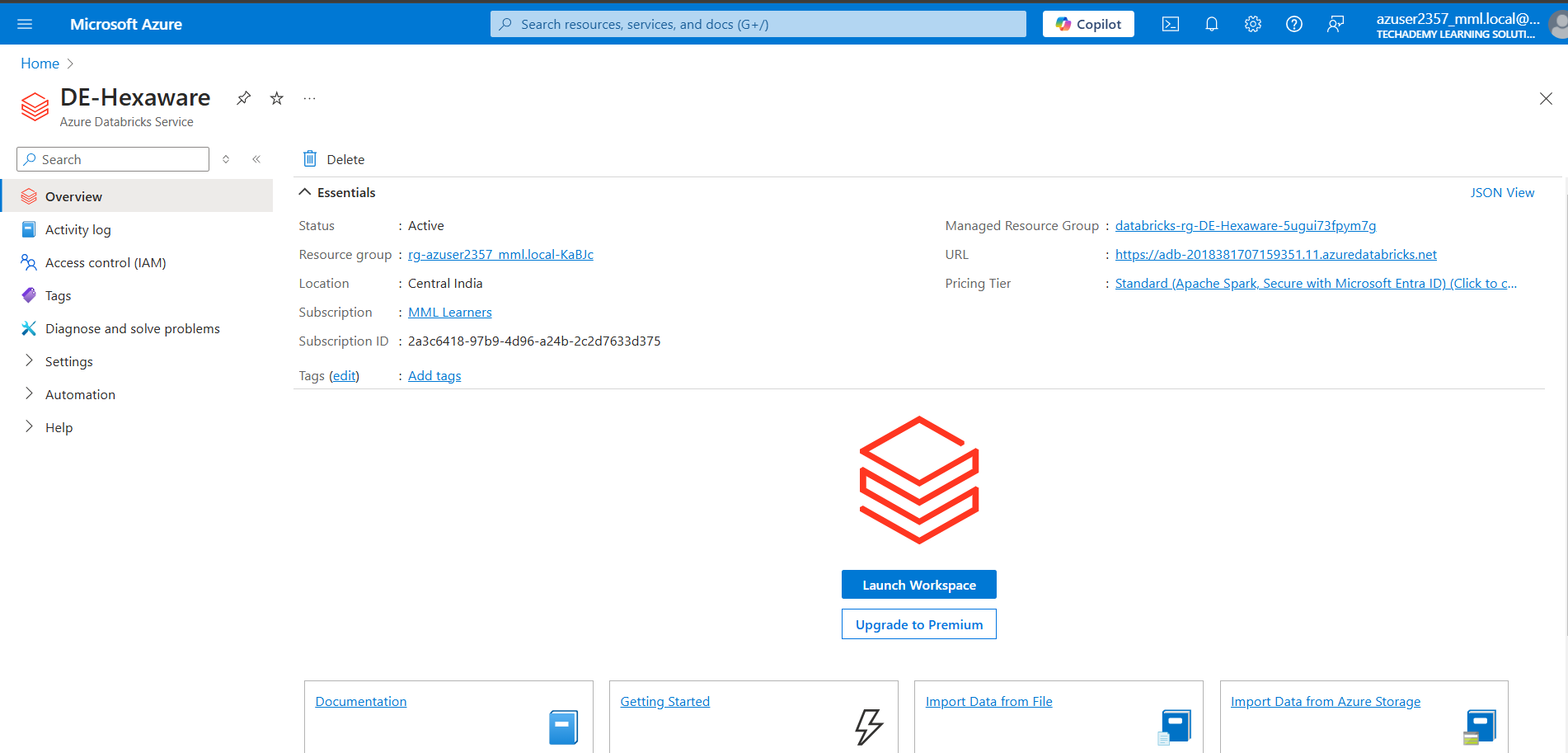
**Hexaware Technologies**

Azure-Case Study

**Setting Up the Workspace and Cluster in Azure Databricks**

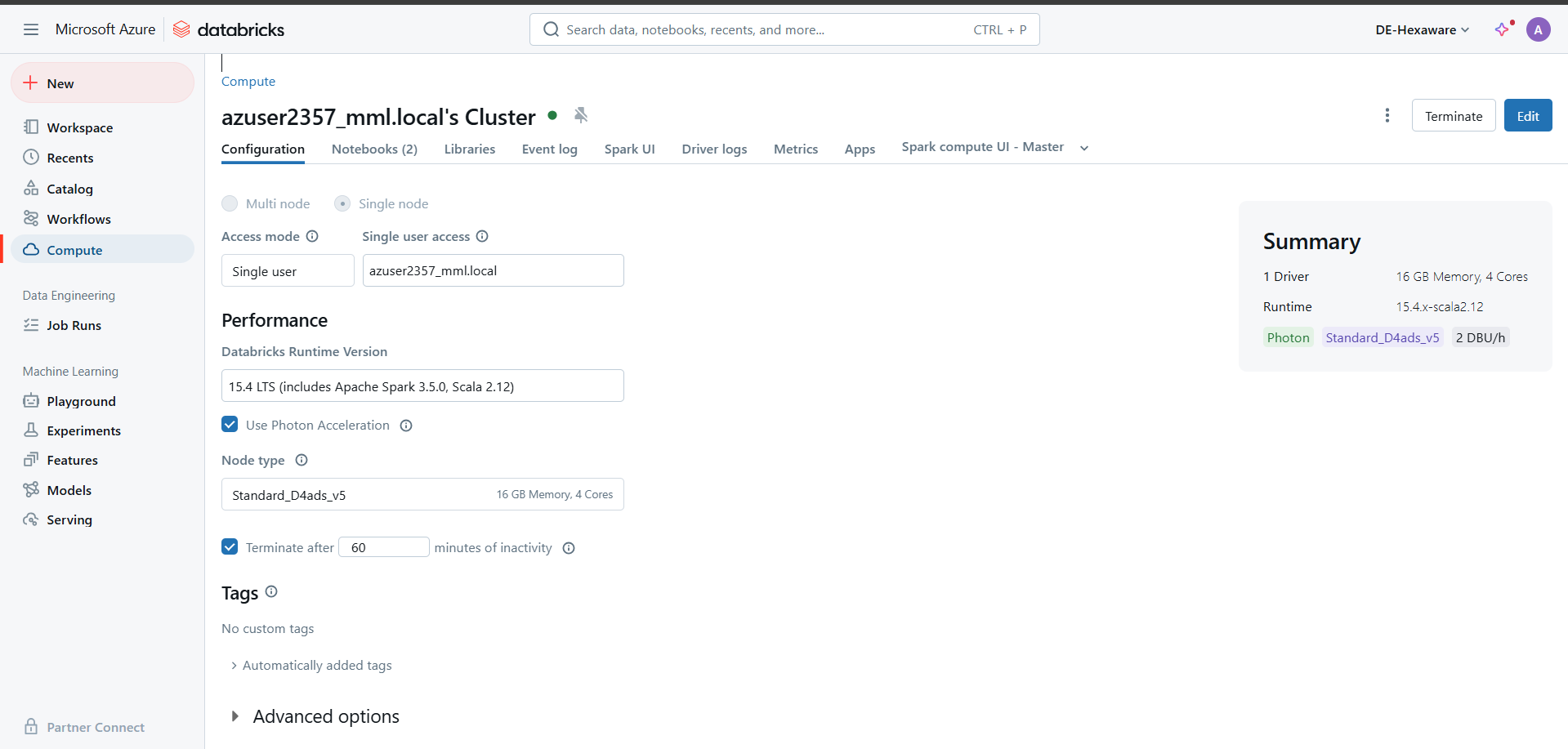
**Step 1: Setting Up the Azure Databricks Workspace**

1. **Accessing the Azure Databricks Portal**:
   * Log in to the **Azure portal** at <https://portal.azure.com>.
   * Navigate to the **Azure Databricks service**. create a new Databricks workspace by selecting “Create a resource” and searching for **Databricks**.
2. **Creating the Databricks Workspace**:
   * Once inside the Azure Databricks portal, click on **Launch Workspace** to enter the Databricks environment.



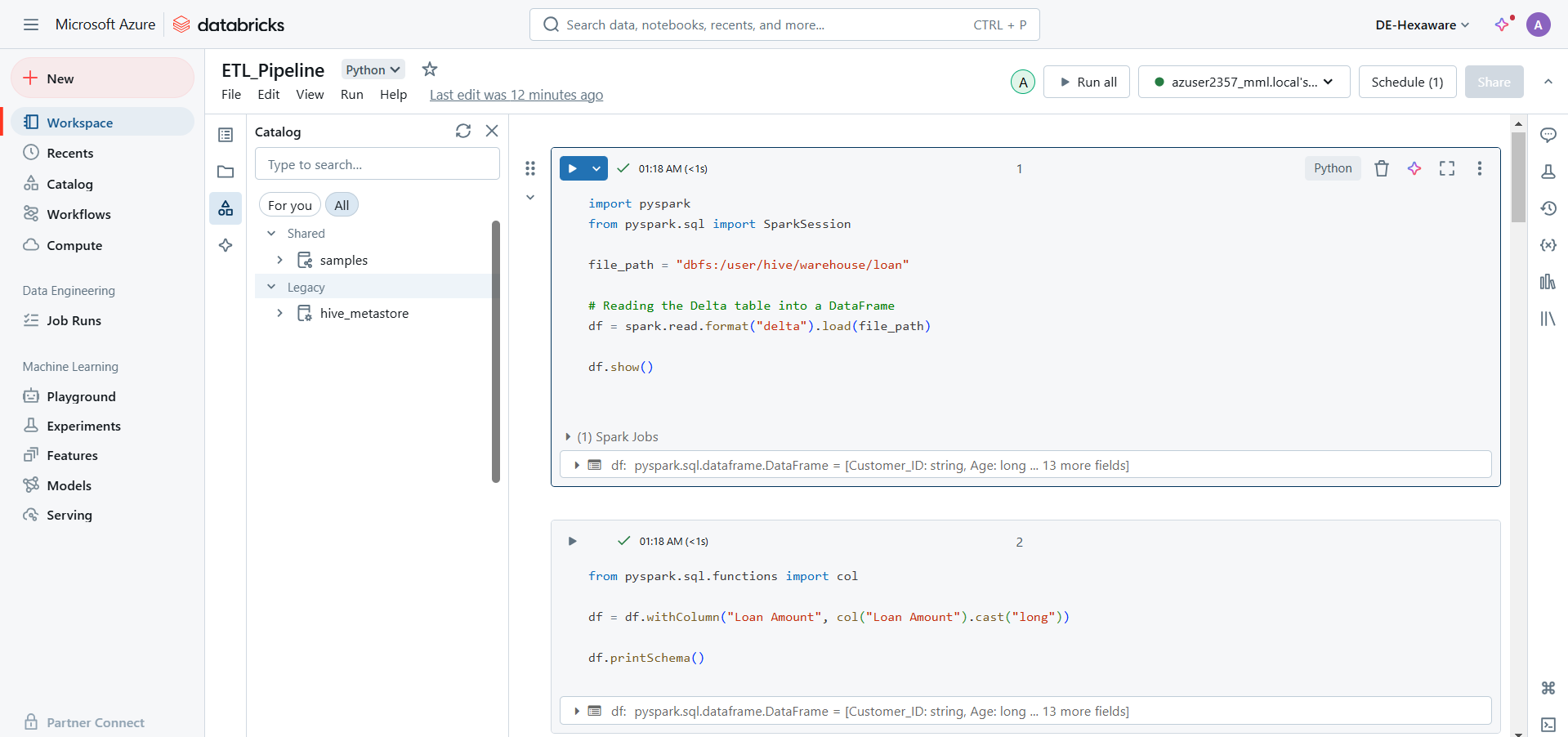
**Step 2: Creating the Cluster**

* + Click on **Create Cluster** to start the cluster creation process.
  + Choose a **Cluster Name** that reflects the purpose of the cluste.
  + Select a **Databricks runtime version**
  + Set the **Cluster Mode** to “Standard” for normal use or “High Concurrency” if required for shared workloads.
  + Select the **Worker type** and **Driver type**. For basic tasks, you can start with smaller VM sizes and increase them based on the workload.



**Step 3: Creating a Notebook**

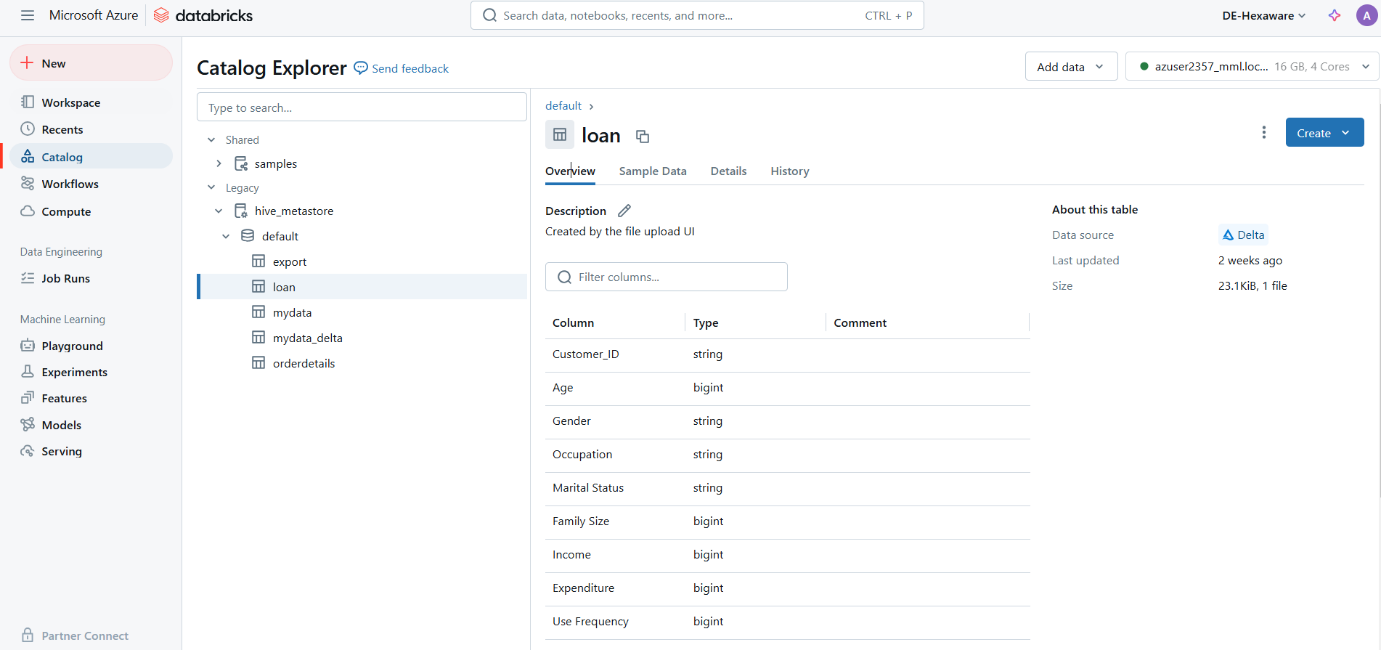
* + Click on **Create** in the upper left corner and select **Notebook**.
  + Provide a **name** for the notebook.
  + Select the **default language** for the notebook. For this project, choose **Python**.
  + Choose the cluster you created earlier to attach the notebook to.



**Create a notebook with ETL queries**

**Step 1: Create a Delta Table**

* Open the Databricks catalog.
* Create a new table.
* Drop the CSV file into the table to register it as a Delta table.

****

**Step 2: Importing Required Libraries and Reading the Delta Table**

* In this step, we import necessary libraries like **pyspark.sql** and the Delta format to enable data processing and transformations. We then read the Delta table we created in Step 1 into a DataFrame.

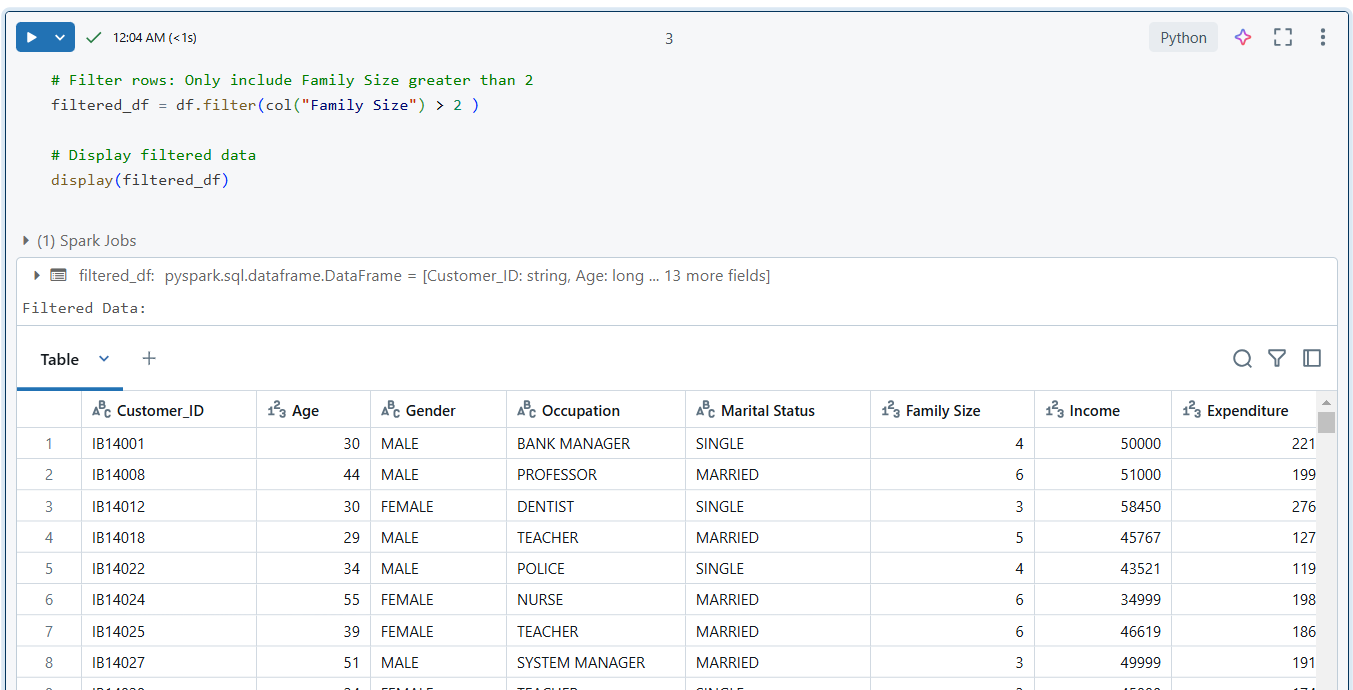


**Step 3: ETL Process**

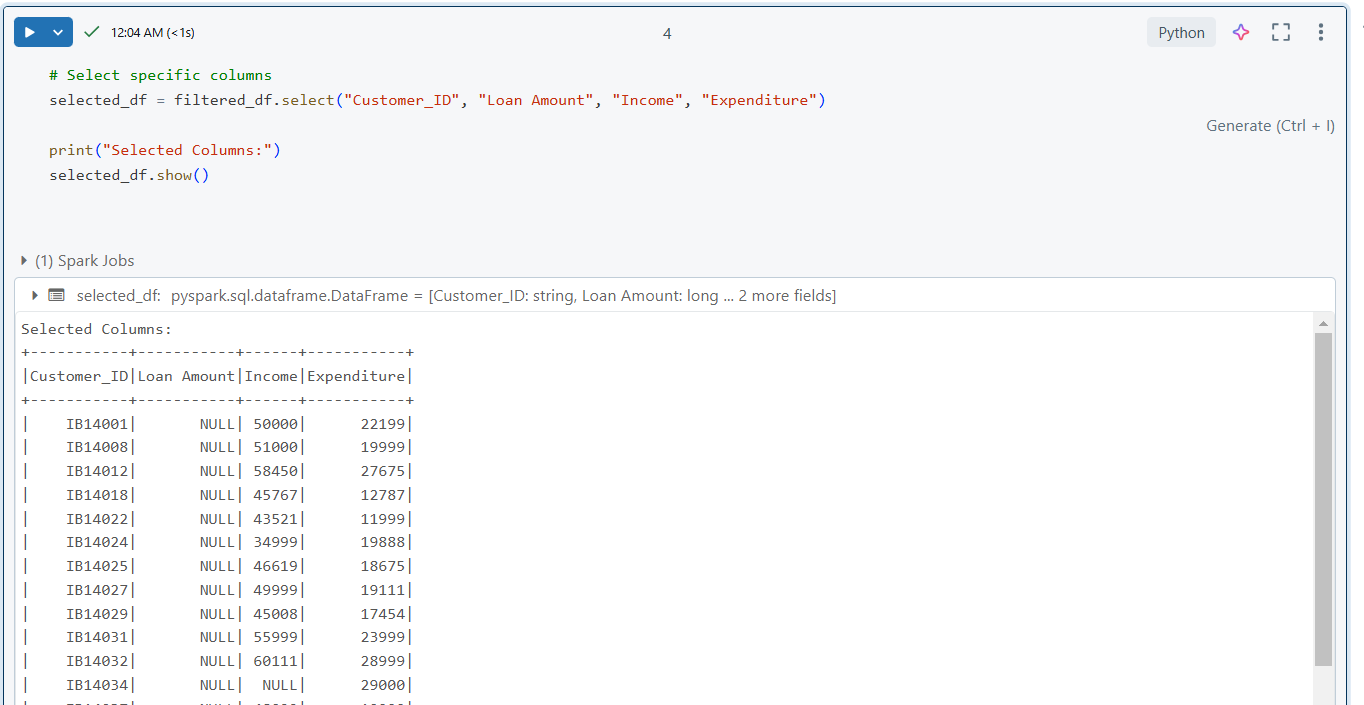
In this step, we perform the **ETL (Extract, Transform, Load)** process to clean and prepare the data for further analysis or saving. The ETL process involves three main stages:

1. **Extract**: We read the dataset from the Delta table.
2. **Transform**: We apply various data transformations to clean the data and make it more usable. In this example:

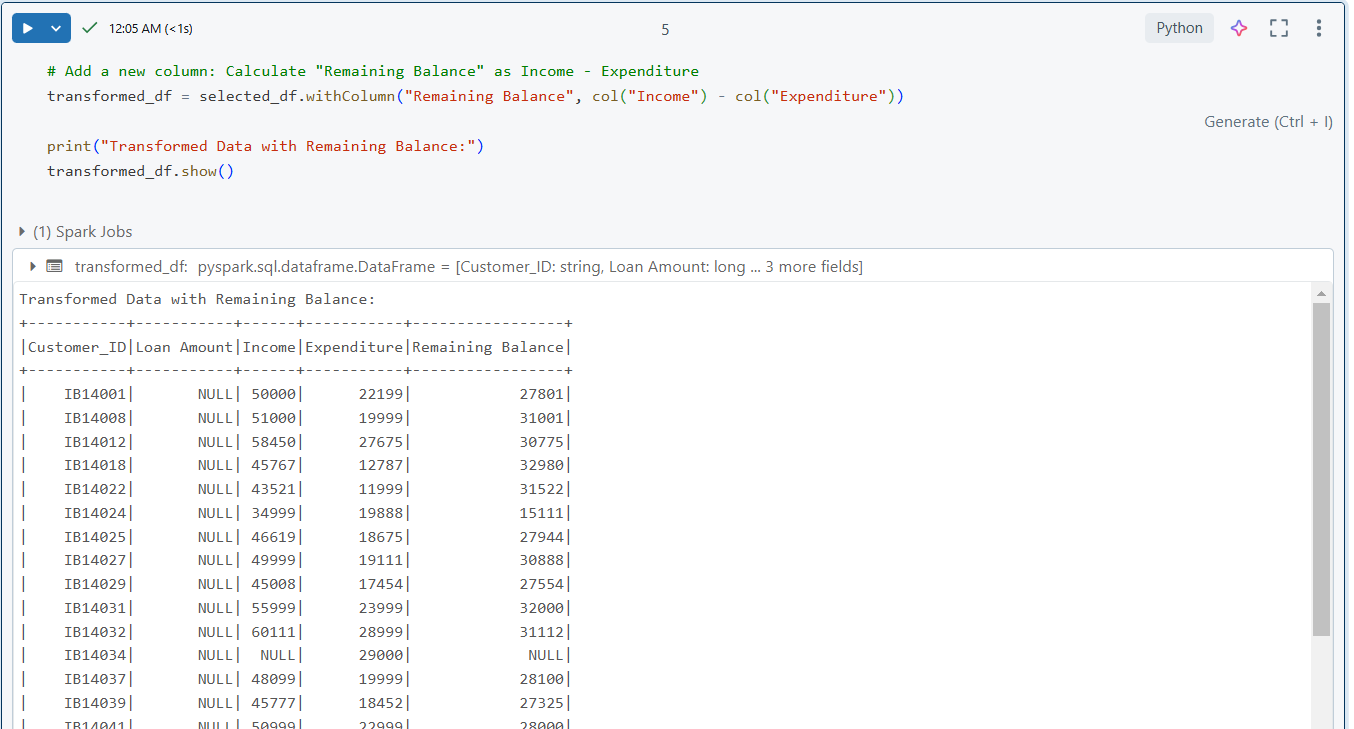
* **Filtered Data**: Displays only the rows where the Family Size exceeds 2.



* **Selected Columns**: Shows only the necessary columns (Customer\_ID, Loan Amount, Income, Expenditure).

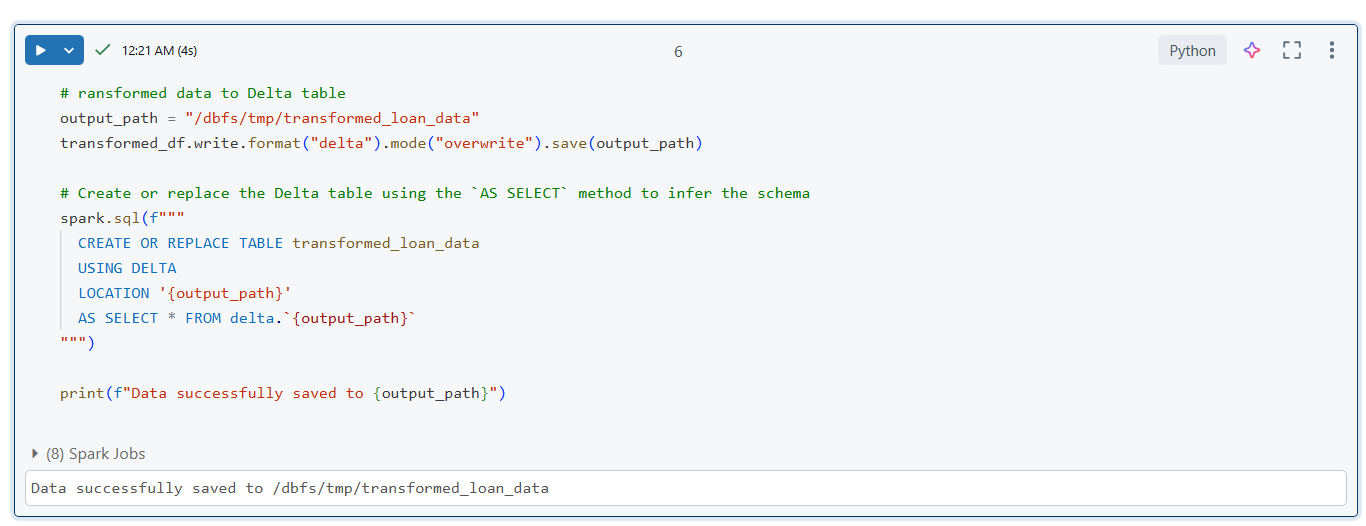


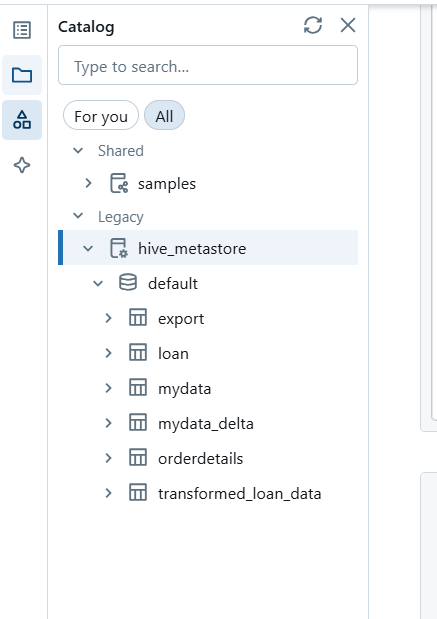
* **Adding a Calculated Column:**



**Step 4: Load the Transformed Data**

* After transforming the data, we **load** it back into a Delta table. This step ensures that the transformed data is saved and can be used for further analysis.

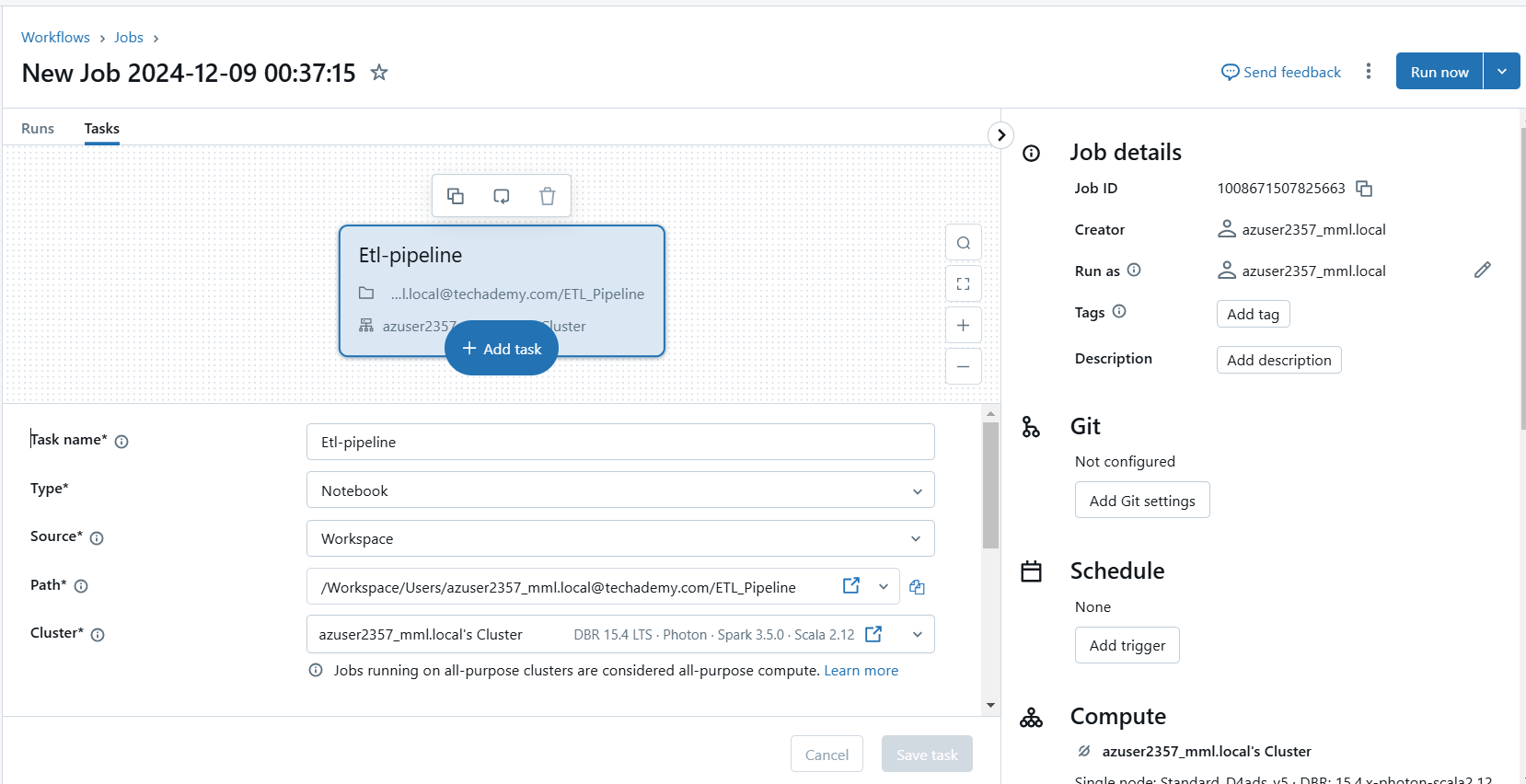




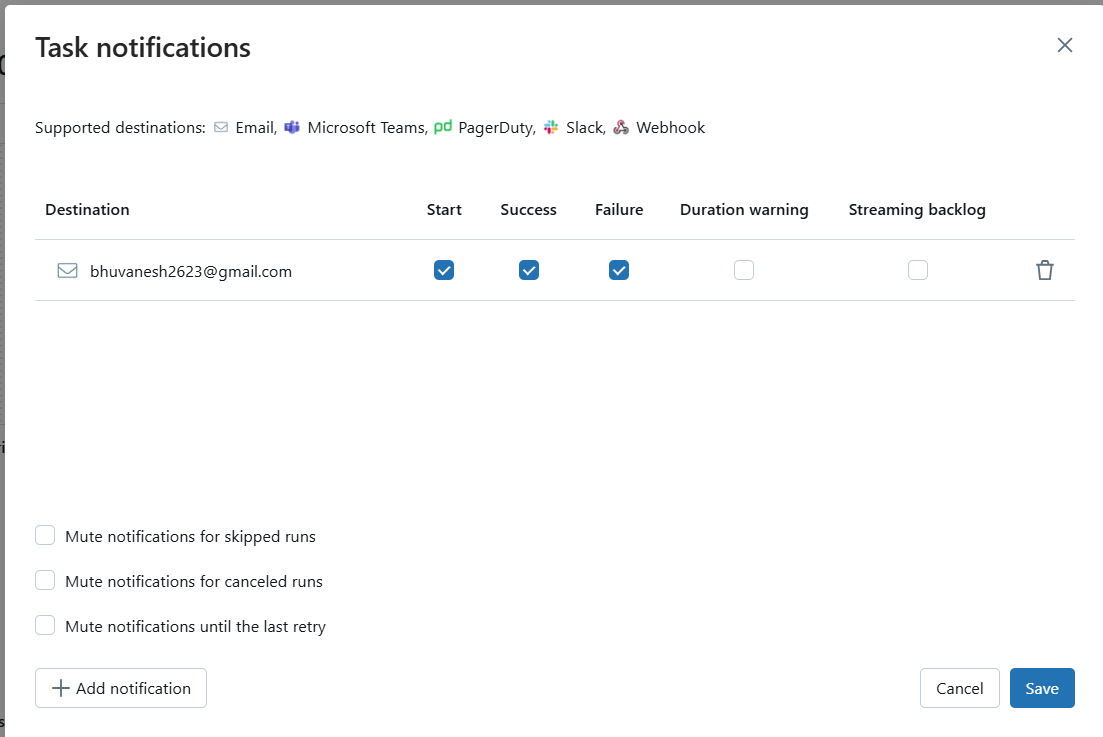
**Run the notebook from workflow pipeline in azure databricks workspace**

**Step 1: Create a Job to Run the Notebook**

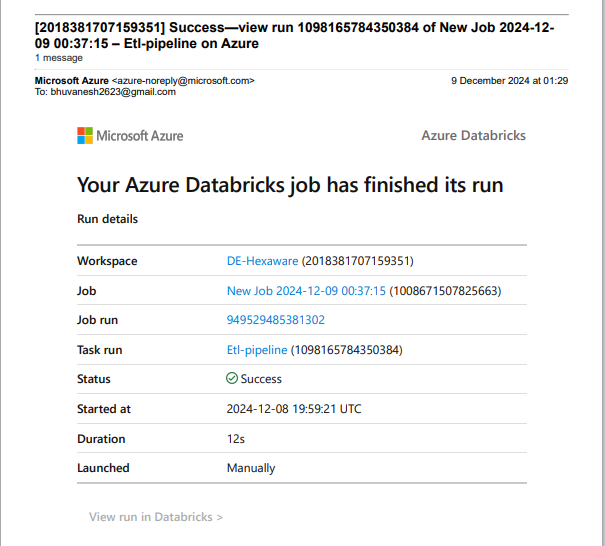
* Click on **"Jobs"** in the left navigation panel in Azure Databricks.
* Click **"Create Job"**.
* Enter a job name.
* Under **"Task"**, select the notebook you want to run.
* Choose the cluster to run the job (either an existing cluster or create a new one).



* Under **"Notifications"**, add your email to receive job status notifications.



* You will receive a **confirmation email** once the job is created



**Step 2: Run the Job and Check Status**

* Click **"Run"** to start the job.
* The job will take a few seconds to run.
* If the job runs successfully, the status will show **"Success"**.

