Teradata Vantage Plugins for Dataiku Data Science Studio

Table of Contents

I. Introduction

Dataiku Data Science Studio (DSS) is a collaborative platform that enables teams of people with different data expertise, such as data engineers, data scientists and analysts, to work together efficiently. Dataiku DSS provides a set of built-in recipes or operations that can be applied to transform or analyze a dataset. It also allows users to create their own recipes in Python, SQL or R. The DSS plugins are custom reusable recipes that can only be written in Python.

The present guide outlines installation and usage of 2 DSS plugins that enable you to interact with Teradata Vantage systems; namely, the Teradata Vantage Analytic Functions Plugin, and the SCRIPT Table Operator Plugin.

Teradata Vantage Analytic Functions Plugin

The Teradata Vantage Analytic Functions Plugin for Dataiku DSS integrates about 180 of the Vantage Machine Learning Engine (MLE) analytic functions, by providing a user-friendly, easy-to-use, no-SQL interface for the functions in the Dataiku DSS environment. The Vantage analytic functions can be accessed through the [`+RECIPE`] menu of the FLOW view of a Dataiku project, and are grouped into nine categories:

- Time Series, Path and Attribution Analysis
- Ensemble Methods
- Text Analysis
- Naïve Bayes
- Graph Analysis
- Association Analysis
- Statistical Analysis
- Cluster Analysis
- Data Transformation

The Teradata Vantage Analytic Functions Plugin for Dataiku DSS integrates about 180 of the Vantage Machine Learning Engine (MLE) analytic functions, by providing a user-friendly, easy-to-use, no-SQL interface for the functions in the Dataiku DSS environment. The Vantage analytic functions can be accessed through the [+RECIPE] menu of the FLOW view of a Dataiku project, and are grouped into nine categories:

SCRIPT Table Operator Plugin

The SCRIPT Table Operator Plugin allows the execution of R or Python scripts inside the Teradata Database. The plugin will take an R or Python script within a DSS notebook, or an R or Python script uploaded to the plugin and install the scripts and other related files (i.e. saved models in RDS or pickle files) on the Advanced SQL Engine.

Similar to the Teradata Vantage Analytic Functions Plugin, the SCRIPT Table Operator Plugin translates the user-requested tasks in the plugin into SQL queries, which are then sent to a connected Vantage system to set up and invoke the SCRIPT Table Operator.

II. Requirements

1. **Dataiku Data Science Studio version 5.1.2 or later**

   Dataiku DSS enterprise edition is required to import datasets from Vantage tables. Dataiku offers both downloadable and online options which can be obtained from the Dataiku website. The downloadable option can be configured to use the free or the enterprise edition, while the online option only comes in enterprise edition with free trial for a period of 14 days. A comparison between the two editions can be seen in the features table for Dataiku DSS Editions.

   Teradata Vantage Analytic Functions plugin has been tested on Dataiku DSS version 5.1.2.

2. **Plugin**

   To use the Teradata Vantage Analytic Functions plugin, you need the compressed file "TeradataVantageFunctionsPlugin.zip" that contains the Teradata Vantage Analytic Functions plugin software and metadata.

   To use the SCRIPT Table Operator plugin, you need the compressed file "TeradataVantageScriptTOPlugin.zip" that contains the SCRIPT Table Operator plugin software and metadata.

3. **Access Credentials**

   To use the plugins, you will need 2 different kinds of credentials, that is, one set for DSS and a second one for Vantage. Specifically:

   i. Dataiku DSS user credentials allow a user to login to a DSS instance. Your DSS server administrator can provide you with these credentials.

   ii. Vantage credentials allow a user to connect to the Advanced SQL Engine Database of a Vantage system, and, with appropriate permissions, read and write tables into the Advanced SQL Engine. Your Vantage database administrator (DBA) can provide you with credentials and suitable permissions for one or more databases on a Vantage system.

   Use your DSS user credentials to log on to a DSS instance, and then use your Vantage credentials to establish a connection between DSS and a Vantage system. Section III ("Creating A Vantage Connection") provides instructions on how to setup a DSS connection to a Vantage Advanced SQL Engine Database. It is suggested to create one connection per each database for which you intend to store output tables in.

4. **Teradata JDBC Driver**

   The Teradata JDBC Driver 16.20 or later is required to establish a connection between DSS and a Vantage System.

5. **Teradata Vantage Version 1.1** Both plugins require a connection to a Teradata Vantage system that minimally comprises of a Advanced SQL Engine. **Teradata Vantage Analytic Functions Plugin** The Teradata Vantage Analytic Functions Plugin for Dataiku DSS integrates about 180 of the Teradata Vantage Analytic Functions Plugin further requires a Vantage System v.1.1. For this plugin, if your Vantage v.1.1 system only has a Advanced SQL Engine, then only the analytic functions built into the Advanced SQL Engine will be available to the plugin, and namely, the following functions:

   - Attribution

- nPath
- Sessionize
- DecisionTreePredict
- DecisionForestPredict
- GLMPredict
- SVMSparsePredict
- NaiveBayesPredict
- NaiveBayesTextClassifierPredict The Machine Learning and Graph engine is required to completely leverage all capabilities of the Teradata Vantage Analytic Functions Plugin. **SCRIPT Table Operator Plugin**

  To use the SCRIPT Table Operator Plugin with a Vantage system Advanced SQL Engine and execute R and Python scripts in the Advanced SQL Engine nodes, one of the following R and/or Python bundles need to be installed directly on each node of the Advanced SQL Engine:

PID	Product name	Database version
9687-2000-0120	R Interpreter and Add-on Pkg on Teradata NewSQL	16.20
9687-2000-0121	R Interpreter and Add-on Pkg on Teradata Database	15.10, 16.10
9687-2000-0122	Python Interpreter and Add-on Pkg on Teradata NewSQL	16.20
9687-2000-0124	Python Interpreter and Add-on Pkg on Teradata Database	15.10, 16.10

  Moreover, your DBA must grant you in advance the additional following privileges:

- `EXECUTE` Function privilege on `TD_SYSFNLIB.SCRIPT` This is needed in order to invoke the SCRIPT Table Operator.
- `EXECUTE` privilege on the functions `SYSUIF.INSTALL_FILE` , `SYSUIF.REMOVE_FILE` , and `SYSUIF.REPLACE_FILE` .

III. Creating a Vantage Connection

1. Follow the instructions in the Dataiku Reference Document for Installing Database Drivers. In summary, one needs to execute from the command line of a DSS server:

   a. Stop the Data Science Studio server, where `DATA_DIR` is the data directory where Data Science Studio is installed:

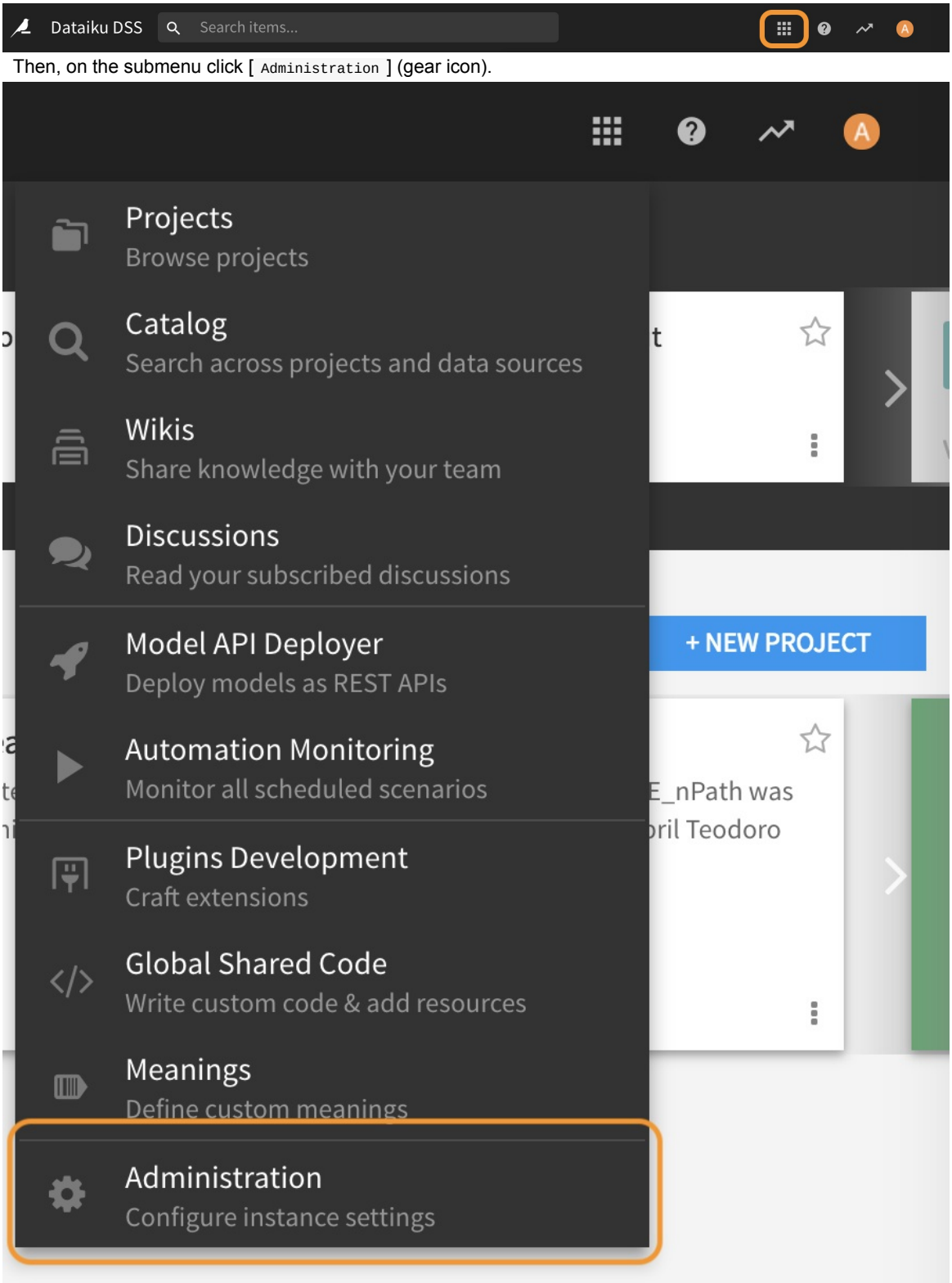
   ```
   DATA_DIR/bin/dss stop
   ```

   b. Copy the Teradata JDBC driver to the `DATA_DIR/lib/jdbc` directory.

   c. Restart Data Science Studio

   ```
   DATA_DIR/bin/dss start
   ```

2. Access Dataiku DSS on a browser. Then, on the Dataiku DSS home page click on Apps.



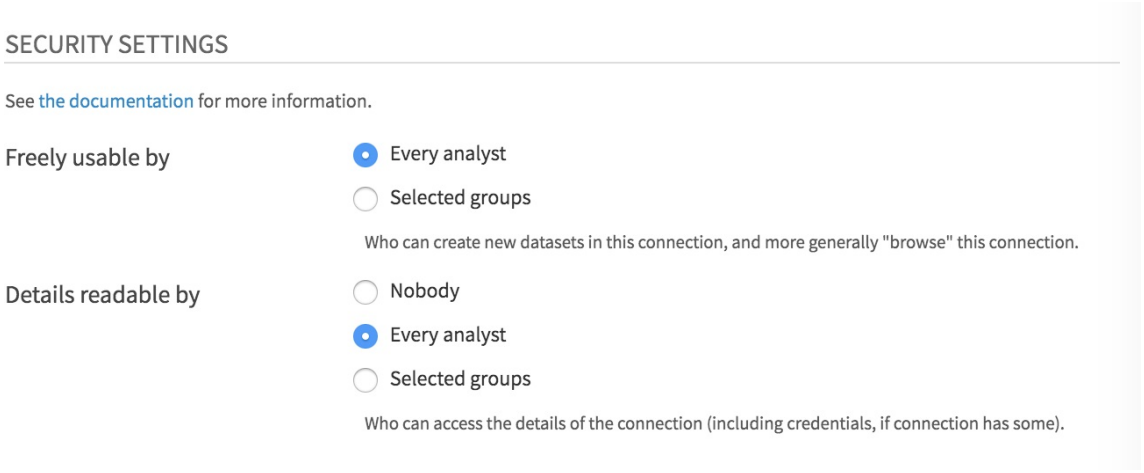
Then, on the submenu click [`Administration`] (gear icon).

Alternatively, you can go to `http://dataikuhost:port/admin/` .

3. On the DSS settings page, go to the [`Connections`] tab. Click on [`NEW CONNECTION`]. Choose [`Teradata`] among the options that will be presented.

4. Fill up the fields as needed:
   Basic Params Host: < `database.host.name` >
   User: < `Username` >
   Password: < `User_password` >
   Default Database: < `default_database` >
   Advanced JDBC properties:
   ```
   CHARSET: UTF8
   TMODE: TERA
   ```
   All other fields can be left as-is.

5. Modify “Details readable by” to either `Every Analyst` or `Selected Groups` .



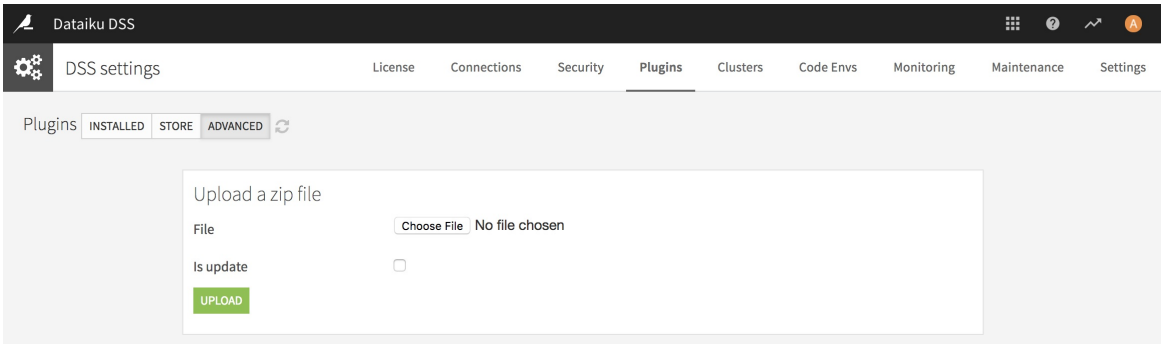
6. Click on the [`Test`] button to verify that connection details provided are valid.

7. Finally, click on the [`Save`] button.

IV. Teradata Vantage Analytic Functions Plugin Installation

The steps to install the Teradata Vantage Analytic Functions or the SCRIPT Table Operator plugins are as follows:

1. Assume that the zip file of the plugin you want to install is stored in your local filesystem.
2. In DSS Settings page (accessible through the Admin Tools button), select the [`Plugins`] tab, then select the [`ADVANCED`] option.



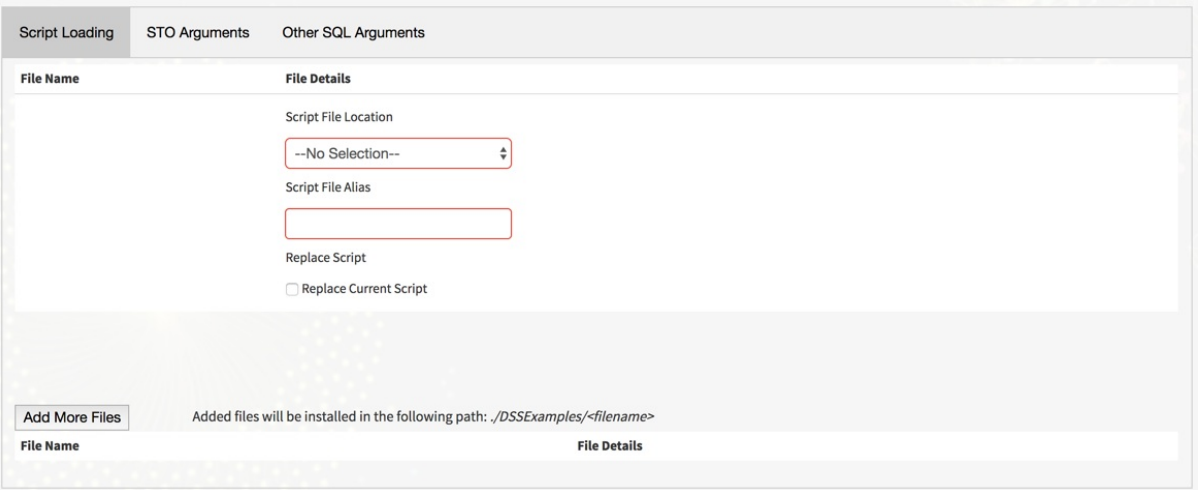
3. Click on [`Choose File`] and navitage to the location of the plugin zip file in your local filesystem.
4. If a previous installation of the Teradata Vantage Analytic plugin exists, check "Is update".
5. Click on the [`UPLOAD`] button.
6. When the upload succeeds, click on the [`Reload`] button, or do a hard refresh (Ctrl + F5) on all open Dataiku browsers for the change to take effect.
7. Repeat process, if you want to install a different plugin.

V. Using the Teradata SCRIPT Table Operator Plugin

This section assumes that a Dataiku DSS project already exists and input datasets have already been imported. Note that recipes need a non-empty dataset as input to run.

There are three (3) main tabs containing arguments used to install/replace the script files on the Advanced SQL Engine Database and/or invoke the SCRIPT Table Operator (STO).

Script Loading

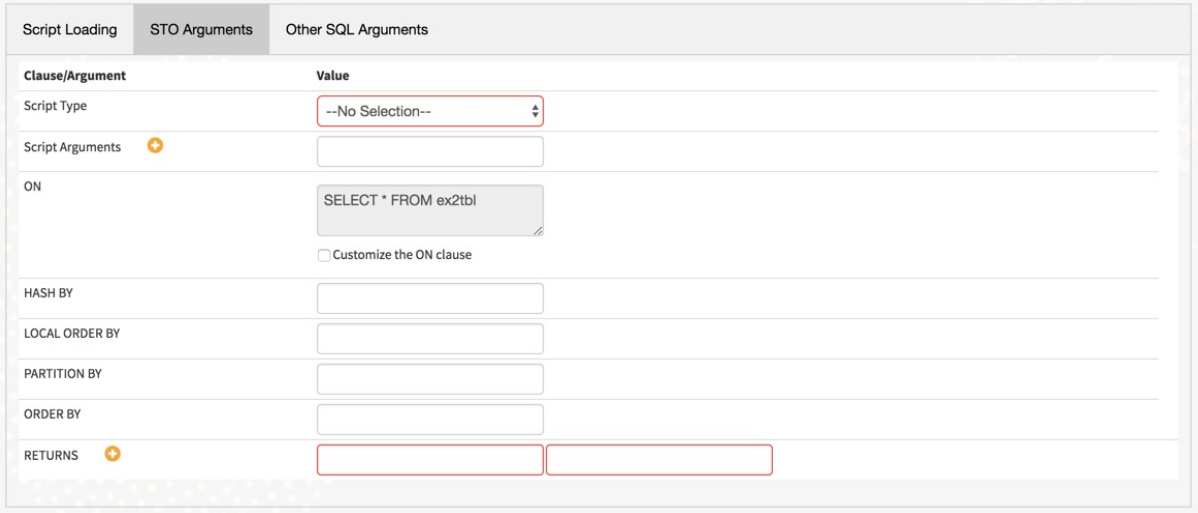


)

- Script File Name
  - The name of the script file to be uploaded.
  - This is the main script used in the SCRIPT Table Operator.
  - Depending on the selected Script File Location this input changes:
    - If the script is on the Vantage Server – A text input field is provided to enter a file name.
    - If the script is in the DSS Managed Folders and DSS Notebooks – A drop-down box containing a list of the files under their respective locations is provided.
  - The Script File Name will not appear until the Script File Location is selected.
- Script File Location
  - The location of the script to be installed, either on the Vantage server, a DSS Jupyter Notebook, or a DSS Managed Folder
- Script File Alias
  - The file alias to be used in the SQL statement
  - This is mainly used by the SCRIPT Installation/Replace process in the metadata tables.
- Script File Address
  - The fully qualified file location on the Vantage Server
  - This only appears if the selected option for Script File Location is “Vantage Server”
- Add More Files
  - This button allows the user to have additional files installed in the Vantage Advanced SQL Engine.
  - There is a file path specified to the right of the button in which the additional files are installed.
    - This may normally be used in instances where the user’s main script references an additional file.
- Additional Files:
  - File Name

        - This is the file name of an additional file.
        - Similar to the Script File Name it is a Text Field for files located in the Vantage Advanced SQL Engine and a drop-down box if DSS Managed Folder is selected as the File Location
    - File Location
        - The location of the file to be installed, either on the Vantage server or a DSS Managed Folder
    - File Address
        - The fully qualified file location on the Vantage server
        - Similar to the Script File Address this only appears when "Vantage Server" is selected as the file location.
    - File Format
        - Specifies whether the additional file to be installed is a BINARY or TEXT file.

SCRIPT Table Operator Arguments

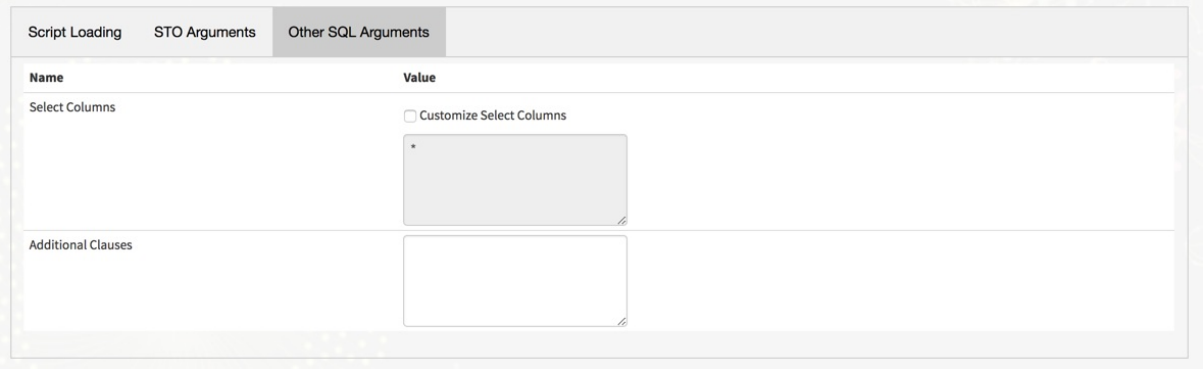


)

- Script Type
    - The type of script to be used typically Python or R'
    - Script Command
        - This is a Text area where the user can enter a custom Script Command.
        - This argument only appears if the selected Script type is “Other”.
- Script Arguments
    - The arguments for the script, place one argument per box. Click on the (+) button to add more arguments'
- ON
    - The ON Clause used as the input data for the script
    - If UNMODIFIED the clause defaults to “SELECT * FROM {input_table}”
- Customize the ON clause
    - A checkbox which specifies whether the ON clause should be modified.
- HASH BY
    - A HASH BY clause will cause the rows in the ON clause to be redistributed to AMPs based on the hash value of the column(s) specified'
- PARTITION BY
    - A PARTITION BY clause will cause the STO to execute against specific groups (partitions) based on the column(s) specified

- ORDER BY
  - 'An ORDER BY clause specifies the order in which values in a group (partition) are sorted
- LOCAL ORDER BY
  - A LOCAL ORDER BY clause orders the rows qualified on each AMP
- RETURNS
  - RETURNS NAME
    - The first column under returns
    - Specifies the name of the column(s) to be returned by the STO'
  - RETURNS TYPE
    - The second column under returns
    - Specifies the data type of the column(s) to be returned by the STO

OTHER SQL Arguments



)

- Select Columns
  - Specifies the contents of a user customized SELECT statement (data to be returned by the query)
  - Default is to SELECT all column(s) in the RETURNS clause
- Customize Select Columns Checkbox
  - Determines whether the SELECT (output) columns (data to be returned by the query) should be modified.
- Additional Clauses
  - Specifies any additional clauses to the output such as a HAVING or QUALIFY clause

Running the SCRIPT Table Operator Plugin

After setting up the arguments, click on the [`RUN`] button to run the SCRIPT Table Operator.


)