Lab 3 - Data exploration and transformation in Azure Databricks

This lab teaches how to use various Apache Spark DataFrame methods to explore and transform data in Azure Databricks. You will learn how to perform standard DataFrame methods to explore and transform data. You will also learn how to perform more advanced tasks, such as removing duplicate data, manipulate date/time values, rename columns, and aggregate data.

After completing this lab, you will be able to:

- Use DataFrames in Azure Databricks to explore and filter data
- Cache a DataFrame for faster subsequent queries
- Remove duplicate data
- Manipulate date/time values
- Remove and rename DataFrame columns
- Aggregate data stored in a DataFrame

Lab setup and pre-requisites

Before starting this lab, ensure you have successfully completed the setup steps to create your lab environment. Additionally, you require an Azure Databricks cluster, which you should have created in Lab 1. If you did not complete lab 1 (or you have deleted your cluster), the instructions below include steps to create one.

Exercise 1 - Working with DataFrames

In this exercise, you'll use some Databricks notebooks to learn fundamentals concepts and techniques for working with DataFrames.

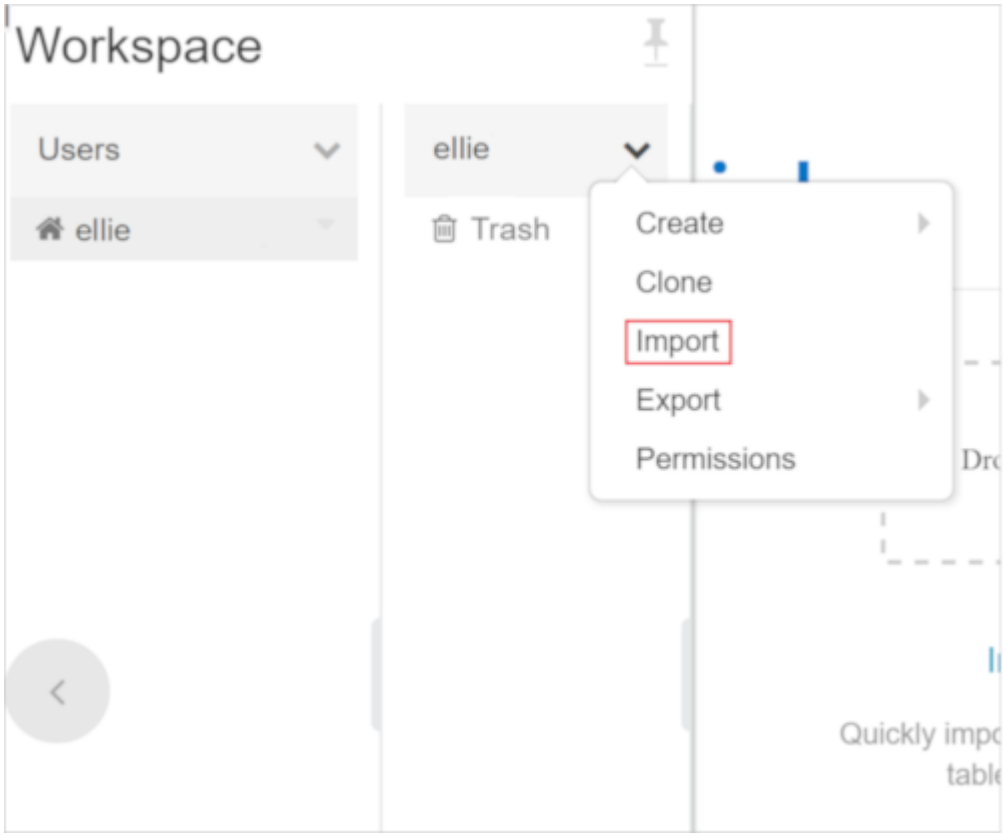
Task 1: Clone the Databricks archive

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

2. In the left pane, select **Compute**. If you have an existing cluster, ensure that it is running (start it if necessary). If you don't have an existing cluster, create a single-node cluster that uses the latest runtime and **Scala 2.12** or later.

3. When your cluster is running, in the left pane, select **Workspace** > **Users**, and select your user name (the entry with the house icon).

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



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

```
https://github.com/MicrosoftLearning/DP-203-Data-Engineer/raw/master/Allfiles/microsoft-learning-paths-databricks-notebooks/data-engineering/DBC/04-Working-With-Dataframes.dbc
```

6. Select **Import**.

7. Select the **04-Working-With-Dataframes** folder that appears.

Task 2: Run the *Describe a DataFrame* notebook

1. Open the **1.Describe-a-dataframe** notebook.
2. Attach your cluster to the notebook before following the instructions and running the cells it contains. Within the notebook, you will:

- Develop familiarity with the DataFrame APIs
- Learn how to work with **SparkSession** and **DataFrame** (aka ***Dataset[Row]***) classes.
- Learn how to use the **count** action.

Task 3: Run the *Working with DataFrames* notebook

1. In your Azure Databricks workspace, in the **04-Working-With-Dataframes** folder, open the **2.Use-common-dataframe-methods** notebook.

2. Attach your cluster to the notebook before following the instructions and running the cells within. Within the notebook, you will:

- Develop familiarity with the DataFrame APIs
- Use common DataFrame methods for performance
- Explore the Spark API documentation

Task 4: Run the *Display Function* notebook

1. In your Azure Databricks workspace, in the **04-Working-With-Dataframes** folder, open the **3.Display-function** notebook.

2. Attach your cluster to the notebook before following the instructions and running the cells within. Within the notebook, you will:

   - Learn to use the following transformations:
     - limit(..)
     - select(..)
     - drop(..)
     - distinct()`
     - dropDuplicates(..)
   - Learn to use the following the actions:
     - show(..)
     - display(..)

Task 5: Complete the *Distinct Articles* exercise notebook

1. In your Azure Databricks workspace, in the **04-Working-With-Dataframes** folder, open the **4.Exercise: Distinct Articles** notebook.

2. Attach your cluster to the notebook before following the instructions and running the cells within. In this notebook, you read Parquet files, apply necessary transformations, perform a total count of records, then verify that all the data was correctly loaded. As a bonus, you can try defining a schema that matches the data and update the read operation to use the schema.

   > 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.

Exercise 2 - Working with DataFrames advanced methods

This exercise builds on the Azure Databricks DataFrames concepts learned in the previous lab above by exploring some advanced methods data engineers can use to read, write, and transform data using DataFrames.

Task 1: Clone the Databricks archive

1. In your Databricks workspace, in the left pane, select **Workspace** and navigate your home folder (your username with a house icon).

2. Select the arrow next to your name, and select **Import**.

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

```
https://github.com/MicrosoftLearning/DP-203-Data-
Engineer/raw/master/Allfiles/microsoft-learning-paths-databricks-
notebooks/data-engineering/DBC/07-Dataframe-Advanced-Methods.dbc
```

4. Select **Import**.
5. Select the **07-Dataframe-Advanced-Methods** folder that appears.

Task 2: Run the *Date and Time Manipulation* notebook

1. In your Azure Databricks workspace, in the **07-Dataframe-Advanced-Methods** folder, open the **1.DateTime-Manipulation** notebook.
2. Attach your cluster to the notebook before following the instructions and running the cells within. You will explore more of the **sql.functions** operations as well as date & time functions.

Task 3: Run the *Use Aggregate Functions* notebook

1. In your Azure Databricks workspace, in the **07-Dataframe-Advanced-Methods** folder, open the **2.Use-Aggregate-Functions** notebook.
2. Attach your cluster to the notebook before following the instructions and running the cells within. Within the notebook, you will learn various aggregate functions.

Task 4: Complete the *De-Duping Data* exercise notebook

1. In your Azure Databricks workspace, in the **07-Dataframe-Advanced-Methods** folder, open the **3.Exercise-Deduplication-of-Data** notebook.
2. 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.

Important: Shut down your cluster

1. After you've completed the lab, in the left pane, select **Compute** and select your cluster. Then select **Terminate** to stop the cluster.