# Exercise: Deduplication of data

In this unit, you need to complete the exercises within a Databricks Notebook.

To begin, you need to have access to an Azure Databricks workspace. If you do not have a workspace available, follow the instructions below.

## Unit Pre-requisites

**Microsoft Azure Account**: You will need a valid and active Azure account for the Azure labs. If you do not have one, you can sign up for a free trial

- If you are a Visual Studio Active Subscriber, you are entitled to Azure credits per month. You can refer to this link to find out more, including how to activate and start using your monthly Azure credit.
- If you are not a Visual Studio Subscriber, you can sign up for the FREE Visual Studio Dev Essentials program to create Azure free account.

## Create the required resources

To complete this lab, you will need to deploy an Azure Databricks workspace in your Azure subscription.

## Deploy an Azure Databricks workspace

1. Click the following button to open the Azure Resource Manager Template (ARM) template in the Azure portal. Deploy Databricks from the ARM Template

2. Provide the required values to create your Azure Databricks workspace:

- **Subscription**: Choose the Azure Subscription in which to deploy the workspace.
- **Resource Group**: Leave at Create new and provide a name for the new resource group.
- **Location**: Select a location near you for deployment. For the list of regions supported by Azure Databricks, see Azure services available by region.
- **Workspace Name**: Provide a name for your workspace.
- **Pricing Tier**: Ensure `premium` is selected.

3. Accept the terms and conditions.

4. Select Purchase.

5. The workspace creation takes a few minutes. During workspace creation, the portal displays the Submitting deployment for Azure Databricks tile on the right side. You may need to scroll right on your dashboard to see the tile. There is also a progress bar displayed near the top of the screen. You can watch either area for progress.

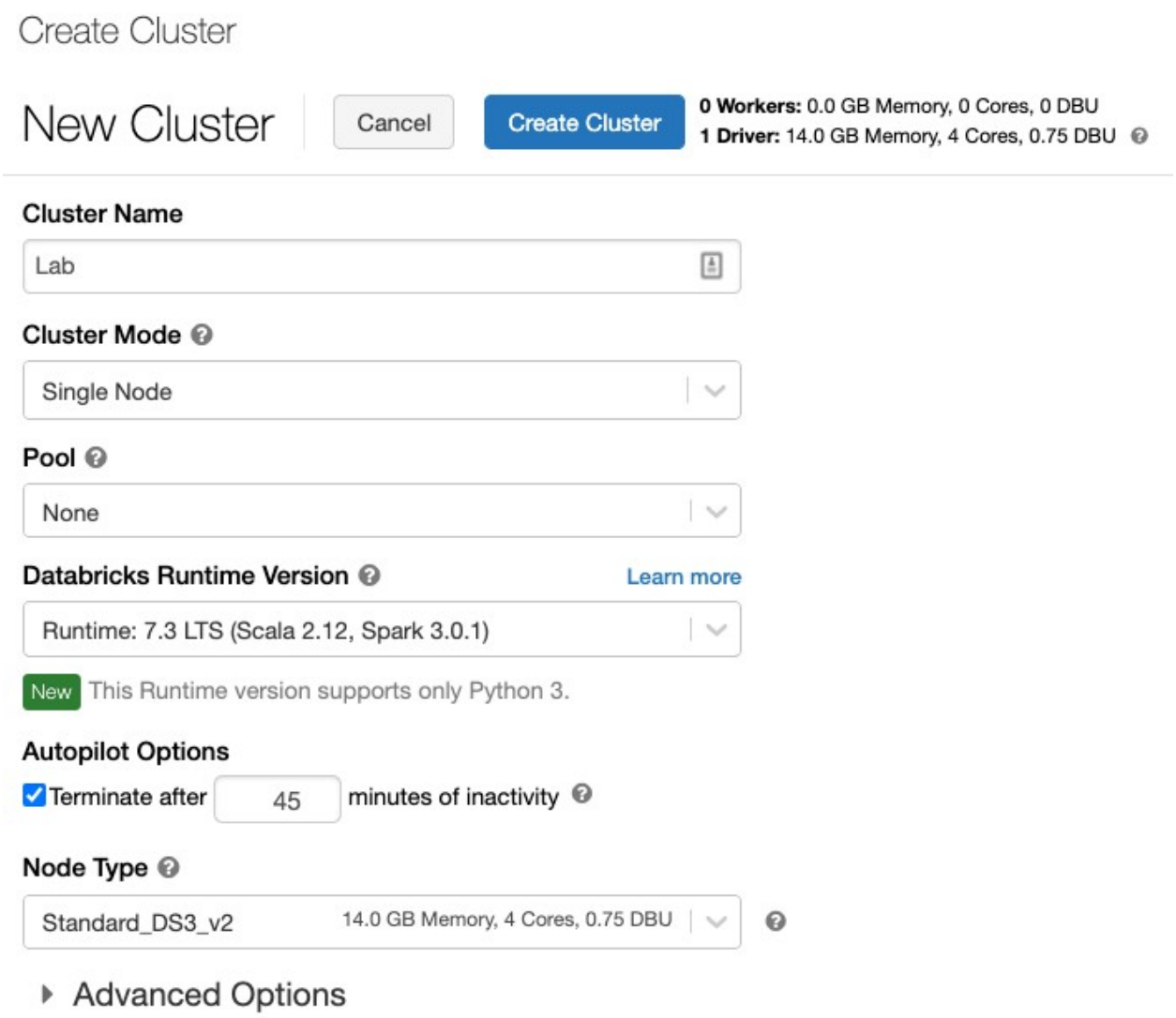
## Create a cluster

1. When your Azure Databricks workspace creation is complete, select the link to go to the resource.

2. Select **Launch Workspace** to open your Databricks workspace in a new tab.

3. In the left-hand menu of your Databricks workspace, select **Clusters**.

4. Select **Create Cluster** to add a new cluster.




The create cluster page.
5. Enter a name for your cluster. Use your name or initials to easily differentiate your cluster from your coworkers.

6. Select the **Cluster Mode**: **Single Node**.

7. Select the **Databricks RuntimeVersion**: **Runtime: 7.3 LTS (Scala 2.12, Spark 3.0.1)**.

8. Under **Autopilot Options**, leave the box **checked** and in the text box enter `45`.

9. Select the **Node Type**: **Standard_DS3_v2**.

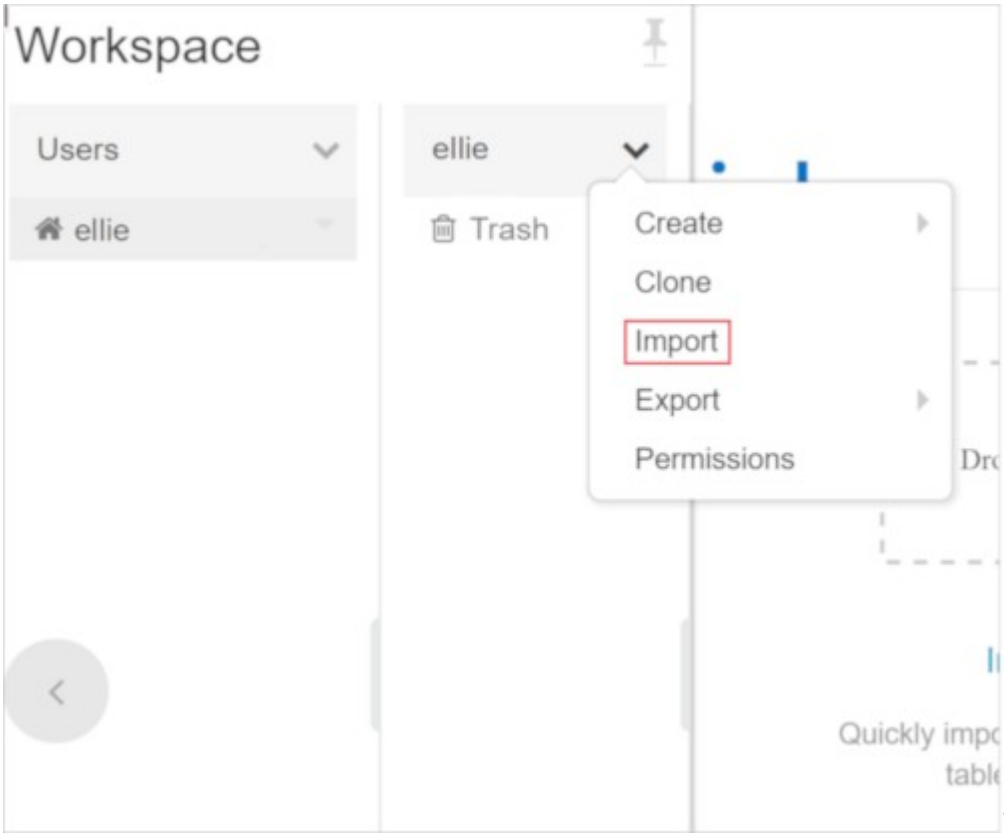
10. Select **Create Cluster**.

# Clone the Databricks archive

1. If you do not currently have your Azure Databricks workspace open: in the Azure portal, navigate to your deployed Azure Databricks workspace and select **Launch Workspace**.

2. In the left pane, select **Workspace** > **Users**, and select your username (the entry with the house icon).

3. In the pane that appears, select the arrow next to your name, and select **Import**.



The menu option to import the archive.

4. In the **Import Notebooks** dialog box, select the URL and paste in the following URL:



https://github.com/solliancenet/microsoft-learning-paths-databricks-notebooks/blob/master/data-engineering/DBC/03-Reading-and-writing-data-in-Azure-Databricks.dbc?raw=true

5. Select **Import**.

# Complete the following Notebook:

In your Azure Databricks workspace, open the **07-Dataframe-Advanced-Methods** folder that you imported within your user folder.

Open the **3.Exercise-Deduplication-of-Data** notebook. Make sure you attach your cluster to the notebook before following the instructions and running the cells within.

The goal of this exercise is to put into practice some of what you have learned about using DataFrames, including renaming columns. The instructions are provided within the notebook, along with empty cells for you to do your work. At the bottom of the notebook are additional cells that will help verify that your work is accurate.

**Note:** You will find a corresponding notebook within the **`Solutions`** subfolder. This contains completed cells for the exercise. Refer to the notebook if you get stuck or simply want to see the solution.

After you've completed the notebook, return to this screen, and continue to the next step.