

Dataiku DSS
Teradata Vantage
Analytic Functions Plugin
User Guide

Table of Contents

I. Introduction	1.1
II. Requirements	1.2
III. Creating a Teradata Connection	1.3
IV. Teradata Vantage Analytic Plugin Installation	1.4
V. Teradata Vantage Analytic Plugin Usage	1.5
VI. Limitations	1.6
Authors	1.7

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 analyse a dataset. It also allows users to create their own recipes in Python, SQL or R. Custom reusable recipes for Dataiku are called plugins and can only be written in Python.

Dataiku provides a platform that allows to visualize and re-run workflows. In a Dataiku project, one can easily visualize how data flows across tables and recipes.

Teradata Vantage Analytic Functions Plugin for Dataiku DSS integrates around 150 Vantage Analytic Machine Learning functions to Dataiku data science studio. Machine Learning 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 provides a user interface-based way of building Vantage Analytic queries to be sent to a Teradata database. Input and output managed datasets are located in the connected Teradata database. All analytic queries are also executed in-database.

II. Requirements

1. Dataiku Data Science Studio version 4.0.1 or later

Dataiku DSS enterprise edition is required to import datasets from Teradata tables. Dataiku offers both downloadable and online options which can be obtained from their company [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 4.0.1.

2. Teradata Vantage Analytic Functions Plugin

TeradataVantagePlugin.zip contains the Teradata Vantage Analytic Functions plugin program and metadata. Please see Appendix A section to obtain a copy of this plugin.

3. Access Credentials

The first set of credentials required is the Teradata Credentials which allow the user to read and write tables into an Teradata Database. These credentials are used as input to the Dataiku-Teradata connector. Section 3 provides instructions on how to setup a Dataiku connection to an Teradata database. It is suggested to create one connection per database where one intends to store output tables.

The next set is the Dataiku User Credentials which allow the user to login to Dataiku DSS. Section 4 outlines the steps in creating a user in Dataiku.

4. Teradata JDBC Driver

The Teradata JDBC Driver is required to establish a connection between an Teradata Database and Dataiku. A copy of the jar file for the driver is appended to this document (Appendix B).

III. Creating a Teradata Connection

1. Follow the instructions in the Dataiku Reference Doc for Installing Database Drivers. In summary, one needs to:

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

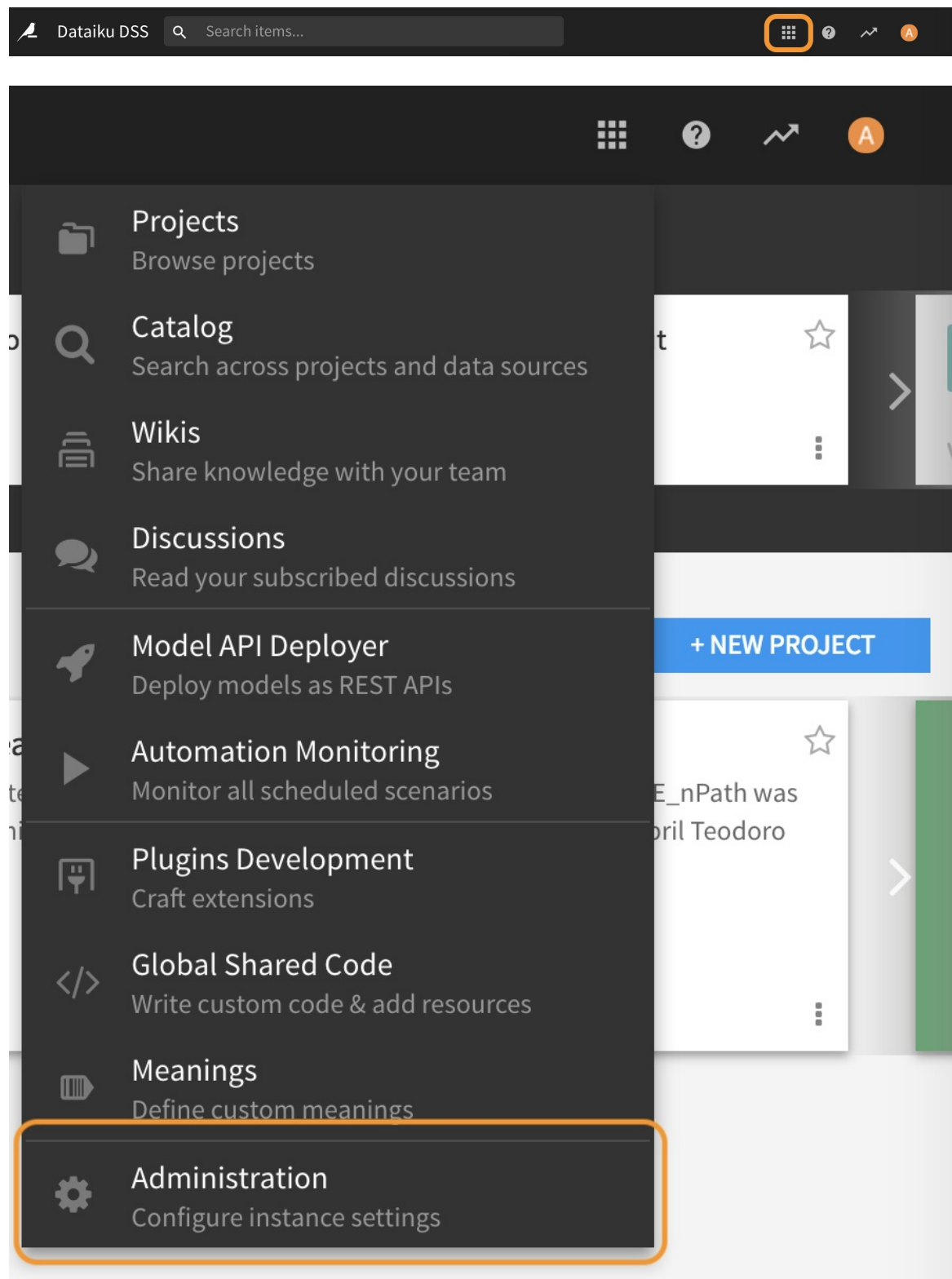
```
DATA_DIR/bin/dss stop
```

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

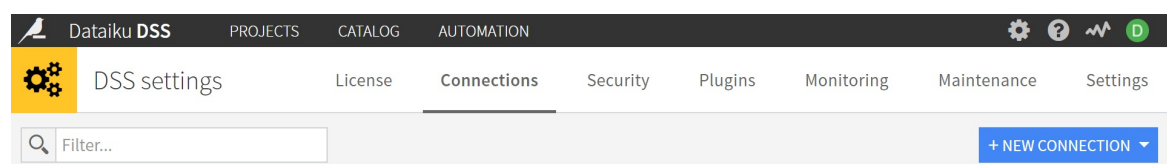
- c. Restart Data Science Studio

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

2. In 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. In the DSS settings page, Connections tab, Click on NEW CONNECTION. From the options that will be presented, choose Teradata.



4. Fill up the fields as needed:

Basic Params Host: database.host.name **User:** Username

Password: User's password

Default Database: default_database

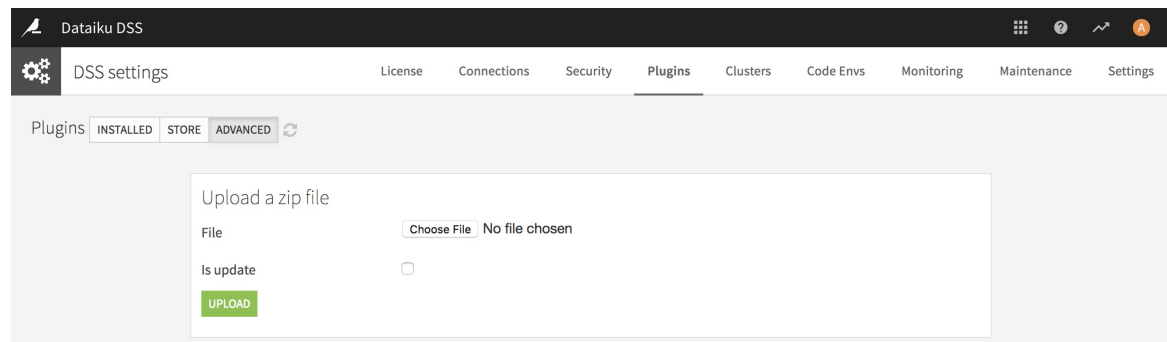
Advanced JDBC properties: CHARSET: UTF8 TMODE: TERA or ANSI

All other fields can be left as-is.

1. Click on Test button to verify that connection details provided are valid.
2. Finally, click on Save button.

IV. Teradata Vantage Analytic Functions Plugin Installation

1. In DSS Settings page (accessible through Admin Tools button), select Plugins tab, then select ADVANCED option.



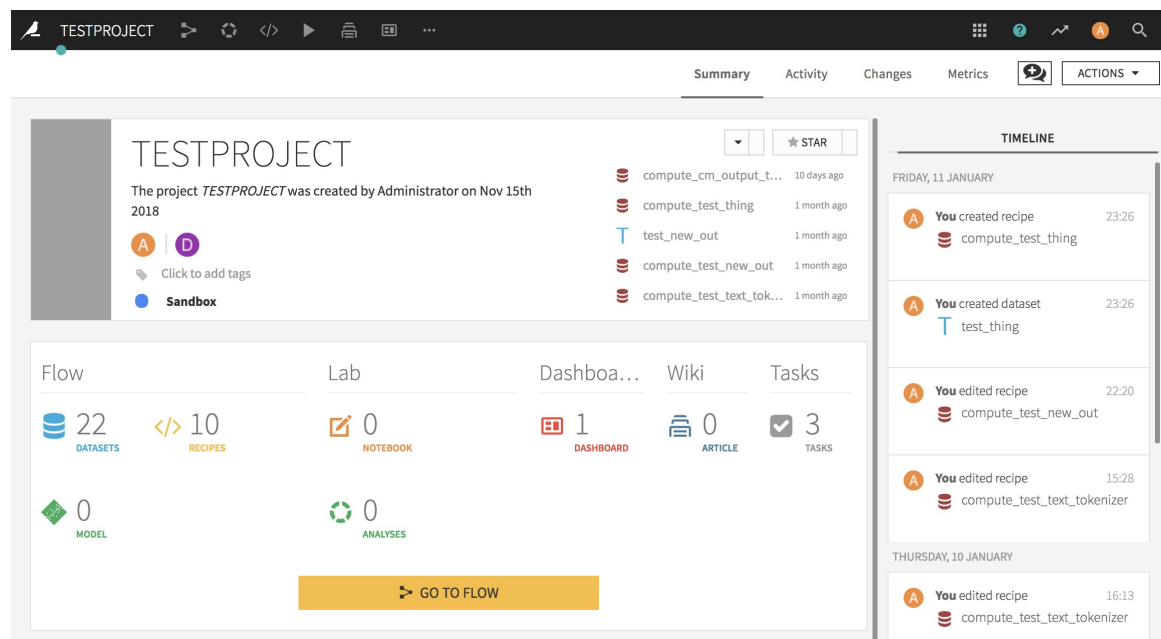
2. Click on Choose File and browse to the location of the Teradata Vantage Analytic plugin zip file in your local filesystem.
3. If a previous installation of the Teradata Vantage Analytic plugin exists, check "Is update".
4. Click on UPLOAD button.
5. When upload succeeds, click on Reload button, or do a hard refresh (Ctrl + F5) on all open Dataiku browsers for the change to take effect.

V. Teradata Vantage Analytic Functions plugin Usage

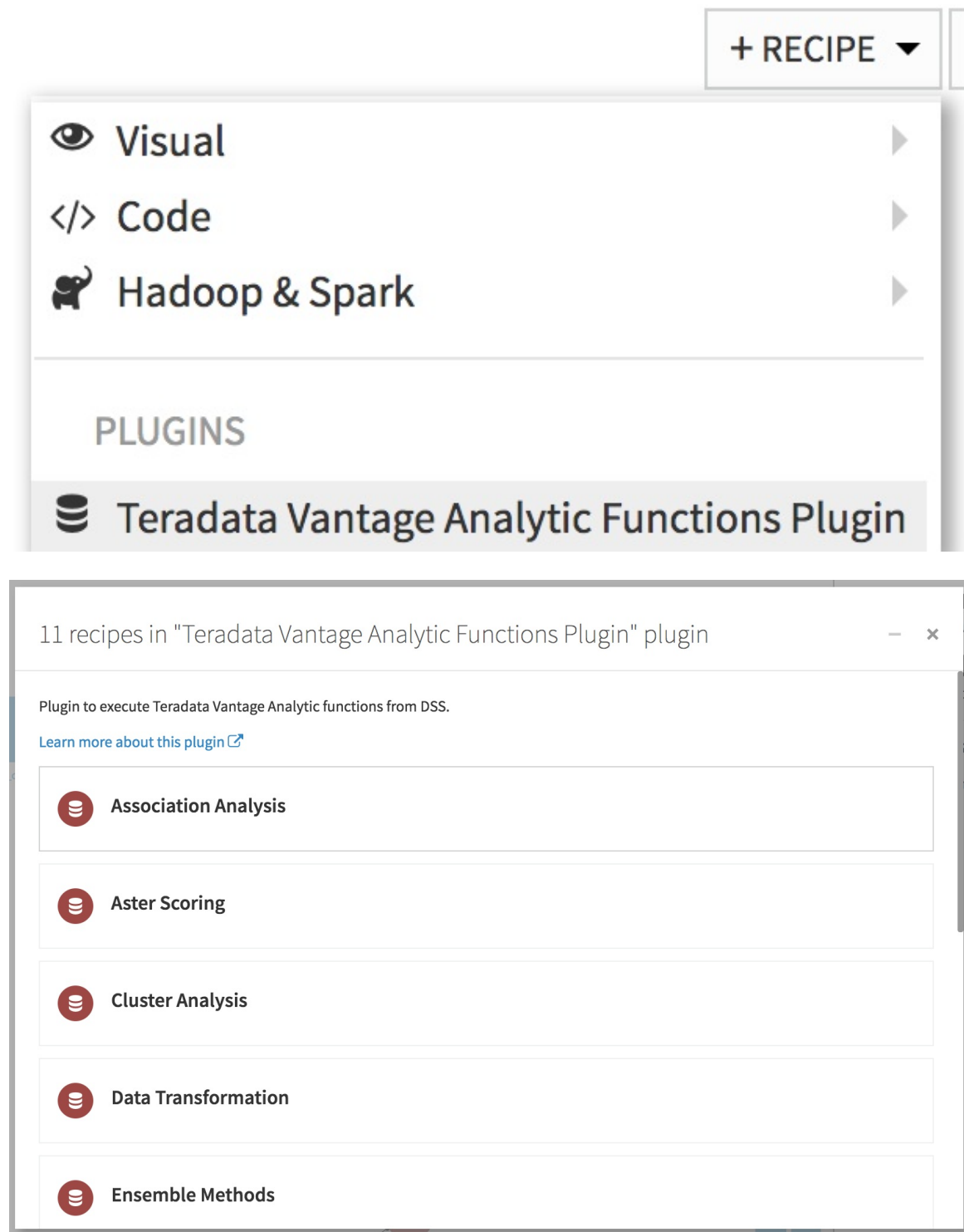
Usage

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

1. Go to the flow view of the Dataiku project, where the recipe is to be created, by clicking on the GO TO FLOW button or by clicking on the flow icon in the project menu.



2. In the Flow view, under Recipe, select desired recipe under the Teradata Vantage Analytic Functions plugin. The recipe names correspond to the different categories of Teradata Vantage Analytic Functions.



3. In New custom recipe popup, specify the input and output datasets. There can be more than one input dataset, as in the case of multiple-input analytic functions. The same is also the case for ML Functions with multiple output datasets. The output dataset will be stored in the database and schema corresponding to the connection selected in the Store into field. Click on CREATE button when done.

Custom recipe "Text Analysis"

Inputs

Search

+ T acc_output

+ T cm_output_test

+ T complaints

+ T count_output

+ T iris_category_expect_predict

+ T test_HMMDecoder

+ T test_new_complaints

+ T text_contents

Outputs

Add new dataset

Name

Store into

dssUser_TERA

CREATE DATASET

NEW DATASET | USE EXISTING

CANCEL

CREATE

4. In the recipe settings, one can select the most suitable function for manipulating or analysing the input dataset. Configure the chosen analytic recipe (input tables, partition and order attributes, and arguments). Required and optional fields are separated into tabs.

11

Recipe settings

Function Name Ldainference ▼

Description This function is used to output the topic distribution for each document in inputtable. Inputtable contains the documents to be inferred and the modeltable is the output of LdaTrainer. The result is stored in outputtable.

Required Arguments

Optional Arguments


Name	Value
Inputtable	complaints_testtoken ▼
Modeltable	ldamodel ▼
Outputtable	ldaout2
Docidcolumn	doc_id (int) ▼
Wordcolumn	token (string) ▼

5. The SQL Clauses tab allows the user to modify the query to be performed.

Required Arguments

Optional Arguments

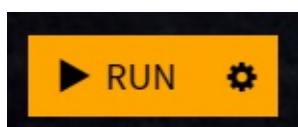
SQL Clauses

Name	Value
Modify Select Columns of Output Query	<input type="checkbox"/> Customize Select Columns <div>*</div>
Additional Clauses 	

Modify Select Columns of Output Query allows the user to modify the SELECT clause of the query. *Additional Clauses* allows the user to append additional SQL clauses to the query such as WHERE, ORDER BY, GROUP BY, and other similar clauses..

```
SELECT {modified select} FROM function_name(
  ...
)
{additional clauses}
```

1. Click on the RUN button or save the recipe settings for later use.



Usage Notes

1. Functions with multiple output datasets will normally require an output dataset for the functions' output message/result alongside any other output tables/datasets specified in the recipe. Please note that the output dataset/s name/s should also match the name within the recipe's settings.

VI. Limitations

1. For analytic functions that take in output table names as arguments and where the select query produces only a message table indicating the name of the output model/metrics table, it is the responsibility of the user to specify output table names that are not the same with that of an existing table. Some analytic functions provide an option to delete an already existing output table prior to executing an algorithm, others do not. If the former is the case, Teradata Database throws an 'already exists' exception.
2. The appended version of the Dataiku DSS Teradata Vantage Analytic Functions plugin was tested on Teradata 16.20. Earlier or later function versions may require a different set of function metadata.
3. The plugin currently only supports Teradata Database datasets as input and output.
4. Functions with any OUTPUT TABLE type arguments will require the user to add an output dataset for the SELECT statement results of the query and any additional output tables. Please refer to the Teradata Vantage Machine Learning Engine Analytic Functions Usage Guide to learn about the output tables of each function.
5. MapReduce Function pairs are currently limited to a select few: ApproxDCount, ApproxPercentile, Correlation, PCA, and Naive Bayes. In order to use these functions, please call their corresponding Map Functions on the function selection box and it will display the arguments for both functions.
6. Usage of certain functions may feature some inaccuracies, or may not work at all. The functions are as follows:
 - Statistical Analysis
 - Approximate Percentile Map/Reduce
 - Correlation Map/Reduce
 - Cox Hazard Ratio
 - Cross Validation
 - Distribution Match Reduce
 - Text Analysis
 - Text Tokenizer
 - Named Entity Finder Evaluator Map/Reduce
 - Time Series
 - Time Series Orders
 - Shapelet Supervised
 - ARIMA
 - Ensemble
 - AdaBoost Predict

Authors

DONNA APRIL TEODORO

Data Science Practice

Global Delivery Center – Manila

donnaapril.teodoro@teradata.com

KEVIN CONTRERAS

Data Science Practice

Global Delivery Center – Manila

kevin.contreras@teradata.com

JOSEPH ALVIN DE JESUS

Data Science Practice

Global Delivery Center – Manila

josephalvin.dejesus@teradata.com