**Case Study**

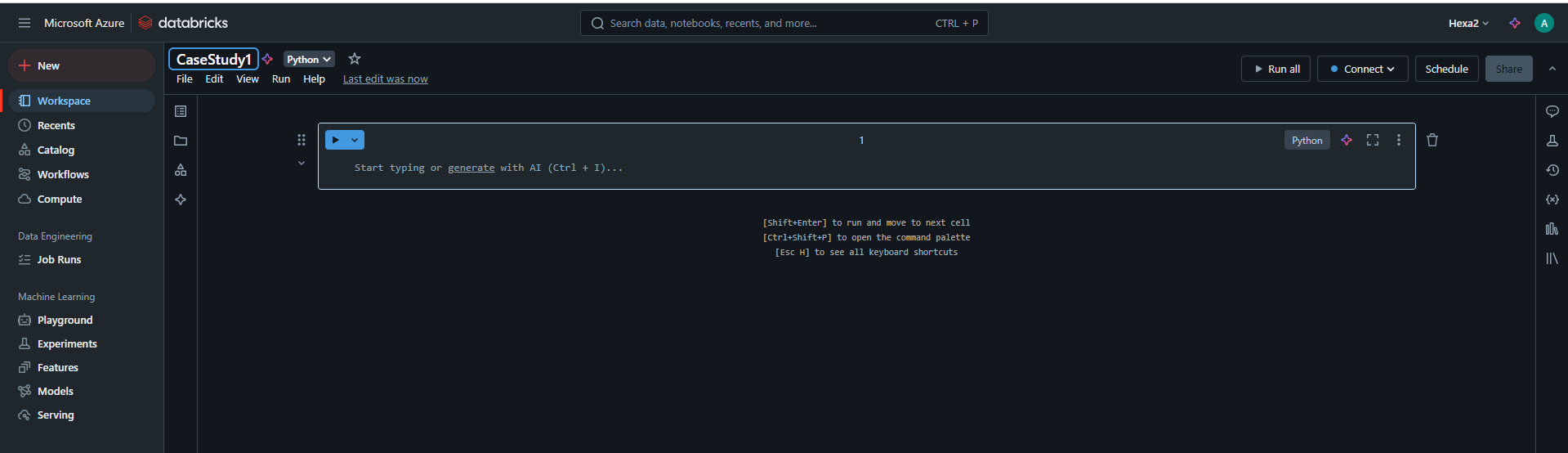
**Done By: S.Yazhini**

**Create an ETL pipeline of ingestion & transform and load queries on any data set and initiate the pipeline from workflow using notebook**

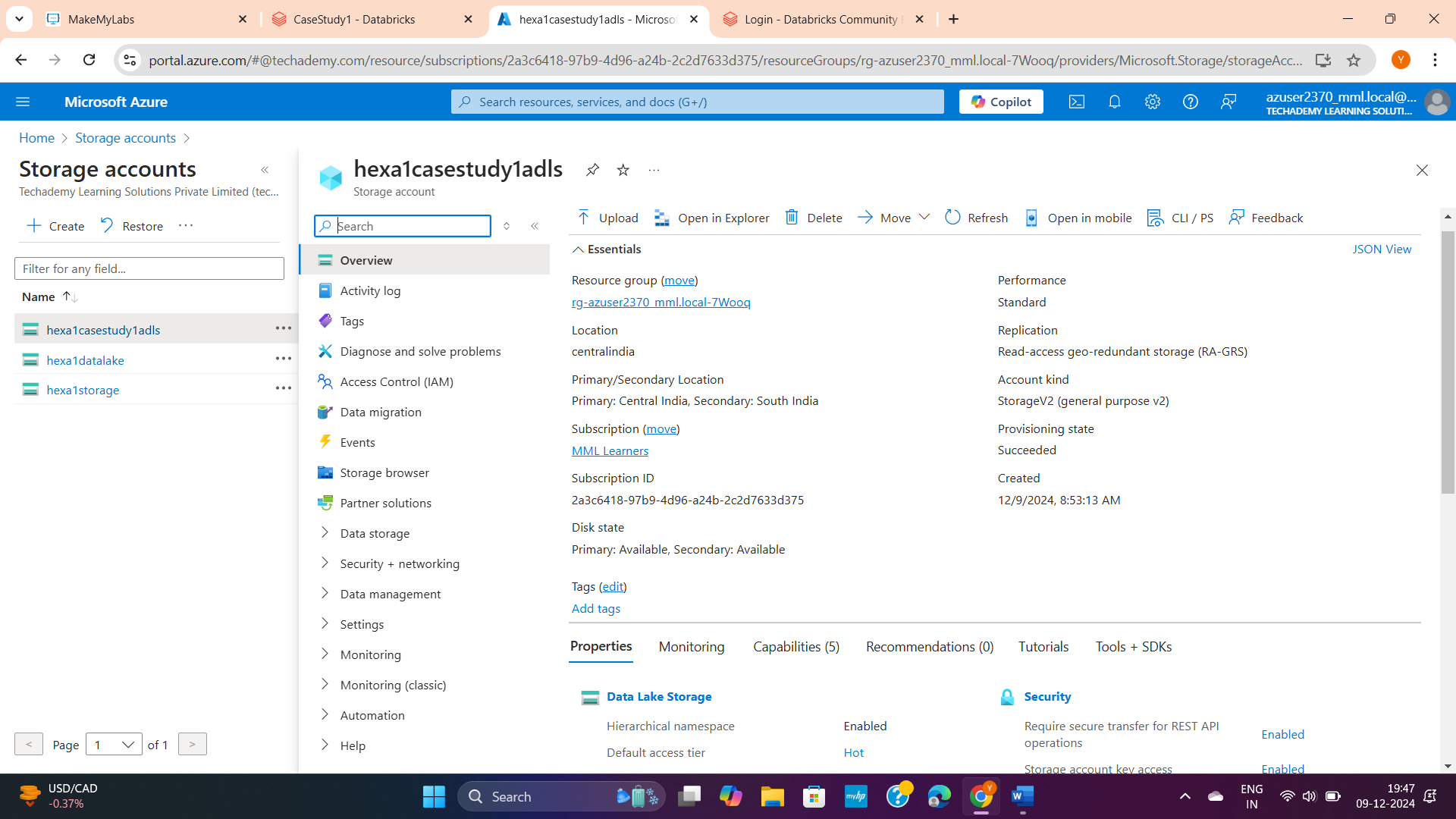
**Part1: create a notebook with ETL queries**

**Part2: Run the notebook from workflow pipeline in azure Databricks workspace**

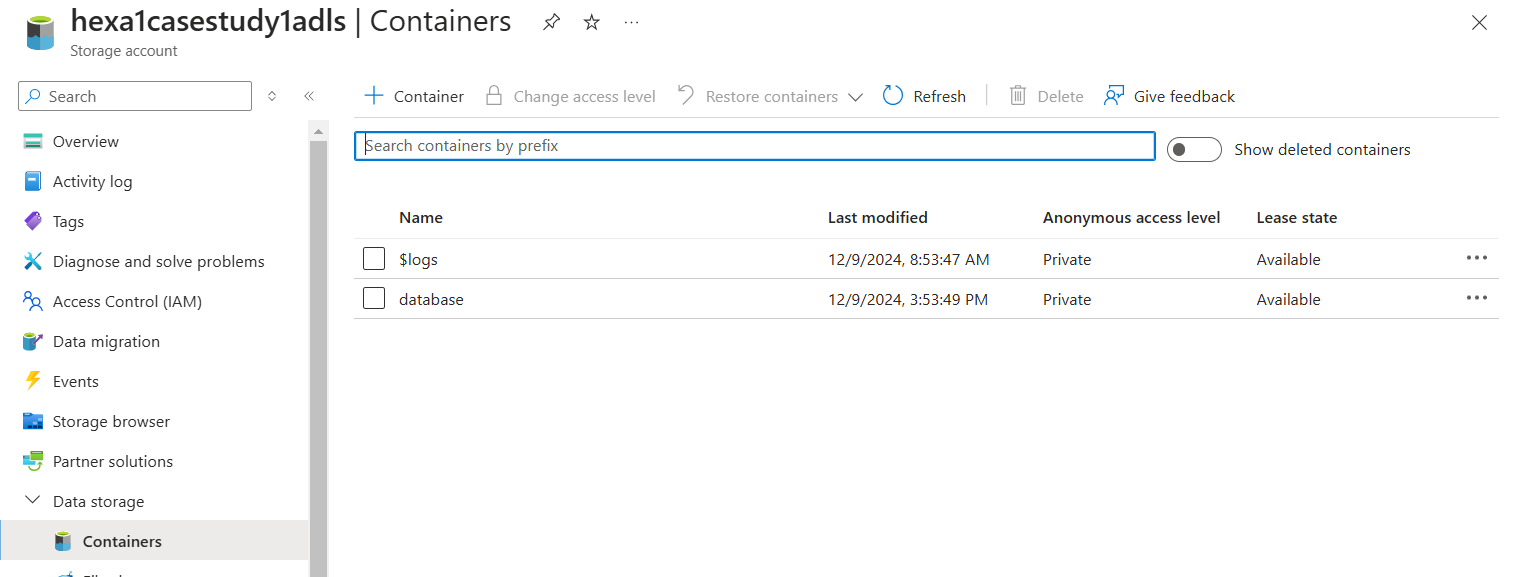
**Notebook created in Azure DataBricks:**

****

**Data Lake Storage account created (ADLS):**



**Containers:**

****

**Database used:**

****

**ETL operations:**

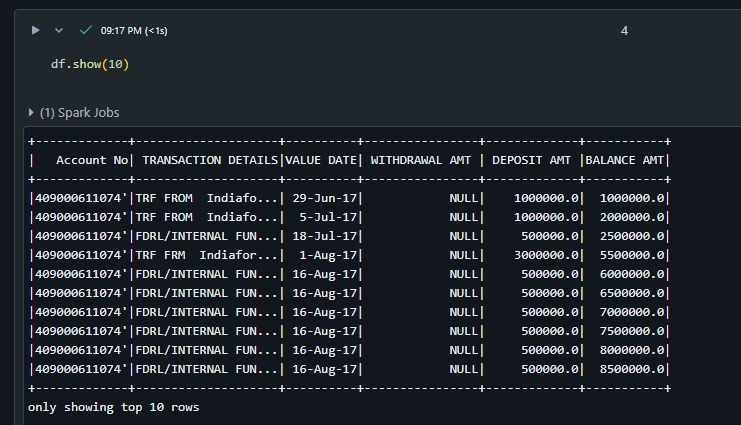
1. **Extracting database file from ADLS:**

Syntax:

df = spark.read.csv("abfss://<container\_name>@<storage\_acc\_name>.dfs.core.windows.net/<file\_name>", header=True, inferSchema=True) – by this we are trying to Load the CSV file from ADLS

df.createOrReplaceTempView("txn\_view") – will create a temporary view of our table

****

****

1. **Transforming data:**

**Replacing invalid characters in column names:**

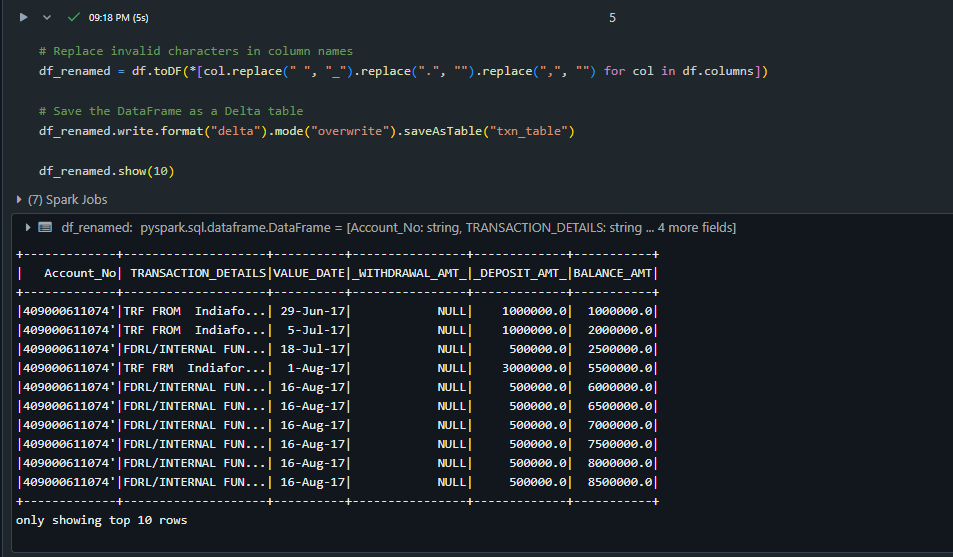
Eg: replacing a column name from “Account No” to “Account\_No”

Syntax used:

df\_renamed = df.toDF(\*[col.replace(" ", "\_").replace(".", "").replace(",", "") for col in df.columns])

# Save the DataFrame as a Delta table

df\_renamed.write.format("delta").mode("overwrite").saveAsTable("txn\_table")

****

**Converting all the null value to zero:**

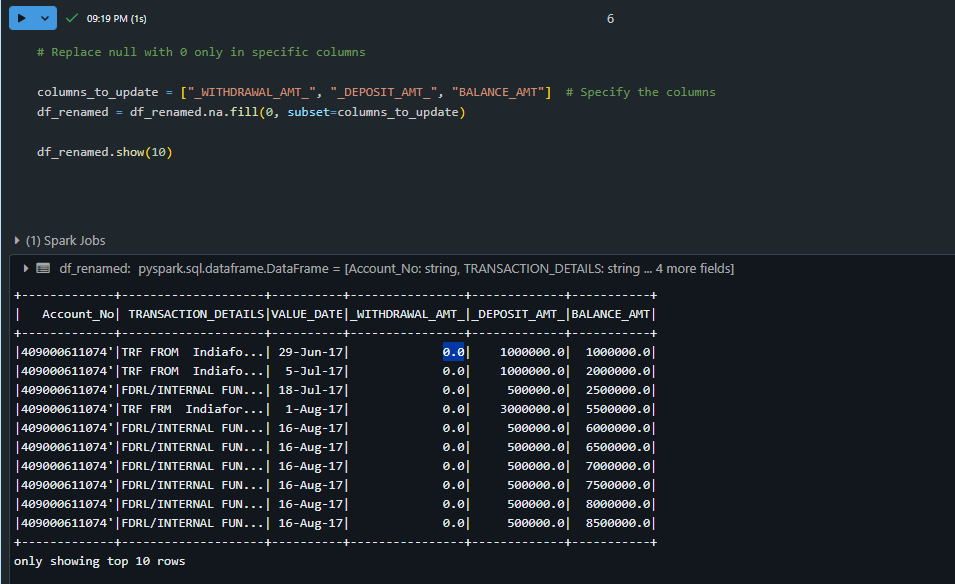
Syntax:

# Replace null with 0 only in specific columns

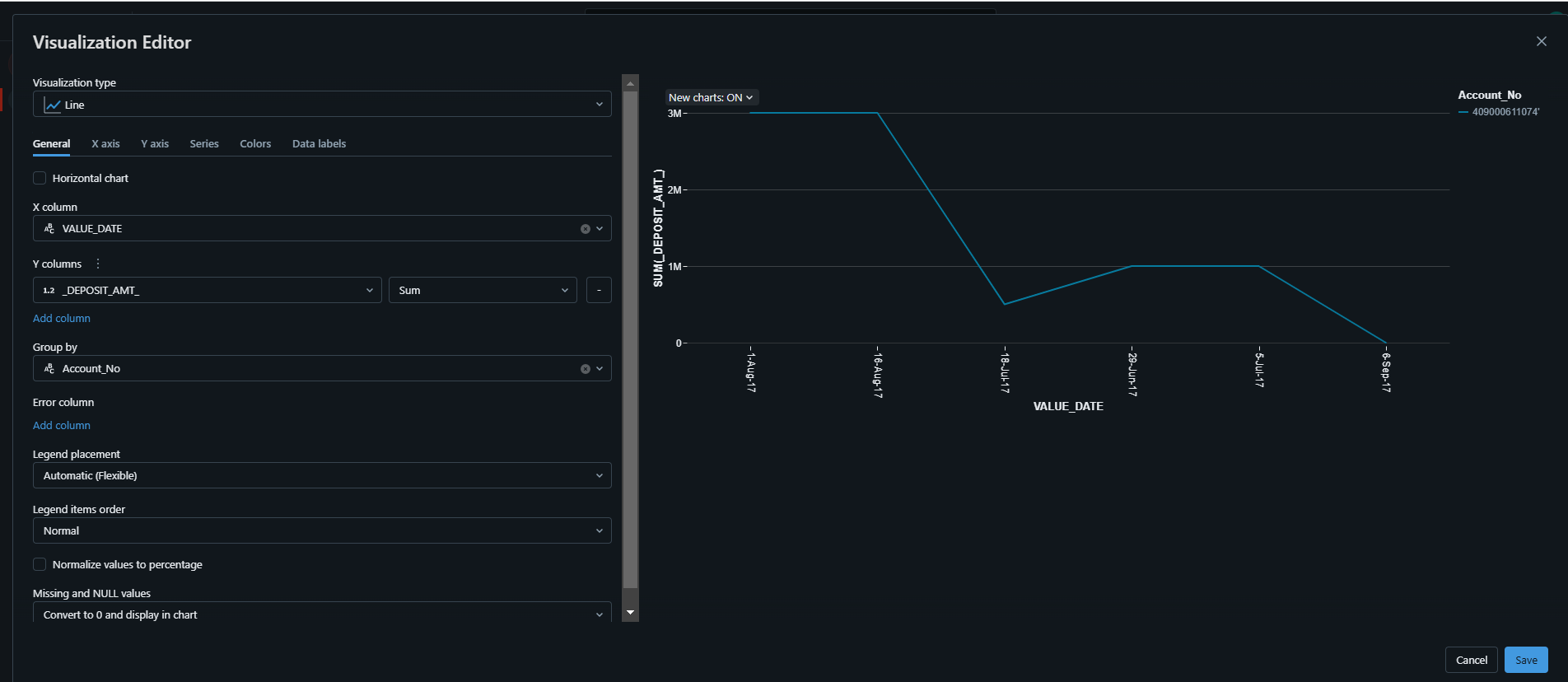
columns\_to\_update = ["col1", "col3"] # Specify the columns

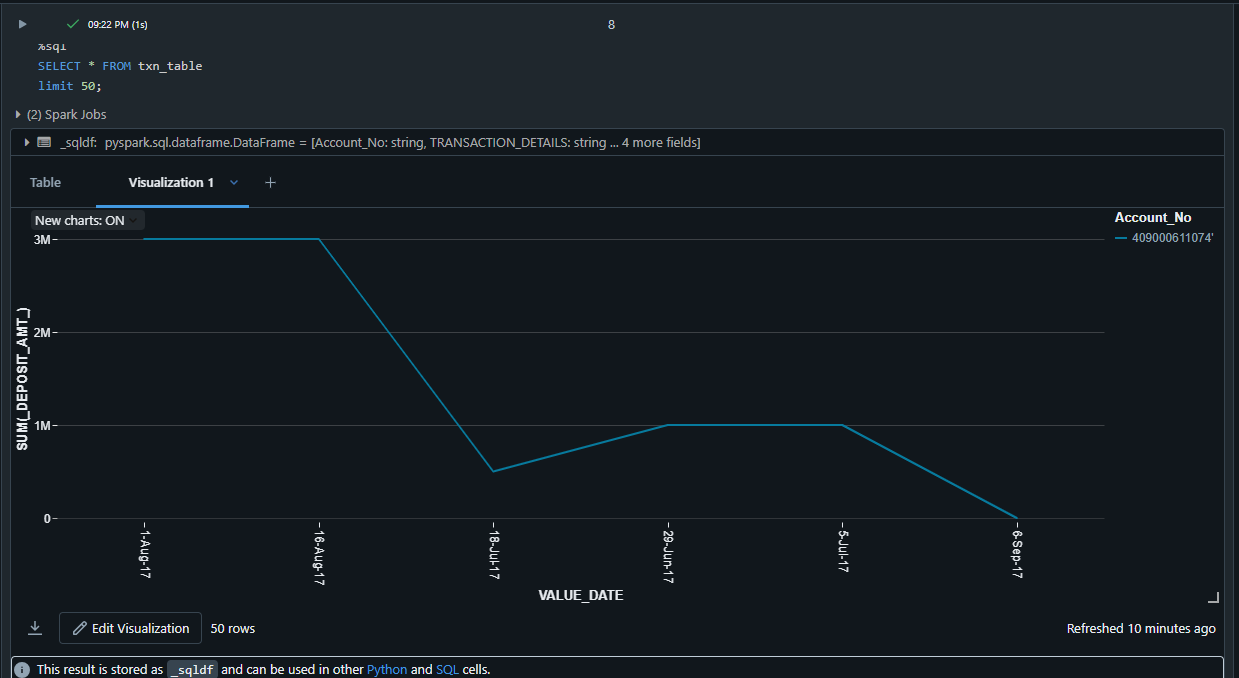
df = df.na.fill(0, subset=columns\_to\_update)

df.show()



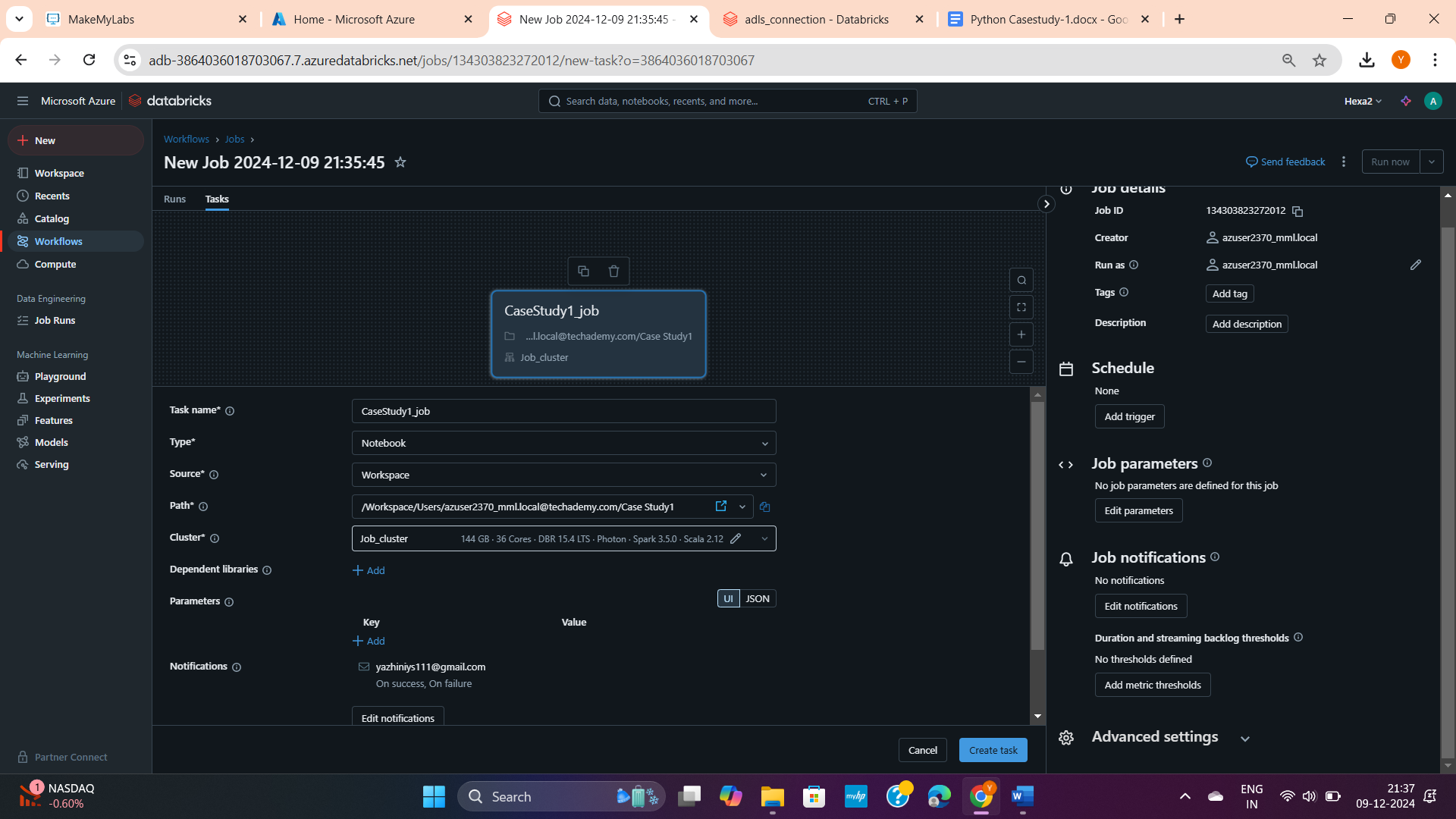
1. **Loading and Visualisation:**



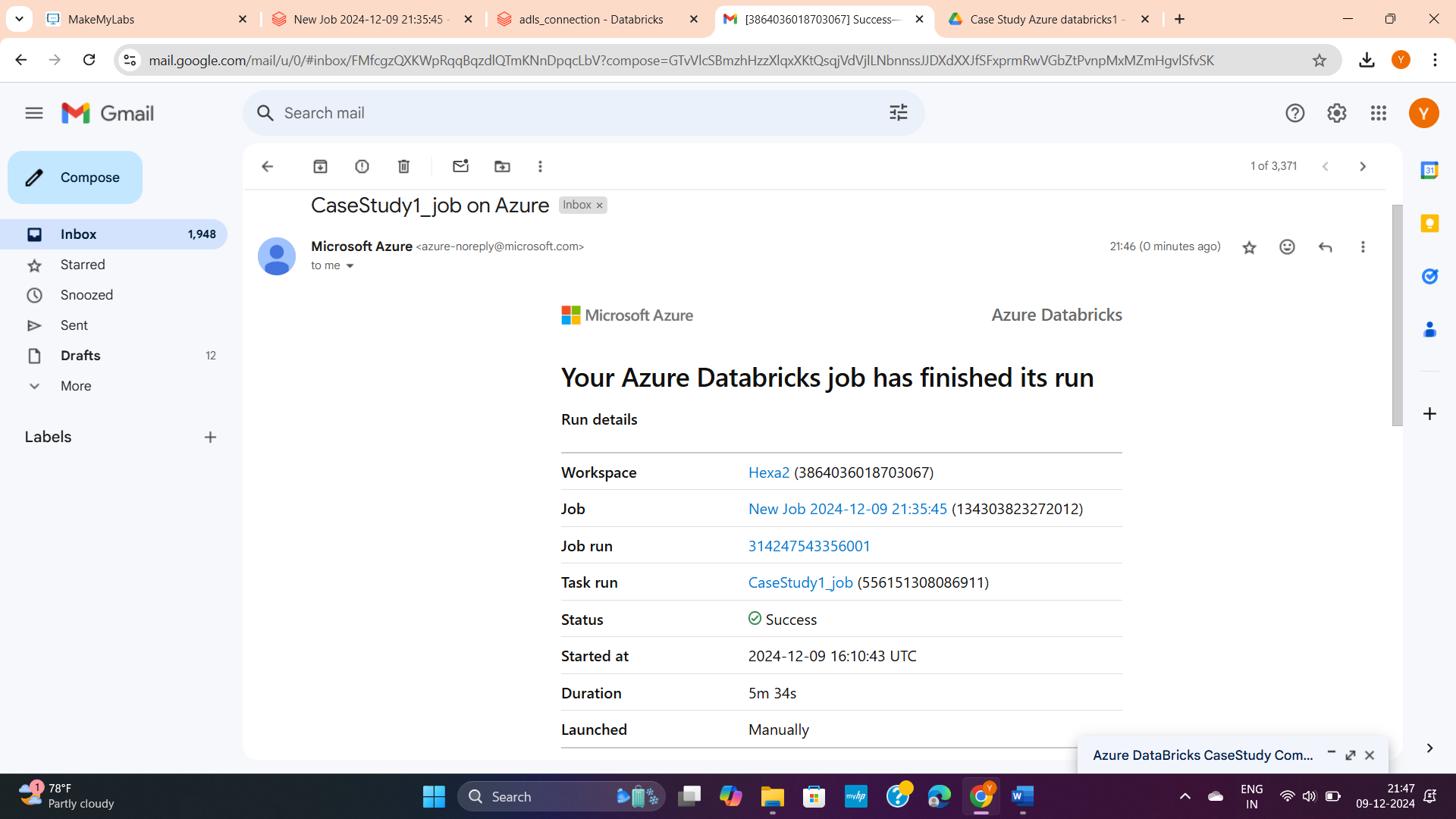


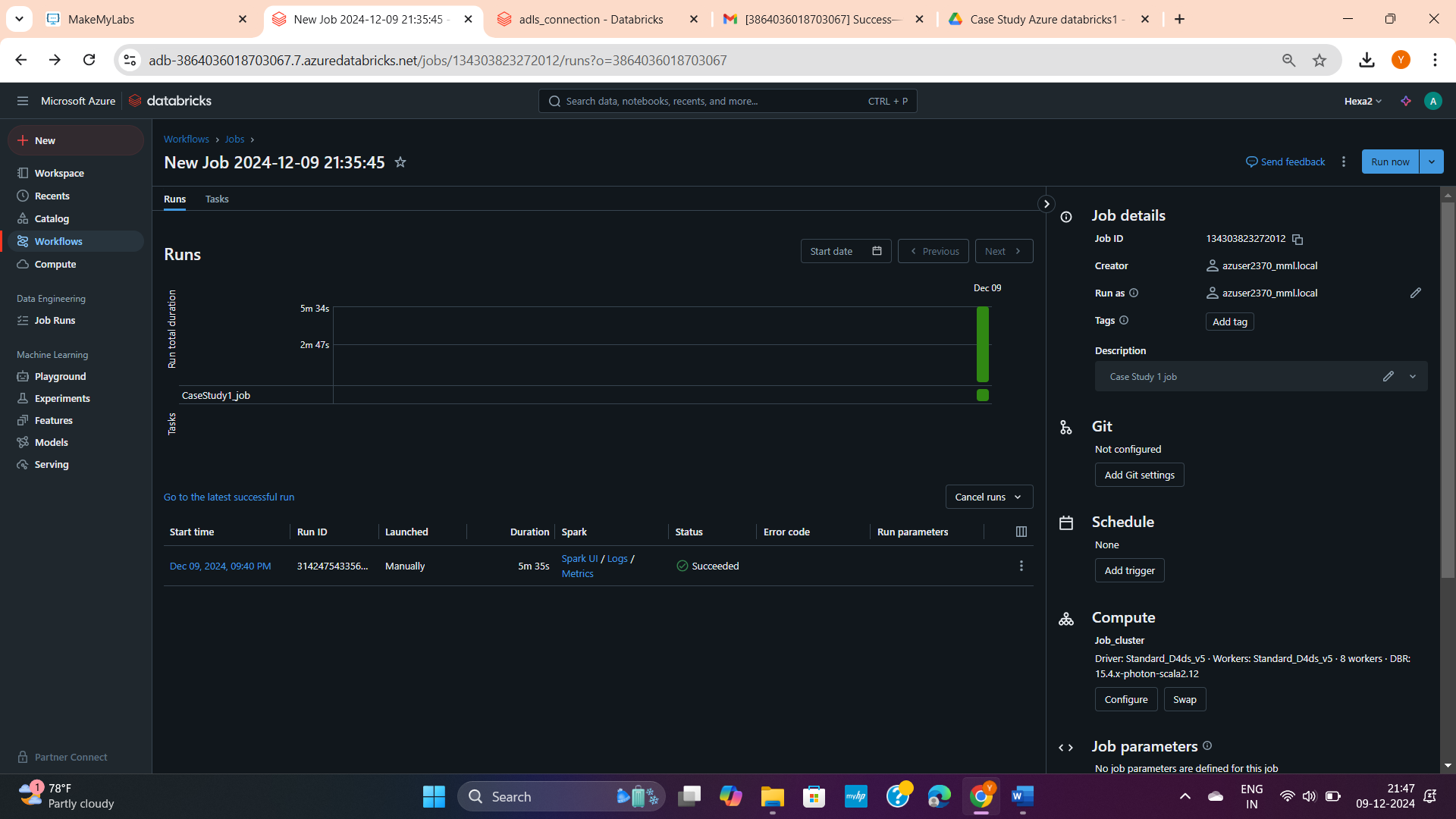
**Part2: Run the notebook from workflow pipeline in azure Databricks workspace**

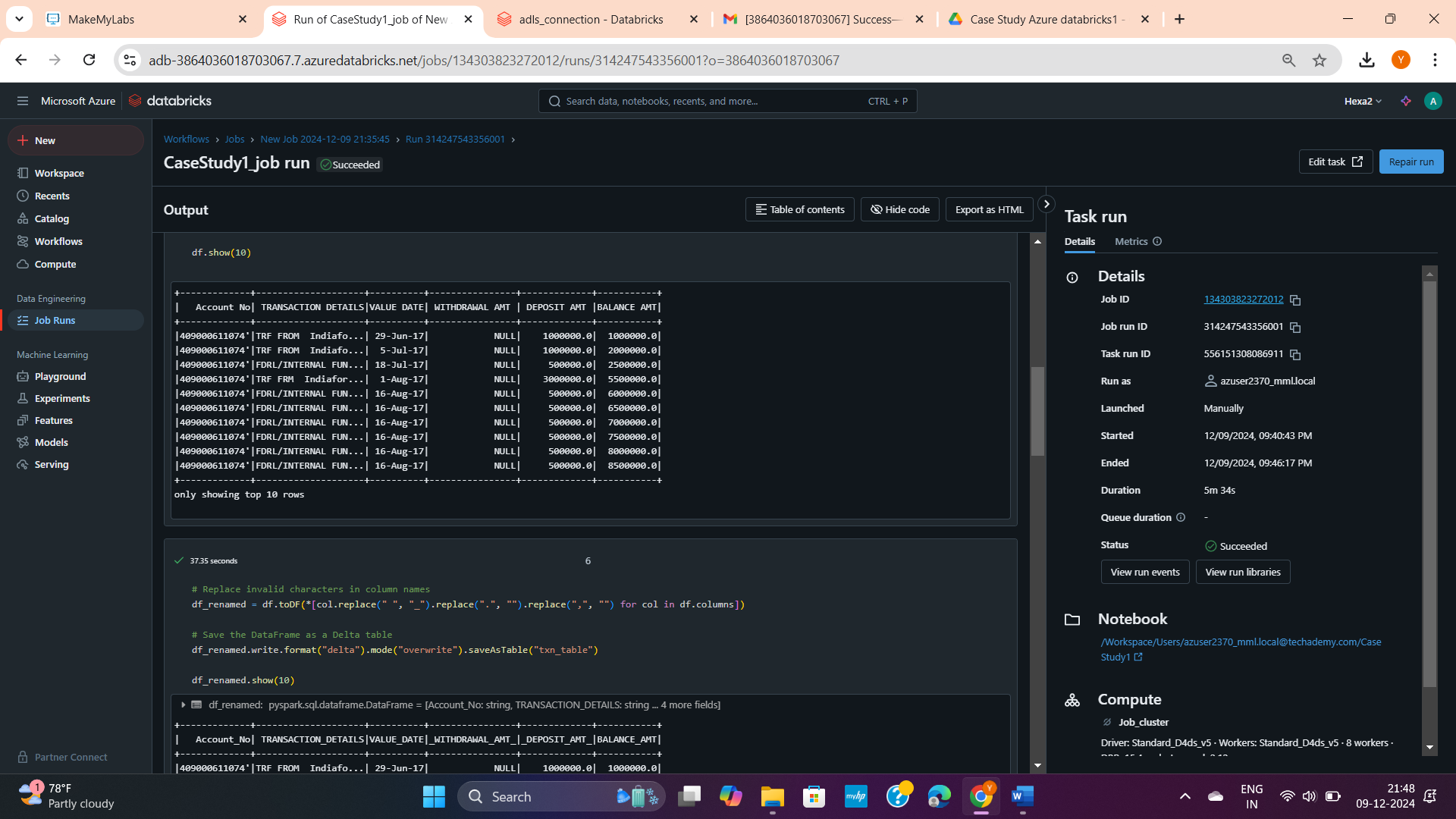
**Creating job:**



**Job run successful:**







**-------------------------------------Thank you----------------------------------------**