Dataiku DSS

Aster Analytics Plugin User Guide

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 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.

Aster Analytics Plugin for Dataiku DSS integrates around 150 Aster Analytics SQL-MR functions to Dataiku data science studio. SQL-MR 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

Aster Analytics Plugin provides a user interface-based way of building SQL-MR queries to be sent to an Aster database. Input and output managed datasets are located in the connected Aster database. All SQL-MR 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 Aster 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 Aster Analytics plugin has been tested on Dataiku DSS version 4.0.1.

2. **Aster Analytics Plugin**

   AsterAnalytics.zip contains the Aster Analytics 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 Aster Credentials which allow the user to read and write tables into an Aster Database. These credentials are used as input to the Dataiku-Aster connector. Section 3 provides instructions on how to setup a Dataiku connection to an Aster database. It is suggested to create one connection per database schema 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. **Aster JDBC Driver**

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

III. Creating an Aster 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.

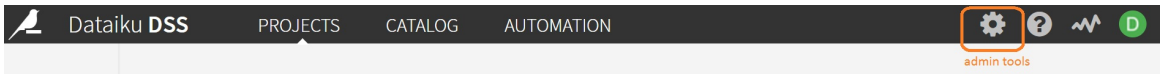
   ```
   DATA_DIR/bin/dss stop
   ```

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

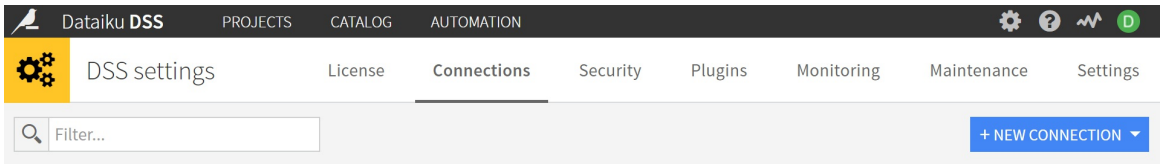
   c. Restart Data Science Studio

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

2. In the Dataiku DSS home page, click on Admin Tools (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 Other SQL databases.



4. Fill up the following fields as appropriate:

   JDBC driver class: com.asterdata.ncluster.Driver
   JDBC URL: jdbc:ncluster://:/
   JDBC properties: user: username password: password **SQL dialect (experimental)**: Aster Data **Naming rules for new datasets**: Schema: schema_name

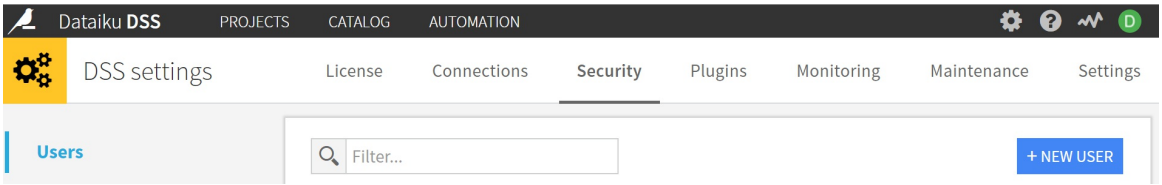
5. Click on Test button to verify that connection details provided in step 3 are valid.

6. Finally, click on Save button.

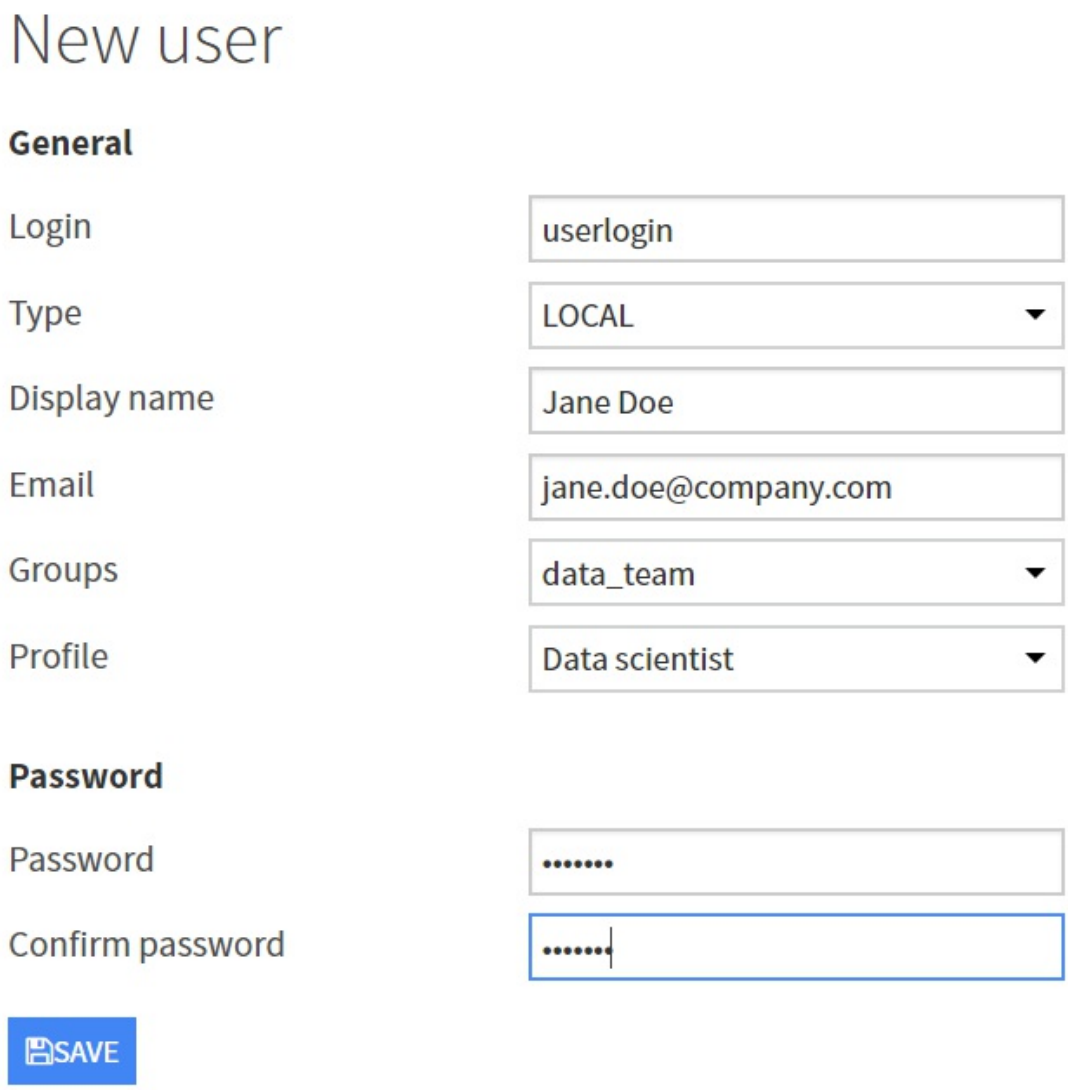
IV. Creating a User

Note: This section assumes that you are logged in as a user with Administrator privileges.

1. Click on Admin Tools (gear icon), and choose the Security tab. Users section automatically gets selected. Alternatively, you can go to http://:/admin/security/users/



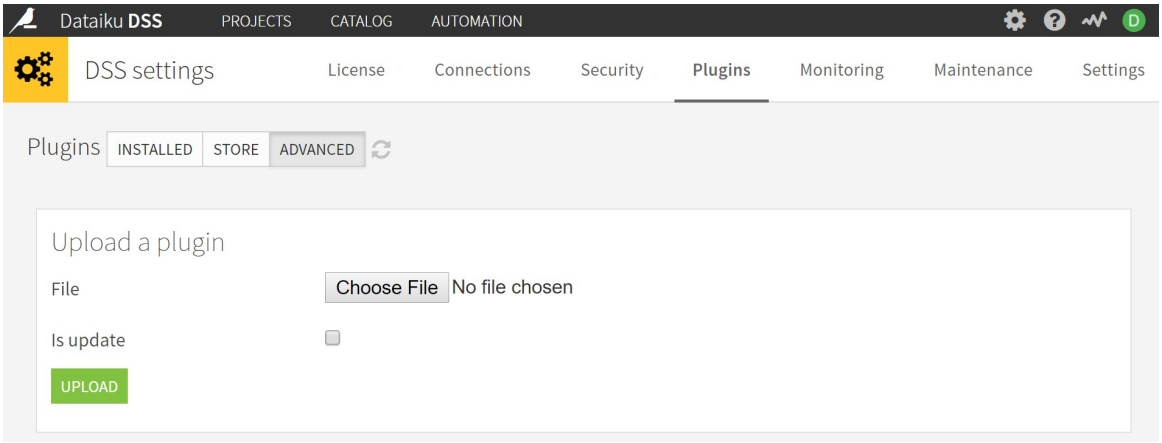
2. Click on New User button. In the New User panel that will appear, fill up the general and password fields. Select the appropriate group for the new user. By default, Dataiku DSS has three groups with different privileges: administration, data_team, and readers. New groups can also be added in the Groups pane of the Security tab. Set the new user profile to Data Scientist. This will allow him/her to run code-based recipes (Python, R…)



3. Finally, click on Save button.

V. Aster Analytics 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 Aster Analytics zip file in your local filesystem.

3. If a previous installation of Aster Analytics plugin exists, check Is update.

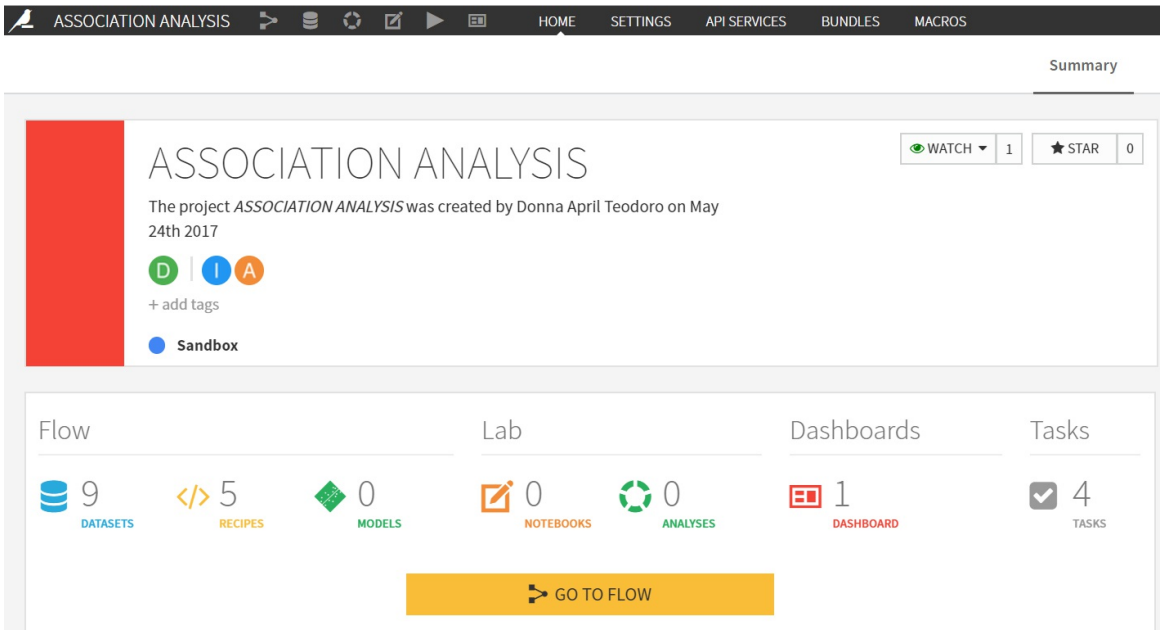
4. Click on UPLOAD button.

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.

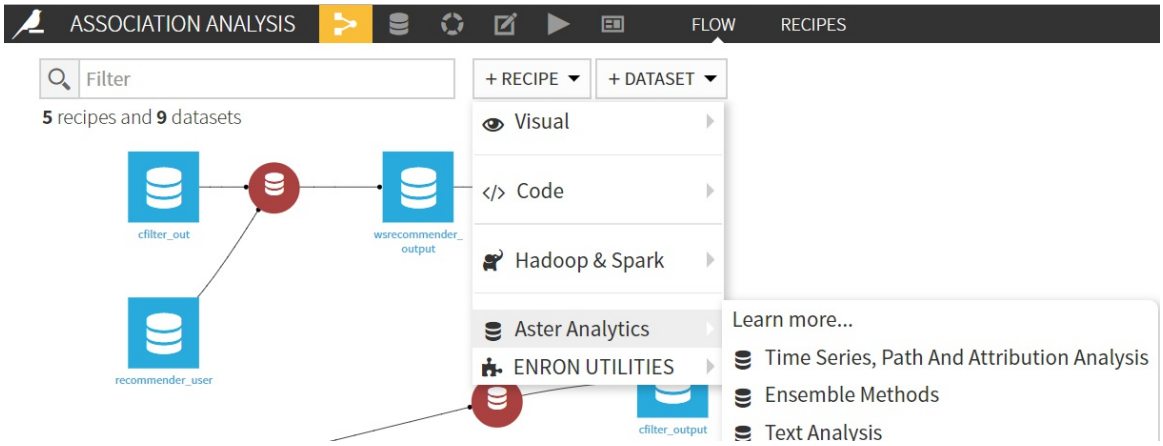
VI. Aster Analytics Plugin Usage

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.

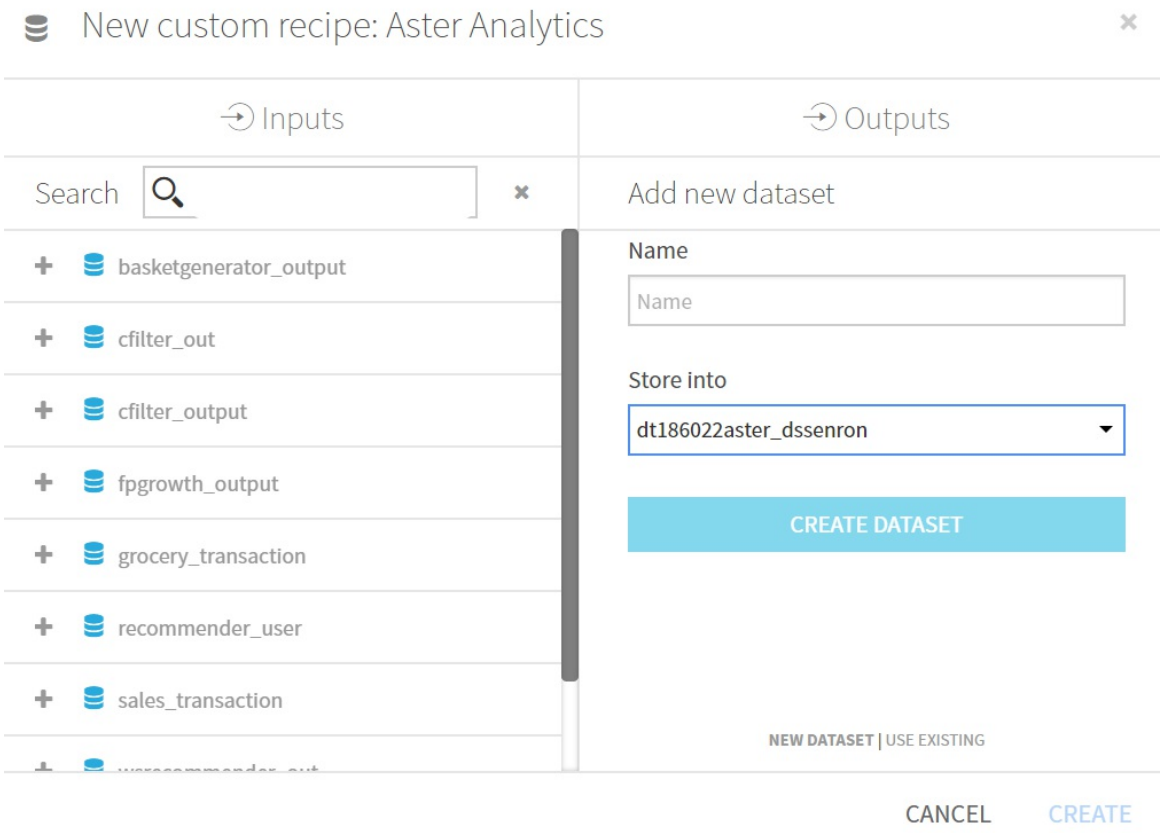
1. Go to the flow view of the Dataiku project where the recipe is to be created by clicking on the GO TO FLOW button or by clicking on the flow icon in the project menu.



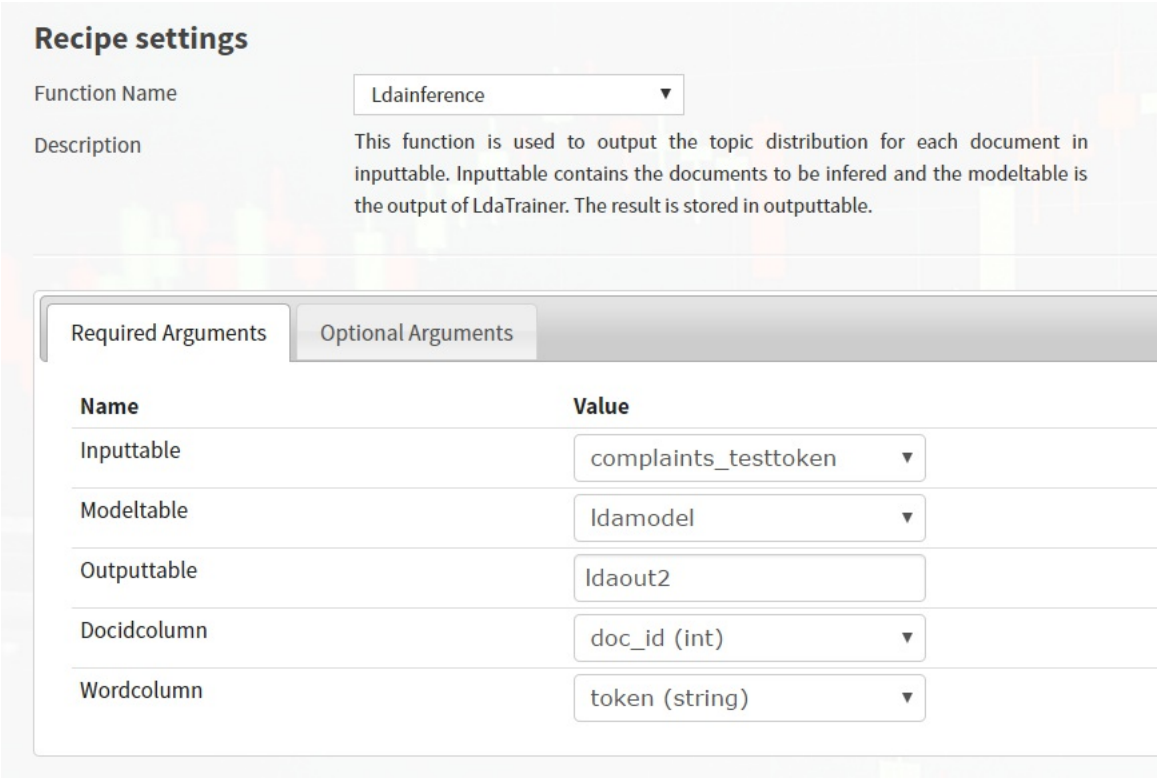
2. In the Flow view, under Recipe, select desired recipe under Aster Analytics plugin. The recipe names correspond to the different categories of Aster Analytics functions.



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 SQL-MR functions. 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.



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



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

VII. Limitations

1. As of this writing, Aster is not an officially supported database in Dataiku. Some tables, those with special characters in column names in particular, may not be properly imported into datasets. This is not a major problem since all SQL-MR functions are performed in the Aster environment and not in the Dataiku virtual machine.
2. DISTRIBUTE BY HASH error is encountered when creating new datasets manually in Dataiku. To work around this issue, go to the Settings page of the newly created dataset. In the SQL section, set table creation mode to Manually define. Then, set table creation SQL to CREATE TABLE "" ($DKU_CREATE_TABLE_FIELDS) DISTRIBUTE BY HASH() This error is not encountered when creating output datasets through the Create Recipe popup.
3. For SQL-MR 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 SQL-MR 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, Aster Database throws an ‘ already exists’ exception.
4. The appended version of the Dataiku DSS Aster Analytics plugin was tested against Aster 6.20 SQL-MR functions. Earlier or later function versions may require different functions metadata.
5. In case HTTP error 413 (Request entity too large) or HTTP error 414 (Request URI too long) is encountered after reloading a saved recipe in Dataiku versions earlier than 4.0.4, one can edit Dataiku’s code itself. This is an issue that will be resolved in Dataiku’s next release. To fix this without updating Dataiku, open INSTALL_DIR/frontend/static/dataiku/js/mainpack.js. Locate the callPythonDo function, and change the HTTP method associated with it from GET to POST.
6. As of this writing, Aster Analytics plugin could only refer to SQL-MR functions installed in the public schema.

Authors

DONNA APRIL TEODORO
Data Science Practice
Global Delivery Center – Manila
donnaapril.teodoro@teradata.com

JOSEPH ALVIN DE JESUS
Data Science Practice
Global Delivery Center – Manila
josephalvin.dejesus@teradata.com

KEVIN CONTRERAS
Data Science Practice
Global Delivery Center – Manila
kevin.contreras@teradata.com