

Dataiku DSS  
Teradata Vantage  
Analytic Functions Plugin  
User Guide

---

# Table of Contents

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

# 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.

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 MLE 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

In the background of the Teradata Vantage Analytic Functions Plugin user interface, the plugin essentially translates its user input into SQL queries that are sent to the NewSQL Engine of a connected Vantage system via JDBC. This way, all analytic queries are executed in-database, while also all input and output managed datasets are physically located in the database of the NewSQL Engine on the connected Vantage system.

## 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. Teradata Vantage Analytic Functions Plugin

The compressed file "`TeradataVantagePlugin.zip`" contains the Teradata Vantage Analytic Functions plugin program and metadata.

### 3. Access Credentials

The first set of credentials required is the Vantage Credentials which allow the user to read and write tables into a NewSQL Engine Database. These credentials are used as input to the Dataiku-Vantage connector. Section 3 provides instructions on how to setup a Dataiku connection to a NewSQL Engine 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 a Vantage system and Dataiku.

## 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:

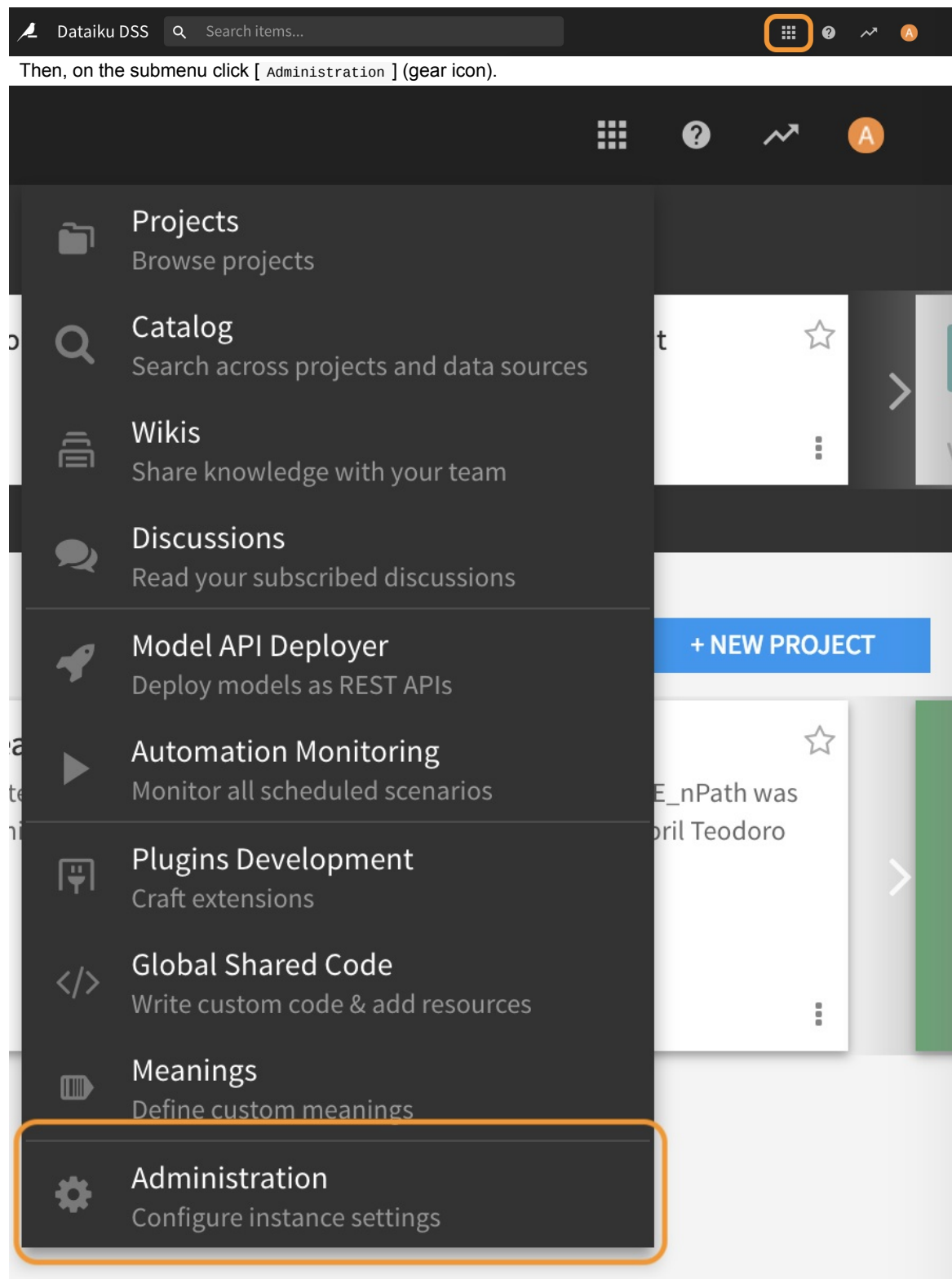
```
DATA_DIR/bin/dss stop
```

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

- c. Restart Data Science Studio

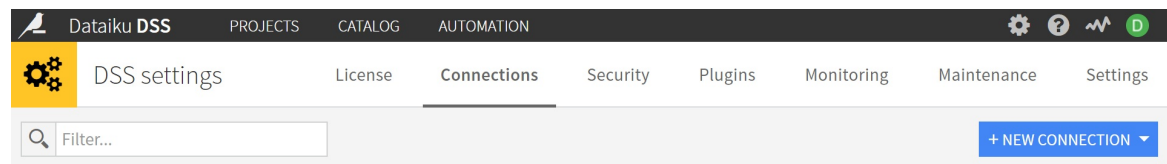
```
DATA_DIR/bin/dss start
```

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



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` **OR** `TMODE: ANSI`

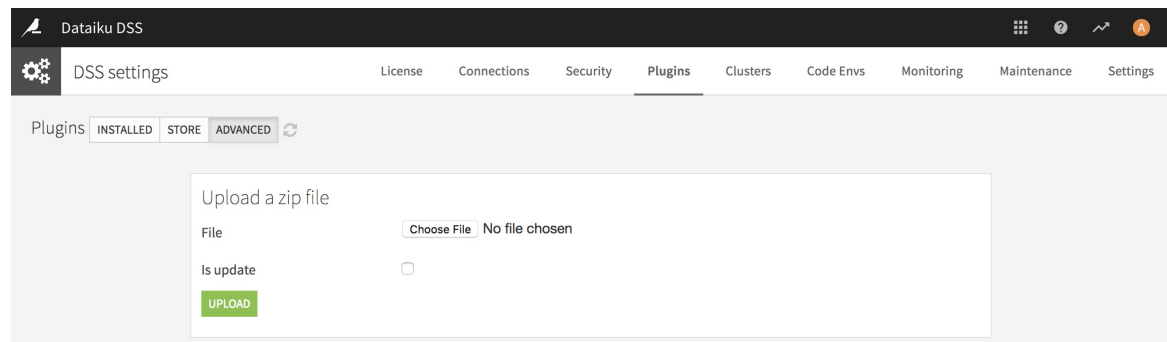
All other fields can be left as-is.

1. Click on the [ Test ] button to verify that connection details provided are valid.
2. Finally, click on the [ save ] button.

## IV. Teradata Vantage Analytic Functions Plugin Installation

---

1. In DSS Settings page (accessible through the Admin Tools button), select the [ `Plugins` ] tab, then select the [ `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 the [ `UPLOAD` ] button.
5. 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.

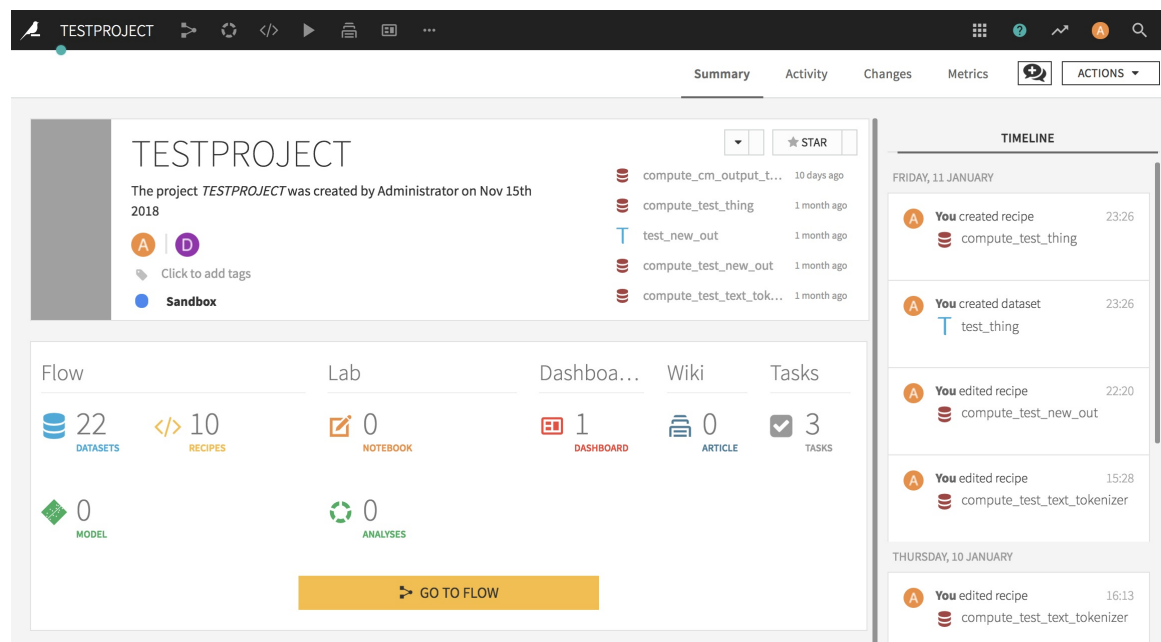


## V. Teradata Vantage Analytic Functions plugin Usage

### 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, click on the [ +RECIPE ] button, then select the [ Teradata Vantage Analytic Functions Plugin ] and further desired the recipe.


+ RECIPE ▼

 Visual ▶

 Code ▶

 Hadoop & Spark ▶


PLUGINS


 **Teradata Vantage Analytic Functions Plugin**


11 recipes in "Teradata Vantage Analytic Functions Plugin" plugin — ×


Plugin to execute Teradata Vantage Analytic functions from DSS.


[Learn more about this plugin](#) ↗

 Association Analysis

 Aster Scoring

 Cluster Analysis

 Data Transformation

 Ensemble Methods

3. In the [ 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 MLE 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 DATASET ] button when done.

The screenshot displays a window titled "Custom recipe 'Text Analysis'". It is divided into two main sections: "Inputs" and "Outputs".

**Inputs Section:**

- A search bar with the text "c" and a magnifying glass icon.
- A list of datasets with a plus icon and a blue 'T' icon next to each name:
  - acc\_output
  - cm\_output\_test
  - complaints
  - count\_output
  - iris\_category\_expect\_predict
  - test\_HMMDecoder
  - test\_new\_complaints
  - text\_contents

**Outputs Section:**

- A header "Add new dataset" with a close icon.
- A "Name" field with a placeholder "Name".
- A "Store into" dropdown menu currently showing "dssUser\_TERA".
- A blue button labeled "CREATE DATASET".
- A link "NEW DATASET | USE EXISTING" at the bottom.

At the bottom right of the window are two buttons: "CANCEL" and "CREATE".

4. In the recipe settings, one can select the most suitable function for the manipulation or analysis of the input dataset. Configure the chosen analytic recipe by specifying parameters such as the input tables, partition and order attributes, and arguments. A recipe's required and optional fields are separated into different tabs.

### 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	<span>complaints_testtoken ▼</span>
Modeltable	<span>ldamodel ▼</span>
Outputtable	<span>ldaout2</span>
Docidcolumn	<span>doc_id (int) ▼</span>
Wordcolumn	<span>token (string) ▼</span>

5. The [ SQL Clauses ] tab allows the user to explicitly modify the query to be executed.

Required Arguments

Optional Arguments

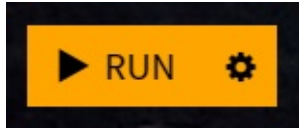
SQL Clauses

Name	Value
Modify Select Columns of Output Query	<input type="checkbox"/> Customize Select Columns <div>* [Text Area]</div>
Additional Clauses <span>+</span>	<div>[Text Field]</div>

The field next to " Modify Select Columns of Output Query " enables the user to modify the SELECT clause of the query. The field next to " Additional Clauses " enables the user to append additional SQL clauses to the query such as WHERE, ORDER BY, GROUP BY, and other similar clauses. These fields have equivalent effects as if the query were modified as:

```
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 **different** from those of the existing tables.

Some analytic functions provide an option to delete an already existing output table prior to executing an algorithm, but others do not. In the former case, the NewSQL Engine throws an " `Already exists` " exception.

1. The appended version of the Dataiku DSS Teradata Vantage Analytic Functions plugin has been tested to work with the MLE analytic functions on Vantage 1.1 systems. Earlier or later analytic function versions may require a different set of function metadata.
2. The plugin currently only supports NewSQL Engine datasets as input and output.
3. 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 Function](https://docs.teradata.com) Reference documentation page at docs.teradata.com to learn about the output tables of each function.
4. MapReduce Function pairs are currently limited to the following a select few: `ApproxDCount` , `ApproxPercentile` , `Correlation` , `PCA` , and `Naïve 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.
5. 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