Arbutus Connectors

Databricks
CONFIGURATION GUIDE

ARBUTUS
Powerful Analytics Simplified

Arbutus Connectors

Contents

Arbutus Connector – Databricks

A. Introduction

The purpose of this Guide is to provide assistance with configuring the Arbutus Databricks Connector using the ODBC Data Source Administrator. The configuration process can involve several technical steps that require a good understanding of IT systems and database management.

To make the most of this guide, it's advisable to have a good understanding of database connectivity, driver installation, and system settings. The ODBC Data Source Administrator, which is used as part of the configuration process, allows for the setup and management of data sources, enabling applications to access data from various database systems.

Due to the complexity and potential impact of these configurations, it is recommended that only those individuals with IT or database expertise undertake this task. In addition, it should also be understood that each client's network environment is different. A one-size-fits-all approach is rarely effective, as what works well in one environment may not be suitable in another.

B. About Databricks

Databricks is a cloud-based platform for big data processing and analytics, integrating with Apache Spark to simplify building, managing, and scaling data pipelines, machine learning models, and analytics workflows. It is widely used for data engineering, data science, and AI tasks, with collaborative features and seamless integration with cloud services like AWS, Azure, and Google Cloud.

In Databricks, data is stored in cloud storage (AWS, Azure, or Google Cloud) and supports both structured and unstructured data. Databases can be defined to organize tables, where data is stored and queried. While Databricks is not a traditional relational database, it supports relational-like operations using SQL and Spark SQL for creating tables, defining schemas, and running queries.
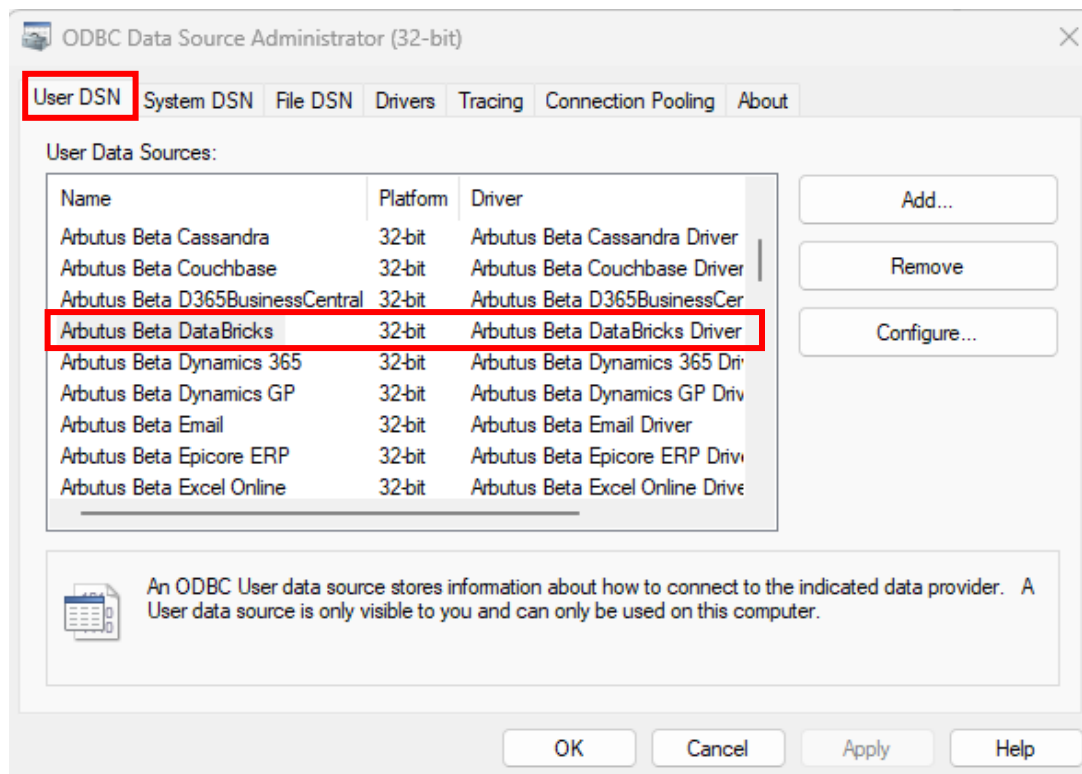
Arbutus Connectors

C. Determining if the Connector exists prior to configuring

Installation of the Arbutus Databricks Connector is done at the time of installing the Arbutus software. For more information on this, please see the **Overview Guide Document.**

Once the Connector has been installed, the next step is to configure it.

Prior to configuring it, you can check to see if the Connector has been installed by opening the **32-bit ODBC Data Source Administrator**, pictured below, and clicking the **User DSN** tab. Included below is information on how you can access the **ODBC Data Source Administrator**.

Arbutus Connectors

- If the Arbutus Databricks Connector appears in the list, it can be considered as installed.
- If it is not listed, it is likely that you did not select it during the installation or modification of the Arbutus software. In this case, it is recommended to reinstall the Arbutus software and choose the **Modify** option when prompted. For more details, please refer to the **Overview Guide Document**.

Below is the file path to access and run the **ODBC Data Source Administrator** application:

C:\Windows\SysWOW64\odbcad32.exe

Alternative, you can also try locating and opening the **ODBC Data Source Administrator** application by doing a search on your desktop application.

D. Configuring the Connector after it has been installed

Once you have verified that the Arbutus Connector has been installed, it is time to configure it.

This process is done using the **ODBC Data Source Administrator**. It can be described as "**editing the DSN configuration**".

DSN, Drivers, and Data Sources
What is a DSN? DSN stands for Data Source Name, and is a unique name used to create a data connection to a database using open database connectivity (ODBC).   A DSN is a data structure that contains the information required to connect to a database. It is essentially a string that identifies the source database, including the driver details, the database name, and often authentication credentials and other necessary connection parameters. DSNs facilitate a standardized method for applications to access databases without needing hard-coded connection details, enhancing flexibility and scalability in database management.

Arbutus Connectors

- *Drivers* are the components that process ODBC requests and return data to the application. If necessary, drivers modify an application's request into a form that is understood by the data source. The **Drivers** tab in the **ODBC Data Source Administrator** dialog box lists all drivers installed on your computer, including the name, version, company, file name, and file creation date of each driver.

- *Data sources* are the databases of files accessed by a driver and are identified by a data source name (DSN). You use the ODBC Data Source Administrator to add, configure, and delete data sources from your system.

All ODBC connections require that a DSN be configured to support the connection. When a client application wants to access an ODBC-compliant database, it references the database using the DSN.

The types of DSNs are:
- **User DSN** – User DSNs are local to a computer and can be used only by the current user. They are registered in the HKEY_Current_USER registry subtree.

- **System DSN** – System DSNs are local to a computer rather than dedicated to a user. The system or any user with privileges can use a data source set up with a system DSN. System DSNs are registered in the HKEY_LOCAL_MACHINE registry subtree.

- **File DSN** – File DSNs are file-based sources that can be shared among all users who have the same drivers installed and therefore have access to the database. These data sources need not be dedicated to a user nor be local to a computer. File data source names are identified by a file name with a .dsn extension.

User and system data sources are collectively known as *machine* data sources because they are local to a computer.

Each of these DSNs has a tab in the **ODBC Data Source Administrator** dialog.

The Arbutus ODBC Driver for Databricks enables real-time access to Databricks data, directly from any applications that support ODBC connectivity, the most widely supported interface for connecting applications with data.

Arbutus Connectors

Follow these steps to edit the DSN configuration and configure the Connector.

1. First open the **ODBC Data Source Administrator**.
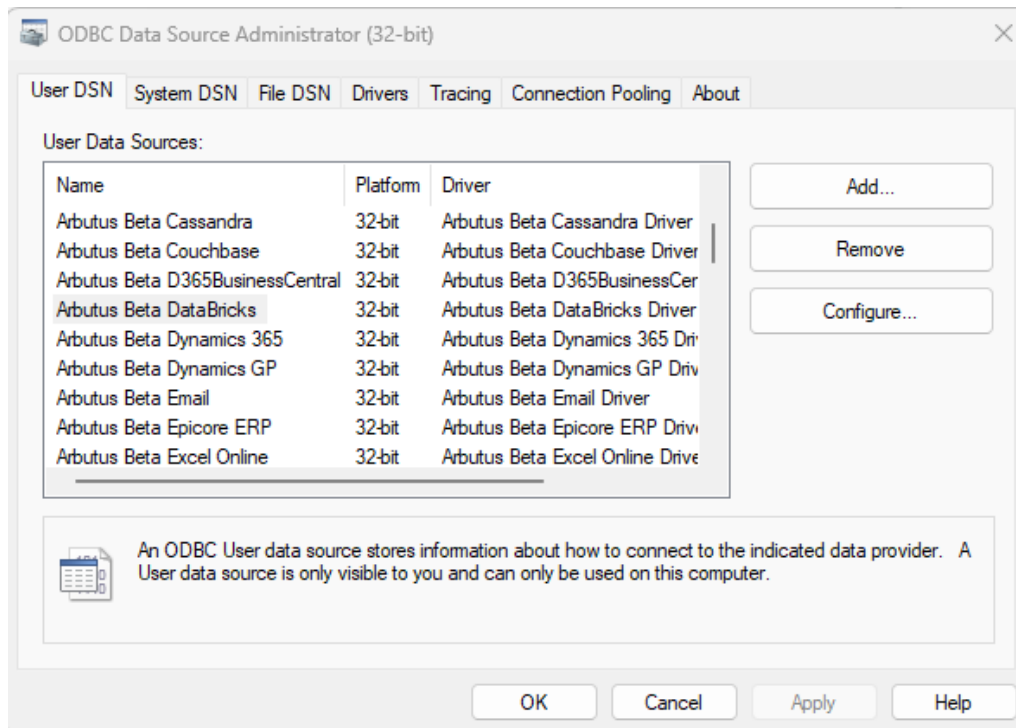
2. Click the **User DSN** tab.

    Selected data connectors are installed as **User DSN's** in Window's 32 Bit **ODBC Data Source Administrator**.

    Also, each of the data connector's names is prefaced with Arbutus, for example, **Arbutus Databricks.**

3. Select the Arbutus Connector, in this case it is **Arbutus Databricks**.
4. Click **Configure**.

Arbutus Connectors

This opens the **Arbutus Databricks Driver – DSN Configuration** dialog.

E. Editing the DSN properties – the Basic and Advanced tabs
With the DSN Configuration dialog open, the next step is to edit the DSN properties, where necessary, in the **Basic** and **Advanced** tabs. For example, editing the **Auth Scheme properties** (per screenshot below) to ensure correct authentication to the server is applied.
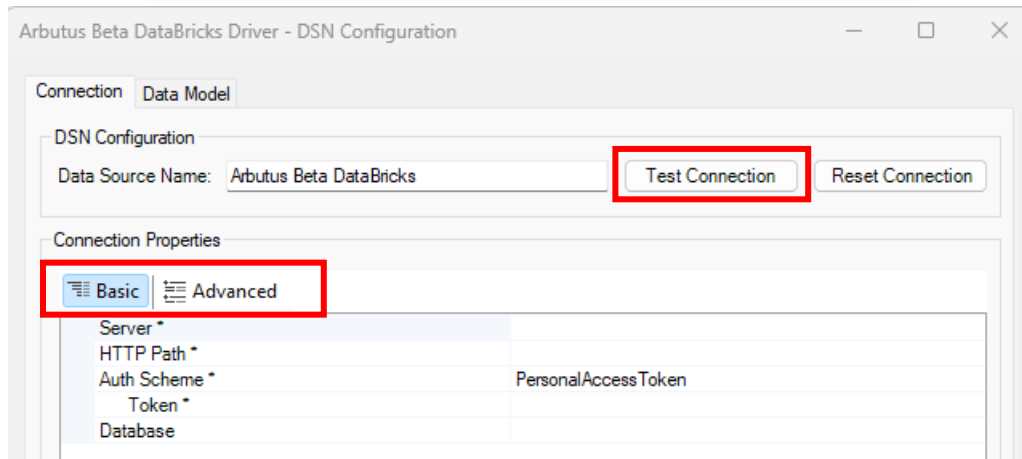
E1. Editing the DSN properties in the Basic tab
The properties listed in the **Basic** tab are typically the ones that are most commonly used, and as such are designed to be more user-friendly and straightforward, allowing you to quickly make changes without needing in-depth technical knowledge.

Arbutus Connectors

Once you have completed editing the properties in the **Basic** tab, you can go ahead and try testing the connection to the Databricks system by clicking the **Test Connection** button, as highlighted in the screenshot below.

In the **Basic** tab, there are **four** main properties to review:

1. **Server** – this is the host name or IP address of the server hosting the Databricks database.

2. **HTTP Path** – this is the path component of the URL endpoint. This property is used to specify the path component of the URL endpoint.

   This property can be found by following the path: Databricks main page -> Compute(in left panel) -> {your Cluster} -> Advanced options(in Configuration tab) -> JDBC/ODBC - HTTP Path.

Arbutus Connectors

3. **Auth Scheme** **–** click the dropdown to select from the list the appropriate type of authentication to use when connecting to Databricks. The options available for selection are as follows:

   o PersonalAccessToken – select this if you need to connect to Databricks programmatically using a token, e.g., scripts, scheduled tasks, without requiring user interaction.

   You can also select this option when you need to provide temporary access to Databricks resources, as tokens can be set with specific expiration dates.

   Another instance is if you don't want to use username/password authentication, and you don't have Single-Sign-On (SSO) configured.

   Personal access tokens are easy to generate and manage, making them suitable for development and testing environments. However, for production environments, consider using more secure methods like OAuth or service principals.

   Selecting **PersonalAccessToken** requires you to specify the following:

   ▪ Token – this is the token used to access the Databricks server. The token can be obtained by navigating to the User Settings page of your Databricks instance and selecting the Access Tokens tab.

   The default value is **PersonalAccessToken**.

   o Basic – select this when you need a straightforward method to authenticate using a username and password/token. As such, your database does not use Windows Single-Sign-On (SSO) or Kerberos authentication.

Arbutus Connectors

However, Basic authentication sends credentials in an unencrypted form, which can be a security risk in production environments. For better security, consider using more secure methods like OAuth or personal access tokens.

Selecting **Basic** requires you to specify the following:
- User – this is the username used to authenticate with Databricks.

- Token – this is the token used to access the Databricks server. The token can be obtained by navigating to the User Settings page of your Databricks instance and selecting the Access Tokens tab.

o AzureServicePrincipal – select this when you need to authenticate a non-human entity, e.g., an application, service, or scheduled job, that needs to connect to Databricks without a user logging in.

A **Service Principal** is essentially an identity for an application or service to authenticate and access Azure resources (like Databricks) securely without requiring user credentials.

Typically, this is used in scenarios where you need a non-interactive or programmatic connection to Databricks.

Service Principal does **not** use Single-Sing-On (SSO). It is an identity created for a service or application to authenticate to Azure resources, so it doesn't rely on user-based login methods. Nor does it use Multi-Factor-Authentication (MFA). MFA typically requires user interaction, but since a Service Principal is not tied to a user and operates automatically, it doesn't support MFA.

If you're setting up a connection for a user manually, go with **Azure AD**. If it's for an automated process or service, go with **Azure Service Principal**.

Selecting **AzureServicePrincipal** requires you to specify the following:

- Azure Tenant ID – this is the Tenant Id of your Microsoft Azure Active Directory.

- Azure Client Id – this is the application (client) Id of your Microsoft Azure Active Directory application. The application(client) can be registered following the AuthScheme -> AzureServicePrincipal.

- Azure Client Secret – this is the application (client) secret of your Microsoft Azure Active Directory application. The application(client) can be registered following the AuthScheme -> AzureServicePrincipal.

- Azure Subscription Id – this is the Subscription Id of your Azure Databricks Service Workspace.

- Azure Resource Group – this is the Resource Group name of your Azure Databricks Service Workspace.

- Azure Workspace – this is the name of your Azure Databricks Service Workspace.

o AzureAD – select this if you want to authenticate using your **Azure Active Directory** credentials. This is ideal when you, as a user, are directly logging in to Databricks through the ODBC connection, and you want to use your **personal Azure AD account** for authentication.

It is a good choice for scenarios where a **human user** is involved and wants to use their regular Azure AD login (email and password).

Arbutus Connectors

When using Azure AD, you can leverage Single-Sign-On (SSO). This means if you're already logged into your Azure account, you won't need to re-enter your credentials to authenticate with Databricks. As well, Multi-Factor Authentication (MFA) can be enforced on Azure AD accounts. If your organization requires MFA for accessing Azure resources, you will be prompted for MFA when you try to authenticate via Azure AD.

If you're setting up a connection for a user manually, go with **Azure AD**. If it's for an automated process or service, go with **Azure Service Principal**.

Selecting **AzureAD** requires you to specify the following:
- Azure Tenant – this is Microsoft Online tenant being used to access data. If not specified, your default tenant is used.

- OAuth Client ID – this is client Id assigned when you register your application with an OAuth authorization server.

  OAuth Client Id is one of a handful of connection parameters that need to be set before users can authenticate via OAuth.

- Callback URL – this is the OAuth callback URL to return to when authenticating. This value must match the callback URL you specify in your App settings.

  During the authentication process, the OAuth authorization server redirects the user to this URL. This value must match the callback URL you specified when you created your custom OAuth application.

4. **Database** – this is the name of the Databricks database.

Arbutus Connectors

E2. Editing the DSN properties in the Advanced tab

This tab includes more detailed and technical properties. It is intended for those users who need more control over the configuration and are comfortable with more complex options. The **Advanced** tab often includes properties that can fine-tune the behaviour of the system feature.

If you have already completed editing the properties in the **Basic** tab, as required, you do not necessarily need to also complete editing the properties in the **Advanced** tab. Instead, once you have completed editing the properties in the **Basic** tab, you may opt to proceed to testing the connection to the Databricks system by clicking the **Test Connection** button.
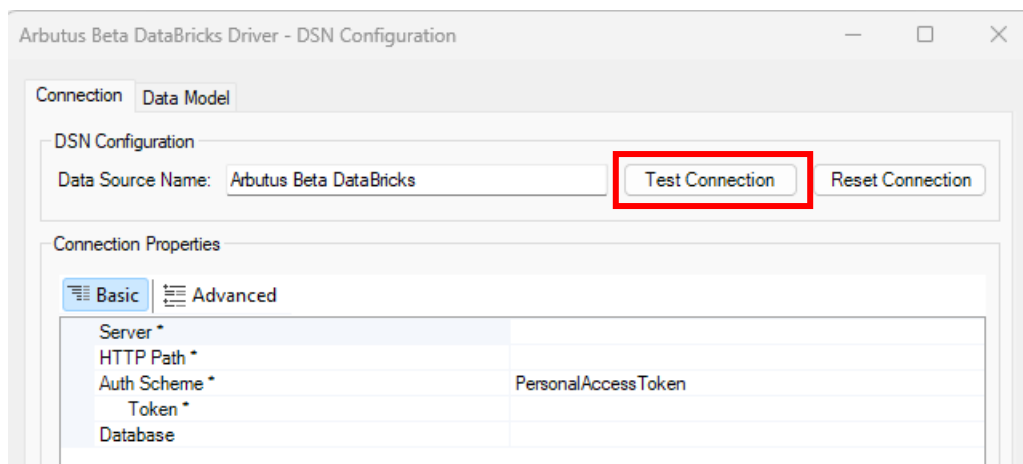
There are a lot more properties included for editing in the **Advanced** tab.

However, it is useful to know that each property does provide a short description of it and as such serves as a guide in terms of what to edit and/or enter. This short description can be seen at the bottom of the **DSN Configuration** dialog box, as seen in the screenshot below.

Arbutus Connectors
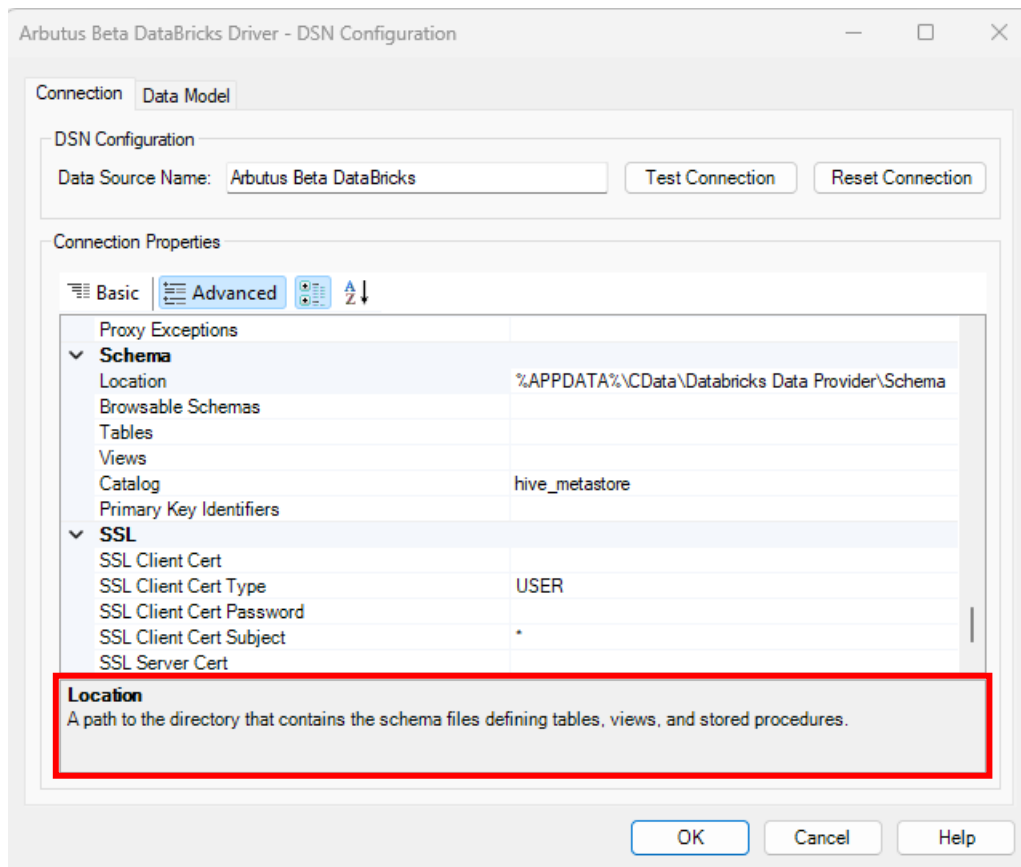
If it is deemed necessary to complete some/all the properties in the **Advanced** tab, it is recommended that you refer to the description shown for any of the properties being edited and/or entered.

If required, more information on the properties listed in the **Advanced** tab can also be provided.

F. Other questions and/or request for assistance

There may be times when you need to consult with the technical support team at Arbutus Software. If so, please send an email request to support@ArbutusSoftware.com.

For more information, please refer to the CONTACT US page on our website.