are installed with the Notebooks in TIBCO Data Science - Team Studio installation.

Package name	Version number
Jupyterhub	1.5.0
Notebook	6.4.8

Package name	Version number
Numpy	1.22.2
Pandas	1.3.5
scikit-learn	1.0.2
scipy	1.8.0
seaborn	0.11.2
tensorflow	2.7.0

TIBCO Data Science - Team Studio Configuration

Use the property files in this section to configure your deployment of TIBCO Data Science - Team Studio. These files are in the Chorus container. You can configure the workspace environment using the settings, files, and properties described in this section.

TIBCO Data Science - Team Studio Deploy Properties

Some of the TIBCO Data Science - Team Studio workflow engine configurations are defined in the file `deploy.properties` available in Alpine container.

The `deploy.properties` file is in the location `$ALPINE_HOME/ALPINE_DATA_ REPOSITORY/configuration`.

Tomcat settings

Use the key `alpine.cataline.opts` to set the JVM settings for the TIBCO Data Science - Team Studio services and other tomcat settings.

```
alpine.catalina.opts=-server -Xms4096M -Xmx8192M -XX:PermSize=1024M -
XX:MaxPermSize=1024M -DREST_ENABLED=true -
Djava.security.egd=file:/dev/./urandom
```

The Properties File

Much of the TIBCO Data Science - Team Studio configuration is defined by properties in the `chorus.properties` file, described below.

The file is located in the `<installation directory>/shared/` directory; for example, `/usr/local/chorus/shared/chorus.properties`.

In the same directory as `chorus.properties` is a file named `chorus.properties.example`. This file contains all of the properties that can be set in the `chorus.properties` file. Review this file to learn about the configuration options for TIBCO Data Science - Team Studio.

The properties in the actual configuration file, `chorus.properties`, might be a subset of the properties in `chorus.properties.example`. You can include any attribute you find in `chorus.properties.example` in your `chorus.properties` configuration file.

To see a detailed breakdown of options you can configure in `chorus.properties`, see TIBCO Data Science - Team Studio Configuration Properties.

You must restart TIBCO Data Science - Team Studio for changes you make in the `chorus.properties` file to take effect.

Configuring Indexing Frequency for Database Instances

The `reindex_datasets_interval_hours` configuration property in the `chorus.properties` file is used to set how frequently data sets are reindexed.

The following example sets data sets to be reindexed once a day (every 24 hours).

```
reindex_datasets_interval_hours= 24
```

TIBCO Data Science - Team Studio Configuration Properties

Use the file `chorus.properties` in Chorus container to configure the TIBCO Data Science - Team Studio.

The configuration file and its companion example file are available in the directory `<installation directory>`/shared/. The example file `chorus.properties.example`, contains examples of all of the properties that you can set in `chorus.properties`. You can use this example file to learn about the possible properties you can set in `chorus.properties`. You can include any attribute from the example file in the configuration file.

> ℹ **Note:** Restart TIBCO Data Science - Team Studio for the changes you make in `chorus.properties` to take effect.

chorus.properties options

Property	Default setting	Description
java_options	-Djava.library.path =$CHORUS_HOME/vendor /hadoop/lib/ -server - Xmx4096m XX:MaxPermSize=128	Command-line options to include on the Java command line when running the TIBCO Data Science - Team Studio application.
workflow.url	http://localhost:8070	The URL of the TIBCO Data Science - Team Studio server, which is installed with TIBCO Data Science - Team Studio. The URL can be an IP address or fully-qualified machine name. Whichever is used, it should be reachable from a browser. If you must change the port number after installing TIBCO Data Science - Team Studio, be sure to change the port number

Property	Default setting	Description
		in the TIBCO Data Science - Team Studio Tomcat server configuration file, $CHORUS_HOME/alpine/apache-tomcat-7.x.x/conf/server.xml. Look for the `<Connector>` element with attribute `protocol="HTTP/1.1"` under the `<Service name="Catalina">` element.
`smtp.address`	`localhost`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users. Sets the network address of the SMTP service.
`smtp.port`	`587`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users. Sets the port for the SMTP service.
`smtp.user_name`	`USER_NAME`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users.
`smtp.password`	`PASSWORD`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users.
`smtp.authentication`	`login`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users.

Property	Default setting	Description
`smtp.enable_starttls_auto`	`false`	Configures the SMTP connection that TIBCO Data Science - Team Studio uses to deliver email notifications to users.
`mail.enabled`	`FALSE`	If `true`, TIBCO Data Science - Team Studio delivers job completion and failure notifications to users by email.
`mail.from`	`FROMNAME <noreply@chorus.com>`	Sets the `from` header in the email message.
`mail.reply_to`	`REPLY NAME <noreply@chorus.com>`	Sets the `reply_to` header in the email message.
`sandbox_recommended_size_in_gb`	`5`	The sandbox-related setting. The default unit is in GB. **Note:** This value provides a visual indicator that indicates when a workspace's sandbox exceeds the recommended size.
`worker_threads`	`200`	Configures the thread pool size of web server and worker processes.
`webserver_threads`	`800`	The number of web server threads determines the maximum number of simultaneous web requests.
`database_threads`	`1200`	The number of worker threads determines the maximum number of asynchronous jobs, such as table copying or

Property	Default setting	Description
		importing, that can be run simultaneously.
		Each web or worker thread can use its own connection to the local PostgreSQL database. Therefore, the sum of `worker_threads` and `webserver_threads` must be less than the `max_connections` configured in `postgresql.conf`. The `max_connections` parameter can be based on the operating system's kernel shared memory size. For example, on OS X this parameter defaults to 20.
`session_timeout_minutes`	480	The default session timeout time. The number of minutes you can be inactive before you are logged out.
`clean_expired_api_tokens_interval_hours`	24	renamed in 6.2 from `clean_expired_sessions_interval_hours`.
`delete_unimported_csv_files_interval_hours`	1	
`delete_unimported_csv_files_after_hours`	12	
`instance_poll_interval_minutes`	5	
`reindex_search_data_interval_hours`	24	
`reindex_datasets_`	24	Sets the frequency for data set

Property	Default setting	Description
`interval_hours`		reindexing.
`reset_counter_cache_ interval_hours`	24	
`file_download.name_ prefix`	n/a	This optional string is prefixed on all generated file names. For example, if a user downloads a dataset, the name of the file downloaded is the specified prefix, followed by the dataset name and then the `.csv` extension. Only the first 20 characters of the prefix are used.
`file_sizes_mb.workfiles`	10	Maximum upload work file size.
`file_sizes_mb.csv_ imports`	100	Maximum size for imported files.
`file_sizes_mb.user_icon`	5	Maximum size for the user icon.
`file_sizes_mb.workspace_ icon`	5	Maximum size for the workspace icon.
`file_sizes_mb.attachment`	10	Maximum size for file attachments.
`logging.syslog.enabled`	false	If `true`, logs are written to syslog rather than to files.
`logging.loglevel`	info	The minimum severity of messages to log. Can be `debug`, `info`, `warn`, `error`, or `fatal`.
`oracle.enabled`	TRUE	Enables use of Oracle databases.

Property	Default setting	Description
gpfdist.ssl.enabled	false	To enable data movement between databases, gpfdist must be installed and running on the TIBCO Data Science - Team Studio host. Two gpfdist processes must be started with different ports pointing to the same directory. An SSL certificate must be installed on all segment servers.
gpfdist.url	sample-gpfdist-server	
gpfdist.write_port	8000	
gpfdist.read_port	8001	
gpfdist.data_dir	/tmp	
tableau.enabled	TRUE	If false, Tableau is disabled even if other Tableau parameters are specified.
tableau.url	>ip address>	The URL of the Tableau server. The URL can be an IP address or a fully-qualified computer name. Whichever is used, it should be reachable from a browser.
tableau.port	8000	The Tableau server port. **Note:** This port must be opened on your Tableau server.
tableau.sites	marketing,sales	The list of Tableau sites. TIBCO Data Science - Team Studio

Property	Default setting	Description
		supports this parameter starting in version 5.3. If this option is not present, TIBCO Data Science - Team Studio publishes to the default Tableau site.
`newrelic.enabled`	`false`	Enables New Relic application performance monitoring. See `http://newrelic.com` for more information.
`newrelic.license_key`	`NEWRELIC_LICENSE_KEY`	
`default_preview_row_ limit`	`500`	The maximum number of preview rows.
`execution_timeout_in_ minutes`	`60`	The workfile execution timeout in minutes.
`visualization.overlay_ string`	`n/a`	This optional string is displayed on all visualizations, both when displaying and when saving. Only the first 40 characters of the prefix are used.
`database_login_timeout`	`10`	Database connection timeout, in seconds. If you are using Google BigQuery as a data source and you are copying large amounts of data between databases, consider increasing this value so the operation does not fail unexpectedly.
`jdbc_schema_ blacklist.postgresql`	`[information_schema, pg_catalog]`	Specifies a list of PostgreSQL schemas that are excluded from

Property	Default setting	Description
		display, index, and search. (That is, they are effectively excluded from TIBCO Data Science - Team Studio).
jdbc_schema_blacklist.sqlserver	[db_accessadmin, db_backupoperator, db_datareader, db_datawriter, db_ddladmin, db_denydatareader, db_denydatawriter, db_owner, db_securityadmin, dbo, INFORMATION_SCHEMA, sys]	Specifies a list of SQL Server schemas that are excluded from display, index, and search. (That is, they are effectively excluded from TIBCO Data Science - Team Studio).
jdbc_schema_blacklist.teradata	[All, Crashdumps, DBC, dbcmngr, Default, EXTUSER, LockLogShredder, PUBLIC, SQLJ, SysAdmin, SYSBAR, SYSLIB, SYSSPATIAL, SystemFe, SYSUDTLIB, Sys_Calendar, TDPUSER, TDQCD, TDStats, tdwm, TD_SYSFNLIB, TD_SYSXML]	Specifies a list of Teradata schemas that are excluded from display, index, and search. (That is, they are effectively excluded from TIBCO Data Science - Team Studio).
job_timeout_in_minutes	60	When the value is set to 0, the feature is bypassed. Else, the worker process resets to **scheduled** after the set time has elapsed.

Property	Default setting	Description
		Note: 1. If the set value is greater than the scheduled interval (for example, a job runs every 1 hour but `job_timeout_in_minutes` is set to `85`), the job does not **unstuck**. 2. The `job_timeout_in_minutes` should be greater than the longest running scheduled job (for example, any enabled job that takes 1.5 hours to run, the `job_timeout_in_minutes` should be greater than `90`). 3. When a job status is reset to **scheduled**, the process spawned by the job is not ended. The job continues to run until it completes or an error occurs, even though the user interface shows that the job is not running.

Server Ports

The TIBCO Data Science - Team Studio application is a server running on a host machine at a specified port.

TIBCO Data Science - Team Studio Chorus server accesses several other servers running on the same or different hosts, including the embedded TIBCO Data Science - Team Studio

alpine server, Solr server, database servers, and other data sources, mail services, authentication servers (LDAP), and so on.

> ℹ **Note:** Each server on a host machine must be listening on a unique port.

Your installation of TIBCO Data Science - Team Studio Chorus server can include an embedded third-party TIBCO Data Science - Team Studio alpine server to manage workflows if the TIBCO Data Science - Team Studio license your organization purchased provides for it. If this is the case, then you must consider the following issues.

- The TIBCO Data Science - Team Studio Chorus and alpine servers must be configured to listen at different ports and they each must be configured to connect to each other at the correct port.

- This configuration is done automatically during installation, but if you reconfigure either server's port, you must change the port in the other's configuration to match the new port.

The following tasks show examples of changing the properties and configurations for the Chorus and embedded alpine services.

Configuring the HDFS Directory and Permissions for Results File Storage

This procedure describes how to create an HDFS directory for a user, set user permissions, and set permissions to temp directories. Perform the following steps when it is connected to Spark cluster with YARN.

> ℹ **Note:** For the `New Workflow` engine, perform the step 1 and step 2. For the `Legacy Workflow` engine, perform the step 1 to step 5.

Procedure

1. Create an HDFS Directory for user "`tds`". This directory is used to cache the uploaded .jar files such as `spark-assembly.jar`.

2. Provide the user with read, write, and execute permissions for the `/user/tds` directory.

3. The staging directory is typically set as `/user`. If it is not, create a directory using the modified `/<staging directory>/tds`.

4. To run Pig jobs, the TIBCO Data Science - Team Studio application attempts to create a folder `/user/<username>` as the Active Directory user. By default, the permissions are set to hdfs:supergroup:drwxr-xr-x, which prevents TIBCO Data Science - Team Studio from creating that folder. To grant write access to that folder to the Active Directory users who are running the TIBCO Data Science - Team Studio application, change permissions to drwxrwxr-x or drwxrwxrwx.

5. Set permissions to temporary directories.

   To run YARN, Pig, and similar jobs, each individual user might need to write temp files to the temporary directories. There are many Hadoop temp directories such as `hadoop.tmp.dir` and `pig.tmp.dir`, all of which are based on the `/tmp` directory by default. Therefore, the `/tmp` directory must be writable by everyone to enable them to run different jobs. Additionally, it must be executable by everyone to enable them to re curse the directory tree. Set the `/tmp` permissions using the following command:

   ```
   hadoop fs -chmod +wx /tmp
   ```

TIBCO Data Science - Team Studio Related HDFS Configuration

TIBCO Data Science - Team Studio uses several temp directories in HDFS. These directories and files are created with HDFS, YARN, and other users when it is connected to EMR 5.35.x.

The temp directories must be made accessible to user `Chorus` and other relevant users at the base level. Only individual directories for the corresponding user can be viewed by the specified user. Those directories are:

- Standard output for operators: @default_tmpdir/dsts_out/`<user_name>`/`<workflow_name>`/

- TIBCO Data Science - Team Studio temporary output: @default_tmpdir/dsts_runtime/`<user_name>`/`<workflow_name>`/

- TIBCO Data Science - Team Studio model location: @default_tmpdir/dsts_model/`<user_name>`/`<workflow_name>`/

Set or change the permissions and ownership as follows:

- The `/tmp` directory should be readable and writable.

- The `/tmp/hadoop-yarn` directory should be readable and writable for Spark jobs.

The upgrade options are as follows (choose one):

- Change /tmp/dsts_* directories with full permissions, so everyone can read/write/execute.

- Delete the /tmp/dsts_* directories and let the upgraded TIBCO Data Science - Team Studio application recreate them. If you are using LDAP, the recreated directories have the default structure /tmp/dsts_*/*<LDAP_username>*/workflowname/operator/, and permissions at this directory level can be limited to the `LDAP_username` as desired.

- By default, `@default_tmpdir` is set to `/tmp`. This can be modified for individual workflows using workflow variables. See "Workflow Variables" in *TIBCO® Data Science - Team Studio User Guide*, or for all newly created workflows using Workflow Editor Preferences.

TIBCO Data Science - Team Studio overwrites @default_tmpdir/dsts* files as users re-run workflows. TIBCO Data Science - Team Studio users can clear selected @default_tmpdir/dsts_out files using Clear Temporary Data, see "Clear Temporary Data" in *TIBCO® Data Science - Team Studio Installation and Administration*. Hadoop administrators can safely clear @default_tmpdir/dsts_runtime from HDFS, because this directory is used to store information for which TIBCO Data Science - Team Studio users have chosen the option **Store Results = False**.

> **Note:** Handle @default_tmpdir/dsts_model with caution, because TIBCO Data Science - Team Studio users might need to export models from this directory.

Deleting Temporary Files

You can clean up temporary files periodically using the following settings.

Through settings in the `alpine.config` file located in the `ALPINE_DATA_REPOSITORY/configuration` directory, administrators can set the period for storing temporary files and how often to run the clean-up task.

The following are the configuration settings for the `sys.properties`:

Setting	Default Value	Description
`temporary_file_livetime`	86400000	Specifies the length of time, in milliseconds, to store temporary files. The default: 86400000 ms (24 hours).
`temporary_file_scan_frequency`	86400000	Specifies the length of time, in milliseconds, of the interval between cleaning tasks. The default: 86400000 ms (24 hours).

Updated values take effect when TIBCO Data Science - Team Studio is restarted.

Security

The following topics describe how to complete various security-related tasks in TIBCO Data Science - Team Studio.

Enabling LDAP Authentication

By default, TIBCO Data Science - Team Studio manages users through the PostgreSQL database. However, it can be configured to authenticate against an external LDAP server.

The TIBCO Data Science - Team Studio collaboration framework uses the LDAPv3 server, including Active Directory support, to manage and authenticate users. For more information about the LDAP server, see http://www.ietf.org/rfc/rfc2251.txt.

LDAP provides the following benefits:

- Adding users to TIBCO Data Science - Team Studio: When a user is added, TIBCO Data Science - Team Studio maintains a read-only copy of common user information, such as the user's name and department.

- Authenticating users with LDAP.

Configuring LDAP

Follow these steps to configure LDAP authentication.

- Try connecting to your AD or LDAP installation with a separate LDAP exploration tool to ensure that all configuration properties are correct before you attempt to configure these in TIBCO Data Science - Team Studio.

- Install TIBCO Data Science - Team Studio.

Procedure

1. Login to Chorus container as a system administrator.

2. Edit the file `<installation directory>`/shared/ldap.properties file to configure LDAP.

   Change the default entries of the `ldap.properties` to match your LDAP installation.

   See the `ldap.properties.active_directory` or `ldap.properties.opensource_ldap` files for examples. The following shows an example.

   ```
   LDAP Settings for Active Directory
   # Set this property to true to enable LDAP authentication.
   Default is false.
   ldap.enable = false
   # Host and port for accessing LDAP server.
   ldap.host = localhost
   ldap.port = 389
   # Set this property to use Transport Level Security (TLS) for
   accessing LDAP server. Default is false.
   ldap.start_tls = false
   #LDAP root for search and query
   ldap.base = DC=www,DC=example,DC=com
   # username and password used for binding to LDAP server
   ldap.bind.username = uid=admin,ou=system
   ldap.bind.password = q2W#e4R%
   #-----  Uncomment following properties to enable group membership
   authentication ------#
   # Note that all three entries must either be commented or
   uncommented
   # List of LDAP group names that are used for verifying group
   membership.
   # NOTE: For release 5.3, only one group is supported.
   ```

```
#ldap.group.names = OtherGroup
# Search base for looking up members in the groups above.
#ldap.group.search_base = DC=www,DC=example,DC=com
#Group Filter for Active Directory. This works only for Active
Directory
#ldap.group.filter = (memberOf={0})

#-----------------------------------------------------------------
--------------------------#
# Search base for user authentication
ldap.user.search_base = OU=CorpUsers,DC=www,DC=example,DC=com
#Search filter for user authentication. This works only for
Active Directory
ldap.user.filter = (sAMAccountName={0})
# Mappings of Chorus user properties to LDAP user attributes.
ldap.attribute.uid = sAMAccountName
ldap.attribute.ou = department
ldap.attribute.gn = givenName
ldap.attribute.sn = sn
ldap.attribute.mail = mail
ldap.attribute.title = title If you want to add users from two
different groups (for example, Marketing and Sales) but
supports only one LDAP group, you have two options:
Add a new LDAP group (MarketingSales) to include users from
Marketing and Sales. Then bulk import using the rake command
below from the MarketingSales group.
Disable group search by commenting the lines below from
ldap.properties. Then, as an Admin, manually add each user to.
ldap.group.search_base
ldap.group.filter
ldap.group.namesLDAP Settings for Active Directory
# Set this property to true to enable LDAP authentication.
Default is false.
ldap.enable = false
# Host and port for accessing LDAP server.
ldap.host = localhost
ldap.port = 389
# Set this property to use Transport Level Security (TLS) for
accessing LDAP server. Default is false.
ldap.start_tls = false
#LDAP root for search and query
ldap.base = DC=www,DC=example,DC=com
# username and password used for binding to LDAP server
```

```
ldap.bind.username = uid=admin,ou=system
ldap.bind.password = q2W#e4R%
#-----  Uncomment following properties to enable group membership
authentication ------#
# Note that all three entries must either be commented or
uncommented
# List of LDAP group names that are used for verifying group
membership.
# NOTE: For release 5.3, only one group is supported.
#ldap.group.names = OtherGroup
# Search base for looking up members in the groups above.
#ldap.group.search_base = DC=www,DC=example,DC=com
#Group Filter for Active Directory. This works only for Active
Directory
#ldap.group.filter = (memberOf={0})

#-------------------------------------------------------------------
------------------------#
# Search base for user authentication
ldap.user.search_base = OU=CorpUsers,DC=www,DC=example,DC=com
#Search filter for user authentication. This works only for
Active Directory
ldap.user.filter = (sAMAccountName={0})
# Mappings of Chorus user properties to LDAP user attributes.
ldap.attribute.uid = sAMAccountName
ldap.attribute.ou = department
ldap.attribute.gn = givenName
ldap.attribute.sn = sn
ldap.attribute.mail = mail
ldap.attribute.title = title If you want to add users from two
different groups (for example, Marketing and Sales) but
supports only one LDAP group, you have two options:
Add a new LDAP group (MarketingSales) to include users from
Marketing and Sales. Then bulk import using the rake command
below from the MarketingSales group.
Disable group search by commenting the lines below from
ldap.properties. Then, as an Admin, manually add each user to.
ldap.group.search_base
ldap.group.filter
ldap.group.names
```

3. Restart TIBCO Data Science - Team Studio as follows after making changes to
    `ldap.properties`:

```
$ chorus_control.sh restart
```

4. Bulk imports LDAP users with a rake command. This rake task reads the LDAP configuration from the `ldap.properties` file and imports users from the LDAP group specified in the `ldap.group.names` property into the TIBCO Data Science - Team Studio database.

> **ℹ Note:** Release 5.3 supports just one group.

```
cd  $CHORUS_HOME
export RAILS_ENV=production
export PATH=$PATH:$CHORUS_HOME/current/bin
cd  $CHORUS_HOME/current
rake ldap:import_users
```

LDAP Configuration Properties

The following table lists configuration properties related to LDAP.

For more information, see Enabling LDAP Authentication.

LDAP Parameter	Description
ldap.enable	Boolean value to enable or disable LDAP (`false` by default).
ldap.host	LDAP server IP or host name.
ldap.port	LDAP server port.
ldap.base	LDAP base DN.
ldap.start_tls	Upgrades LDAP connection to a secure connection using TLS (false by default).
ldap.bind.username	Username for LDAP server.
ldap.bind.password	Password for LDAP server.

LDAP Parameter	Description
ldap.group.names	Name of group(s) to search for users. (Optional. LDAP users outside of specified groups cannot be added or authenticated as TIBCO Data Science - Team Studio users.)
ldap.group.search_base	Group DN to search for users. (Optional. LDAP users outside of specified groups cannot be added or authenticated as TIBCO Data Science - Team Studio users.)
ldap.group.filter	Group filter used to search for users. (Optional. LDAP users outside of specified groups cannot be added or authenticated as TIBCO Data Science - Team Studio users.)
ldap.user.search_base	Base DN to search for users when authenticating.
ldap.user.search_filter	Search filter for users when authenticating.
ldap.attribute.uid	LDAP username attribute (uid by default). Required. For Active Directory, this is often sAMAccountName. Another common value is cn.
ldap.attribute.ou	LDAP attribute name for Organizational Unit or Department (ou by default). Maps to the Department in TIBCO Data Science - Team Studio.
ldap.attribute.gn	LDAP attribute name for First name (gn or givenName by default). Maps to the First Name in TIBCO Data Science - Team Studio.
ldap.attribute.sn	LDAP attribute name for Last name. (sn by default). Maps to the Last Name in TIBCO Data Science - Team Studio.
ldap.attribute.mail	LDAP attribute name for e-mail address. (mail by default). Maps to the Email in TIBCO Data Science - Team Studio.
ldap.attribute.title	LDAP attribute name for User's title. (title by default). Maps to the Title in TIBCO Data Science - Team Studio.

Adding LDAP Users

The following procedures demonstrate how to add and remove LDAP users through TIBCO Data Science - Team Studio.

Procedure
1. As an admin user, select **People** from the navigation menu.

2. On the **People** page, click **Add Person**.

3. To search for an LDAP user, type the username, and then click **Check for Account**. The user's information is retrieved from the LDAP server and automatically filled in.

4. Click **Add Person**.

Removing LDAP Users

Users removed from the authorized group on the LDAP server are not automatically removed from TIBCO Data Science - Team Studio. They can no longer log into TIBCO Data Science - Team Studio, but the user account still exists in the application. To remove the account, manually delete it from TIBCO Data Science - Team Studio.

Procedure
1. Select **People** from the navigation menu.

2. Browse to find the account.

3. Select the person to delete, and then click **Delete**.

Troubleshooting LDAP Configuration

Here are solutions for some common issues that you might encounter while enabling LDAP authentication.

LDAP user does not appear in Chorus

Issue: After running rake `ldap:import_users`, you do not see all of the users from your group imported into TIBCO Data Science - Team Studio.

Potential cause: The TIBCO Data Science - Team Studio license allows fewer users than exist in the LDAP group.

Potential resolution: Identify the number of allowed users by checking the TIBCO Data Science - Team Studio license, and then limit the imported users to the number that the license allows. For more information, see the License Agreement for your TIBCO Data Science - Team Studio installation.

Potential cause: The Windows server running LDAP needs to communicate with the Unix servers, but some LDAP users are not listed on the Network Information System (NIS) domain.

Potential resolution: Add the users to the NIS global user list using the LDAP server tool.
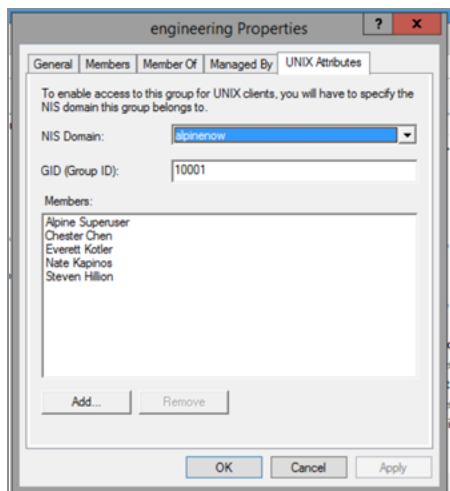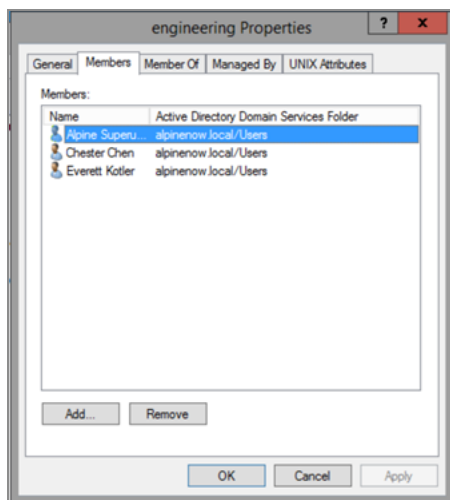
No entry found

Error message: No entry found for user ___ in the LDAP group TIBCO Data Science - Team Studio Users. Contact your system administrator.

Potential cause:

1. Create a group on the LDAP server.

2. Set `ldap.properties` to search that group.

3. Attempt to log in with a user who is not a member of the group.

Potential resolution: Verify that the user is a member of the desired group.

LDAP settings are mis-configured

Error message: LDAP settings are mis-configured. Please contact your System Administrator.

Potential cause: `ldap.properties` file contains parameters that are invalid, such as `ldap.bind.username` or `ldap.bind.password`.

Potential resolution: Validate parameters outside of TIBCO Data Science - Team Studio, and then modify as needed in `ldap.properties`.

LDAP Use Case Scenarios

The following LDAP use case scenarios describe steps for authenticating users who have or do not have group memberships, and how to import users from an LDAP group to TIBCO Data Science - Team Studio.

Scenario 1: LDAP Authentication with Group Membership

This workflow describes the steps for authenticating a user with a group membership.
In this scenario, the first user's membership in a specified group is verified. After successful membership verification, the user is authenticated against the LDAP server with a qualified distinguished name (DN) and a user-supplied password.

Procedure

1. The user provides a username and password to log in.

2. The login request is received by the TIBCO Data Science - Team Studio back end.

3. The TIBCO Data Science - Team Studio back end verifies that the username is for a registered/licensed user. Note that the user's password is not being authenticated at this time. The only thing being verified is that the user is a valid TIBCO Data Science - Team Studio user.

4. If the user is a valid TIBCO Data Science - Team Studio user, the TIBCO Data Science - Team Studio back end sends a query message to the LDAP server to verify the user's group membership. The query parameters are read from the `ldap.properties` file.

5. The query request returns a result that verifies whether the current user is a member of the specified group.

6. If the current user is a member of the specified group, the TIBCO Data Science - Team Studio back end sends an authentication request to the LDAP server. The request parameters consist of a DN and a user-supplied password. The DN is constructed using the parameters from the `ldap.properties` file.

7. The LDAP server authenticates the user and returns.

8. If the user is authenticated successfully, the TIBCO Data Science - Team Studio back end navigates the user to the dashboard page. If the user is not authenticated, an error message is displayed.

Scenario 2: LDAP Authentication without Group Membership

This workflow describes the steps for authenticating a user without group membership. In this scenario, the user is authenticated against the LDAP server with a user ID and a user-supplied password.

Procedure

1. User provides a username and password to log in.

2. Login request is received by the TIBCO Data Science - Team Studio back end.

3. The TIBCO Data Science - Team Studio back end verifies that the username is a registered/licensed user. Note that we are not authenticating the user's password at this time. We are only verifying that the user is a valid TIBCO Data Science - Team Studio user.

4. The TIBCO Data Science - Team Studio back end sends an authentication request to the LDAP server. The request parameters consist of a Distinguished Name (DN) and user supplied password. The DN is obtained by querying the LDAP server.

5. LDAP Server authenticates the user and returns.

6. If the user is authenticated successfully by the LDAP server, the Team Studio back end sends the user to the dashboard page. If the LDAP server fails to authenticate the user, an appropriate error message is displayed to the user.

Scenario 3: Import Users from an LDAP Group to TIBCO Data Science - Team Studio

After installing, customers must run a rake task to import users from an LDAP server to TIBCO Data Science - Team Studio if they authenticate users through a group membership.

Before you begin

Before running the rake command, customers must edit the `ldap.properties` file to fill in values for LDAP server parameters.

Procedure

1. The admin user starts the `rake ldap:import_users` command from the `$CHORUS_HOME/current` directory.

2. The TIBCO Data Science - Team Studio back end starts processing the rake task.

3. The TIBCO Data Science - Team Studio back end reads the LDAP server configuration from the `ldap.properties` file and sends a query to the LDAP server to fetch a list of members for a specified group.

4. The LDAP server returns with a list of members for the specified group.

    For each user in the list:

5. The TIBCO Data Science - Team Studio back end sends a query to the LDAP server to fetch the properties (first name, last name, email address, user ID, and so on) for a specified user.

6. The LDAP server returns a list of properties for the specified user.

7. The TIBCO Data Science - Team Studio back end creates or updates a user in the TIBCO Data Science - Team Studio database.

8. The user is created successfully in the TIBCO Data Science - Team Studio database using the Collaborator role. The admin user can modify the role for each user, for example, by adding the Developer or Admin role.

9. The TIBCO Data Science - Team Studio back end completes the rake task and returns.

Command-line Utilities for Managing the Services

We provide several utilities for system administrators to manage the services of the TIBCO Data Science - Team Studio from the command line, as well as from the application.

Backing up in the previous version of TIBCO Data Science - Team Studio

To protect against data loss and to plan for disaster recovery, add a regular backup procedure for the TIBCO Data Science - Team Studio to your other disaster recovery plans. Perform this task on the server where version 6.6.0 of TIBCO Data Science - Team Studio is installed.

Before you begin
- Make sure that TIBCO Data Science - Team Studio is shut down before you run the backup process.

- Place the **tds-migrate.sh** and **tds-backup66.sh** scripts in the home directory of application user. For example, if the user is `tds`, then copy the scripts to `/home/tds`.

> ⓘ **Note:** The script's owner must be the application user.

To perform the backup in the previous version of TIBCO Data Science - Team Studio, perform the following steps:

Procedure

1. Access the server where the previous version of TIBCO Data Science - Team Studio is installed.

2. Use the following command line to backup the files.

```
./tds-backup66.sh <backup_filename>.tgz
```

For example,

```
./tds-backup66.sh my-backup.tgz
```

Backing up in the current version of TIBCO Data Science - Team Studio

To protect against data loss and to plan for disaster recovery, add a regular backup procedure for the TIBCO Data Science - Team Studio to your other disaster recovery plans. Perform this task on the server where TIBCO Data Science - Team Studio is installed.

Before you begin

- Make sure that the TIBCO Data Science - Team Studio is shut down before you run the backup process.

To perform the backup in the current version of TIBCO Data Science - Team Studio, perform the following steps:

Procedure

1. Access the server where the current version of TIBCO Data Science - Team Studio is installed.

2. Use the following command line to backup the files.

```
tds platform backup
```

The above command creates a backup file in /tmp directory. The file name and path of backup files are displayed on the screen.

Restoring the TIBCO Data Science - Team Studio

You can restore the data and configuration files in the current version of TIBCO Data Science - Team Studio.

> **Note:** Due to the restrictions made in the Ruby file that handles the migration of the PostgreSQL database, the backup and restore from version 6.6.0 to 7.1.0 is not supported. To migrate from version 6.6.0 to 7.1.0, upgrade 6.6.0 to 7.0.x, and then from 7.0.x to 7.1.0.

Perform this task on the server where the data and configuration files of TIBCO Data Science - Team Studio needs to restore.

Before you begin

- You must have a backup that contains your data and configuration files.

- You must have installed the latest version of TIBCO Data Science - Team Studio.

- Make sure that the volumes created by the TIBCO Data Science - Team Studio microservices are clean.

Procedure

1. Use the following command line to restore the backup files.

   ```
   tds platform restore /path/to/<backup_filename>.tgz
   ```

2. Start the TIBCO Data Science - Team Studio.

TIBCO Data Science - Team Studio Log Files

Depending on the log level set in `chorus.properties`, the volume of log files can vary substantially.

Supported log levels:

- debug

- info

- warn

- error

- fatal

Log name	File path	Description
production.log	`<chorus-root>/shared/log/production.log`	Contains information about requests sent to the TIBCO Data Science - Team Studio web server, and various debugging information such as server errors, file not found errors, and permission denied messages.
worker.production.log	`<chorus-root>/shared/log/worker.production.log`	Contains logs for the background worker threads that TIBCO Data Science - Team Studio uses to perform various asynchronous tasks such as database imports and checking instance statuses.
scheduler.production.log	`<chorus-root>/shared/log/scheduler.production.log`	Contains information about jobs that the scheduler issues to different background workers. This mainly shows that a task was scheduled. See

Log name	File path	Description
		worker.production.log for more detailed information about what happened during execution of a task.
solr-production.log	`<chorus-root>/shared/log/solr-production.log`	Contains information about solr search queries issued against TIBCO Data Science - Team Studio.

nginx

`nginx` maintains `access.log` and `error.log` files in `<chorus-root>/shared/log/nginx`.

Download Logs

TIBCO Data Science - Team Studio provides a download of the application logs. When contacting Support, please provide the logs. Having the logs makes it much easier for Support to assist you.
Perform this task on the server where TIBCO Data Science - Team Studio is installed.

Before you begin

> **Important:** You must have administrator privileges to download logs. If you do not have the permissions, contact your administrator.

Procedure

1. In the upper right corner, click your username, and then click **Support**.

2. Click **Download Logs**. The logs download in a .zip file.

There are several types of logs, depending what type of agents you have installed on your TIBCO Data Science - Team Studio instance.

TIBCO Data Science - Team Studio workflow editor installation logs

A log of the installation process is stored here. Send this file along if you have had problems with your installation.

TIBCO Data Science - Team Studio workflow editor logs

TIBCO Data Science - Team Studio workflow editor install logs

This log contains information about the installation or upgrade process.

TIBCO Data Science - Team Studio agent logs

For each TIBCO Data Science - Team Studio agent you have enabled, the log outputs one file per agent. If you know what agents you are using, you can inspect these logs to determine whether there are underlying problems with your data source. These are also very useful in helping Support understand the situation.

Alpine.log

This log contains debug logs and information about your TIBCO Data Science - Team Studio installation. Error information about operators also can be found here.

Config files

These files have information about your preferences and runtime settings for TIBCO Data Science - Team Studio. This includes information such as Spark defaults and operator preferences.

License file

This folder contains the license information for your TIBCO Data Science - Team Studio installation.

Postgres logs

These logs are for the PostgreSQL database that supports the TIBCO Data Science - Team Studio web application. This is different from any PostgreSQL databases that you might be using as a data source.

Properties files

These files are useful for Support to see how your instance is configured. These files contain information about your enabled agents, ports that are in use, Java options, and other configuration information.

Registered HDFS data sources

If you have data sources on HDFS, you can see configuration information about them here.

Monitoring Logs

TIBCO Data Science - Team Studio provides a feature to monitor the logs of all microservices with the help of monitoring tools. This helps in tracking the issues related to New Workflow. You can check the logs of *tds-workflows* and *tds-executions* microservers along with the Spark cluster logs.

For the Legacy Workflow, see Download Logs.
Perform this task on the server where TIBCO Data Science - Team Studio is installed.

Before you begin

> ❗ **Important:** You must have administrator privileges to monitor the logs. If you do not have the permissions, contact your administrator. Based on the requirement, the system administrator can create users and grant them permissions.

The user must install the docker monitoring tools such as Portainer, DataDog, or Prometheus to access the container logs. The user can also access the container logs from the command line using the docker commands. For example: *docker logs -f <container_name>*

To access the monitoring tool, enter */monitoring* at the end of the URL.

Administering TIBCO Data Science - Team Studio

 As administrator, you can configure the TIBCO Data Science - Team Studio experience, manage user permissions, set preferences, and other tasks. This help provides information

on those tasks.

Connecting TIBCO Data Science - Team Studio to Data Sources

Review and follow these steps to connect your installation of TIBCO Data Science - Team Studio to your data sources.
Perform this task on the computer where you have installed TIBCO Data Science - Team Studio.

Before you begin

Test network connectivity and configure the TIBCO Data Science - Team Studio server.

Procedure

1. Enable web sockets.

   Verify that Web Sockets are correctly enabled by using a Web Socket test.

2. Access the cluster nodes, including the NameNode and DataNodes for Hadoop.

   Verify that you can connect to them by using the command $ telnet hostname port.

3. Enable read and write permissions for the appropriate directories, including `/tmp` for Hadoop.

   Verify this step by writing to a file in one of those directories and running a MapReduce job, if applicable.

4. Ensure that the appropriate agent is enabled for your data source.

5. Configure the necessary ports in `$CHORUS_HOME`/shared/ALPINE_DATA_ REPOSITORY/configuration/alpine.conf.

6. If you are using Spark, ensure the following.

   - The Spark host is added in `$CHORUS_HOME`/shared/ALPINE_DATA_ REPOSITORY/configuration/alpine.conf.

     ```
     alpine.spark.sparkAkka.akka.remote.netty.tcp.hostname = IP
     address for Team Studio Server
     ```

- Full communication is open between the TIBCO Data Science - Team Studio server and all cluster nodes.

7. Ensure the TIBCO Data Science - Team Studio server can access the LDAP server if applicable.

   Verify that you can connect by using $ `telnet hostname port`.

What to do next

Connect to either a Database data source, a Hadoop data source, or a TIBCO Data Virtualization data source.

Database Data Sources

You can add a database as a data source in TIBCO Data Science - Team Studio from the sidebar menu by selecting **Data** and then selecting **Add Data Source**.

For each database or JDBC data source, provide the following information about that data source.

> **Note:** The Database data sources are for Legacy Workflows and are not tested with the current version of TIBCO Data Science - Team Studio. We recommend to using the TIBCO® Data Virtualization data sources.

Connect to a JDBC Data Source

You can connect your TIBCO Data Science - Team Studio installation to a JDBC data source.
Perform this task on the computer where TIBCO Data Science - Team Studio is installed.

Before you begin

- Check System Requirements to ensure you are using a supported version of the JDBC data source.

- You must have write access to the computer where TIBCO Data Science - Team Studio server is installed.

- Before you add a JDBC data source to TIBCO Data Science - Team Studio, place the

associated JDBC driver for the data source in the `$CHORUS_HOME/shared/ALPINE_ DATA_REPOSITORY/jdbc_driver/Public` and `$CHORUS_HOME/shared/libraries` folders

Procedure

1. Open the Add Data Source dialog.

2. For **Data Source Name**, provide a user-facing name.

   You can provide any useful text.

3. Provide a useful **Description**.

4. For **JDBC URL**, provide the JDBC URL used to connect to the data source.

5. Provide the **Database Account** and **Database Password**.

   These values are your database credentials.

6. Select **Set database credentials as a shared account** if you intend to allow all users to access the data source without using their own credentials.

   Users access the database with your credentials as the data source owner. If you do not select this checkbox, each user must provide credentials for that data source to access it. You can check the box later if you change your mind.

Connect to a Hive JDBC Data Source

This topic describes how to make a Hive data source available as a JDBC connection to TIBCO Data Science - Team Studio.

For information about which Hadoop distributions support Hive as a JDBC data source, see System Requirements.

Place the appropriate Hive JAR files in the `~/ALPINE_DATA_REPOSITORY/jdbc_ driver/Public` and `$CHORUS_HOME/shared/libraries` folders. The list of necessary JARS is as follows:

- `commons-logging-*.jar`
- `hive-common*.jar`
- `hive-exec*.jar`
- `hive-jdbc*.jar`
- `hive-metastore*.jar`
- `hive-service*.jar`
- `httpclient*.jar`
- `httpcore*.jar`
- `libfb303*.jar`
- `libthrift*.jar`
- `log4j*.jar`
- `slf4j-api*.jar`

The * indicates that the version might be different, depending on the vendor. These JARs should all be available from the vendor installation.

Hive JDBC on CDH, HDP, or PHD

Fill in the required fields, marked with an asterisk *.

- **Data Source Name**: Set a user-facing name for the data source. You can choose anything you like.

- **Description**: Enter some optional text with information about this data source.

- **Hadoop Version**: The distribution of Hadoop that is running your Hive server. CDH5, HDP and PHD are supported Hadoop distributions. Note: the JAR files copied into `$CHORUS_HOME/shared/ALPINE_DATA_REPOSITORY/jdbc_driver/Public/` and `$CHORUS_HOME/shared/libraries` must match the Hadoop distribution you select here.

- **JDBC URL**: the JDBC URL used to connect to the data source.

  - **CDH**: The URL should be in the format `jdbc:hive2://SERVER_HOSTNAME:PORT`

  - **HDP**: The URL should be in the format `jdbc:hive2://SERVER_HOSTNAME:PORT/default?stack.name=hdp;stack.version=<hdpversion>`

  - **PHD**: The URL should be in the format `jdbc:hive2://SERVER_HOSTNAME:PORT/default?stack.name=phd;stack.version=<phdversion>`

- **Authentication**: TIBCO Data Science - Team Studio supports standard password authentication and Kerberos authentication. Select the type of authentication that is configured on your Hive server.

- ○ **Database Account** and **Database Password**: If **Account/Password** authentication type is selected, enter the Hive metastore account and password. By default, Hive uses an account of **hive** with the password of **hive**.

- ○ Kerberos: If the Kerberos authentication type is selected, enter the Kerberos Principal and Kerberos Keytab Location. The Kerberos principal must have permission to access the Hive server, and is typically `hive/myHadoopcluster.com@mycompany.com`.

- **Set database credentials as a shared account**: If your authentication type is set to **Account/Password**, the option to share the database credentials is available. If you check **Set database credentials as a shared account**, all users can access the data source without providing their own credentials - they are accessing the database with your credentials as the data source owner. If you do not check this box, each user must enter his or her own credentials for that data source in order to access it. You can change this setting later if you change your mind.

Connect to an Oracle Database

You can add Oracle as a data source, but to do so you must first enable it in the `chorus.properties` file. You might need to add `oracle.enabled=true`. You must also copy the `ojdbc8.jar` file to your `/shared/libraries` folder.

> **i** **Note:** TIBCO Data Science - Team Studio can exclude schemas from displaying. The list of blacklisted schemas is in `chorus.properties` and is editable.

Procedure
1. Follow the steps in Enable Oracle Databases. Supported Oracle versions can be found in System Requirements.

2. Fill in the required fields and, if you like, the optional **Description**.

   - **Data Source Name**: Set a user-facing name for the data source. You can choose anything you like.

   - **Database Name**: Enter the actual name of your Oracle database.

   - **Database Account** and **Database Password**: Provide your credentials for that database.

   - If you check the **Set database credentials as a shared account** box, all users can access the data source without providing their own credentials - they are accessing the database with your credentials as the data source owner. If you do not check this box, each user must enter his or her own credentials for that data source in order to access it. You can change this setting later if you change your mind.

Enable Oracle Databases

This topic describes how to enable an Oracle database before adding Oracle as a data source.

Procedure

1. Place the Oracle client driver JAR, `ojdbc8.jar`, in the following locations:

   ```
   <installation directory>/shared/ALPINE_DATA_REPOSITORY/jdbc_driver/
   <installation directory>/shared/ALPINE_DATA_REPOSITORY/jdbc_
   driver/Public
   <installation directory>/shared/libraries
   ```

   > **ℹ Note:** If the JAR file is not in the correct place, the error "JDBC Driver class not found" appears.

   You can find the Oracle client driver at
   http://www.oracle.com/technetwork/database/enterprise-edition/jdbc-112010-090769.html.

2. Set the following permissions on the file:

   ```
   chmod 644 ojdbc8.jar
   chown chorus:chorusojdbc8.jar
   ```

3. Set `oracle.enabled` to true in the `chorus.properties` file and restart TIBCO Data Science - Team Studio.

Connect to a Greenplum Database

You can connect your TIBCO Data Science - Team Studio installation to a Greenplum database.
Perform this task on the computer where TIBCO Data Science - Team Studio server is installed.

Before you begin

- Check System Requirements to ensure you are using a supported version of Greenplum.

- You must have write access to the computer where TIBCO Data Science - Team Studio server is installed.

- If another TIBCO Data Science - Team Studio user has greater privileges for a shared

account database than you, the owner, that user must provide credentials to see the parts of the database denied to you.

Procedure

1. Open the Add Data Source dialog.

2. For **Data Source Name**, provide a user-facing name.

   You can provide any useful text.

3. For **Database Name**, set the actual database name.

   PostgreSQL is the default because many Greenplum users have a database with that name. If you are adding a database with a different name, provide the name here.

4. Provide the **Database Account** and **Database Password**.

   These values are your database credentials.

5. Select **Set database credentials as a shared account** if you intend to allow all users to access the data source without using their own credentials.

   Users access the database with your credentials as the data source owner. If you do

not select this checkbox, each user must provide credentials for that data source to access it. You can check the box later if you change your mind.

6. Select **Use SSL** if you are using SSL-enabled PostgreSQL and Greenplum

   If you choose this option, communications between TIBCO Data Science - Team Studio and the database are secured using SSL.

7. Provide an entry for the **Description**.

Connect to a Pivotal HAWQ Database

You can connect your TIBCO Data Science - Team Studio installation to a Pivotal HAWQ database.
Perform this task on the computer where TIBCO Data Science - Team Studio server is installed. Supported HAWQ versions can be found in System Requirements.

Before you begin
- Check System Requirements to ensure you are using a supported version of Pivotal HAWQ.

- You must have write access to the computer where TIBCO Data Science - Team Studio server is installed.

- If another TIBCO Data Science - Team Studio user has greater privileges for a shared account database than you, the owner, that user must provide credentials to see the parts of the database denied to you.

Procedure
1. Open the Add Data Source dialog.

2. For **Data Source Name**, provide a user-facing name.

   You can provide any useful text.

3. For **Database Name**, set the actual database name.

   PostgreSQL is the default because many Greenplum users have a database with that name. If you are adding a database with a different name, provide the name here.

4. Provide the **Database Account** and **Database Password**.

   These values are your database credentials.

5. Select **Set database credentials as a shared account** if you intend to allow all users to access the data source without using their own credentials.

   Users access the database with your credentials as the data source owner. If you do not select this checkbox, each user must provide credentials for that data source to access it. You can check the box later if you change your mind.

6. Select **Use SSL** if you are using SSL-enabled PostgreSQL and Greenplum

   If you choose this option, communications between TIBCO Data Science - Team Studio and the database are secured using SSL.

7. Provide an entry for the **Description**.

Connect to an Amazon RedShift Data Source

You can connect TIBCO Data Science - Team Studio to an Amazon RedShift data source. Perform this task on the computer where TIBCO Data Science - Team Studio server is installed.

Before you begin

You must have write access to the TIBCO Data Science - Team Studio server. You must also have access to the Amazon RedShift configuration files.

Procedure

1. Copy the RedShift driver to the following directories. and then change the ownership of these copies to the user who runs TIBCO Data Science - Team Studio (usually user 'Chorus').

   - *$CHORUS_HOME*/shared/*ALPINE_DATA_REPOSITORY*/jdbc_driver/Public

   - *$CHORUS_HOME*/shared/libraries

2. Change the ownership of these copies to the user who runs TIBCO Data Science - Team Studio.

   Usually, that username is 'Chorus'.

3. Create a new redshift directory named *$CHORUS_HOME*/shared/*ALPINE_DATA_REPOSITORY*/jdbc/redshift.

4. Copy the file driver.properties from the directory $CHORUS_HOME/shared/ALPINE_DATA_REPOSITORY/jdbc/default to the newly created redshift directory.

5. Edit the contents of the file *$CHORUS_HOME*/shared/*ALPINE_DATA_REPOSITORY*/jdbc/redshift/driver.properties as follows.

   ```
   # Specify the JDBC class driver for the desired database type.
   # Examples:
   # Oracle = oracle.jdbc.driver.OracleDriver
   # Greenplum = org.postgresql.Driver
   # DB2 = com.ibm.db2.jcc.DB2Driver
   # Netezza = org.netezza.Driver
   # PostgreSQL = org.postgresql.Driver
   # SQLServer = com.microsoft.sqlserver.jdbc.SQLServerDriver
   # MySQL = com.mysql.jdbc.Driver
   # Teradata = com.teradata.jdbc.TeraDriver
   # Vertica = com.vertica.jdbc.Driver
   # Sybase = com.sybase.jdbc2.jdbc.SybDriver
   # Informix = com.informix.jdbc.IfxDriver
   # SAPDB = com.sap.dbtech.jdbc.DriverSapDB
   # InterBase = interbase.interclient.Driver
   # HSqlDB = org.hsqldb.jdbcDriver
   ```

```
# MariaDB = org.mariadb.jdbc.Driver
# MySQL = com.mysql.jdbc.Driver
driverClass=com.amazon.redshift.jdbc41.Driver
```

6. Locate and open for editing the file `additional_jdbc_drivers.rb`.

   The path is similar to path similar to
   /data/chorus/install/releases/5.9.1.0.3973–
   5d95f7c97/components/core/app/mixins/sequel/extensions/additional_jdbc_
   drivers.rb

7. Add a line for the `redshift` class so that the content resembles the following.

```
module Sequel
  module AdditionalJdbcDrivers
    MAP =  {
        mariadb: ->(db) { org.mariadb.jdbc.Driver },
        teradata: ->(db) { com.teradata.jdbc.TeraDriver },
        vertica: ->(db) { com.vertica.jdbc.Driver },
        hive2: ->(db) { org.apache.hive.jdbc.HiveDriver },
        hive: ->(db) { org.apache.hadoop.hive.jdbc.HiveDriver },
        redshift: ->(db) { com.amazon.redshift.jdbc41.Driver }
        }

    MAP.each do |key, driver|
      ::Sequel::JDBC::DATABASE_SETUP[key] = driver
    end
  end
end
```

> **!** **Important:** You must apply this change to the file `additional_jdbc_drivers.rb` again after upgrading TIBCO Data Science - Team Studio.

8. Restart TIBCO Data Science - Team Studio.

9. Open the Add Data Source dialog.

10. Provide the **Data Source Type**, the **Data Source Name**, and (optionally), the **Description**.

11. Set the data connection (**JDBC URL**) using a URL similar to the following.

> ✅ **Tip:** You can copy your RedShift URL from your AWS RedShift page

```
jdbc:redshift://armen-jjredshift.csyb6t8bifc8.us-west-
1.redshift.amazonaws.com:5439/armenjjdb
```

12. Select **Set database credentials as a shared account** if you intend to allow all users to access the data source without using their own credentials.

    Users access the database with your credentials as the data source owner. If you do not select this checkbox, each user must provide credentials for that data source to access it. You can check the box later if you change your mind.

Connect to a BigQuery Data Source

You can connect TIBCO Data Science - Team Studio to a GCP BigQuery data source. Perform this task on the computer where TIBCO Data Science - Team Studio server is installed.

Before you begin

You must have write access to the TIBCO Data Science - Team Studio server. You must also have access to the GCP BigQuery configuration files.

Procedure

1. Copy the BigQuery driver to the following directories. and then change the ownership of these copies to the user who runs TIBCO Data Science - Team Studio (usually user 'Chorus').

   - *$CHORUS_HOME*/shared/*ALPINE_DATA_REPOSITORY*/jdbc_driver/Public

   - *$CHORUS_HOME*/shared/libraries

2. Change the ownership of these copies to the user who runs TIBCO Data Science - Team Studio.

   Usually, that username is 'Chorus'.

3. Create a bigquery directory named *$CHORUS_HOME*/shared/*ALPINE_DATA_ REPOSITORY*/jdbc/bigquery.

4. Copy the file driver.properties from the directory $CHORUS_HOME/shared/ALPINE_ DATA_REPOSITORY/jdbc/default to the newly created bigquery directory.

5. Edit the contents of the file *$CHORUS_HOME*/shared/*ALPINE_DATA_ REPOSITORY*/jdbc/bigquery/driver.properties as follows.

```
# Specify the JDBC class driver for the desired database type.
# Examples:
# Oracle = oracle.jdbc.driver.OracleDriver
# Greenplum = org.postgres#ql.Driver
# DB2 = com.ibm.db2.jcc.DB2Driver
# Netezza = org.netezza.Driver
# PostgreSQL = org.postgresql.Driver
# SQLServer = com.microsoft.sqlserver.jdbc.SQLServerDriver
# MySQL = com.mysql.jdbc.Driver
# Teradata = com.teradata.jdbc.TeraDriver
# Vertica = com.vertica.jdbc.Driver
# Sybase = com.sybase.jdbc2.jdbc.SybDriver
# Informix = com.informix.jdbc.IfxDriver
# SAPDB = com.sap.dbtech.jdbc.DriverSapDB
# InterBase = interbase.interclient.Driver
# HSqlDB = org.hsqldb.jdbcDriver
```

```
# MariaDB = org.mariadb.jdbc.Driver
# MySQL = com.mysql.jdbc.Driver
driverClass=com.simba.googlebigquery.jdbc42.Driver

# BigQuery (like Hive) does not support "schema"."tablename".
# For BigQuery, this entry must be empty string:
identiferQuotation=
# with no whitespace (except newline) or characters after the
equals sign
identifierQuotation=
```

6. Locate and open for editing the file `additional_jdbc_drivers.rb`.

   The path is similar to path similar to
   `/usr/chorus/install/releases/6.3.2.0.8068-7ac910ae3/components/core/app/mixins/sequel/extensions/additional_jdbc_drivers.rb`

7. Add a line for the `bigquery` class so that the content resembles the following.

```
module Sequel
  module AdditionalJdbcDrivers
    MAP =  {
        mariadb: ->(db) { org.mariadb.jdbc.Driver },
        teradata: ->(db) { com.teradata.jdbc.TeraDriver },
        vertica: ->(db) { com.vertica.jdbc.Driver },
        hive2: ->(db) { org.apache.hive.jdbc.HiveDriver },
        hive: ->(db) { org.apache.hadoop.hive.jdbc.HiveDriver },
        bigquery: ->(db) { com.simba.googlebigquery.jdbc42.Driver }
        }

    MAP.each do |key, driver|
      ::Sequel::JDBC::DATABASE_SETUP[key] = driver
    end
  end
end
```

> ❗ **Important:** You must apply this change to the file `additional_jdbc_drivers.rb` again after upgrading TIBCO Data Science - Team Studio.

8. Open the file *$CHORUS_HOME*/shared/*ALPINE_DATA_REPOSITORY*/configuration/alpine.conf and add the following configuration parameter.

```
database.bigquery.batchSize (default = 1000)
```

> ℹ **Note:** Due to idiosyncrasies in the way BigQuery handles batch updates, uploading a narrow table can result in an error. This configuration parameter addresses possible batch size issues.

9. Restart TIBCO Data Science - Team Studio.

10. Open the Add Data Source dialog.

11. Provide the **Data Source Type**, the **Data Source Name**, and (optionally), the **Description**.

12. Set the data connection (**JDBC URL**) using a URL similar to the following.

> ✅ **Tip:** You can copy your BigQuery URL from your GCP BigQuery page

```
jdbc:bigquery://https://www.googleapis.com/bigquery/v2:443;ProjectI
d=teamstudio-user@teamstudio-alpine.iam.gserviceaccount.com
```

13. Specify **Workspace Visibility**.

    A data source with **Limited** visibility must be manually associated with a workspace for members of that workspace to use the data source. By default, this option is set to **Public**.

14. Select **Set database credentials as a shared account** if you intend to allow all users to access the data source without using their own credentials.

    Users access the database with your credentials as the data source owner. If you do not select this checkbox, each user must provide credentials for that data source to access it. You can check the box later if you change your mind.

BigQuery Data Source Connection Tests and Troubleshooting

If you encounter error messages or other problems with your data source connection, check this topic for suggestions for troubleshooting.

BigQuery data source: "Too many queries"

This error results when the batch size is too small. To correct this error, open the file $CHORUS_HOME/shared/ALPINE_DATA_REPOSITORY/configuration/alpine.conf and add the configuration parameter database.bigquery.batchSize (default = 1000).

BigQuery data source: "Cannot parse query"

This error results when the batch size is too big. To correct this error, open the file $CHORUS_HOME/shared/ALPINE_DATA_REPOSITORY/configuration/alpine.conf and add the configuration parameter database.bigquery.batchSize (default = 1000).

Google recommends that BigQuery data source users insert data using the Google API. Using DML issued through JDBC can cause problems because of statement size limitations; therefore, you should avoid using TIBCO Data Science - Team Studio operators to copy data into BigQuery. For more information on BigQuery limitations, see https://cloud.google.com/bigquery/quotas.

Hadoop Data Sources

These topics show you how to add a Hadoop data source from the command line or through the TIBCO Data Science - Team Studio user interface, and how to connect to various data sources.

Adding a Hadoop Data Source from the User Interface

To add an HDFS data source, first make sure the TIBCO Data Science - Team Studio server can connect to the hosts, and then use the Add Data Source dialog to add it to TIBCO Data Science - Team Studio.
Supported Hadoop distributions are listed in TIBCO Data Science - Team Studio System Requirements.

Before you begin

You must have data administrator or higher privileges to add a data source. Ensure that you have the correct permissions before continuing.

Procedure

1. From the menu, select **Data**.

2. Select **Add Data Source**.

3. Choose **Hadoop Cluster** as the data source type.

4. Specify the following data source attributes:

Data Source Name	Set a user-facing name for the data source. This should be something meaningful for your team (for example, "Dev_CDH5_cluster").
Description	Enter a description for your data source.
Hadoop Version	Select the Hadoop distribution that matches your data source.
Use High Availability	Check this box to enable High Availability for the Hadoop cluster.
Disable Kerberos Impersonation	If this box is selected and you have Kerberos enabled on your data source, then the workflow uses the user account configured as the **Hadoop Credentials** here.

	If this box is cleared, the workflow uses the user account of the person running the workflow.
	If you do not have Kerberos enabled on your data source, you do not need to select this box. All workflows run using the account configured as the **Hadoop Credentials**.
NameNode Host	Enter a single active NameNode to start. Instructions for enabling High Availability are in Step 10. To verify the NameNode is active, check the web interface. (The default is `http://namenodehost.localhost:50070/`)
NameNode Port	Enter the port that your NameNode uses. The default port is 8020.
Job Tracker/Resource Manager Host	For MapReduce v1, specify the job tracker. For YARN, specify the resource manager host.
Job Tracker/Resource Manager Port	Common ports are 8021, 9001, 8012, or 8032.
Workspace Visibility	There are two options here: • Public - Visible and available to all workspaces. • Limited - Visible and available only to workspaces they are associated with. To learn more about associating a data source to a workspace, see Data Visibility.
Hadoop Credentials	Specify the user or service to use to run MapReduce jobs. This user must be able to run MapReduce jobs from the command line.
Group List	Enter the group to which the Hadoop account belongs.

5. For further configuration, choose **Configure Connection Parameters**.

6. Specify key-value pairs for YARN on the TIBCO Data Science - Team Studio server. Selecting **Load Configuration** from Resource Manager attempts to populate configuration values automatically.

- `yarn.resourcemanager.scheduler.address`

- `yarn.app.mapreduce.am.staging-dir`

> ℹ **Note:** Be sure the directory specified above in the `staging-dir` variable is writable by the TIBCO Data Science - Team Studio user. Spark jobs produce errors if the user cannot write to this directory.

Required if different from default:

- `yarn.application.classpath`

  - The `yarn.application.classpath` does not need to be updated if the Hadoop cluster is installed in a default location.

  - If the Hadoop cluster is installed in a non-default location, and the `yarn.application.classpath` has a value different from the default, the YARN job might fail with a "cannot find the class AppMaster" error. In this case, check the `yarn-site.xml` file in the cluster configuration folder. Configure these key:value pairs in the UI using the **Configure Connection Parameters** option.

- `yarn.app.mapreduce.job.client.port-range`

  - This describes a range of ports to which the application can bind. This is useful if operating under a restrictive firewall that needs to allow specific ports.

Recommended:

- `mapreduce.jobhistory.address = FQDN:10020`

> ⚠️ **Caution:** Operators that use Pig for processing do not show the correct row count in output if `mapreduce.jobhistory.address` is not configured correctly.

- `yarn.resourcemanager.hostname = FQDN`

- `yarn.resourcemanager.address = FQDN`

- `yarn.resourcemanager.scheduler.address = FQDN:8030`

- `yarn.resourcemanager.resource-tracker.address = FQDN:8031`

- `yarn.resourcemanager.admin.address = FQDN:8033`

- `yarn.resourcemanager.webapp.address = FQDN:8088`

- `mapreduce.jobhistory.webapp.address = FQDN:19888`

7. Save the configuration.

8. To perform a series of automated tests on the data source, click **Test Connection**.

9. Click **Save Configuration** to confirm the changes.

10. When the connectivity to the active NameNode is established above, set up NameNode High Availability (HA) if enabled.

Required:

- `dfs.ha.namenodes.nameservice1`

- `dfs.namenode.rpc-address.nameservice1.namenode<id>` (required for each namenode id)

- `dfs.nameservices`

- `dfs.client.failover.proxy.provider.nameservice1`

Recommended:

- `ha.zookeeper.quorum`

> **(i)** **Note:** Support for Resource Manager HA is available.
>
> To configure this, add `failover_resource_manager_hosts` to the advanced connection parameters and list the available Resource Managers.
>
> If one of the active Resource Managers fails during a job running, you must rerun the job, but you no longer must reconfigure the data source that failed. If one of the active Resource Managers fails while a job is not running, you do not need to do anything. TIBCO Data Science - Team Studio uses another available Resource Manager instead.

Hadoop Data Source Connection Tests and Troubleshooting

You can test Hadoop connections to troubleshoot the connection to the datasource. TIBCO Data Science - Team Studio provides a variety of tests to verify connectivity and troubleshoot problems.
Perform this task on the computer where TIBCO Data Science - Team Studio is installed.

Before you begin
You must have access to the Hadoop data source.

Procedure
1. From the menu, click **Data**.

   The data sources are displayed.

2. On the right side of the display, click **Edit Data Source**.

   The Edit Data Source Configuration dialog is displayed.

3. Click **Test Configuration**.

   The Test Data Source Configuration Settings dialog is displayed.

4. Run the tests.

   You can run all tests at the same time.

   The progress is displayed from this configuration.

What to do next

Click **Details** for any of the tests to review specific test details and error messages.

TIBCO Data Virtualization Data Sources

You can add a TIBCO® Data Virtualization as a data source in TIBCO Data Science - Team Studio. The TIBCO® DV data source can be used with New Workflow and also with the Legacy Workflow using Database (DB) operators.

Before you begin

You must have Data Administrator or higher privileges to add a TIBCO Data Virtualization. Ensure that you have the correct permissions before continuing.

Procedure

Perform the following steps to add a TIBCO Data Virtualization data source in TIBCO Data Science - Team Studio.
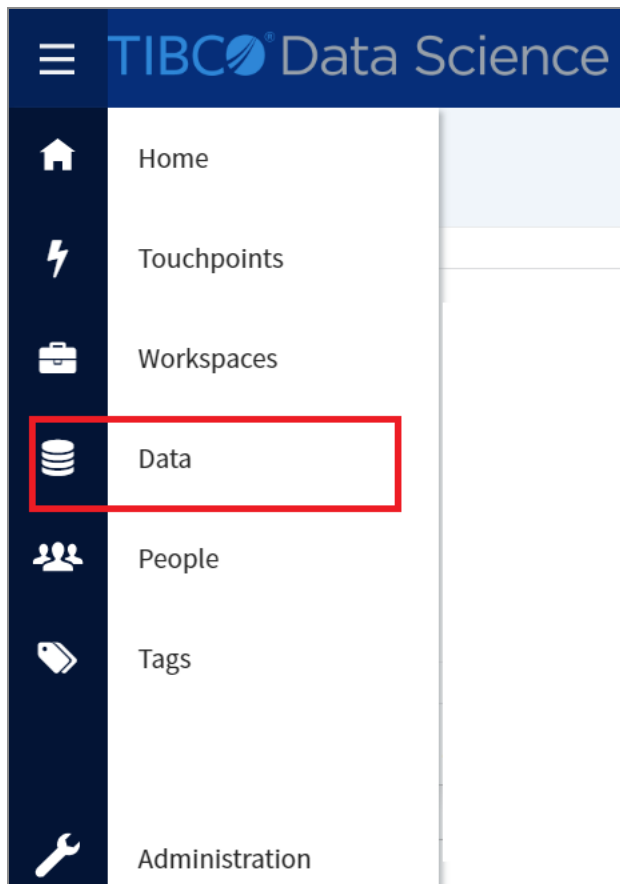
1. From the sidebar menu, select **Data**.

2. On the Data Sources page, click **Add Data Source**.

3. The Add Data Source dialog appears. From the **Data Source Type** dropdown list, select **TIBCO Data Virtualization**.

ADD DATA SOURCE ✖

Data Source Type

🔷 TIBCO Data Virtualization ▼

ℹ️ Data Source and Domain values are appended to the Connection URL. Both unsupportedMode and enableFlood parameters are automatically set to "false." Other query string parameters can be added to Additional Parameters separated by "&." Click Preview Connection URL to view the generated URL.

Data Source Name *

Description

Host *   Port *

9401

Data Source *

Data

Domain *

composite

Additional Parameters

Preview Connection URL

Workspace Visibility ❓

Public ▼

Username *   Password *

☐ Set Data Source credentials as shared ❓

Load Configuration from File…

* Required

Cancel   **Add Data Source**

4. In the **Data Source Name** field, specify a name for the data source. You can provide any useful text.

5. In the **Description** field, provide a useful description of your data source. This field is optional.

6. In the **Host** field, specify the hostname or address to connect to the data source.

7. In the **Port** field, specify the port number where TIBCO Data Virtualization server is running.

8. In the **Data Source** field, specify the source of the data.

9. In the **Domain** field, specify the domain of the data source.

10. In the **Additional Parameters** field, provide the additional parameters information. This field allows you to add additional parameters to the JDBC connection URL, such as adding credentials for a data source, encryption token, and so on. This field is optional.

   > **ⓘ Note:** If the SSL certificate is not installed on the TIBCO Data Virtualization server connected with TIBCO Data Science - Team Studio, then enter `allowHttpInsecure=true` in the **Additional Parameters** field.

   The **Preview Connection URL** displays the full JDBC URL that TIBCO Data Science - Team Studio uses. Based on the user's input, TIBCO Data Science - Team Studio generates this.

11. In the **Workspace Visibility** dropdown list, select the visibility of the workspace. The available options are **Public** and **Limited**.

   > **ⓘ Note:** A data source with **Limited** visibility must be manually associated with a workspace for members of that workspace to use the data source.

12. In the **Username** and **Password** field, enter your database credentials.

13. Select **Set Data Source credentials as a shared** if you intend to allow all users to access the data source without using their own credentials. This is optional.

   Users access the database with your credentials as the data source owner. If you do not select this checkbox, each user need to manually enter the credentials for that data source to access it. You can also check the box later.

14. By using the **Load Configuration from File** option, you to set the values from a file that has been saved from another TIBCO Data Virtualization connection.

15. Click **Add Data Source** to add the data source.

Connecting TIBCO Data Science - Team Studio to Spark Cluster

A Spark cluster is a MapReduce-like cluster computing framework designed for low-latency iterative jobs and interactive use from an interpreter. It provides clean, language-integrated APIs in Scala, Java, Python, and R along with an optimized engine that supports general execution graphs. Additionally, a wide range of advanced tools is supported, such as Spark SQL for processing SQL and structured data, Pandas API on Spark for pandas workloads, MLlib for machine learning, GraphX for processing graphs, and Structured Streaming for incremental computation and stream processing.

A Spark cluster is a combination of a Driver Program, Cluster Manager, and Worker Nodes that work together to complete tasks. You can coordinate processes throughout the cluster using SparkContext. The SparkContext can connect to various cluster managers that distribute resources among applications. Once connected, Spark acquires executors on nodes in the cluster, which are processes that run computations and store data for your application. The executors then receive your application code (specified by JAR or Python files supplied to SparkContext), and then SparkContext sends tasks to the executors for execution.

The following figure helps to visualize the Spark cluster:
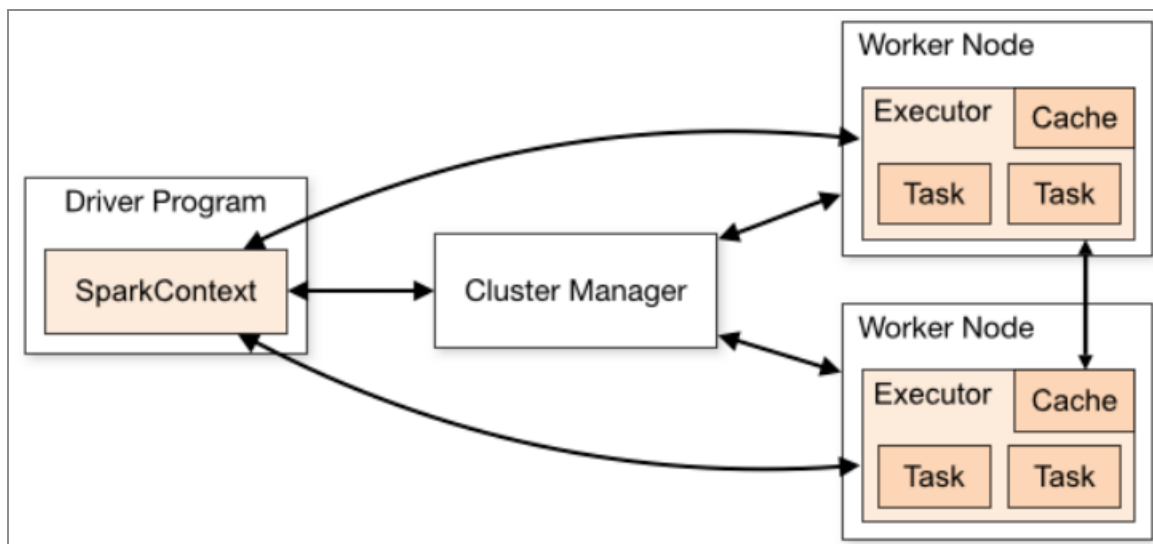
TIBCO Data Science - Team Studio 7.1.0 supports Spark cluster version 3.3 that can only be used with Apache Spark 3.2 or later workflows. These workflows consist of modern operators that use Spark SQL for data processing.

The file storage system lets you store large data sets across an infinite number of servers, rather than storing all the data sets on a single server. The file storage can either be a file system associated with the Spark cluster (local) or outside of the Spark cluster (remote). Apache Spark 3.2 or later cluster doesn't have a dedicated file storage system. In either case, they can only use TIBCO® Data Virtualization as a file storage system.

When a file or data is stored outside of the Spark cluster, then the data is copied to the Spark cluster so that Spark can perform the analytic processing. Whereas, when a file or data is stored in a database connected to TIBCO® DV, the data is already available to the Spark cluster so that Spark can perform the analytic processing. In this case, since there is no copy, transfer, or movement of any data, the performance is optimized.

Spark Cluster Manager

TIBCO Data Science - Team Studio can be connected to several different Spark systems using different Spark cluster managers. The following Spark cluster manager is supported in the 7.1.0 version of TIBCO Data Science - Team Studio:

Standalone cluster manager

There are three possible configurations when using Spark Standalone cluster manager:

1. When TIBCO Data Science - Team Studio, Spark Standalone cluster, and TIBCO® DV are running on the same server. This is the simplest configuration, but since they are sharing the resources, it is not recommended for big data environments.

2. When TIBCO Data Science - Team Studio and Spark Standalone cluster are on the same server, but TIBCO® DV is on a different server. In this case, a network file system is shared between the two servers and mounted at the same mount point.

3. When TIBCO Data Science - Team Studio, Spark Standalone cluster, and TIBCO® DV are on different servers. In this case, the same network file system must be mounted on the TIBCO® DV and Spark cluster servers.

YARN cluster manager

A YARN cluster manager is used by the Cloudera CDH / CDP and Amazon EMR. These are existing clusters that have an associated file system (`HDFS` in the case of Cloudera and `S3`

in the case of EMR). These file systems are the shared file system and TIBCO® DV is connected to this file system.

Adding a Spark Cluster from the User Interface

To add a Spark cluster, first, make sure the TIBCO Data Science - Team Studio server can connect to the hosts.
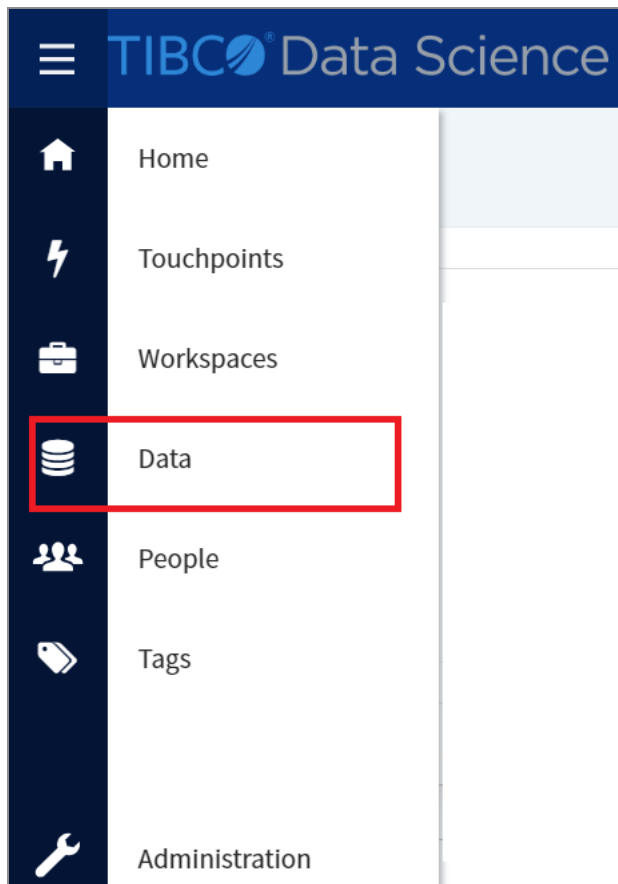
Before you begin

You must have Data Administrator or higher privileges to add a Spark cluster. Ensure that you have the correct permissions before continuing.

Procedure

Perform the following steps to add a Spark cluster in TIBCO Data Science - Team Studio

1. From the sidebar menu, select **Data**.

2. On the Data Sources page, click **Add Data Source**.

3. The Add Data Source dialog appears. From the **Data Source Type** dropdown list, select **Spark Cluster**.

4. In the **Data Source Name** field, specify a user-facing name. You can provide any useful text.

5. In the **Description** field, provide a useful description for your Spark cluster. This field is optional.

6. From the **Cluster Manager Type** dropdown list, select the cluster manager for Spark. The available options are **Apache Spark Standalone** and **YARN**.

i. When **Apache Spark Standalone** is selected, provide the URL for Spark in the **Spark Master URL** field.

ii. When **YARN** is selected, perform the following steps:

    a. In the **YARN Configurations Files**, click **Select Files**, and then browse for the (*yarn-site.xml* and *core-site.xml*) configuration files in your local system.

> ℹ️ **Note:** You can download the (*yarn-site.xml* and *core-site.xml*) files from the YARN clusters. You can also get these files from the cluster admin.
>
> Before uploading, open the *core-site.xml* file and remove the *com.hadoop.compression.lzo.LzoCodec* and *com.hadoop.compression.lzo.LzopCodec* value from the `io.compression.codecs` property.

    b. In the **Hadoop Username** field, enter your Hadoop username.

> ℹ️ **Note:** The username must be available in the `/user` directory of Hadoop cluster. The user must have the read/write permissions. If you face any permission issues, see Configuring the HDFS Directory and Permissions for Results File Storage.

7. For further configuration, choose **Configure Connection Parameters**. The CONFIGURE CONNECTION PARAMETERS dialog appears.

CONFIGURE CONNECTION PARAMETERS	✕
Edit Connection Parameters Individually	Bulk Edit
Key	Value ❌ Add parameter
	Cancel **Save**

    a. Specify key-value pairs for YARN on the TIBCO Data Science - Team Studio server.

    b. To add a new parameter, click **Add parameter**.

    c. To edit the connection parameters in bulk, click **Bulk Edit**.

     d. Click **Save**.

For more information, see the Configuring Connection Parameters.

8. In the **Workspace Visibility** dropdown list, select the visibility of the workspace. The available options are **Public** and **Limited**.

> ℹ️ **Note:** A data source with **Limited** visibility must be manually associated with a workspace for members of that workspace to use the data source. To learn more about associating a data source to a workspace, see Data Visibility.

9. The **Load Configuration from File** allows you to set the values from a file that has been saved from another Spark Cluster connection.

10. Click **Add Data Source** to add the data source.

Configuring Connection Parameters

To run a New Workflow, you should have a shared file system between TIBCO Data Virtualization and Spark cluster. The following shared file system are available:

- NFS shared drive when using the Apache Spark Standalone cluster.

- Amazon S3 bucket when using the EMR cluster.

- HDFS folder when using the Cloudera cluster.

The system administrator should create separate directories for models and outputs on the shared volume. Intermediate results from the Model operators are stored in the *Models* directory and the output tables from the operators are stored in the *Output* directory.

In the Configure Connection Parameters dialog, add the following parameters to store the intermediate results and output of the operators:

Parameter	Description
`tds.datavirt.sharedDataVolumes`	This parameter specifies the directory where the system administrator wants to store the output tables from the operators. You can also provide multiple shared volumes. For more information, see Accessing

Parameter	Description
	Data in TIBCO Data Virtualizationfrom TIBCO Data Science - Team Studio.
tds.runtime.sharedTempVolume	This parameter specifies the directory where the system administrator wants to store the intermediate results from the Model operators. **Note:** If you are using an EMR cluster as a Spark cluster, then you can use the HDFS file system on EMR for better performance.

If you are using,

- Amazon S3 bucket, then the URL should be: *s3a://<directory_path>*

- NFS shared drive, then the URL should be: *file:// <directory_path>*

- HDFS shared drive, then the URL should be: *hdfs://<directory_path>*

For example:

tds.datavirt.sharedDataVolumes = s3a://qat3/output2/

tds.runtime.sharedTempVolume = s3a://qat3/models/

If you are using the EMR as the Spark cluster, then add the following parameters:

Parameter	Description
spark.yarn.populateHadoopClasspath = true	This parameter appends the YARN classpath on the EMR cluster.
spark.hadoop.fs.s3a.aws.credentials.provider = com.amazonaws.auth.InstanceProfileCredentialsProvider	This parameter enables the authentication.
spark.yarn.stagingDir	If you are using the Amazon EMR 6.7 cluster, then the administrator must configure the Hadoop YARN data source with this parameter. This

Parameter	Description
	parameter specifies the URL of the staging directory that the execution submitter uses while submitting the Spark jobs.
	Note: Enter a valid HDFS directory on the EMR cluster and make sure that the Hadoop user configured in the data source has read, write, and run permissions on this directory.
	Example: `spark.yarn.stagingDir=hdfs://<ip_address>/user/foobar/.sparkStaging`

If you are using an Apache Spark Standalone cluster, then add the following parameters:

Parameter	Description
`tds.executions.sparkClusterVersion`	This parameter specify the version of the Spark cluster. Example: **3.2.1**
`spark.dynamicAllocation.disabled`	By default, the platform enables dynamic allocation. If you want to disable the dynamic allocation, then this parameter is set to *true*.

Workflow Editor Preferences

Use the Preferences dialog to modify a variety of Workflow Editor preferences.

> ⚠ **Important:** You must be an administrator to edit preferences.

To open the Preferences dialog, click the **Actions** menu, and then select **Preferences**.
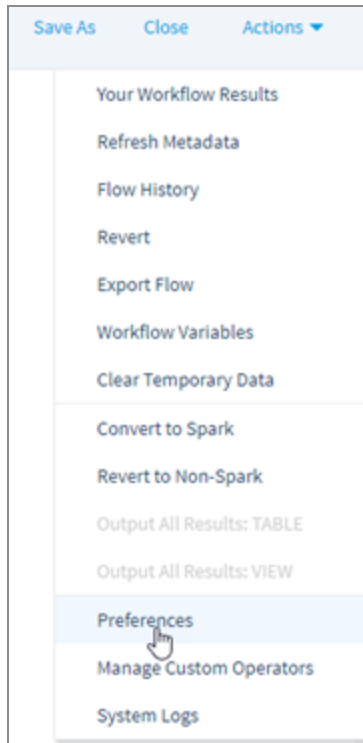
Using the **Preferences** dialog:

For each tab, click **Save** to save changes. If you do not click **Save**, you are prompted to save or revert the changes when you change preference sections.

You can configure the Editor preferences at the system level.

- Edit the `alpine.config` file located in `ALPINE_DATA_REPOSITORY/configuration`.

- Defaults are available in `alpine.config.defaults`.

- Values in `.defaults` are overwritten at system startup, so all customizations should be made in alpine.config.

Algorithm Preferences

Use algorithm preferences to control computation and certain behaviors when running a workflow.

> ❗ **Important:** You must be an administrator to edit preferences.

Procedure

1. To open the Preferences dialog, click the **Actions** menu and then select **Preferences**.

2. Select **Algorithm**.

Distinct Value Count

A general setting for the way the application stores and analyzes the set of possible values that each variable can take on.

In many cases, TIBCO Data Science - Team Studio must be aware of all possible distinct values of a given variable or column. For example, when building a classification model, TIBCO Data Science - Team Studio might have to store the set of classes represented by the distinct values of the dependent variable. Storing these values often require memory, and so we limit the maximum size of memory used to store and analyze these distinct values.

> ⚠️ **Caution:** This setting is widely used throughout the application, so be careful of changing this value, but consider using it when you encounter memory issues.

Summary Statistic Distinct Value Count

A special case of the Distinct Value Count preference.

The Summary Statistics operator must store the distinct values of each column so that it can compute the number of distinct values and the most common values for each column. This can consume a lot of memory, so use this setting to control the amount of memory used by the Summary Statistics operator.

You can also limit the memory that the Summary Statistics operator uses by setting **Calculate Number of Distinct Values** to false in the operator properties.

Decimal Precision Digits

Controls the number of decimal places used to display results throughout the application.

Click **Save** to save changes. Click **Restore** to return to default values.

System Preferences

Use system preferences to change log levels and grant access to logs.

> ! **Important:** You must be an administrator to edit preferences.

Procedure

1. To open the Preferences dialog, click the **Actions** menu and then select **Preferences**.
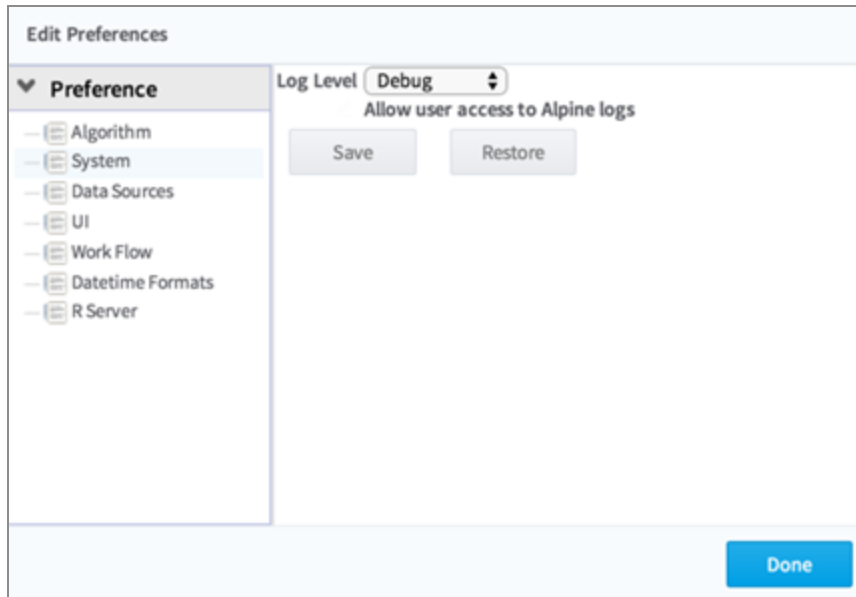
2. Select **System**.

    Log Level

    Sets the log level of the `alpine.log` file to **Info** or **Debug** (for more detail) mode.

    Allow user access to Team Studio logs

    Grants non-admin users access to the system logs (from the UI).

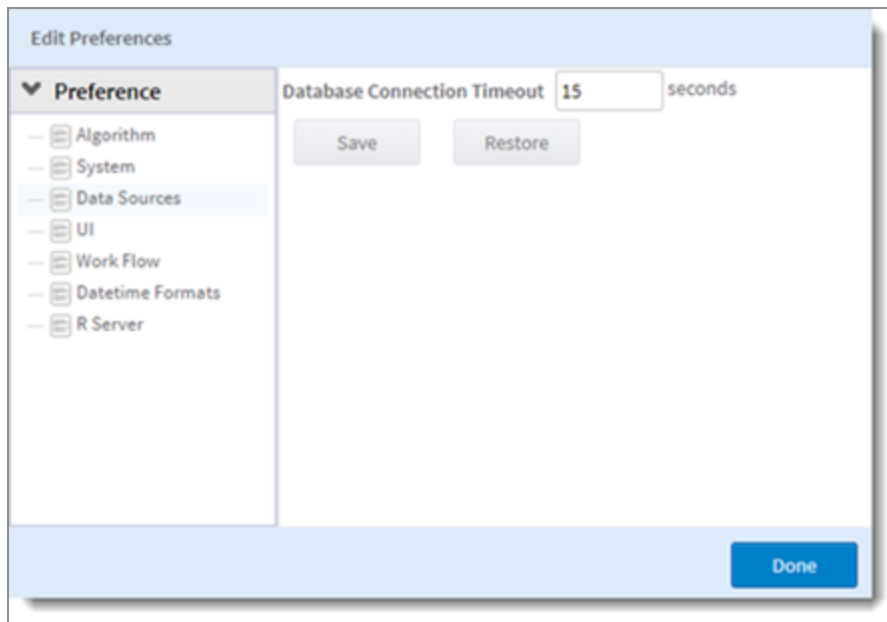   Click **Save** to save changes. Click **Restore** to return to default values.

Data Source Preferences

Use data source preferences to set the amount of time allowed for testing database connections before system timeout. Data source preferences apply to all data source connections used in a workflow.

> ⊘ **Important:** You must be an administrator to edit preferences.

Procedure

1. To open the Preferences dialog, click the **Actions** menu and then select **Preferences**.

2. Select **Data Sources**.

Database Connection Timeout (seconds)

The amount of time allowed for testing database connections before the system reaches timeout (in seconds).

Click **Save** to save changes. Click **Restore** to return to default values.

UI Preferences

Use UI preferences to control various display options.

> ❗ **Important:** You must be an administrator to edit preferences.

Procedure

1. To open the Preferences dialog, click the **Actions** menu, and then select **Preferences**.

2. Select **UI**.

Data Preview Max Rows

Number of rows to display by the **Data Preview** right-click option.

Max Points of Scatter Plot

Number of points to display by scatter plots.

Max Points of Cluster

Number of points to display in Cluster visualizations.

Standard notation threshold

Threshold number of digits a value can have before being displayed using Scientific Notation format.

Decimal Precision

Maximum number of decimals to display for values.

Click **Save** to save changes. Click **Restore** to return to default values.

Work Flow Preferences

Use Work Flow preferences to modify the global temp directory.

> ❗ **Important:** You must be an administrator to edit preferences.

Procedure

1. To open the Preferences dialog, click the **Actions** menu, and then select **Preferences**.

2. Select **Work Flow**.
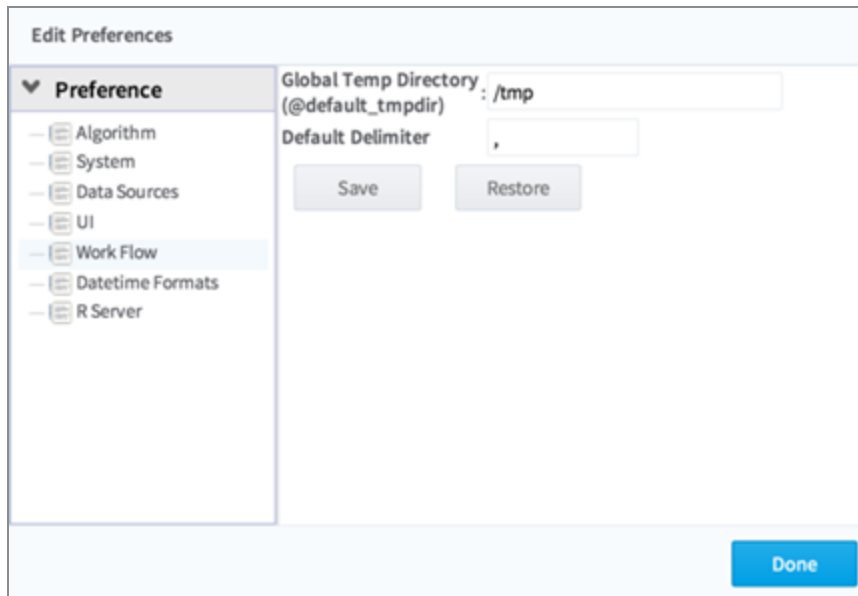
Global Temp Directory

- Sets the *@default_tmpdirworkflow* variable for all flows created in TIBCO Data Science - Team Studio.

- Sets the `/tmp` directory used by the MapReduce code in TIBCO Data Science - Team Studio when running Hadoop operators.

- Ensures that UNIX users running the TIBCO Data Science - Team Studio application have read/write permissions for this tmp directory, as well as the staging directory specified in the `mapred-site.xml` or `conf.xml` files from the cluster.

Default Delimiter

- Sets the default delimiter of HDFS files created by TIBCO Data Science - Team Studio for storing intermediate results.

> **ⓘ** **Note:** The Work Flow preferences setting is used to set the delimiter at a global level across all workflows. However, you can override this global setting at the workflow variable level for a specific workflow. For details, see "Workflow Variables" in *TIBCO® Data Science Team Studio User Guide*.

Click **Save** to save changes. Click **Restore** to return to default values.

Datetime Formats Preferences

Use the Datetime Formats Preferences to modify the appearance of dates and times in the Workflow Editor.

> **❗** **Important:** You must be an administrator to edit preferences.

The Datetime formats are used by the Workflow Editor in the following cases:

- Using the Hadoop File Structure configuration to import `datetime` data from Hadoop File System text files. See the "Hadoop File" operator help in *TIBCO® Data Science Team Studio User Guide* for details.

- Using the Pig Execute operator to run Hadoop Pig script against `datetime` data fields and leverage Pig `Datetime` functions such as `GetMonth(datetime)` or `GetDay (datetime)`. See the "Pig Execute" operator help in *TIBCO® Data Science Team Studio User Guide* for details.

- Using the Variable operator to convert source `datetime` formats into new `datetime` formats or to convert `datetime` data into new data fields with a Pig `DateTime` function. See the "Variable" operator help in *TIBCO® Data Science Team Studio User Guide* for details.

- Using the Set operator to combine two or more data sources that contain `datetime` data types. See the "Set Operations" in *TIBCO® Data Science Team Studio User Guide* for details.

- Using the Null Value Replacement operator to replace null values with a default `datetime` format. See the "Null Value Replacement" operator help in *TIBCO® Data Science Team Studio User Guide* for details.

- Using the Row Filter operator to filter data by `datetime` formats or by a value derived from an applied the `Datetime` Pig function. See the "Row Filter" operator help in

TIBCO® Data Science Team Studio User Guide for details.

Procedure

1. To open the Preferences dialog, click the **Actions** menu, and then select **Preferences**.
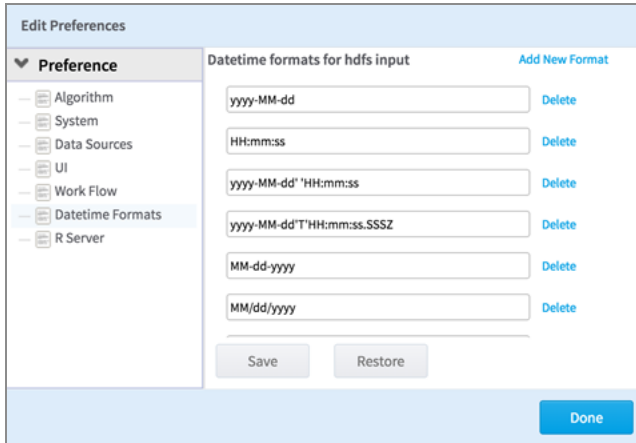
2. Select **Datetime Formats**.

   Adding a New Datetime Format

   To add a new input datetime format, click **Add New Format**.

   Click **Save** to save changes. Click **Restore** to return to default values.

   Custom Datetime Formats

   Datetime data type formats must follow Joda-Time API formatting. For more information about Joda-Time formatting, see Joda Time Formatting.

   Some commonly used Joda-Time pattern letters include:

   - Uppercase Y refers to the Year of the era (>0); lowercase y refers to the year.

   - Uppercase M refers to the Month of the year; lowercase m refers to the minute of the hour.

   - Uppercase D refers to the Day of the year (1-365); lowercase d refers to the day of the month (1-31).

   - Uppercase E refers to the Day of the week in text (Tuesday); lowercase e refers to the numeric day of the week (1-7).

   - Uppercase H refers to the Hour of the day (1-24); lowercase h refers to the clock hour of the half day (1-12).

- Uppercase S refers to the Fraction of a Second; lowercase s refers to the second of the minute (1-60).

The count of pattern letters determines the overall datetime format. For example, YYYY specifies a 4-digit year format, such as 2019.

# Administrator Options in TIBCO Data Science - Team Studio

The Administration page provides a set of tools for administrators to view information about their users, licensing, and running workflows in one place. Administrators can also download logs from the console.
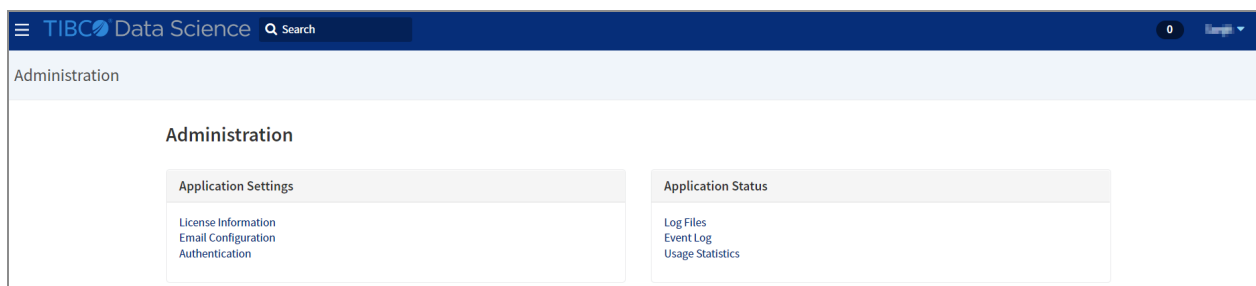
> ❗ **Important:** You must be an Administrator to see this option.

For information about downloading logs, see Download Logs.

To navigate to the Administration page, select **Administration** from the sidebar menu.
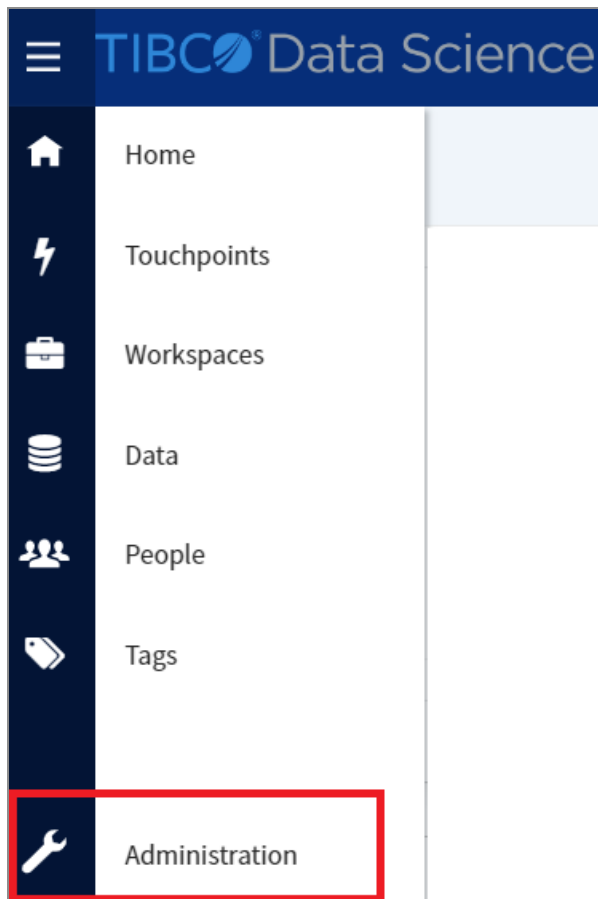
The Administration page appears. You can perform the following actions on the Administration page:

1. **License Information**: You can view the information on license and usage at a glance by selecting **License Information**. The following pieces of information are available.

License Information

Application Version

7.0.0.17813-0e846b7d6

License Expiration

Jan 01, 2030

Licensed Features

Feature	Enabled?
Model Operations	✔
Custom Operators	✔
Jobs & Scheduling	✔
Milestones	✔
Touchpoints	✔
Application API	✔

Licensed MAC Address

!!!!!!!!!!

MAC Address Details

Licensed Address: !!!!!!!!!!

All Available MAC Addresses:
- 02:42:ac:17:00:0b

Any one of the available MAC addresses can be used for the application license.

- **Application Version**: The version of TIBCO Data Science - Team Studio installed.

- **License Expiration**: The validity of the TIBCO Data Science - Team Studio license.

- **Licensed Features**: If the **Enabled?** column is checked, then those features are available in the current TIBCO Data Science - Team Studio installation. If the **Enabled?** column is not checked, then those features are not part of your current license plan.

    ◦ **Model Operations**: The ability to export and import custom models using PMML, Alpine Model (AM), and PFA format.

    ◦ **Custom Operators**: The ability to use and create Custom Operators. For more information, see "Custom Operators" in *TIBCO® Data Science - Team Studio Development Kit*.

    ◦ **Jobs & Scheduling**: The ability to schedule jobs such as loading or analyzing data.

    ◦ **Milestones**: The ability to set up milestones and track progress in TIBCO Data Science - Team Studio Connect.

    ◦ **Touchpoints**: The ability to create and use Touchpoints. For more information, see "Touchpoints" in the *TIBCO® Data Science Team Studio User Guide*.

    ◦ **Application API**: The ability to build access and use the TIBCO Data Science - Team Studio API. For more information, see "TIBCO Data Science - Team Studio API" in the *TIBCO® Data Science - Team Studio Development Kit*.

  - **Licensed MAC Address**: The MAC address that this version of TIBCO Data Science - Team Studio is registered with.

2. **Email Configuration**: You can configure information to use for emailing people about their TIBCO Data Science - Team Studio notifications. All new users to TIBCO Data Science - Team Studio receive a welcome email, and you can configure who receives notifications for finished workflows or jobs. Use this dialog to set up your SMTP server.

   For more information, see Email Configuration.

3. **Authentication**: This option is used by an administrator to configure SAML or select an LDAP or TIBCO Data Science - Team Studio authentication method.

   See Enabling LDAP Authentication for details.

4. **Log Files**: Administrators can access log files for the application that helps in

troubleshooting and maintenance. In addition, these files are very helpful to provide when filing a support request. For more information on the log files, see Download Logs.

5. **Event Log**: You can view and download a full event log for audit purposes. Information is tracked over time for action type (for example, upgrading a work file version), with details regarding the action performed by the user, the IP address, and the names of any work files, data sources, or jobs that were associated with the action. To download all the logs, click **Download Full Event Log**.

Event Log

Download Full Event Log

Event ID	Action	Actor ID	Actor User Name	Actor Sign In IP	Target	Target ID	Target Name	Time
9984	WorkFlowUpgradedVersion	3		127.0.0.1	Workfile	1726	PCA Shared	2022-11-02 12:25:47 UTC
9983	JobSucceeded	18		127.0.0.1	Job	8	test_legacy_1_job	2022-11-02 12:17:01 UTC
9982	WorkfileResult	18		127.0.0.1	Workfile	653	Randon_Sampling_old	2022-11-02 12:16:17 UTC
9981	WorkfileCreated	3		127.0.0.1	Workfile	1726	PCA Shared	2022-11-02 12:13:22 UTC
9980	JobSucceeded	18		127.0.0.1	Job	3	test_modern_job	2022-11-02 12:10:05 UTC
9979	WorkfileResult	18		127.0.0.1	Workfile	1148	join_modern_workflow_1	2022-11-02 12:09:21 UTC

6. **Usage Statistics**: Administrators can view usage statistics for users in TIBCO Data Science - Team Studio. The login data includes username, email, and creation date, as well as total logins, activity in the last 30 days, and the latest login date.

For more information, see Usage Statistics.

Email Configuration

The Administrator configures email notifications from TIBCO Data Science - Team Studio using the Email Configuration dialog.

Procedure

1. To navigate to the email configuration settings, click **Administration** from the sidebar menu, and then click **Email Configuration**.

Email Configuration   Send Test Email

Email Enabled   ✔   Edit Configuration

Email Message Settings
From Address   alpinenotif <alpinenotif@alpinenow.com>
Reply-to Address   dont_need_no_stinkin_emails@alpinenow.com

SMTP server configuration
Address   smtp.gmail.com
Port   587
HELO domain   gmail.com
Authentication   login
Username   alpinenotif@alpinenow.com
Password   ·············
STARTTLS Auto   ✔
OpenSSL Verify Mode   none

2. To edit settings, click **Edit Configuration**. The Editing email configuration dialog appears.

Email enabled	Select the checkbox to enable the email notification.
From Address	Specify the email address from which emails are sent.
Reply-to Address	Specify the email address that the reply email is sent by the user when they respond to an email notification.
Address	Specify the address of the SMTP server.
Port	Specify the Port of the SMTP server.
HELO Domain	HELO is an SMTP command sent by an email client when connecting to an email server. The command tells the server that the client wants to initiate an email transaction. It is followed by the client's domain name. For example, if you are using a Gmail SMTP server, the HELO Domain is gmail.com.
Username	Specify the username for the email service.
Password	Specify the password for the email service.

Authentication	Authentication method for sending passwords in the emails sent to new users. The following settings are available. • **Plain**: Send the password as non-secure plain text. • **Login**: Send the password Base64-encoded. • **Cram MD5**: Combines a Challenge/Response mechanism to exchange information and a cryptographic Message Digest 5 algorithm to hash important information.
STARTTLS Auto	When enabled, it automatically detects whether STARTTLS is enabled in your SMTP server and, if so, uses it.
OpenSSL Verify Mode	When using TLS, you can set how OpenSSL checks the certificate. This setting is useful if you need to validate a self-signed certificate or a wildcard certificate, or both. The following options are available: • **None** • **Peer** • **Client Once** • **Fail if no peer cert**

> ⚠ **Caution:** If the **Username** and **Password** are not set, SMTP authentication is not enabled. An SMTP server that does not require authentication is vulnerable to malicious activity. Proper network security, including host whitelisting and firewall and routing rules, should be ensured.

3. Click **Update** to save the changes.

4. After the configuration are updated, restart the server.

5. To test the new settings, click **Send Test Email**.

Usage Statistics

TIBCO Data Science - Team Studio provides login and activity statistics for users. This helps administrators understand how their system is being used.

Procedure

1. From the sidebar menu, click **Administration**.

2. To see the latest activity aggregates for TIBCO Data Science - Team Studio users, from the **Application Status** list, select **Usage Statistics**.

   The **Usage Statistics** page appears. This screen displays information about TIBCO Data Science - Team Studio user activity. The top line, highlighted yellow, shows usage aggregations across all users in your instance of TIBCO Data Science - Team Studio. Each following line shows statistics specific to an individual user.

Usage Statistics

User Name	User Email	User Created Date	Total Logins	Actions in the Last 30 Days	Last Login
—	—	—	1791	2631	2022-10-21 03:29:45 UTC
sroy	sanroy@tibco.com	2022-07-28 07:56:45 UTC	76	52	2022-10-21 03:29:45 UTC
ajangam	ajangam@tibco.com	2022-07-28 07:53:15 UTC	66	264	2022-10-21 02:55:34 UTC
siteadmin	siteadmin@example.com	2022-07-27 12:10:25 UTC	314	502	2022-10-20 22:24:45 UTC
mleehong	mleehong@tibco.com	2022-09-12 17:47:34 UTC	18	17	2022-10-20 19:32:22 UTC
yhu	yhu@tibco.com	2022-07-29 16:02:35 UTC	52	50	2022-10-20 13:05:07 UTC
gpaolini	gpaolini@tibco.com	2022-09-19 19:52:07 UTC	24	116	2022-10-20 10:06:20 UTC
ckilner	ckilner@tibco.com	2022-09-26 17:15:07 UTC	5	69	2022-10-19 20:47:10 UTC
vmanamala	vmanamal@tibco.com	2022-09-19 20:10:37 UTC	8	10	2022-10-19 18:37:05 UTC
drope	drope@tibco.com	2022-07-27 21:17:29 UTC	24	11	2022-10-19 18:06:17 UTC
aasoni	aasoni@tibco.com	2022-09-01 12:30:37 UTC	40	572	2022-10-19 04:57:25 UTC
bgokhale	bgokhale@tibco.com	2022-08-01 13:19:10 UTC	18	45	2022-10-19 03:30:57 UTC
afaskowitz	afaskowi@tibco.com	2022-09-19 20:15:28 UTC	14	37	2022-10-18 16:19:54 UTC

- **User Name**: TIBCO Data Science - Team Studio username.

- **User Email**: Email address associated with the user account.

- **User Created Date**: Creation date of the user.

- **Total Logins**: Total count of logins since the creation date of the user.

- **Actions in the Last 30 Days**: Rolling count of activities over the past 30 days.

- **Last Login**: Date of the latest login by the user.

Logins are tracked based on browser sessions when users go through the TIBCO Data Science - Team Studio authentication process, Single-Sign-On, or LDAP authentication.

Activities are counted for each action that a user takes while logged in to TIBCO Data Science - Team Studio, including things like saving a workflow, creating a new

workspace, changing a data source, and more. This is an aggregated form of the Event Log, highlighted in Administrator Options in TIBCO Data Science - Team Studio.

Data Visibility

Data visibility in TIBCO Data Science - Team Studio is a system for managing what data sources users can see and access within workspaces. The goal is to provide more granular control and security by allowing users to only know about certain data in the application.

The TIBCO Data Science - Team Studio data visibility system has four tenets. All of them come together to form a robust and cohesive data visibility offering.

Browsing Datasets In Your Workspace

As an administrator you can select and view the contents of any data source that you have the permissions to view.
Perform this task from the **Data Sources** section.

Before you begin
You must have administrative credentials, and you must have permissions to see the data source. These can be shared credentials, individual credentials, or workspace credentials.

- Shared credentials- If the data source has shared credentials, then everyone can access it using the same set of credentials. You can see the tables and use them in TIBCO Data Science - Team Studio workflows.

- Individual credentials- If you have individual credentials for this data source, you are prompted to enter them. Everyone that accesses the data source must use his or her credentials. If you do not have credentials, or if you do not know your data source credentials, then contact the person in charge of your data source.

- Workspace credentials- As a data administrator, you can specify credentials for the entire workspace. Everyone in the workspace shares the set of credentials, but those credentials apply only to the active workspace.

> ℹ️ **Note:** Workspace credentials take precedence over individual credentials if both exist in this workspace.

Procedure

1. To associate more data sources to this workspace, click **Associate a Data Source**.
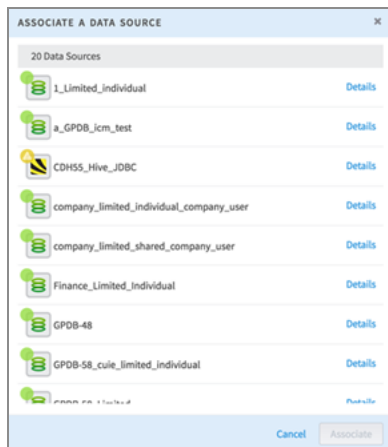

   A dialog is displayed containing the data sources not yet associated. You can also view details about the data sources including the owner and the host address.

2. Select the data source that you want to add, and then click **Associate**.

Browsing Datasets In the Entire Application

From the Data section, you can see a list of all data sources connected to the application, and you can add more data sources.

From the sidebar menu, click **Data**.
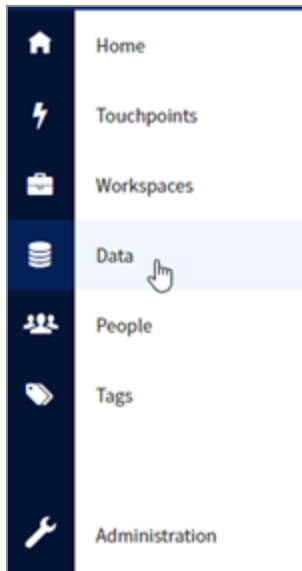
Result

From this section of the application, you can also control options such as data source visibility and permissions.

Controlling Data Source Visibility

Data sources can be global; that is, they can be designated as "public" or they can be scoped as "limited", restricting their visibility and available to only the workspaces with which they are associated.

Perform this task from the main Data Sources section of the application, accessible from the **Data** sidebar menu. You cannot edit data sources from a workspace.

Before you begin

You must have Data Administrator credentials to control data source visibility.

Procedure

1. Select the data source.

2. In the contextual sidebar, click **Edit Data Source**.

3. In the **Workspace Visibility** dropdown list box, select either **Public** or **Limited**.

> ℹ️ **Note:** A data source set to **Limited** must be associated manually with a workspace for members of that workspace to use the data source.

4. Click **Save Configuration**.

Controlling Data Source Permissions

Besides changing data source visibility, you can also change the level of permissions on the data source.
A user might be able to see a data source in the Data Sources section in their workspace, but they won't be able to access it unless they have permission.

Perform this task from the main Data Sources section of the application, accessible from the **Data** sidebar menu. You cannot edit data sources from a workspace.

Before you begin
The permission options range from shared by everyone to workspace-specific credentials to individual credentials for each user. Depending on your use case, you can change this permission scheme at any time.

Procedure
1. Select the source from the data primary navigation.

2. Click **Edit**.

   In the following example, the data source has Individual credentials and only one

user account has access.
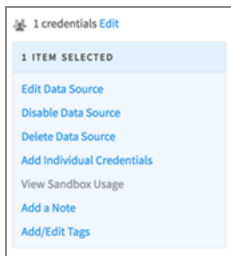
3. In the Data Source Permissions dialog, perform one of the following tasks.
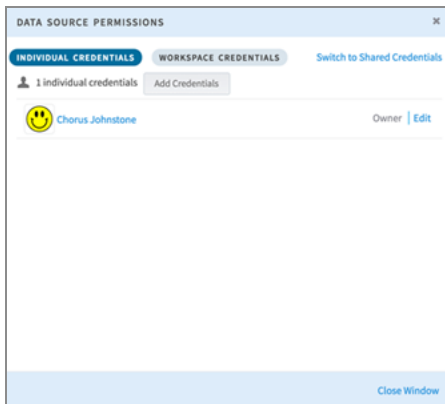
- View the individual credentials, as well as credentials granted to any workspaces.

- Click **Add Credentials** to add new user accounts for individual credentials.

> **Note:** These credentials must also be present on the data source in order to validate properly.

- Switch to shared credentials (which means that one set of credentials is used by everyone), by clicking **Switch to Shared Credentials**.

Adding Data to a Workspace

For limited-visibility data sources, you must associate them with each workspace you want them to be accessible to.

Before you begin

You must be either a data administrator or an application administrator.

Procedure

1. From the sidebar menu, click **Data**.

2. Select the data source.

3. Click **Associate Data Source to a Workspace**.

   If this option is not displayed, ensure that the data source you have selected is Limited visibility. If it is Public, the data source is already available to all workspaces.

   > **Note:** Associating a data source to a workspace makes it visible, but users must still have proper credentials to access the data within. You can set this up in one of three ways.
   >
   > - **Shared credentials**: Everyone uses the same set of credentials-this is the most broad scope.
   >
   > - **Workspace credentials**: Everyone in the workspace uses the same set of credentials. See Controlling Data Source Permissions.
   >
   > - **Individual credentials**: Each member that wants to access the data source must have credentials for that data source.

Data Source Associations

In TIBCO Data Science - Team Studio, you must associate data sources with workspaces.

This process is similar to (but not exactly like) how datasets are associated with workspaces. Associated data sources are displayed in a new workspace section labeled **Data Sources**.

- All public data sources are automatically associated to workspaces.

- Limited data sources must be manually associated to a workspace.

Workfiles (workflows, SQL, and notebooks) can use only the data sources available to the workspace. Datasets in the workspace must come from data sources available to that workspace.

For more information, see "Data Sources" in the *TIBCO® Data Science Team Studio User Guide*.
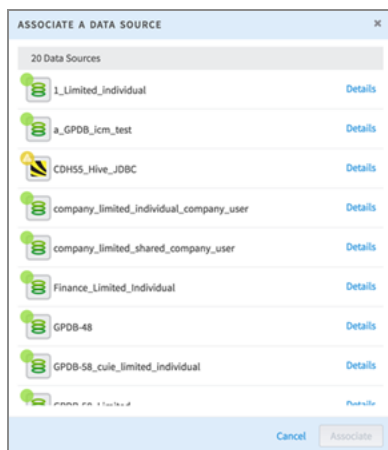
Associating a Data Source

Follow these steps to associate a data source to a workspace in TIBCO Data Science - Team Studio.

Before you begin

You must be a member of the workspace and have the Data Administrator role.

Procedure

1. Navigate to a workspace.

2. Select the **Data Sources** section.

3. Click **Associate a Data Source**, and then from the dialog, select a data source.

4. Click **Details** to see more information about its data source.

5. After choosing the data source, click **Associate** to save the changes.

Data Source Credentials

A data source credential option is available in TIBCO Data Science - Team Studio that provides limited visibility.

You can use this credential option when you need to specify a data source that is not just "individual" and is not "shared by all." You can provide authorization credentials that specify only that workspace.

Each workspace can have its own set of shared credentials. These credentials are not shared globally across the application, but are scoped to activity just within the workspace. This feature is available for database data sources only.

> **ℹ** **Note:** Workspace credentials override any existing credentials when accessing the data source in that workspace.
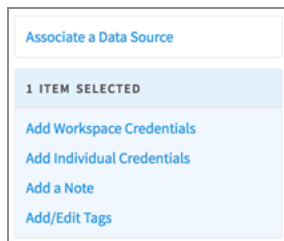
From the **Data Sources** section of the workspace, specify **Add Workspace Credentials** or **Add Individual Credentials**.

Data Administrators

The Data administrator has permissions to see all data sources and manage the data source associations for workspaces. However, this role does not have full application administration permissions.

You can find the Data Administrator role in the Roles section of your profile. You can have only one administration role at a time. Like the Application Administrator role, it is not controlled by licensing limits, so you can have as many administrators in TIBCO Data Science - Team Studio as you need.

To learn more about the Data Administrator role, see TIBCO Data Science - Team Studio Licensing.

Data Source States

Data source states allow users to have more control over their data sources. Some of the features are:

- Users can save incomplete data source configurations and come back to them later.

- Users can disable a data source from the TIBCO Data Science - Team Studio web application.

- Users can view the status of a data source at a glance.

Stateful Data Sources

Data sources within TIBCO Data Science - Team Studio have states that reflect their status in the system. There are four different states:

- Online

- Offline

- Incomplete

- Disabled

These states have an impact on whether a user can browse them or use them in workflows or sandboxes.

Online State

A data source becomes active when a user enters correct connection information and a successful connection is made. An active data source can be used and TIBCO Data Science - Team Studio indexes it. In the list of data sources, this is displayed with a green status marker and the following tooltip:
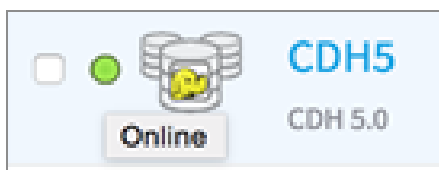
Offline State

This data source was once normal, active, and correctly configured. However, for some reason, the data source is no longer connecting correctly. The data source might still be indexed by TIBCO Data Science - Team Studio, but might not be up to date. Verify that the data source is connected before using it in a workflow.

Incomplete State

An incomplete data source is one that a user has begun to add, but has not yet filled out every parameter and has not yet connected to the underlying data source. For that reason, the data source is still in a draft state. Because users might not have all the information needed to configure a new data source at one time, this state allows users to save progress and return to it later. The data source is not indexed or available until it is properly configured.

To save a data source that is incomplete or has invalid information, click **Save**. The following dialog is displayed.
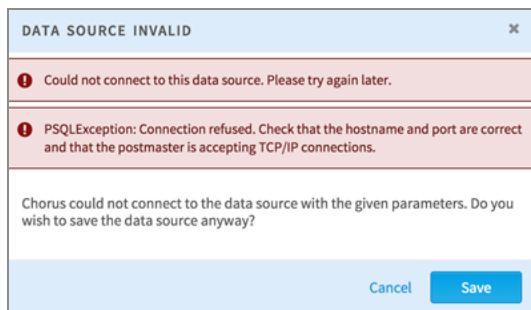
If you click **Save** in this dialog, the data source is in the Incomplete state and is not usable until a complete connection occurs. If you click **Cancel**, you are redirected to the **Add Data Source** screen to edit the configuration.

> **Note:** For database data sources, the user must fill out all fields with red asterisks. The user can save the data source configuration with invalid data, rendering it Incomplete.
>
> For Hadoop data sources, the user must fill out all fields with red asterisks, but can save the configuration while leaving the extra connection parameters section empty.

After all of the required fields are completed, the data source can be validated and used. In the list of data sources, this is displayed as plain, read-only text and a red square:
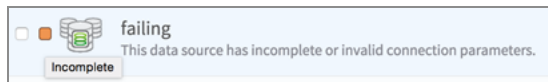

Disabled State

A disabled data source is one that has been active and correctly configured, but has been manually disabled by an application administrator. A disabled data source cannot be used and cannot be browsed by application users. TIBCO Data Science - Team Studio does not index or check the status of disabled data sources.

An administrator can reactivate a disabled data source by selecting **Enable Data Source** from the sidebar menu. When the data source is reactivated, TIBCO Data Science - Team Studio attempts to test it and validate the configuration. If this works normally, it becomes Active. If the data source does not validate, it remains disabled.

In the list of data sources, a disabled data source displays as plain read-only text, without a link. It displays a gray circle as a status marker:
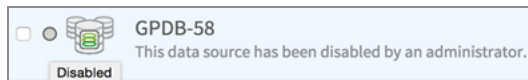
Data Source States in Workflows

A data source's state affects how workflows appear to the user.

When selecting a workflow, users can now see the status of the underlying data source. For example:
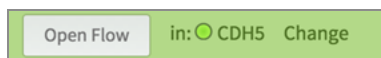
The icon changes depending on the data source status.

> ⚠️ **Caution:** Existing workflows with an underlying incomplete or disabled data source cannot be opened.

Manage TIBCO Data Science - Team Studio Users

The **People** page displays a complete list of TIBCO Data Science - Team Studio users, developers, and administrators.

On this page, you can do the following:

- See a list of all users, developers, and administrators, sorted by last name (change the default to sort by first name).

- Select a username to go to that user's page for more information.

> ℹ **Note:** Only an administrator can add or delete a user.

Managing User Profiles

Users can manage their profile information details by clicking their names, and then clicking **Your Profile**.

The resulting TIBCO Data Science - Team Studio User Management user interface provides the following options:
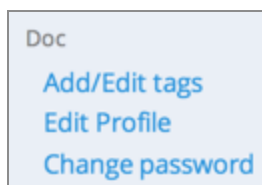
Before you begin

You must be an administrator to change user profiles.

Procedure

1. To associate specific topics that are discovered in a search, select **Add/Edit Tags**.
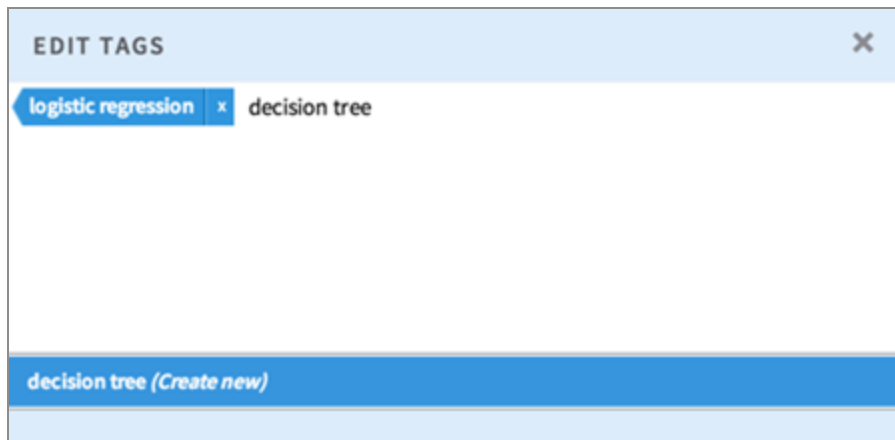
2. To change basic profile information, select **Edit Profile** and edit the following properties as necessary:

   Account Information

   - **User Name**: The unique name for each user.

   Application Roles

   - Administrators can give other users administrator and/or developer privileges.

   Personal Information

   - You can edit the fields **First Name**, **Last Name**, **Email**, **Title**, **Department**, and **Description**. **First Name**, **Last Name**, and **Email** are required fields.

   You can subscribe to receive email notifications.

3. Click **Save Changes** to keep your changes, or click **Cancel** to revert to the previous version.

4. To reset your password, select **Change password**.

Add a New Person

This topic describes how to add a new user to the **People** page, which includes a complete list of TIBCO Data Science - Team Studio users, developers, and administrators.

Before you begin

You must be an administrator to add a new person.

Procedure

1. Open the **People** page and click **Add Person**.

   The **New Person** page is displayed.

2. Fill in the required fields, including a **Username** and **Password** for the new person. The account owner can change the password later.

   > ℹ️ **Note:** If TIBCO Data Science - Team Studio is configured with LDAP/AD (see the Enabling LDAP Authentication), there are no password fields.

   Multiple administrators can exist within TIBCO Data Science - Team Studio, and an administrator can make any other user an administrator. To do this, select the **Application Administrator** checkbox.

   A person can have one of four roles within the application. Each role has a different permissions level. Analytics Developers have the highest permissions level, and Business Users have the lowest. For more information on the application roles, see User Roles.

3. Click **Add Person**.

   The new account is created.

4. To view the list of people, select **People** from the Quick menu on the left sidebar.

   The Contextual Sidebar includes the following actions.

   - **Add Tags** or **Edit Tags** (any user)

   - **Edit Profile** (the selected user or an administrator only)

- **Delete Person** (an administrator only)

- **Change password** if not using LDAP (the selected user or an administrator only)

User Roles

User roles describe a person's expertise and role in the project. Each role enables a set of licenses that correspond to an administration or application role. This helps with team communication and messaging of what each person works on. You can have only one application and one administration role at a time. The available roles are:

Application Role

Determines the user permissions to work with the analytic workflows and data. The total number of application roles depends on the application license. The available application roles are:

Role	Definition
Analytics Developer	Analyzes the sets of data and finds common patterns, themes, and trends using the latest machine-learning techniques. The analytics developer focus on developing and implementing analytical solutions. They have the broadest permission. Analytical developers collaborate closely with data analysts and business users to comprehend their needs and transform them into practical analytical solutions.
Data Analyst	Examine large sets of data to extract meaningful insights and make data-driven recommendations. They employ various analytical tools to carry out tasks like data cleaning, data exploration, statistical analysis, and data visualization. Data analysts collaborate closely with business users to comprehend their goals and offer practical information to support decision-making.
Collaborator	Refers to a team member who actively participates in the data analysis process. Individuals from several departments or job roles who provide domain-specific knowledge or skills are considered collaborators. They provide input on data requirements, offer insights and perspectives, and work with data analysts to ensure that the analysis aligns with the

Role	Definition
	organization's goals.
Business User	A professional who applies data and analytics to their roles to help them make informed decisions. Business users rely on the insights provided by data analysts and analytical developers to comprehend market trends, customer behavior, and operational performance. These roles often collaborate with one another and work together to use data for better business outcomes. They have the fewest permissions.

Administration Role

Determines the user permissions to manage the items in the application. You can have only one administration role at a time. The available administration roles are:

Role	Definition
Application Administrator	Responsible for managing and maintaining the application. Their primary focus is on the operational aspects of applications, ensuring their availability, performance, and security. The number of application administrators depends on the licensing limits.
Data Administrator	Manages the data source associations for workspaces, ensuring data quality, security, and compliance. There are no limits on the number of data administrators.

Establish Your Identity

TIBCO Data Science - Team Studio is intended to expose you to the knowledge and expertise of colleagues in your organization you do not work closely with. After a TIBCO Data Science - Team Studio administrator creates your credentials, opens your User Profile and provides information about yourself to help that exposure and your interactions.

As a user, you can edit your profile in one of two ways:

- From the contextual sidebar of the **People** page, click **Edit Profile**.

- On the home page, on the global navigation bar, click your name, and then click **Your Profile**.

Procedure

1. When your profile opens in the main panel, in the contextual sidebar, click **Edit Profile**. Your profile page appears. You can edit nearly every aspect of it, including changing your password (you cannot, however, change the username assigned by the administrator who added you). For example, you might add a picture and fill out the relevant fields with information you believe could be useful to others in your organization.

2. When you are finished, click **Save Changes**.

   Your updated profile is visible to you and other accounts.

TIBCO Documentation and Support Services

For information about this product, you can read the documentation, contact TIBCO Support, and join TIBCO Community.

How to Access TIBCO Documentation

Documentation for TIBCO products is available on the Product Documentation website, mainly in HTML and PDF formats.

The Product Documentation website is updated frequently and is more current than any other documentation included with the product.

Product-Specific Documentation

The documentation for this product is available on the TIBCO® Data Science - Team Studio Product Documentation page.

How to Contact Support for TIBCO Products

You can contact the Support team in the following ways:

- To access the Support Knowledge Base and getting personalized content about products you are interested in, visit our product Support website.

- To create a Support case, you must have a valid maintenance or support contract with a Cloud Software Group entity. You also need a username and password to log in to the our product Support website. If you do not have a username, you can request one by clicking **Register** on the website.

How to Join TIBCO Community

TIBCO Community is the official channel for TIBCO customers, partners, and employee subject matter experts to share and access their collective experience. TIBCO Community offers access to Q&A forums, product wikis, and best practices. It also offers access to extensions, adapters, solution accelerators, and tools that extend and enable customers to gain full value from TIBCO products. In addition, users can submit and vote on feature

requests from within the TIBCO Ideas Portal. For a free registration, go to TIBCO Community.

Legal and Third-Party Notices

SOME CLOUD SOFTWARE GROUP, INC. ("CLOUD SG") SOFTWARE AND CLOUD SERVICES EMBED, BUNDLE, OR OTHERWISE INCLUDE OTHER SOFTWARE, INCLUDING OTHER CLOUD SG SOFTWARE (COLLECTIVELY, "INCLUDED SOFTWARE"). USE OF INCLUDED SOFTWARE IS SOLELY TO ENABLE THE FUNCTIONALITY (OR PROVIDE LIMITED ADD-ON FUNCTIONALITY) OF THE LICENSED CLOUD SG SOFTWARE AND/OR CLOUD SERVICES. THE INCLUDED SOFTWARE IS NOT LICENSED TO BE USED OR ACCESSED BY ANY OTHER CLOUD SG SOFTWARE AND/OR CLOUD SERVICES OR FOR ANY OTHER PURPOSE.

USE OF CLOUD SG SOFTWARE AND CLOUD SERVICES IS SUBJECT TO THE TERMS AND CONDITIONS OF AN AGREEMENT FOUND IN EITHER A SEPARATELY EXECUTED AGREEMENT, OR, IF THERE IS NO SUCH SEPARATE AGREEMENT, THE CLICKWRAP END USER AGREEMENT WHICH IS DISPLAYED WHEN ACCESSING, DOWNLOADING, OR INSTALLING THE SOFTWARE OR CLOUD SERVICES (AND WHICH IS DUPLICATED IN THE LICENSE FILE) OR IF THERE IS NO SUCH LICENSE AGREEMENT OR CLICKWRAP END USER AGREEMENT, THE LICENSE(S) LOCATED IN THE "LICENSE" FILE(S) OF THE SOFTWARE. USE OF THIS DOCUMENT IS SUBJECT TO THOSE SAME TERMS AND CONDITIONS, AND YOUR USE HEREOF SHALL CONSTITUTE ACCEPTANCE OF AND AN AGREEMENT TO BE BOUND BY THE SAME.

This document is subject to U.S. and international copyright laws and treaties. No part of this document may be reproduced in any form without the written authorization of Cloud Software Group, Inc.

TIBCO, the TIBCO logo, the TIBCO O logo, Spotfire, and Alpine Data Labs are either registered trademarks or trademarks of Cloud Software Group, Inc. in the United States and/or other countries.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

All other product and company names and marks mentioned in this document are the property of their respective owners and are mentioned for identification purposes only. You acknowledge that all rights to these third party marks are the exclusive property of their respective owners. Please refer to Cloud SG's Third Party Trademark Notices (https://www.cloud.com/legal) for more information.

This document includes fonts that are licensed under the SIL Open Font License, Version 1.1, which is available at: https://scripts.sil.org/OFL

Copyright (c) Paul D. Hunt, with Reserved Font Name Source Sans Pro and Source Code Pro.

Cloud SG software may be available on multiple operating systems. However, not all operating system platforms for a specific software version are released at the same time. See the "readme" file for the availability of a specific version of Cloud SG software on a specific operating system platform.