Perform basic data transformations

Task 1. Clone the Databricks archive

1. From the Azure portal, go to your Databricks workspace and select Launch workspace.
2. In the left pane, select Workspace, select Users, and then select your username (the entry with the house icon).
3. In the pane that appears, select the downward-pointing chevron next to your name, and then select Import.
4. In the Import Notebooks pane, select URL, and paste in the following URL:

   https://github.com/MicrosoftDocs/mslearn-perform-basic-data-transformation-in-azure-databricks/blob/master/DBC/05.1-Basic-ETL.dbc?raw=true

5. Select Import.
6. A folder named after the archive should appear. Select that folder. The folder contains one or more notebooks that you'll use in completing this lab.

Task 2. Complete the following notebooks

To complete the labs, continue working within your Azure Databricks workspace and open the new 05.1-Basic-ETL folder. Within the folder, you'll find Python, Scala, and Spark subfolders.

1. Choose the folder for the language you prefer to use, open the corresponding folder, and then open the notebook.
2. Follow the instructions within the notebook until you've completed the entire notebook. Then continue with the remaining notebooks in order:
   a. 01-Course-Overview-and-Setup: This notebook gets you started with your Databricks workspace.
   b. 02-ETL-Process-Overview: This notebook contains exercises to help you query large data files and visualize your results.
   c. 03-Connecting-to-Azure-Blob-Storage: You do basic data aggregation and joins in this notebook.
   d. 04-Connecting-to-JDBC: This notebook lists the steps for accessing data from various sources by using Databricks.
   e. 05-Applying-Schemas-to-JSON: In this notebook, you learn how to query JSON and hierarchical data with DataFrames.
   f. 06-Corrupt-Record-Handling: This notebook lists exercises that help you create an Azure Data Lake Storage Gen2 storage account and use Databricks DataFrames to query and analyze this data.
   g. 07-Loading-Data-and-Productionalizing: Here you use Databricks to query and analyze data stores in Azure Data Lake Storage Gen2.
   h. Parsing-Nested-Data: This notebook is in the Optional subfolder. It includes a sample project you can explore later.