| Name | Manoj Mani |
| --- | --- |
| Date | 04/01/2024 |
| Email Id | jonam1012@gmail.com |
| Type | Daily Assignment |
| Topic | Setting pipeline for ETL |

ETL stands for Extract, Transform, Load, and it refers to the process of extracting data from source systems, transforming it to specific needs, and loading it into a target or destination system.

**Extract:**

* Extracting data involves retrieving information from one or more source systems.

**Transform:**

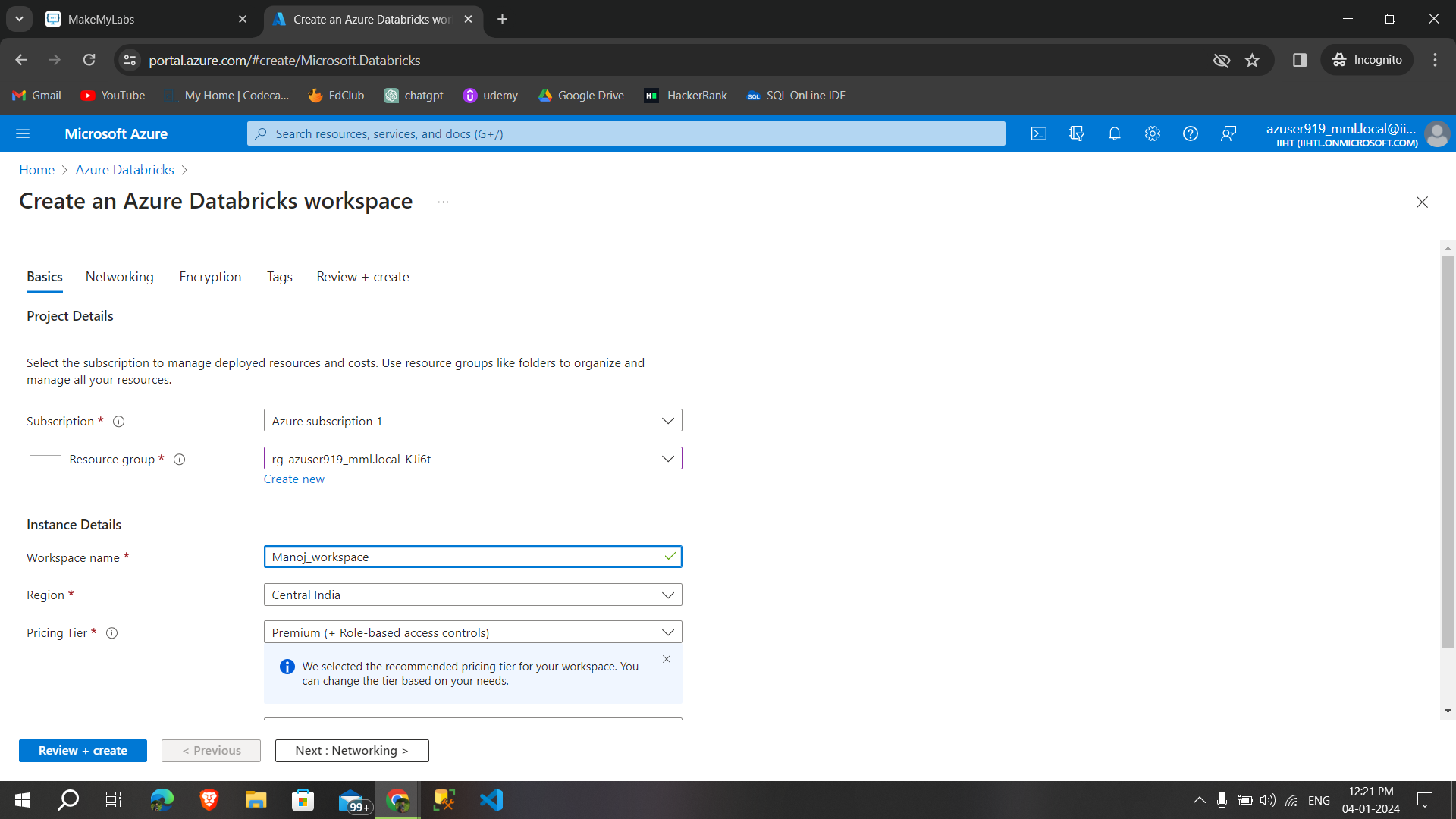
* Transforming data involves applying various operations to convert and clean the extracted data.
* This step ensures that the data is in a format suitable for analysis and reporting.

**Load:**

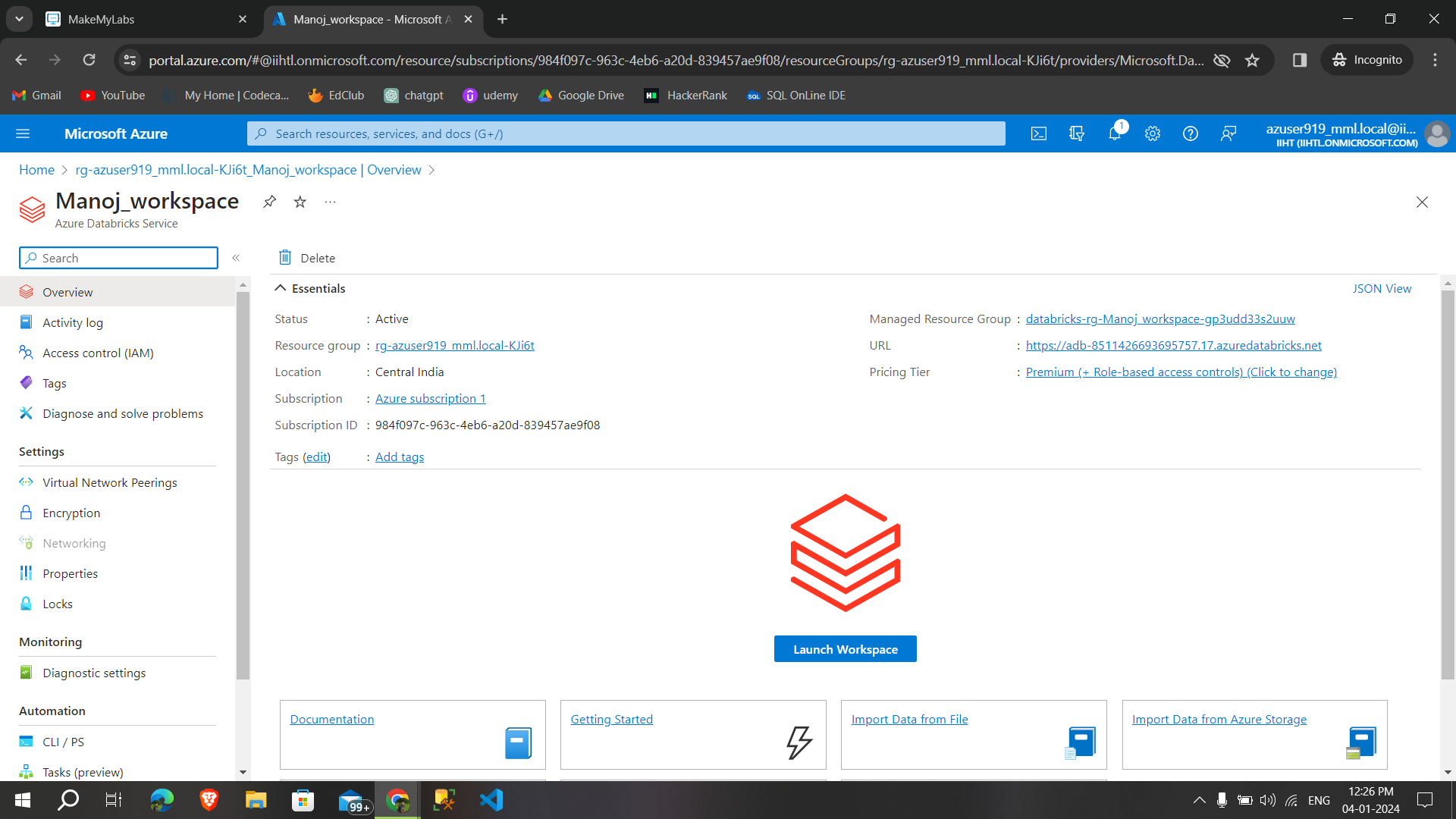
* Loading data is the process of transferring the transformed data into a target system, often a data warehouse, data mart, or any other storage solution.

Performing ETL workload on Azure databricks

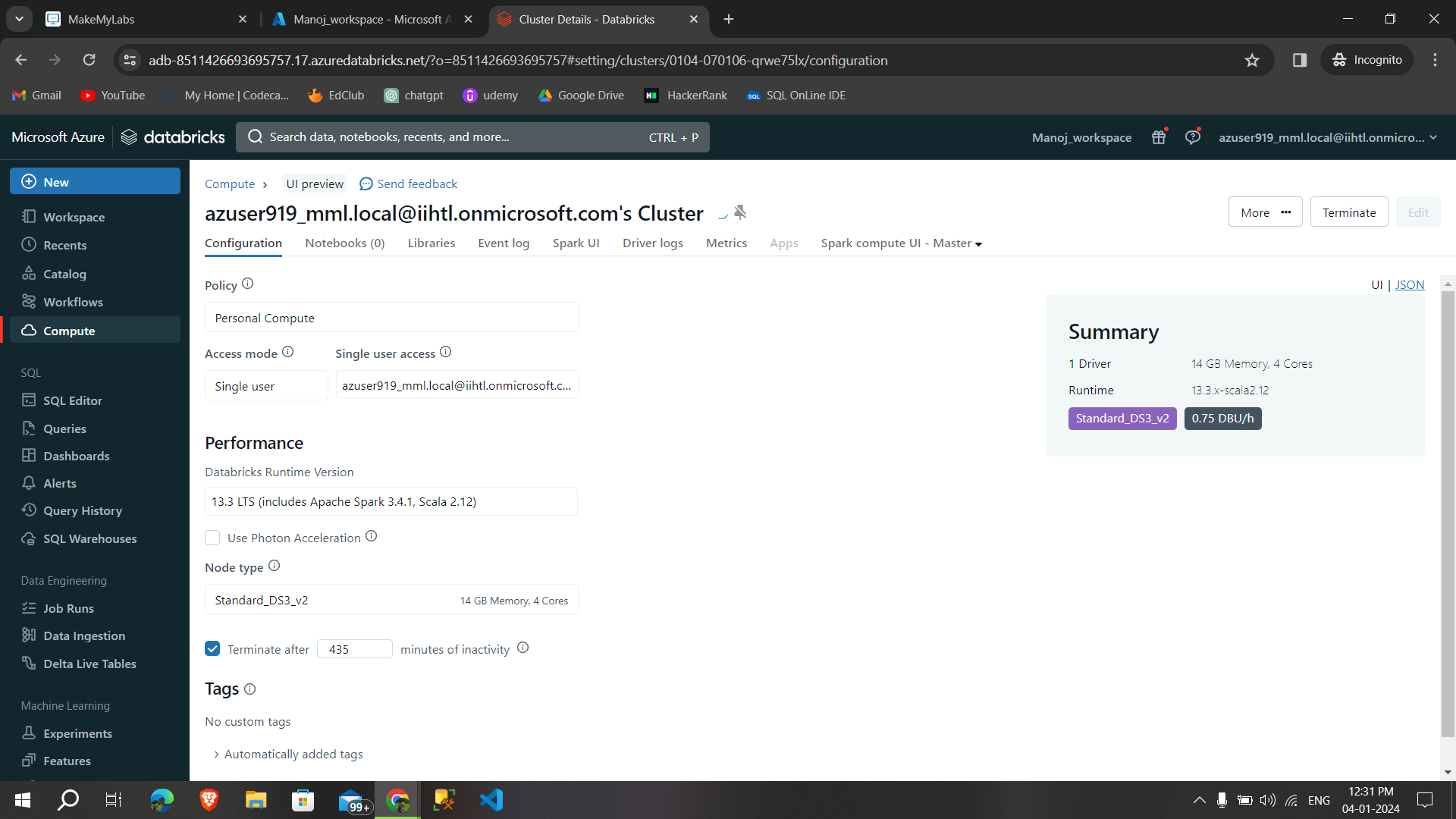
**Step 1**: I created a azure databricks workspace called **Manoj\_workspace**



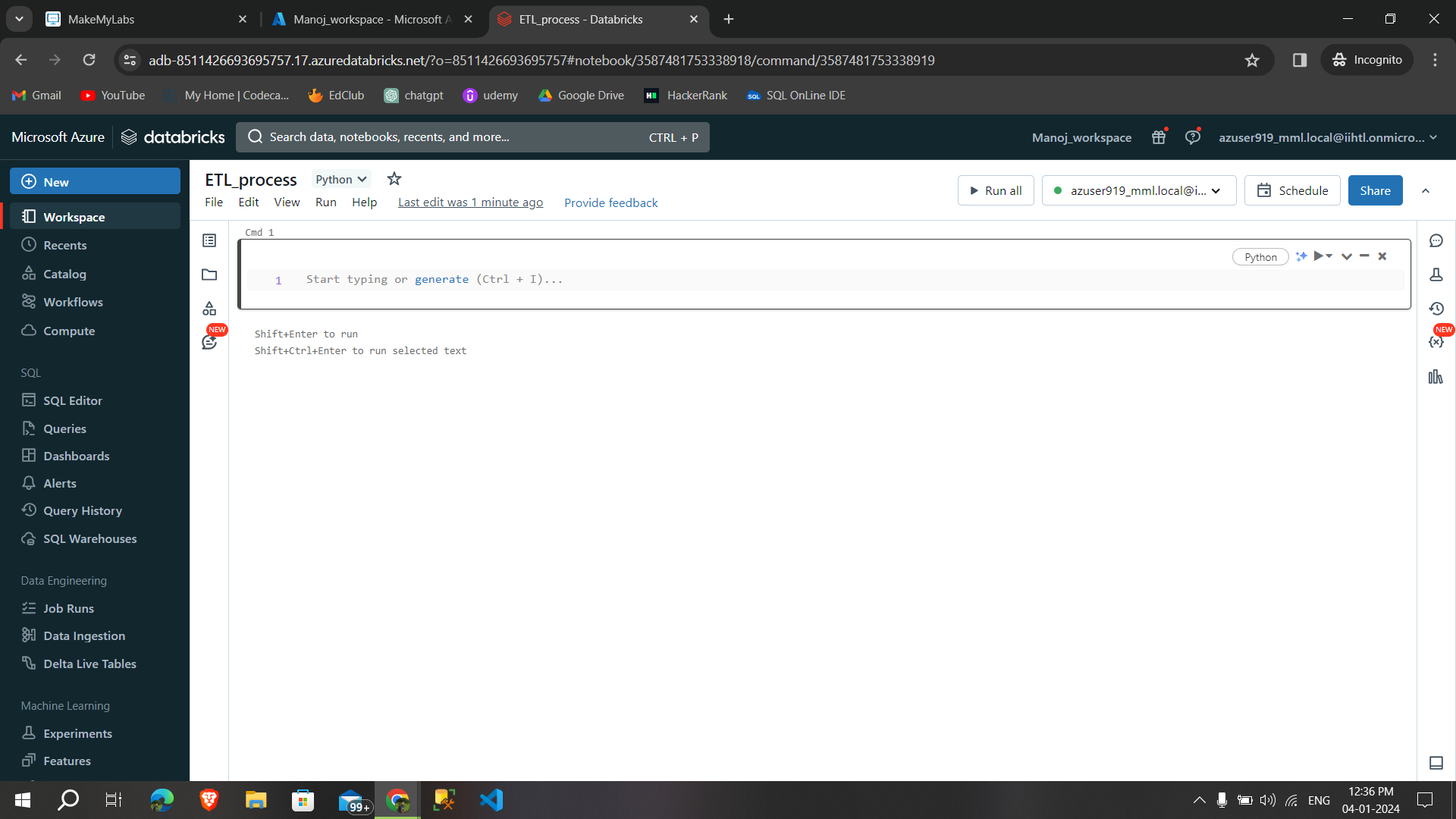
**Step 2:**After successfully created a workspace launched the **workspace to create a cluster**



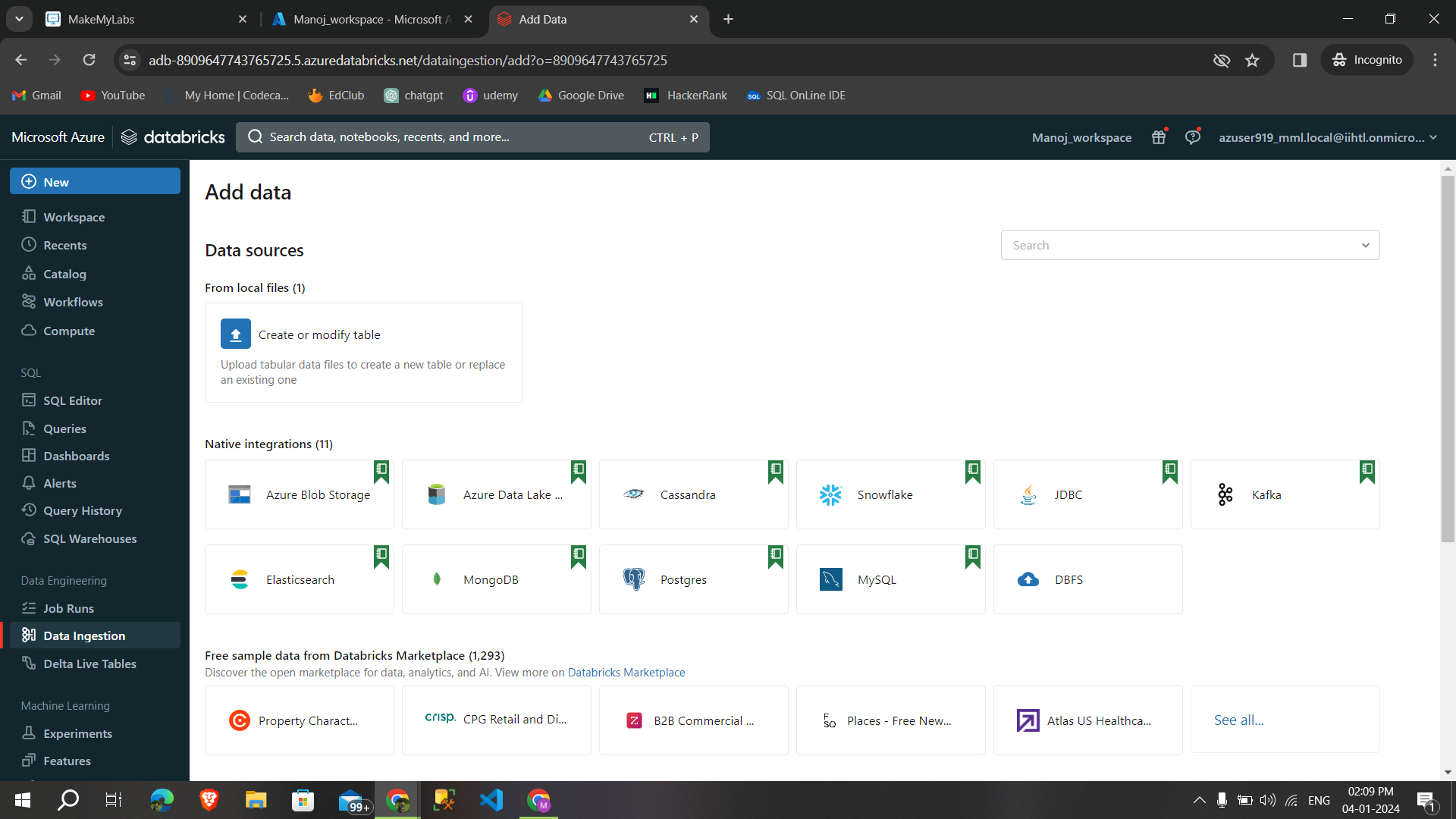
**Step 3:**Successfully created a cluster and **create a notebook**



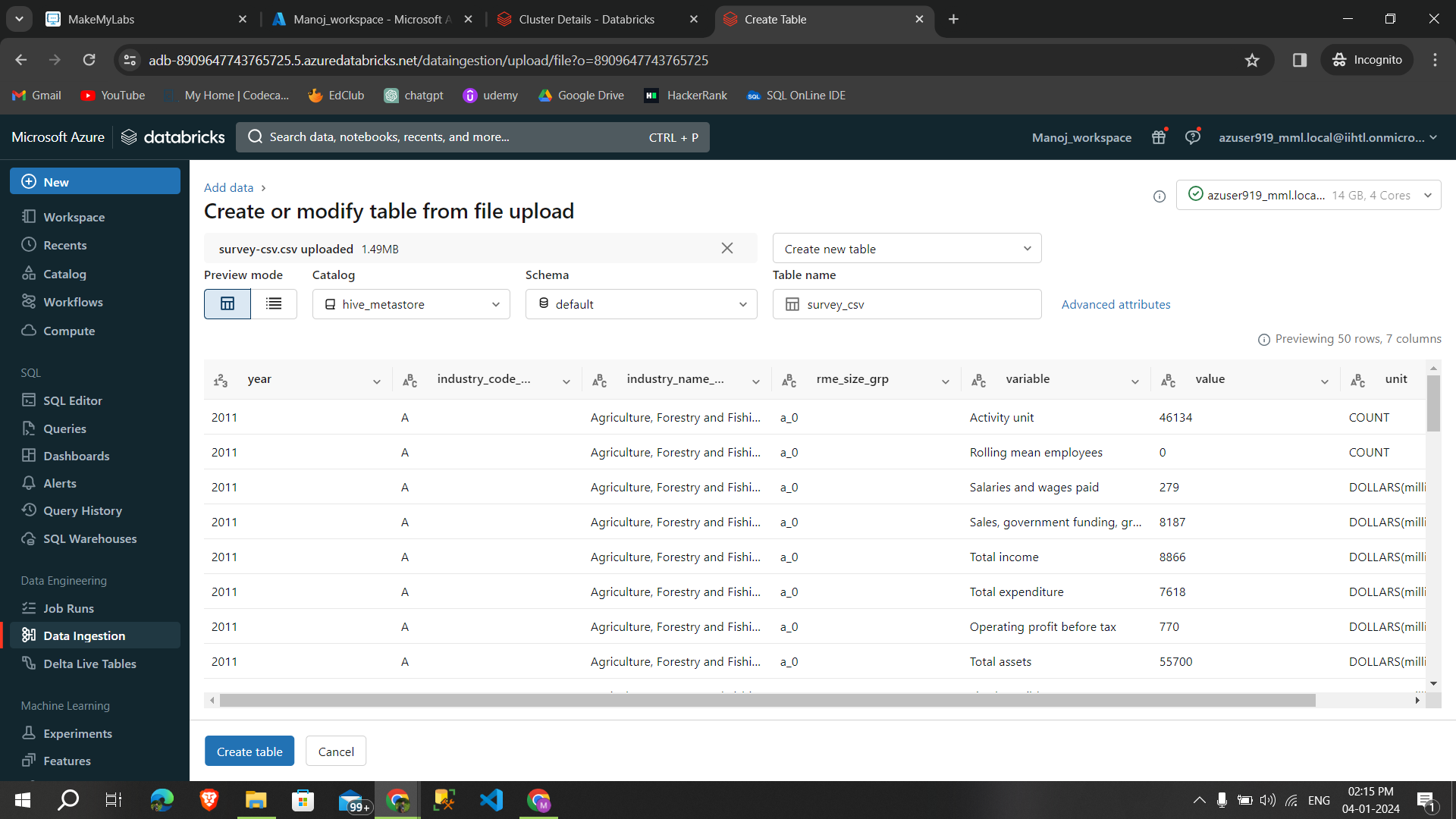
**Step 4:**created a notebook called **ETL\_process**



**Step 5:**After that selecting data ingestion to **ingest the data from the source**

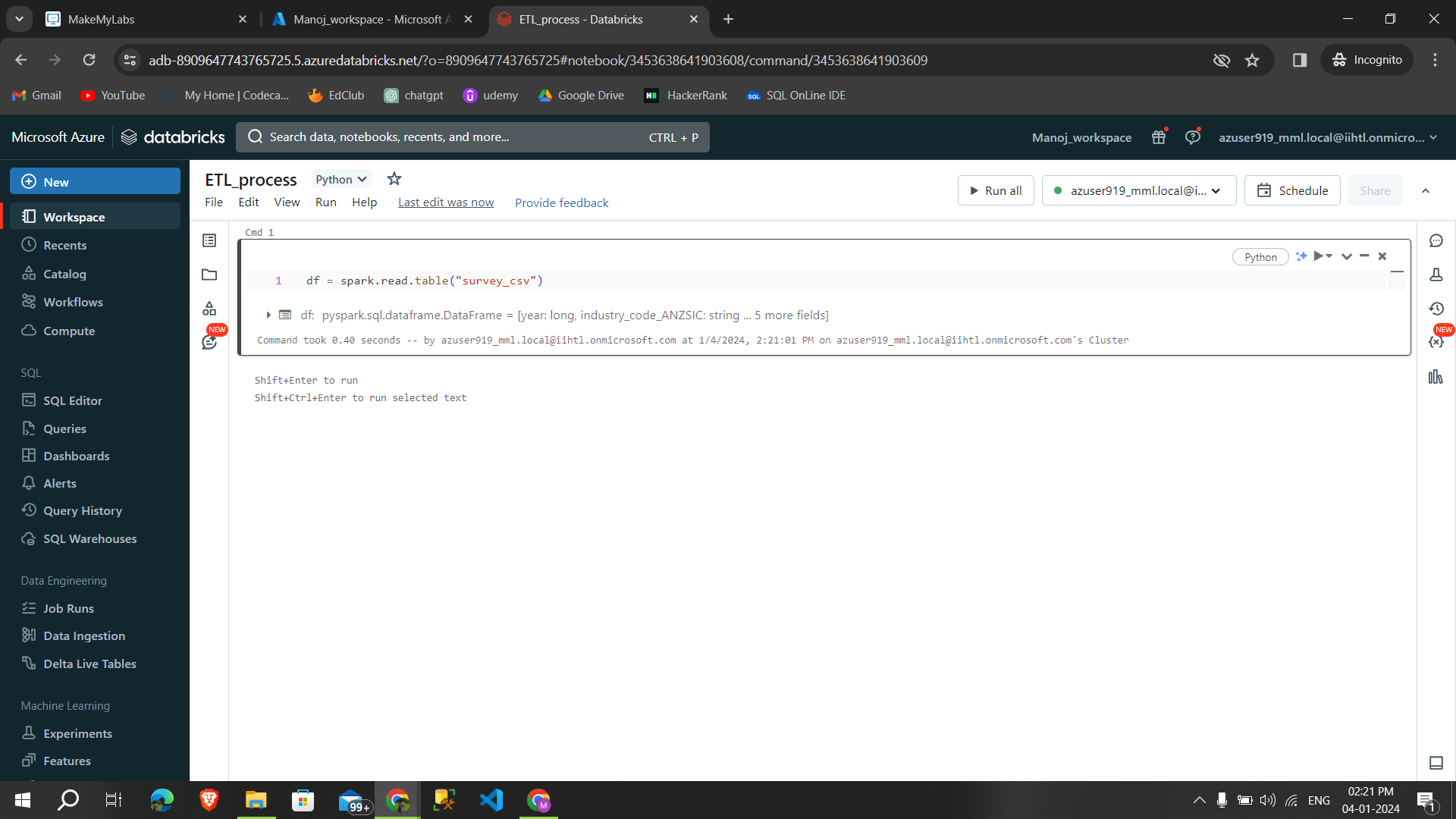


**Step 6:**Then i uploaded a csv file **from local to azure databricks** called survey-csv and created a table on the particular cluster(extracting process) **Data ingesting**

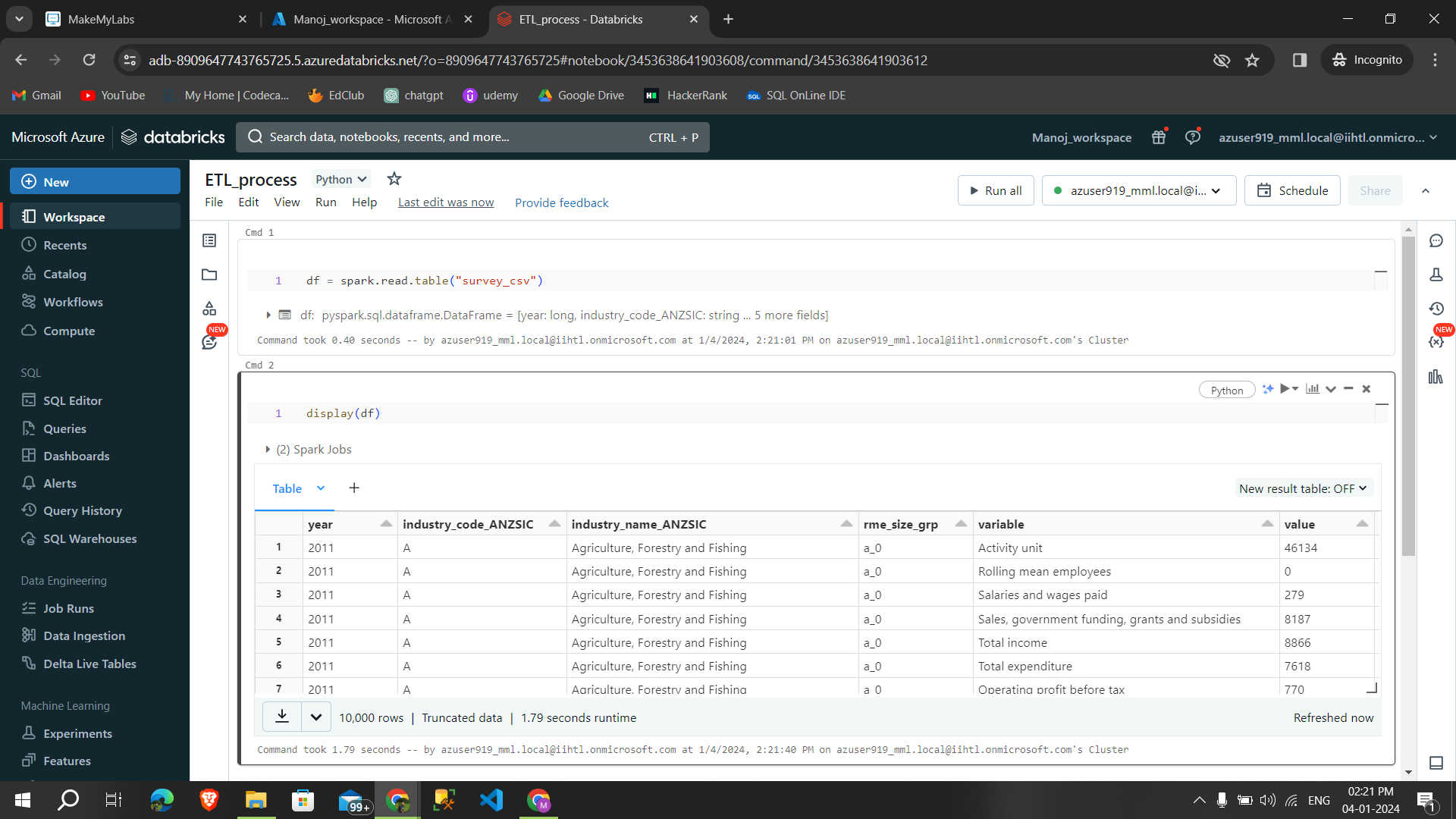


**Step 7:**After that reading that file from notebook by using the following command

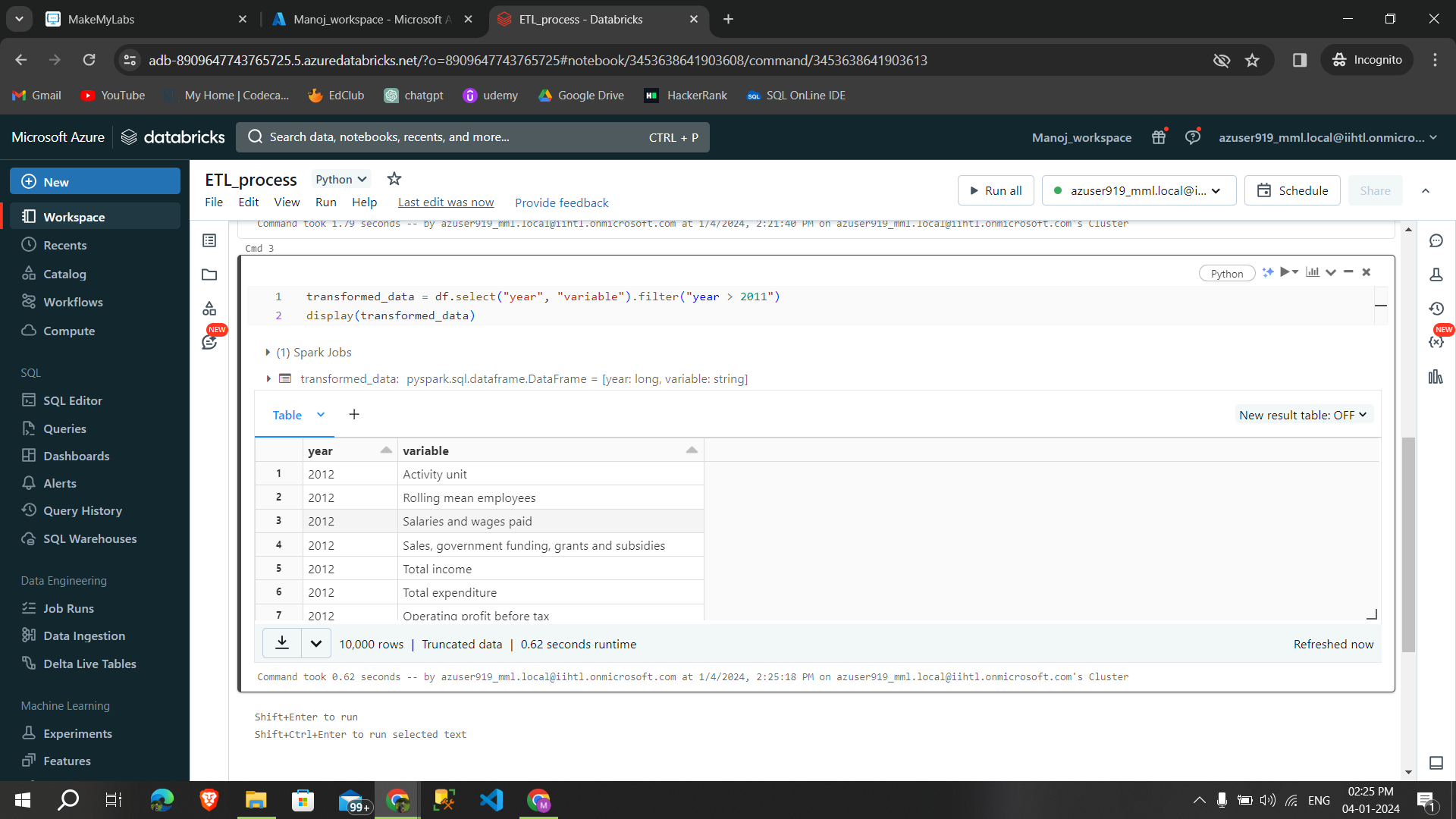
**df = spark.read.table(“survey-csv”)**



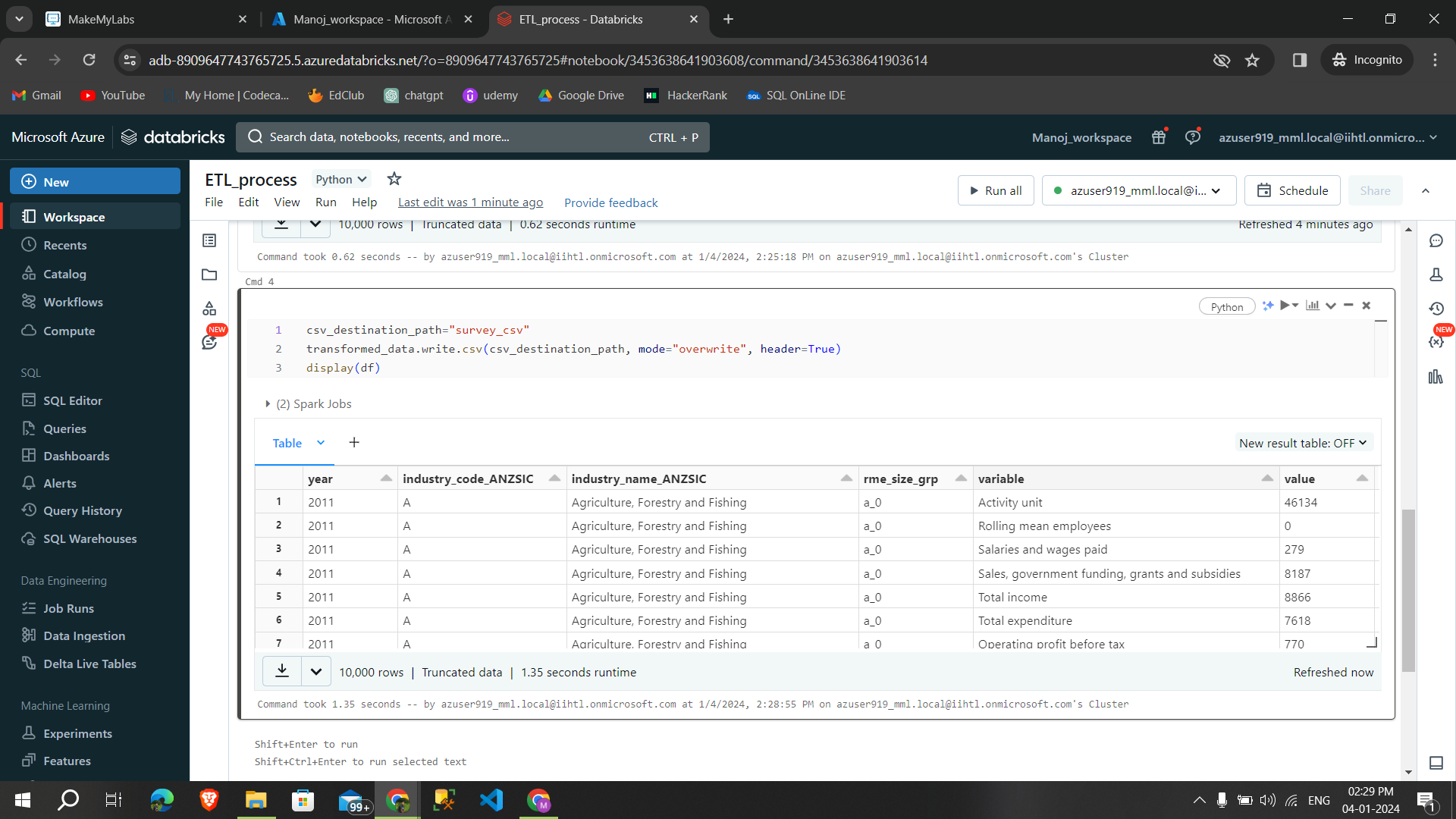
**Step 8:** Then displaying the table by using **display(df)**



**Step 9:**Transforming the data into the year greater than 2011 by using the filter command and displaying it(**transforming process**) based on needs,



**Step 10:** Then loading the transformed data into the same csv file **using overwrite (loading)**



We can also schedule the job and monitor them

**Schedule the Job:**

* If you want to run the ETL process at **regular intervals,** you can schedule it as a Databricks Job using the **Jobs UI.**

**Monitor and Debug:**

* Use Databricks' monitoring tools to **track the progress of your job** and identify any issues.
* The Databricks **UI provides visualizations and metrics** to help you monitor the **performance of your Spark job.**

These are steps to Setting up pipeline for ETL