# Basic ETL Process with AWS S3 and Microsoft Azure

## Prerequisites

Before running the code, ensure you have access to the following:

- A Databricks account with the ability to create and run notebooks.

- AWS S3 Bucket with your dataset stored in CSV format.

- Microsoft Azure Storage account with your dataset stored in CSV format.

## Setting Up the Environment

1. Databricks Workspace: Log in to your Databricks workspace.

2. Create a New Notebook: Import the Python scripts into new notebooks in your Databricks workspace.

## Running the Code

AWS S3

1. AWS Credentials: Ensure your AWS credentials (Access Key ID and Secret Access Key) are stored in a CSV file in the Databricks FileStore.

2. Open the AWS S3 Notebook: In your Databricks workspace, open the notebook created for the AWS S3 Python script.

3. Run All Cells: Execute all cells in the notebook. The code will:

- Mount the AWS S3 bucket to Databricks FS.

- Perform basic data transformations.

- Save the transformed data back to S3 in Parquet format.

4. Verify the Output: Check the mounted S3 bucket path to ensure the `New\_circuits` Parquet file has been created.

Microsoft Azure

1. Azure Blob Storage Configuration: Make sure the Azure Storage account key is correctly configured in the script.

2. Open the Microsoft Azure Notebook: In your Databricks workspace, open the notebook created for the Microsoft Azure Python script.

3. Run All Cells: Execute all cells in the notebook. The code will:

- Access the Azure Blob Storage.

- Perform similar data transformations.

- Save the transformed data to a new location in Azure Blob Storage in Parquet format.

4. Verify the Output: Check the new Blob Storage location to ensure the Parquet file has been created.

## Understanding the Code

- The code performs basic ETL tasks: extracting data from CSV files stored in cloud storage, transforming the data (e.g., renaming columns), and loading the transformed data into a new location in Parquet format.

- Both scripts utilize Databricks' Spark DataFrame API for data manipulation.

## Troubleshooting

- Authentication Issues: Ensure your cloud storage credentials are correct and have the necessary permissions.

- File Not Found: Verify the paths to your datasets in both AWS S3 and Azure Blob Storage are correct.

## For more detailed explanations and additional configurations, refer to the AWS and Azure documentation respectively.