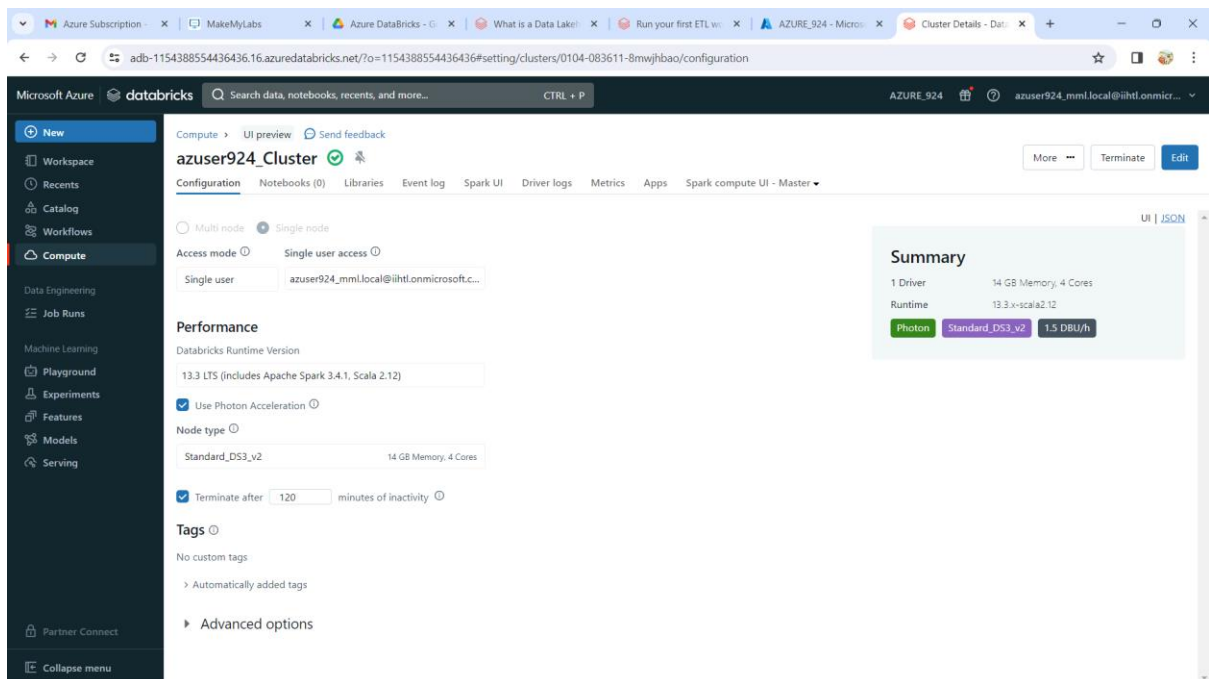
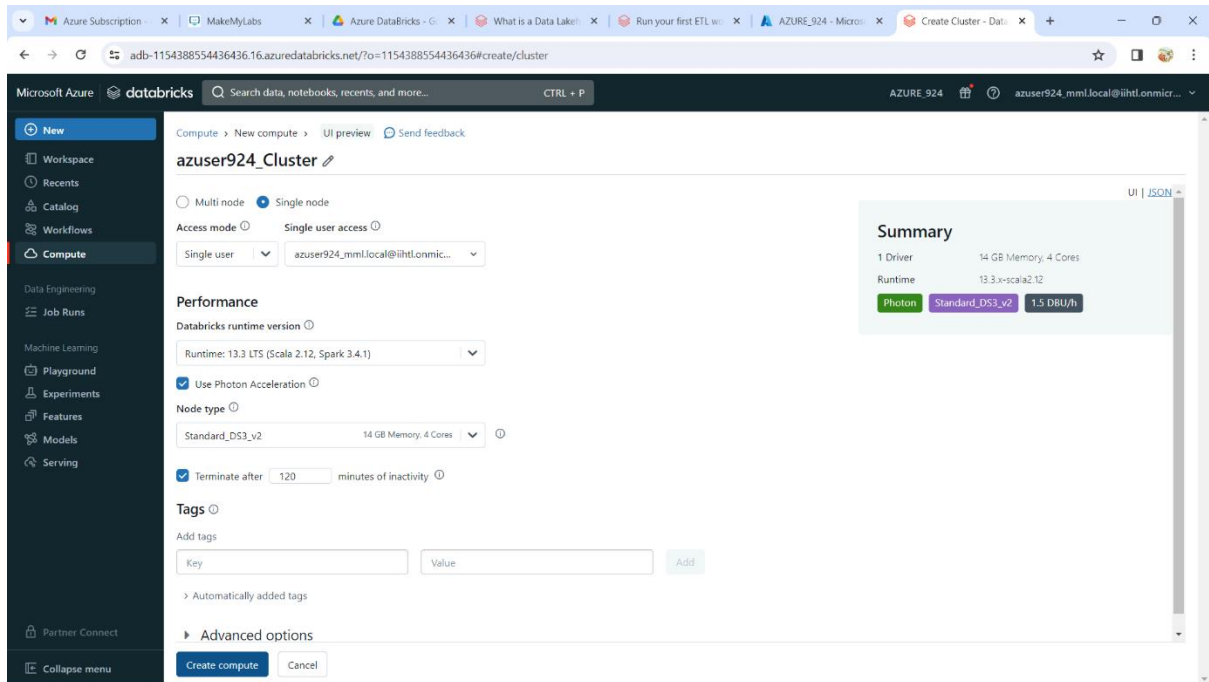


# ASSESSMENT

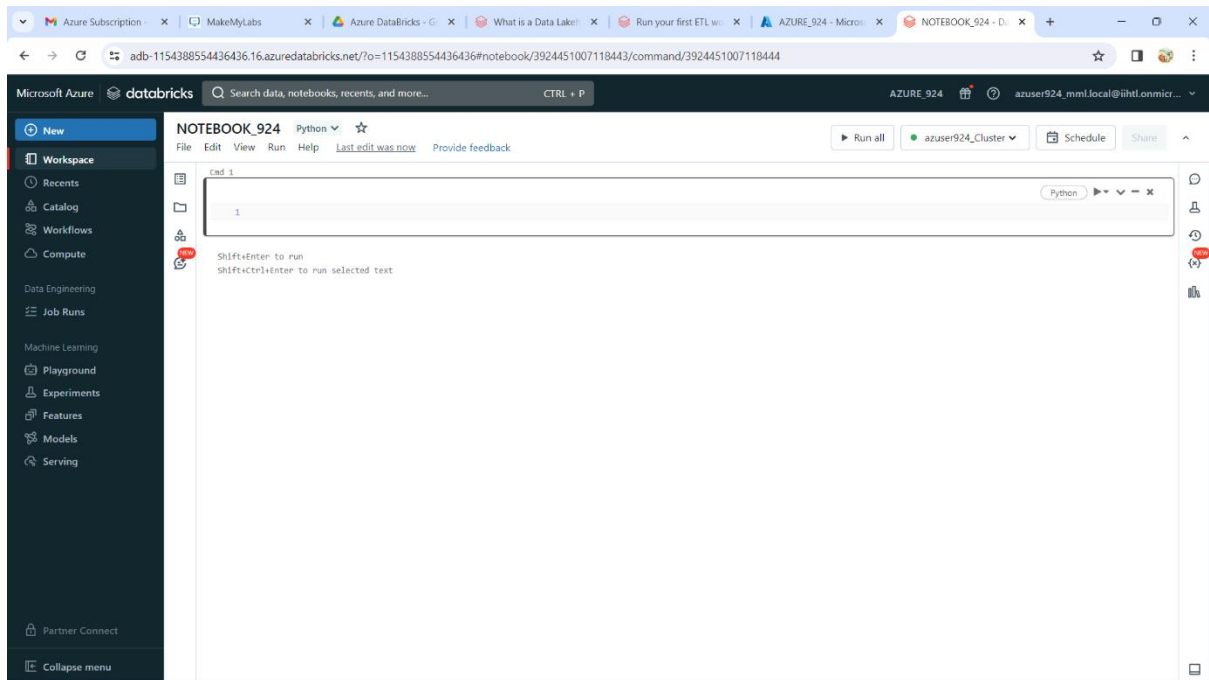
NAME-SWARNA LATHA GUVVALA  
DATE-04/01/2024

## RUN YOUR FIRST ETL WORKLOAD ON AZURE DATABRICKS:

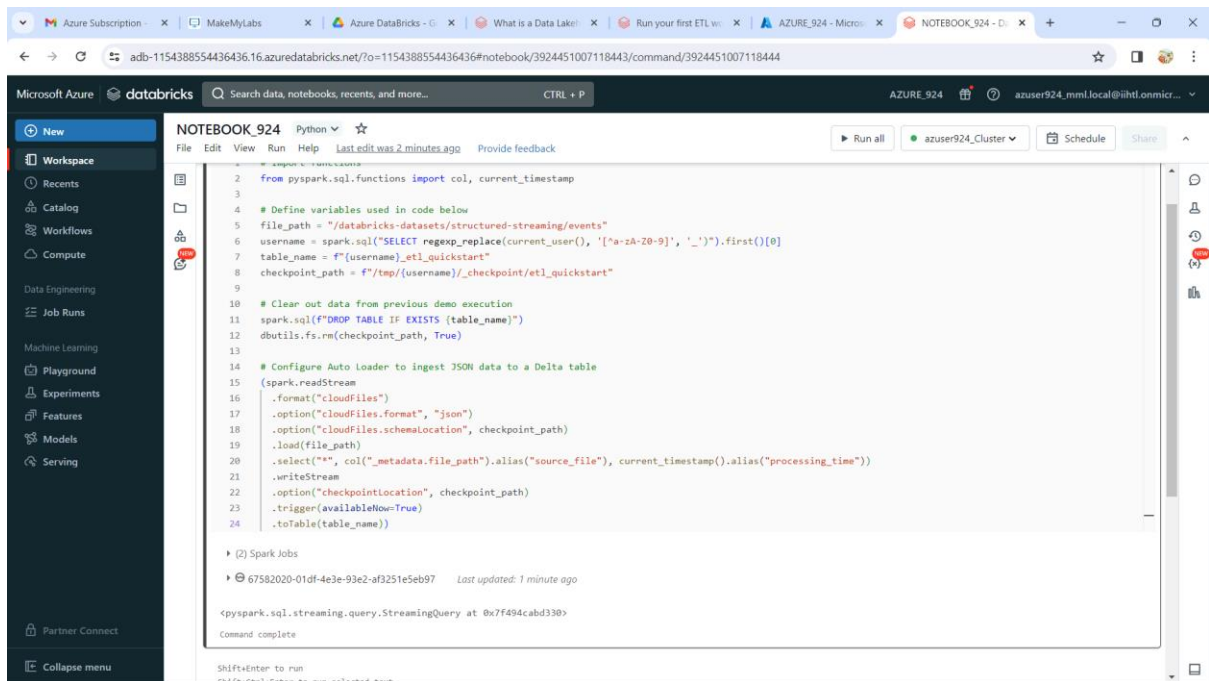
### Step 1: Create a cluster



## Step 2: Create a Databricks notebook



## Step 3: Configure Auto Loader to ingest data to Delta Lake



## Step 4: Process and interact with data

The screenshot shows the Databricks Notebook interface for 'NOTEBOOK\_924'. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows, Compute, Data Engineering, Job Runs, Machine Learning, Playground, Experiments, Features, Models, and Serving. The main editor area displays a Python code cell with the following code:

```
10 .format("cloudFiles.json")
11 .option("cloudFiles.format", "json")
12 .option("cloudFiles.schemaLocation", checkpoint_path)
13 .load(file_path)
14 .select("col('metadata.file_path').alias('source_file'), current_timestamp().alias('processing_time'))
15 .writeStream
16 .option("checkpointLocation", checkpoint_path)
17 .trigger(availableNow=True)
18 .toTable(table_name)
```

Below the code, the output shows '(2) Spark Jobs' and a job ID '67582020-01df-4e3e-93e2-af3251e5eb97'. The command is marked as 'Command complete'.

A second code cell is shown with the following code:

```
1 df = spark.read.table(table_name)
```

The output for this cell shows 'df: pyspark.sql.dataframe.DataFrame = [action:string,time:string... 3 more fields]' and 'Command took 0.51 seconds'.

The screenshot shows the Databricks Notebook interface for 'NOTEBOOK\_924' after the streaming query has been executed. The main editor area displays the same Python code as the previous screenshot, but the output now shows '(1) Spark Jobs' and a job ID '1469665104'. The command is marked as 'Command complete'.

A third code cell is shown with the following code:

```
1 display(df)
```

The output for this cell shows a table with the following data:

	action	time	_rescued_data	source_file	processing_time
1	Open	1469665104	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
2	Close	1469665106	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
3	Close	1469665108	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
4	Open	1469665110	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
5	Open	1469665110	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
6	Open	1469665113	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
7	Open	1469665118	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z

The table is truncated to 10,000 rows. The command took 1.71 seconds.

## Step 5: Schedule a job

The screenshot shows the Databricks Notebook interface with a job named "FIRST ETL\_924" being scheduled. The job is configured with the following settings:

- Job name:** FIRST ETL\_924
- Schedule:** Manual (selected), Every Day at 14:19, Asia/Calcutta
- Cluster:** azuser924\_Cluster (14 GB - 4 Cores - DBR 13.3 LTS - Photon - Spark 3.4.1 - Scala 2.12)
- Parameters:** + Add
- Alerts:** azuser924\_mml.local@ihl1.onmicrosoft.com (checked for Start, Success, Failure)

The job is currently in a "Paused" state. The notebook code shows a Spark SQL query that reads a table named "table\_name" and displays the results. The table has 10,000 rows and a runtime of 1.71 seconds.

action	time	_rescu
1 Open	1469665104	null
2 Close	1469665106	null
3 Close	1469665108	null
4 Open	1469665108	null
5 Open	1469665110	null
6 Open	1469665113	null
7 Open	1469665118	null

The screenshot shows the Databricks Notebook interface with the job "FIRST ETL\_924" running successfully. The job is now in a "Running" state. The notebook code shows a Spark SQL query that reads a table named "table\_name" and displays the results. The table has 10,000 rows and a runtime of 1.71 seconds.

action	time	_rescu	source_file	processing_time
1 Open	1469665104	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
2 Close	1469665106	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
3 Close	1469665108	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
4 Open	1469665108	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
5 Open	1469665110	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
6 Open	1469665113	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z
7 Open	1469665118	null	/databricks-datasets/structured-streaming/events/file-45.json	2024-01-04T08:43:04.878Z