Vertica Knowledge Base Article

Vertica Integration with Dataiku DSS: Connection Guide

Document Release Date: 8/30/2018

MICRO FOCUS | VERTICA

Legal Notices

Warranty

The only warranties for Micro Focus International plc products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Micro Focus shall not be liable for technical or editorial errors or omissions contained herein.

The information contained herein is subject to change without notice.

Restricted Rights Legend

Confidential computer software. Valid license from Micro Focus required for possession, use or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Copyright Notice

Trademark Notices

Adobe™ is a trademark of Adobe Systems Incorporated.

Microsoft® and Windows® are U.S. registered trademarks of Microsoft Corporation.

UNIX® is a registered trademark of The Open Group.

This product includes an interface of the 'zlib' general purpose compression library, which is Copyright © 1995-2002 Jean-loup Gailly and Mark Adler.

Contents

Vertica Integration with Dataiku: Connection Guide

For a PDF version of this document, click here.

About Vertica Connection Guides

Vertica connection guides provide basic information about setting up connections to Vertica from software that our technology partners create. These documents provide guidance using one version of Vertica and one version of the third party vendor's software. Other versions of the third-party product may work with Vertica. However, we may not have tested those other versions.

Vertica and Dataiku: Versions Tested

Software	Version
Partner Product	Dataiku Data Science Studio 4.3.3
Vertica Client	Vertica 9.1.1-0 JDBC Driver
Vertica Server	Vertica Analytic Database 9.1.1-0

Dataiku Overview

Dataiku Data Science Studio (DSS) is an analytic workbench that allows data scientists to build an end-to-end workflow that transforms raw data into visualizations of predictions. For more information, view a sample use case that shows how Dataiku Data Science Studio used Medicare data stored in Vertica for analysis and prediction.

Install Dataiku

Dataiku Data Science Studio is a web-based application available for Linux. A beta version is available for Mac OS X but is not recommended for a production environment. Data Science Studio uses the JDBC driver to connect to Vertica and is compatible with Chrome and Firefox.

Before you install Dataiku Data Science Studio, review the requirements for installing on Linux.

Download the latest version of Dataiku Data Science Studio that corresponds to your Linux distribution and architecture. After the download is complete, follow the instructions for installation.

Install the Vertica Client Driver

Before you can connect to Vertica using Dataiku Data Science Studio, you must download and install the Vertica JDBC client driver. Follow these steps:

1. Navigate to the Vertica Client Drivers page.
2. Download the JDBC driver for your version of Vertica.

   > Note For details about client driver and server version compatibility, see the Vertica documentation.

3. Before installing the driver, you must stop Data Science Studio.

   Navigate to the directory where Data Science Studio is installed, which by default is DATA_DIR. Stop the application using the following command:

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

4. Place the client .jar file in the Data Science Studio directory for external libraries as follows:
   a. Locate the Vertica JDBC `.jar` file from the driver location. For example:

      ```
      /opt/vertica/java/lib/vertica-jdbc-9.X.X-0.jar
      ```

      Replace X.X with the version of your Vertica database.

   b. Copy the .jar file into the DATA_DIR/lib/jdbc folder. For example, on Linux Centos with a user called Dataiku

      ```
      $ /home/dataiku/dataiku-dss-2.0.1/DATA_DIR/lib/jdbc/vertica-jdbc-
      9.X.X.0.jar
      ```

      Replace X.X with the version of your Vertica database.

      > Note Do not modify the CLASSPATH.

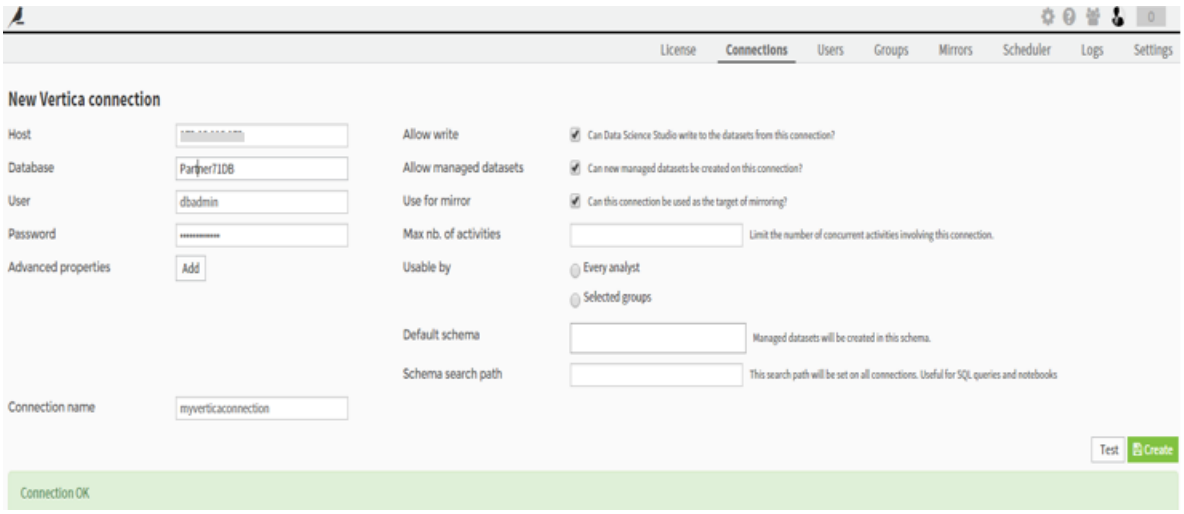
5. Restart Data Science Studio with the following command:

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

Connect to Vertica from Dataiku

1. Open Dataiku from your web browser.
2. Click **Create a New Project**.
3. In the upper right corner of the screen, click the gear button.
4. Click **Connections** > **New Connection** and select **HP Vertica**.

5. Enter your connection information. Data Science Studio automatically tests your connection. The following fields are required:
   - Host
   - Database
   - User
   - Password
   - Connection name

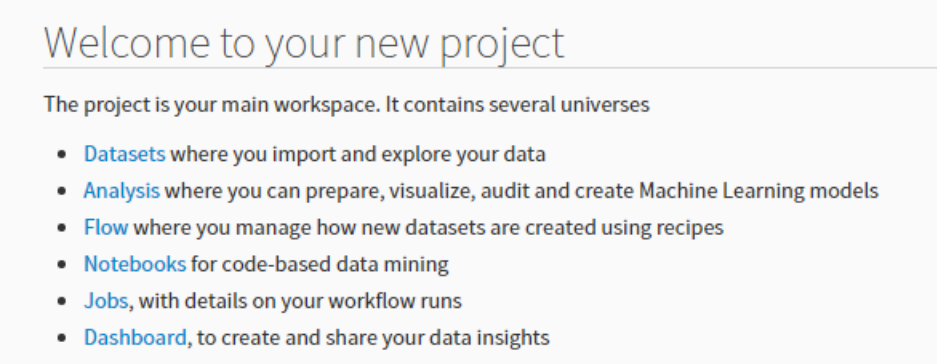


6. Click **Create**.
7. Use this connection to explore data stored in Vertica.

Creating a Dataset

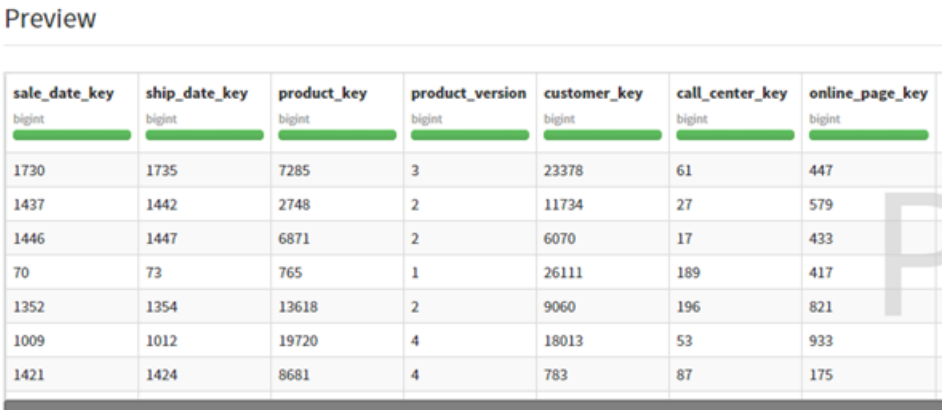
After you have an established connection, follow these steps to create a dataset:

1. From your project screen, click **Datasets**.



2. Click the **New Dataset** icon.
3. From the drop-down menu, select **HP Vertica**.

4. On the **Connection** tab, enter the following required fields:
   - **Connection**: Your connection to HP Vertica
   - **Mode**: Choose **connect to a table** or **write a query**
   - **Table**: Table name
   - **Schema**: Schema name
5. Click **Test** to see a preview of the data.



6. Enter a dataset name and click **Create**.

Data Type Limitations

Dataiku supports and correctly displays all Vertica data types. However, you might see the following behavior when you preview the data:

- Dataiku truncates CHAR, VARCHAR, and LONG VARCHAR values with more than 32,767 characters to 32,767 characters.
- Dataiku might not support TIMETZ and TIMESTAMPTZ values.
- BINARY, VARBINARY, and LONG VARBINARY values are displayed in hexadecimal format.

You might see the following behavior when you load data into Vertica:

- Empty values are loaded as NULL.
- All date values must have a time zone. Date values that are not assigned a time zone default to UTC.
- TIMETZ values might be loaded on the client time zone.
- BINARY, VARBINARY, and LONG VARBINARY values are loaded in the VARCHAR hexadecimal format.
- If you have a string that is longer than 16,200 characters, change the Table Creation Mode (located in **Settings** > **Advanced**) from **Automatically generate** to **Manually define** to load all the characters.

- Interval values are loaded as VARCHAR. To change the value, change the Table Creation Mode (located in **Settings** > **Advanced**) from **Automatically generate** to **Manually define** and change the value to Interval.

For More Information

- Dataiku
- Dataiku Tutorials
- Vertica Community Edition
- Vertica User Community
- Vertica Documentation