TIBCO® Data Science Team Studio Release Notes

Version 6.6.0

June 2021

Contents

TIBCO® Data Science Team Studio Release Notes

The release notes for this product version are provided to inform you of new features, known issues, and issues from previous releases that have been closed.

These release notes are for TIBCO® Data Science Team Studio version 6.6.0.

Team Studio is an environment built to facilitate the analytics process at all levels of the business, from defining a problem to driving business action.

New Features

The following new features were added to version 6.6.0 of Team Studio.

Support for TIBCO Data Virtualization

Team Studio has been tested with TIBCO Data Virtualization for connecting to Amazon EMR S3 views, Cloudera Data Platform (CDP 7) views, and Snowflake views and tables. This release provides limited support for database operators. You should be able to connect Team Studio to other data sources through TIBCO Data Virtualization, but other connections have not been certified. For more information on tested operators, see TIBCO Data Virtualization Compatibility in the *TIBCO® Data Science Team Studio User's Guide*.

To use Team Studio with TIBCO Data Virtualization, the administrator must download and follow the installation instructions for the integration package, which is available on the TIBCO Exchange.

User interface improvements

- The Add a Note dialog box for notifying workspace users has a simplified design.
- The Select Columns dialog box (Distinct operator) now has an **Add All** button to add all columns at once.
- The Ordered Columns dialog box (Reorder Columns operator) now has an **Add All** button to add all columns at once.

Improvement to using Parquet files

- Parquet files that contain decimal type columns are now supported.
- Team Studio can accept partitioned parquet files as inputs to a workflow.

Multiple file selection

When you upload files to Hadoop, you can now select multiple files in the same operation.

Export to ModelOps

PMML versions of models can now be exported to TIBCO ModelOps.

Support for running temporary functions in Big Query using SQL Execute

Team Studio now supports running temporary functions in Big Query using SQL Execute. See the help for SQL Execute for more information.

Improvements to the PCA operator

Various reliability, scalability, and usability improvements have been made to the PCA operator.

New operators

You can use the following new operators in your workflows. See the help for these operators in the *TIBCO® Data Science Team Studio User's Guide*.

Operator name	Description
Wide Data Variable Selector - Correlations Wide Data Variable Selector - Chi Square / Anova	From a very large data set (that is, one whose variables number in the thousands or millions), these operators produce a new data set with correlations and significance statistics for each predictor (X) variable against a user-specified dependent (Y) variable. • For Wide Data Variable Selector - Correlations, your dependent variable must be continuous. • For Wide Data Variable Selector - Chi Square / Anova, your dependent variable must be categorical.
Times Series SAX Encoder	This operator produces a new data set with one or more columns that contains the time series ID and a discretized string representation of the original time series.

Changes in Platform Support

This topic lists changes in platform support in version 6.6.0 of Team Studio.

For a complete list of all the currently supported platforms, see the System Requirements.

Support for RHEL 7.8 and 7.9

Team Studio now supports running on Red Hat Enterprise Linux versions 7.8 and 7.9.

Changes in Functionality

There are no changes in functionality in version 6.6.0 of Team Studio.

Migration and Compatibility

If you are migrating from a previous release to version 6.6.0 of Team Studio, contact Team Studio Support for assistance.

There are no migration or compatibility issues for this release of Team Studio.

Deprecated and Removed Features

The following features have been deprecated or removed from version 6.6.0 of Team Studio.

Deprecated Features

Hadoop data sources

Support for Hortonworks HDP Hadoop is deprecated and no longer supported.

Databases

Support for directly connecting to the following databases is deprecated and no longer supported. Connecting to these databases using TIBCO Data Virtualization should work.

- Oracle_JDBC
- Teradata
- MySQL
- Oushu

- MariaDB

Operators

The PCA Apply operator is deprecated.

Browser support

Support for the Firefox browser is deprecated and is no longer supported.

Removed Features

Python

Support for Python 2 has been removed in this release.

Closed Issues

Version 6.6.0 of Team Studio contains assorted bug fixes.

Key	Summary
DSC-22897	Summary Statistics now shows the relevant statistics (min, max, and so on) for `datetime` data types.
DSC-25382	The schema from the Pivot operators is now dynamically passed to downstream operators after step-running the operator.
DSC-25394	For Random Sampling, the tooltip now explains that the random seed must be between 0 and 1 (inclusive).
DSC-25396	The alias extensions in the Hadoop Join operator have been simplified to be more tolerant to upstream changes.
DSC-25399	In the Join operator, columns are now selectable with a single click.
DSC-25426	Metadata is now passed correctly when a dataset from PostgreSQL is imported into a Python Notebook.
DSC-25538	Team Studio Jupyter notebooks now correctly infer the data types and handle missing/NaN values when you import or export data to or from a database data source.
DSC-25612	When you download the results of an operator, help text is now available for **Number of Lines**.
DSC-25672	Documentation is updated to note that the value of lambda in regularized regressions can be any number greater than zero.
DSC-25694	Importing a data set with missing values into a BigQuery data source now correctly sets those cells to NULL.
DSC-25739	The list of files in the data tab of the workflow editor is now sorted properly. It no longer shows upper-case file names first.

Known Issues

The table lists known issues in version 6.6.0 of Team Studio.

Key	Summary
DSC-25884	**Containerized notebooks not accessible by Team Studio** The DockerSpawner version of Python Notebooks prevents notebooks from being accessed from Team Studio UI. **Workaround** Use the LocalSpawner version of Python Notebooks for Team Studio. You can do this by choosing the LocalSpawner during the installation of the Python Notebooks server.
DSC-25838	**Parquet library conflict** For EMR data sources, operators cannot write or infer parquet files successfully if there is a parquet library conflict. **Workaround** In this case, use Avro or CSV as the output format from operators.
DSC-25835	**Summary** R Execute does not work for EMR S3 and CDP7. **Workaround** None.
DSC-25810	**Summary** The column names for TIBCO Data Virtualization (TDV) metadata tables are stored in the order in which they are modified, so when Team Studio queries the metadata table, the column names can be displayed in the wrong order in its user interface. **Workaround** You can correct this problem by republishing the view/table in TDV with the correct column order.
DSC-25806	**Summary** Python Execute with Notebooks is not supported for TDV Object Storage. **Workaround** None.

Key	Summary
DSC-25791	**Summary** The **Classifier** and **Predictor** operators do not work for the TIBCO Data Virtualization (TDV) EMR S3 integration and the CDP HDFS integration. **Workaround** None.
DSC-25785	**Summary** Notebook can fail when it writes to TIBCO Data Virtualization S3. **Workaround** When you insert into a TIBCO Data Virtualization data source from a Python notebook, make sure that the **ChorusCommander** field **tdv_scripts_location** matches the one set in your `alpine.config` file. By default, **cc.tdv_scripts_location** is set to `"Util"`. For example, if your `startBatch`, `addBatch`, and `loadBatch` scripts are in a TIBCO Data Virtualization schema called Utilities, then use `cc.tdv_scripts_location='Utilities'` before calling `cc.write_output_table`.
DSC-25783	**Summary** Copy Between Databases is not supported for TIBCO Data Virtualization Object Storage. **Workaround** None.
DSC-25780	**Summary** Creating a view using SQL Execute operator with TIBCO Data Virtualization S3 **Workaround** Only SQL that is supported by the TIBCO Data Virtualization JDBC driver works with the SQL Execute operator. For more information, see the TIBCO Data Virtualization reference at https://docs.tibco.com/pub/tdv/8.4.0/doc/pdf/TIB_tdv_8.4.0_ReferenceGuide.pdf
DSC-25776	**Summary** For the TIBCO Data Virtualization (TDV) data source, the **Copy to Database** operator option **Copy mode > Parallel** does not work. The operator displays an explicit error message, specifying that the user should use the Simple mode for copying. **Workaround** Use **Simple** mode for copying with a TDV data source.

Key	Summary
DSC-25774	**Summary** The **Copy to Hadoop** operator does not support Appending to an existing Parquet or Avro file on HDFS. This process is not supported by the Hadoop API used by Team Studio. **Workaround** None.
DSC-25747	**Summary** The Pivot operator choice **Use Array** does not work with a TIBCO Data Virtualization data source because TIBCO Data Virtualization does not support array types. **Workaround** None.
DSC-25664	**Summary** The **Copy to Database** operation is not supported for TIBCO Data Virtualization Object Storage. **Workaround** None.
DSC-25657	**Summary** An exported workflow lost many operators from the original workflow. **Workaround** The Team Studio administrator, using ssh access, can edit the file `workfileid.inf` to change the version number in the file to the latest version of the work file. You can find this file under `/usr/chorus/shared/ALPINE_DATA_REPOSITORY/flow/Public/`. The latest version of the file can be found at `/usr/chorus/shared/ALPINE_DATA_REPOSITORY/flow_version directory`.
DSC-25472	**Summary** PySpark notebook that uses avro can fail. **Workaround** You must change the PySpark initialization to `os.environ['PYSPARK_SUBMIT_ARGS'] = "--master yarn-client --num-executors 1 --executor-memory 1g --packages org.apache.spark:spark-avro_2.11:2.4.3 pyspark-shell"`. Older PySpark notebooks accessing avrò files must be migrated to use this change for the notebooks to work with the new notebook server. This is a known issue with Spark. See https://issues.apache.org/jira/browse/SPARK-27623 for more information.

Key	Summary
DSC-23858	**Summary** After you update the email configuration through the user interface, you must restart the server for the configuration changes to take effect. **Workaround** None.
DSC-23804	**Summary** For the data source format BigQuery, a Naive Bayes analysis fails if a datetime column is included in the data. **Workaround** None.
DSC-23670	**Summary** Copy to Hadoop and Copy to Database operators do not work in parallel mode between GPDB5 and Hadoop Data Sources. **Workaround** None
DSC-23629	**Summary** For BigQuery data sources, using DML issued through JDBC can cause problems because of statement size limitations; therefore, you should avoid using Team Studio operators to copy data into BigQuery. **Workaround** Google recommends that BigQuery data source users insert data using the Google API. For more information on BigQuery limitations, see https://cloud.google.com/bigquery/quotas.
DSC-23587	**Summary** When you use a Google BigQuery data source, Copy Between Databases does not work with a Greenplum database. **Workaround** None.
DSC-23585	**Summary** When you use a Google BigQuery data source in parallel mode, Copy to Database and Copy to Hadoop do not work. **Workaround** Use simple mode to copy the data to and from a BigQuery data source.

Key	Summary
DSC-23121	**Summary** The Hadoop Hive 1.1 client libraries that Team Studio is built with do not yet support TLS/SSL encryption when combined with Kerberos authentication. **Workaround** Do not enable Hive TLS/SSL encryption when Kerberos authentication is also enabled for your Hadoop cluster.
DSC-22930	**Summary** The compression format Snappy is not supported by the file format CSV. **Workaround** For this file format, use another compression format.
DSC-22859	**Summary** A Spark normalization run on a very wide data set fails due to an issue in Spark versions prior to 2.3. **Workaround** Use the MR version on a very wide dataset, or set the option to not use Spark.
DSC-22778	**Summary** In the Hadoop Join operator, clicking a check box to add or remove a column appears to not work. **Workaround** Click twice to select the check box for adding or removing columns in a Hadoop Join operator. (This problem is due to a limitation in the user interface framework.)
DSC-17859	**Summary** Avro file formats do not support null values and can cause errors when running downstream operators. **Workaround** Clean your data prior to using the Avro format by using the **Null Value Replacement** operator.
DSC-342	**Summary** Non-Spark operators can fail for a newly-added Active Directory user. **Workarounds** • The user can use the Spark versions of the operators in the workflow. • The administrator can issue a `kinit` command for the mapred user on the cluster.

Key	Summary
DSC-341	**Summary** Hive does not handle new lines in text. **Workaround** Clean all text data prior to processing.

TIBCO Documentation and Support Services

How to Access TIBCO Documentation

Documentation for TIBCO products is available on the TIBCO Product Documentation website, mainly in HTML and PDF formats.

The TIBCO Product Documentation website is updated frequently and is more current than any other documentation included with the product. To access the latest documentation, visit https://docs.tibco.com.

Product-Specific Documentation

The following documents for this product can be found in the TIBCO Documentation Library.

- *TIBCO® Data Science Team Studio Release Notes*
- *TIBCO® Data Science Team Studio System Requirements*
- *TIBCO® Data Science Team Studio Version and Licensing*
- *TIBCO® Data Science Team Studio Installation and Administration*
- *TIBCO® Data Science Team Studio User's Guide*
- *TIBCO® Data Science Team Studio Development Kit*

How to Contact TIBCO Support

You can contact TIBCO Support in the following ways:

- For an overview of TIBCO Support, visit http://www.tibco.com/services/support.
- For accessing the Support Knowledge Base and getting personalized content about products you are interested in, visit the TIBCO Support portal at https://support.tibco.com.
- For creating a Support case, you must have a valid maintenance or support contract with TIBCO. You also need a user name and password to log in to https://support.tibco.com. If you do not have a user name, you can request one by clicking Register on the website.

System Requirements for TIBCO® Data Science Team Studio

For information about the system requirements for Team Studio, see *TIBCO® Data Science Team Studio System Requirements*.

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 https://community.tibco.com.

Legal and Third-Party Notices