Register and Scan Hive Metastore source (preview)

This article outlines how to register a Hive Metastore database in Purview and set up a scan.

Supported capabilities

The Hive Metastore source supports Full scan to extract metadata from a Hive Metastore database and fetches Lineage between data assets.

Prerequisites

1. Set up the latest self-hosted integration runtime. For more information, see Create and configure a self-hosted integration runtime.
2. Make sure JDK 11 is installed on your virtual machine where self-hosted integration runtime is installed.
3. Make sure "Visual C++ Redistributable 2012 Update 4" is installed on the self-hosted integration runtime machine. If you don't yet have it installed, download it from here.
4. You will have to manually download the Hive Metastore database's JDBC driver on your virtual machine where self-hosted integration runtime is running. For example, if the database used is mssql, make sure to download Microsoft's JDBC driver for SQL Server.
   Note: The driver should be accessible to all accounts in the VM. Do not install it in a user account.
5. Supported Hive versions are 2.x to 3.x.

Feature Flag

Registration and scanning of Hive Metastore source is available behind a feature flag. Append the following to your URL: &feature.ext.datasource={"hive":"true"}

E.g., full URL https://web.purview.azure.com/?feature.ext.datasource={"hive":"true"}

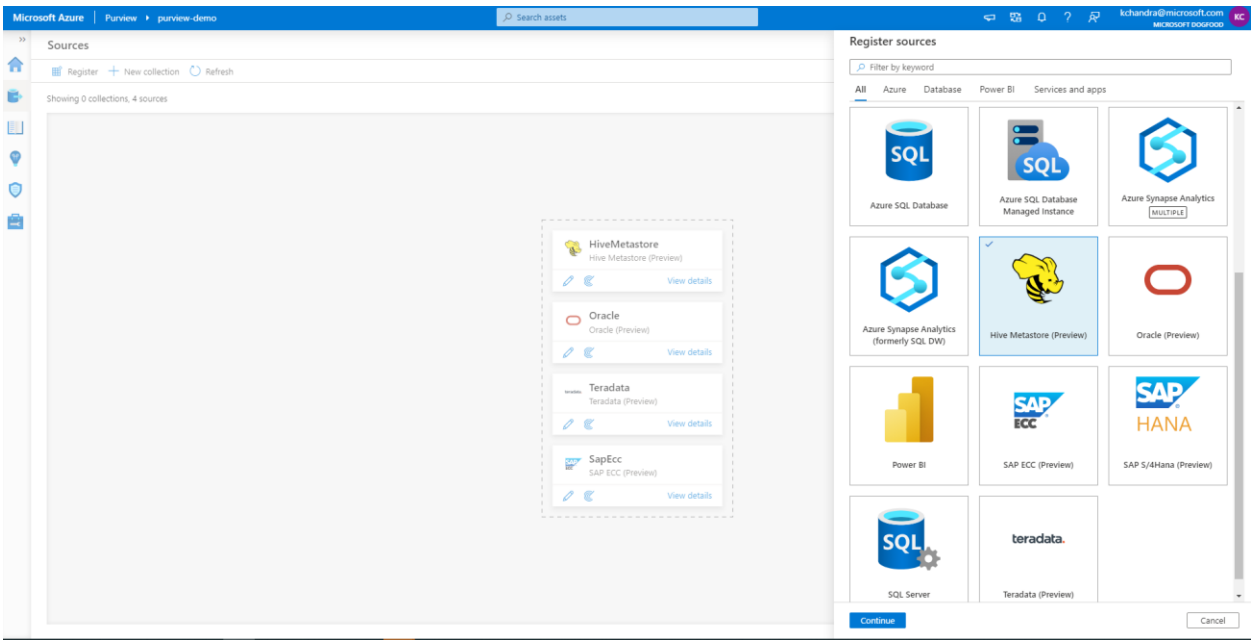
Setting up authentication for a scan

The only supported authentication for a Hive Metastore database is **Basic authentication.**

Register a Hive Metastore database

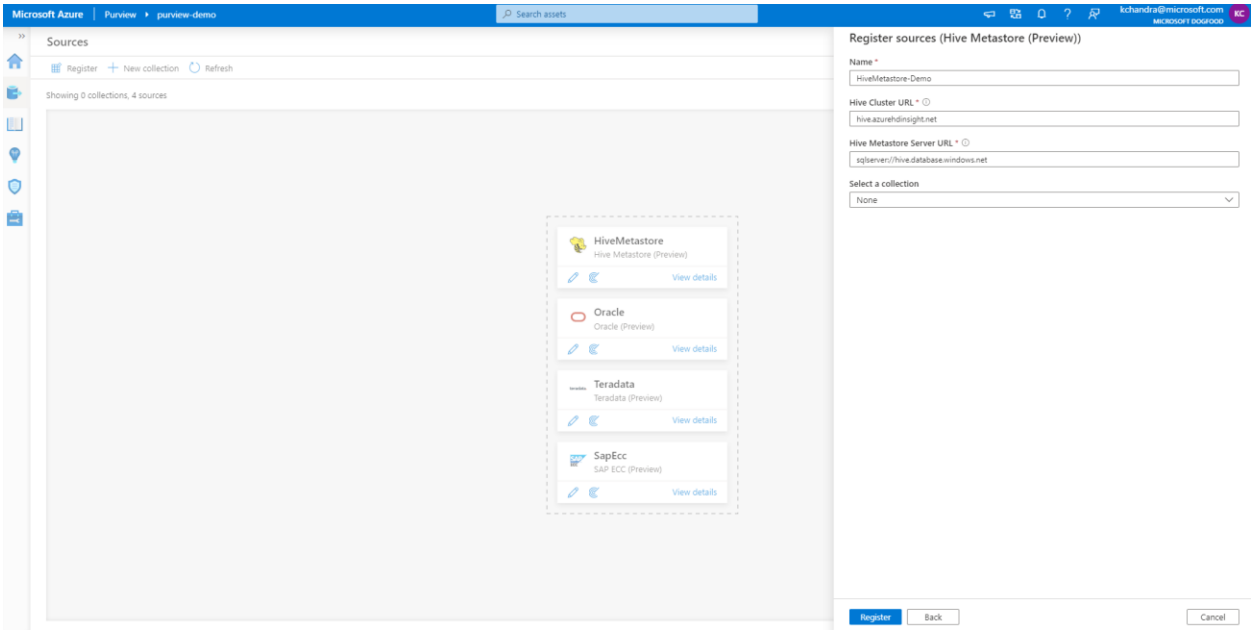
To register a new Hive Metastore database in your data catalog, do the following:

1. Navigate to your Purview account.
2. Select **Sources** on the left navigation.
3. Select **Register.**
4. On Register sources, select Hive **Metastore**. Select **Continue.**



On the Register sources (Hive Metastore) screen, do the following:

1. Enter a **Name** that the data source will be listed within the Catalog.
2. Enter the **Hive Cluster URL.** For example, hive.azurehdinsight.net
3. Enter the **Hive Metastore Server URL.** For example, sqlserver://hive.database.windows.net
4. Select a collection or create a new one (Optional)
5. Finish to register the data source.



Creating and running a scan

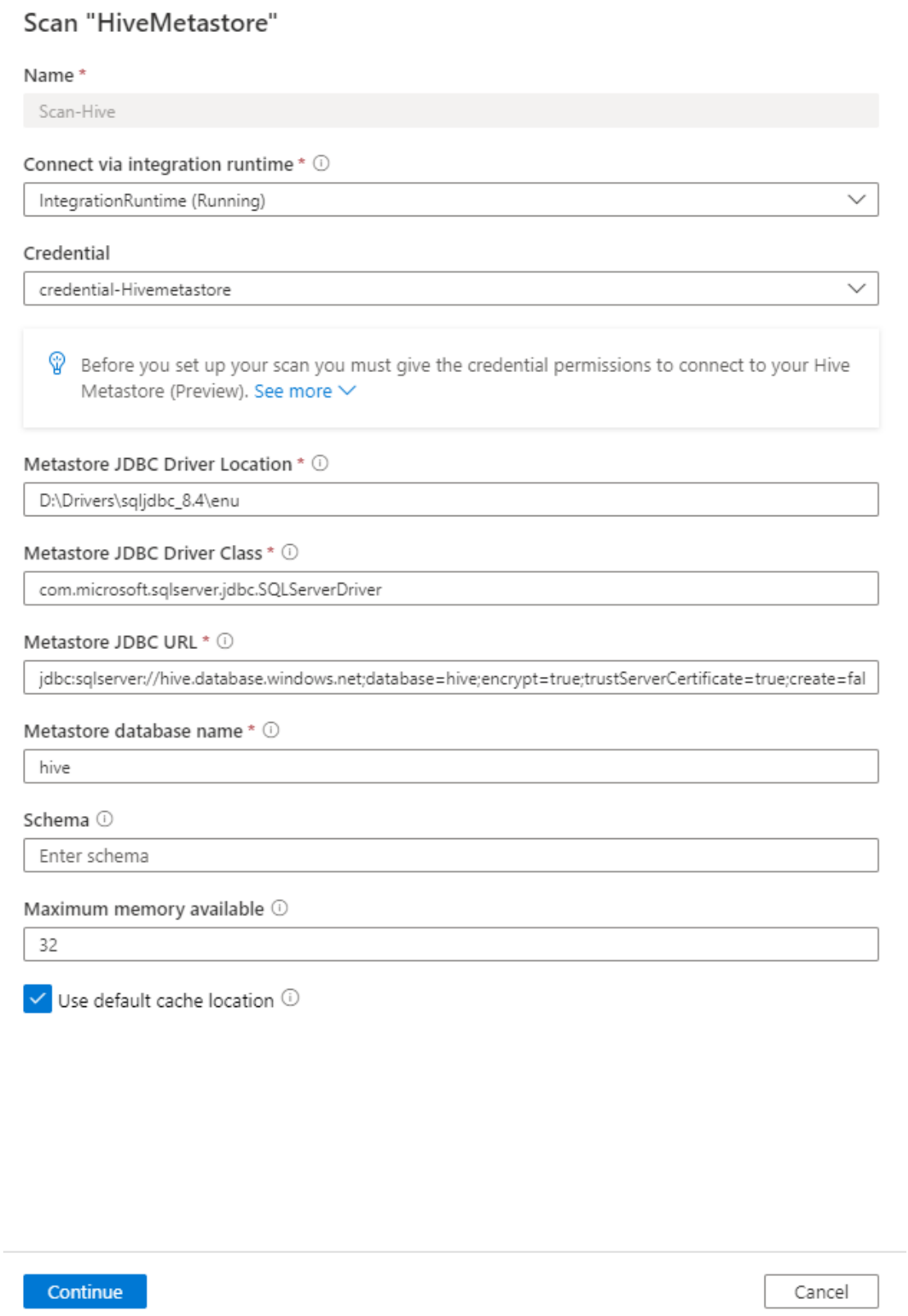
To create and run a new scan, do the following:

1. In the Management Center, click on Integration runtimes. Make sure a self-hosted integration runtime is set up. If it is not set up, use the steps mentioned here to setup a self-hosted integration runtime
2. Navigate to **Sources**.
3. Select the registered **Hive Metastore** database.
4. Select **+ New scan**.
5. Provide the below details:
    a. **Name**: The name of the scan
    b. **Connect via integration runtime**: Select the configured self-hosted integration runtime
    c. **Credential**: Select the credential to connect to your data source.
    Make sure to:
        - Select Basic Authentication while creating a credential.
        - Provide the Metastore username in the User name input field
        - Store the Metastore password in the secret key.

    To understand more on credentials, refer to the link here

    d. **Metastore JDBC Driver Location**: Specify the path to the JDBC driver location in your VM where self-host integration runtime is running. This should be the path to valid JARs folder location.
    e. **Metastore JDBC Driver Class**: Provide the connection driver name. For example, com.microsoft.sqlserver.jdbc.SQLServerDriver
    f. **Metastore JDBC URL**: Provide the Connection URL value and define connection to DB server URL. For example, jdbc:sqlserver://hive.database.windows.net;database=hive;encrypt=true;trustServerCertificate=true;create=false;loginTimeout=300
    g. **Metastore database name**: Provide the Hive Metastore Database name
    h. **Schema**: Specify a list of Hive schemas to import. For example, schema1; schema2. All user schemas are imported if that list is empty. All system schemas (for example, SysAdmin) and objects are ignored by default. When the list is empty, all available schemas are imported.
    Acceptable schema name patterns using SQL LIKE expressions syntax include using %, e.g. A%; %B; %C%; D
        - start with A or
        - end with B or
        - contain C or
        - equal D

    Usage of NOT and special characters are not acceptable.
    i. **Maximum memory available**: Maximum memory (in GB) available on customer's VM to be used by scanning processes. This is dependent on the size of Hive Metastore database to be scanned.



6. Click on **Continue**.
7. Choose your **scan trigger**. You can set up a schedule or ran the scan once.
8. Review your scan and click on **Save and Run**.