Schedule Databricks jobs in a Data Factory pipeline

In this reading you can see the steps involved in the process of scheduling Databricks jobs in a Data Factory pipeline.

Note

You are not required to complete the processes, tasks, activities, or steps presented in this example. Your system set-up may differ from the system set-up in the demonstration in this reading. The various samples provided are for illustrative purposes only and it's likely that if you try this out you will encounter issues in your system.

Azure Data Factory is a cloud-based ETL and data integration service that allows the creation of data-driven workflows for orchestrating data movement and transforming data at scale. With Azure Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.

A data factory can have one or more pipelines. A pipeline is a logical grouping of activities that together perform a task. For example, a pipeline could contain a set of activities that ingest and clean log data, and then kick off a mapping data flow to analyze the log data. The pipeline allows you to manage the activities as a set instead of each one individually. You deploy and schedule the pipeline instead of the activities independently.

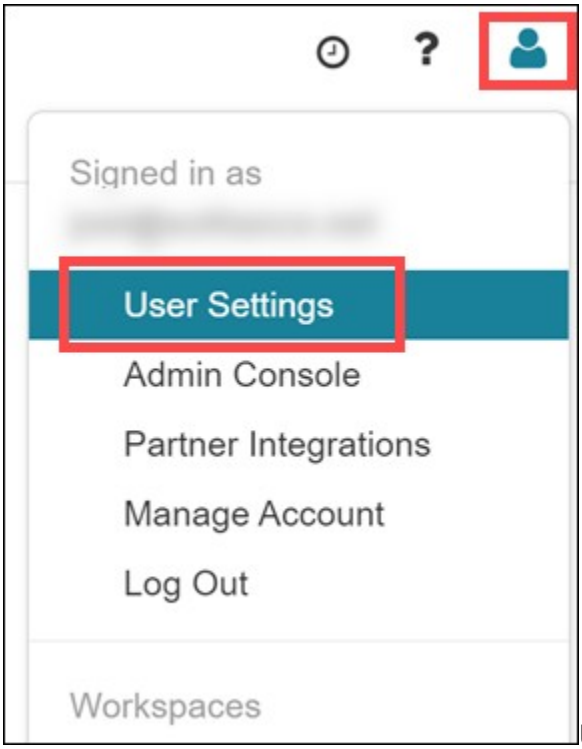
While here we will show scheduling a notebook with the Data Factory UI, you can also schedule `.jar` and `.py` files to take advantage of the much lower cost of Data Engineering vs. interactive clusters. You can find pricing details at this link.

Retrieve Access Token from the Azure Databricks workspace

Go to your Azure Databricks workspace and follow these steps to generate a user access token for the Data Factory pipeline that you will create later:

1. Navigate to your Azure Databricks workspace.

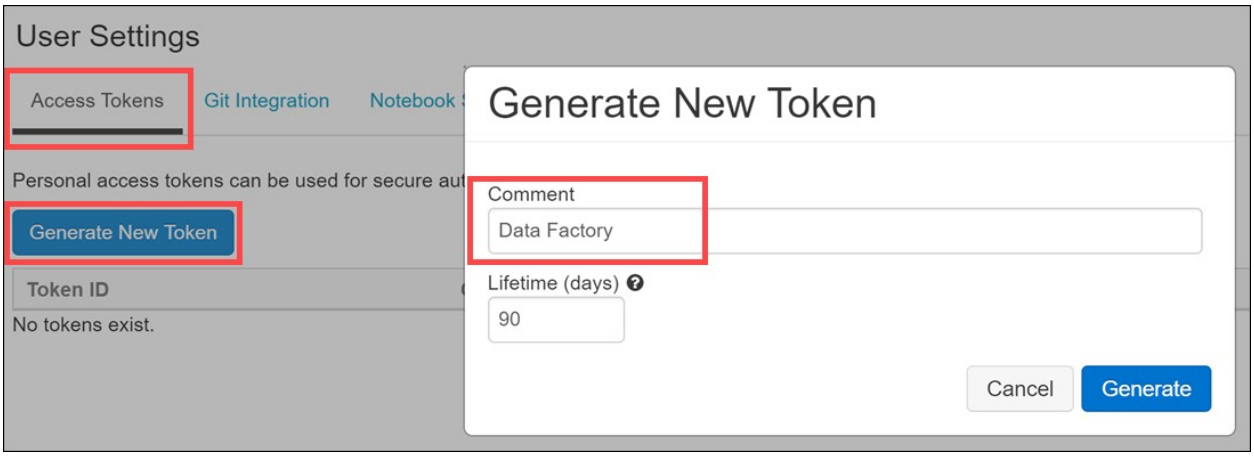
2. Select the **User Settings** option from the user icon on the top-right of the workspace.



User Settings menu link.

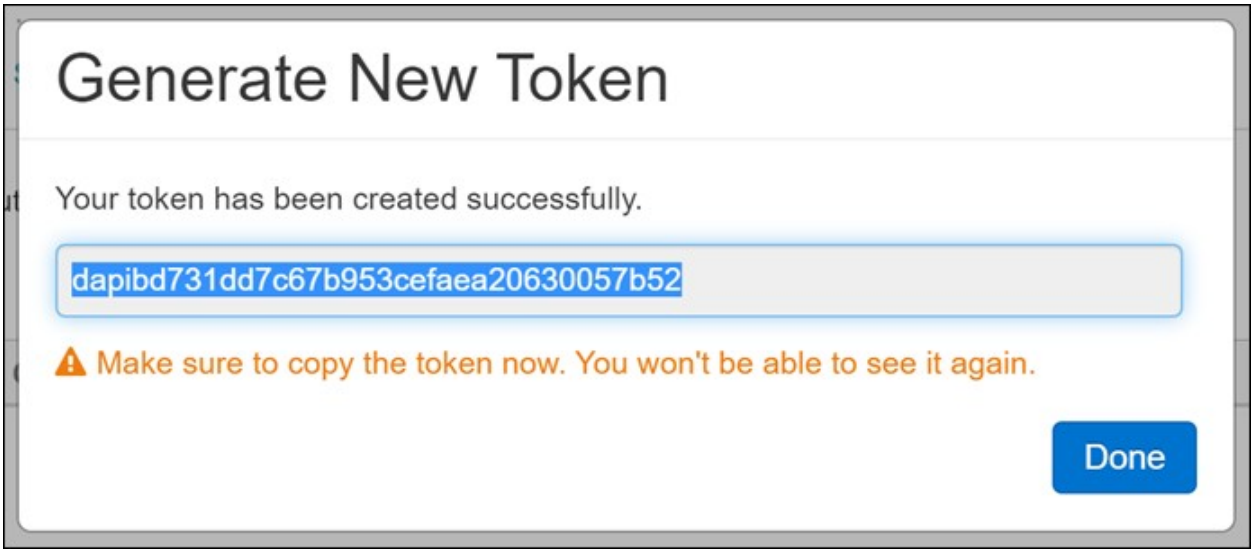
3. Under the Access Tokens tab, select **Generate New Token**.

4. In the Generate New Token dialog, add **Data Factory** for the Comment, then select **Generate**.



Generate New Token form is displayed.

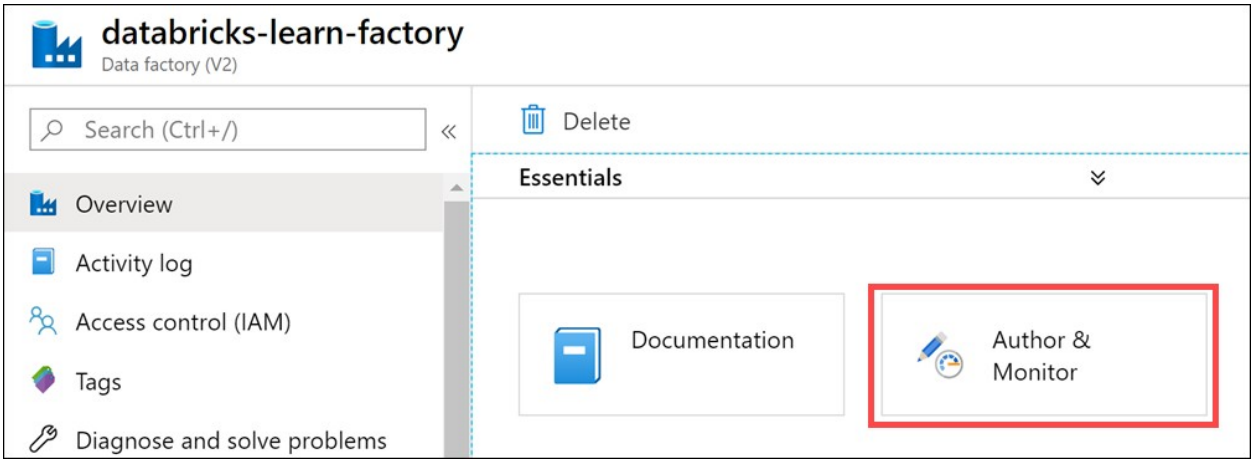
5. Copy the new token and save it to a text editor for later reference. It is important that you record the token as it is only displayed once.



New Token is displayed.

Navigate to Azure Databricks Linked Service

1. Open the Azure Data Factory service in Azure. Select **Author & Monitor** to open Data Factory in a new browser tab.



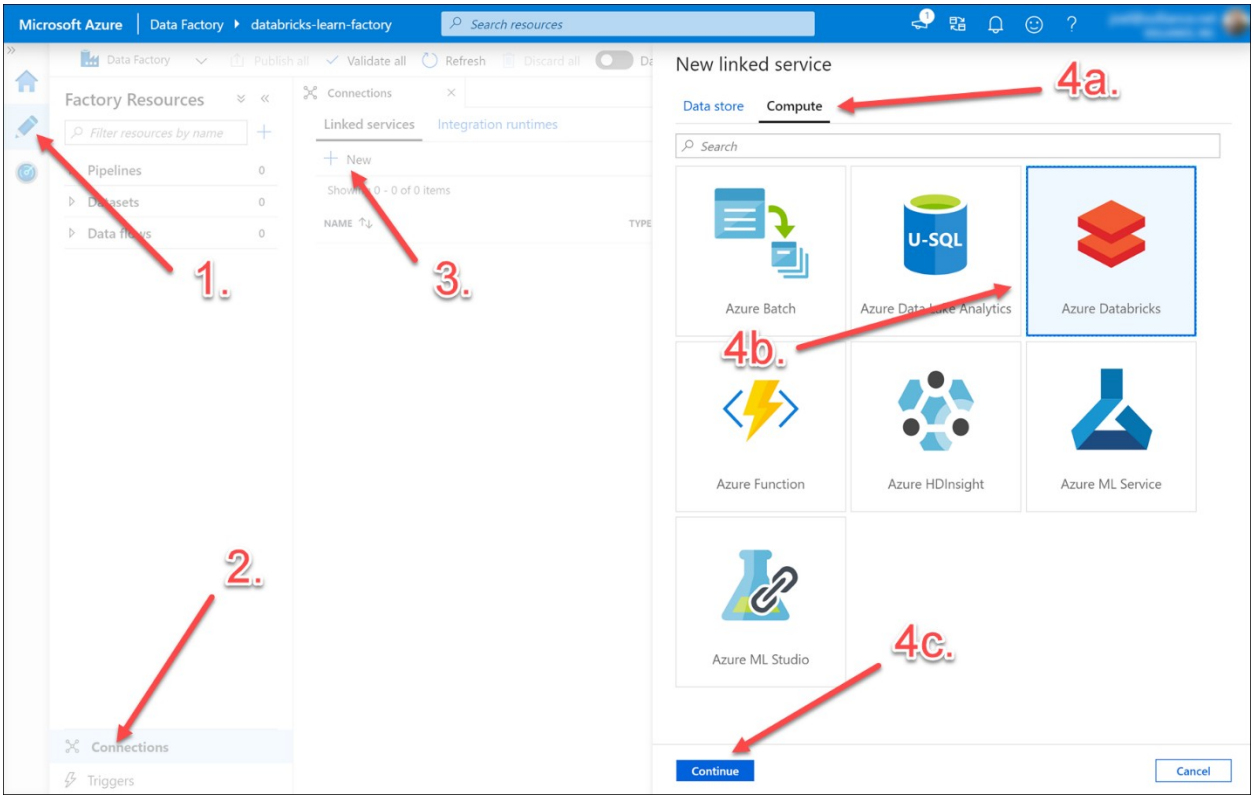
Author and monitor link.
2. On the **Let's get started** page, select the Pencil icon in the left pane to switch to the Author tab.

3. Select **Connections** at the lower left corner of the Factory Resources pane.

4. Select **+ New** in the Connections tab under **Linked Services**.

5. In the New Linked Service window, select the **Compute** tab,

6. Then select the **Azure Databricks** tile and select **Continue**.



New linked service.

Configure Linked Service

Important: *This form should be completed from top to bottom. All fields that are not mentioned below can be left with defaults.*

1. Select your current subscription from the drop down for Azure subscription.

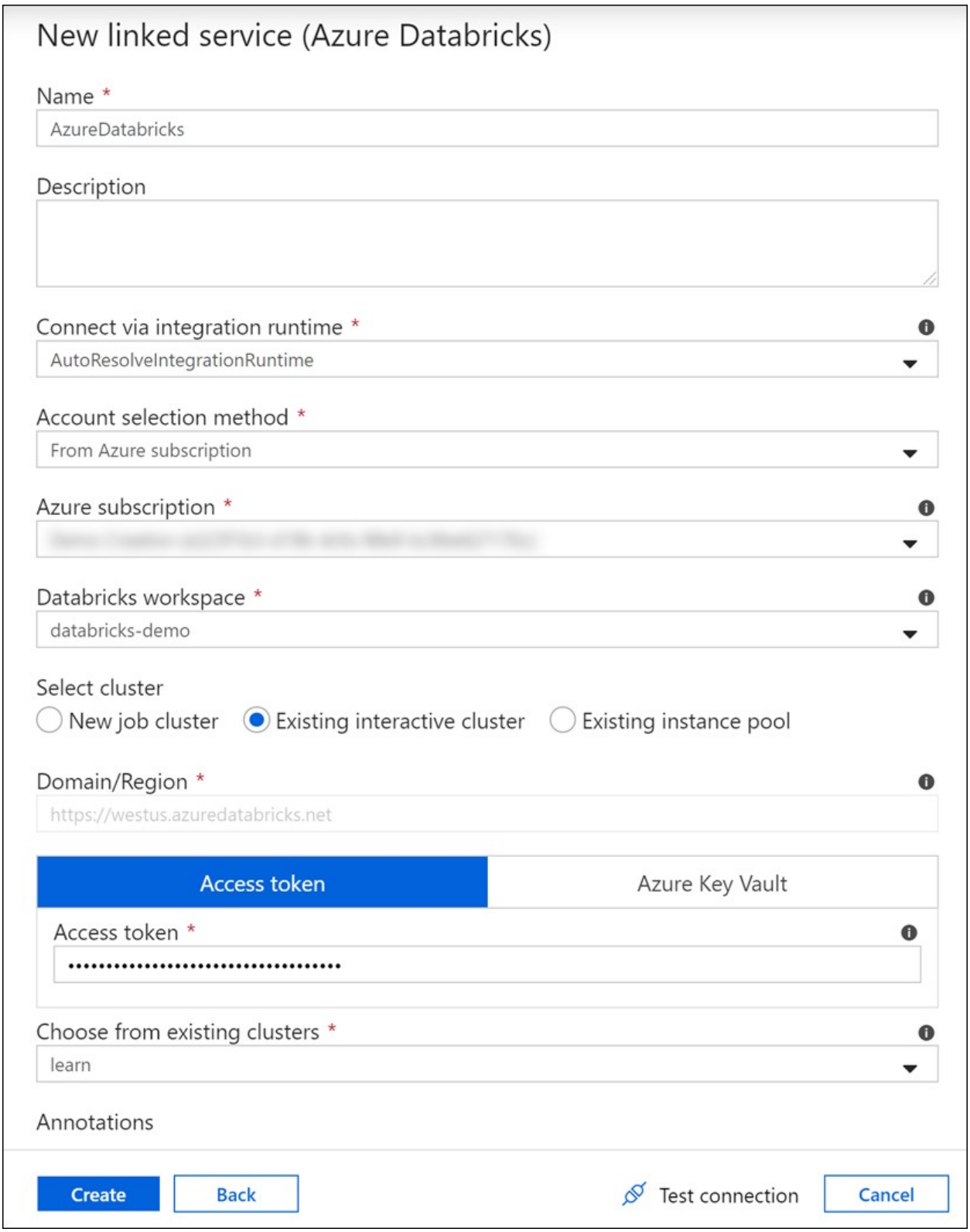
2. Select the Databricks workspace for this module.

3. For Select cluster, select **Existing interactive cluster** (Note that **New job cluster** is preferred for triggered pipelines as they use a lower cost engineering tier cluster).

4. For Access token, paste the access token you created in Step 1.

5. Select the name of your cluster from the drop down list under **Choose from existing clusters**.

6. Select **Create**.



New linked service form.

Create an ADF Pipeline and Add a Databricks Notebook Activity

1. Hover over the number to the right of Pipelines and select the ellipses that appears.

2. Select **New pipeline**.

3. In the Activities panel to the right of the Factory Resources panel, select **Databricks** to expand this section.

4. Drag the Notebook option into the tableau to the right.