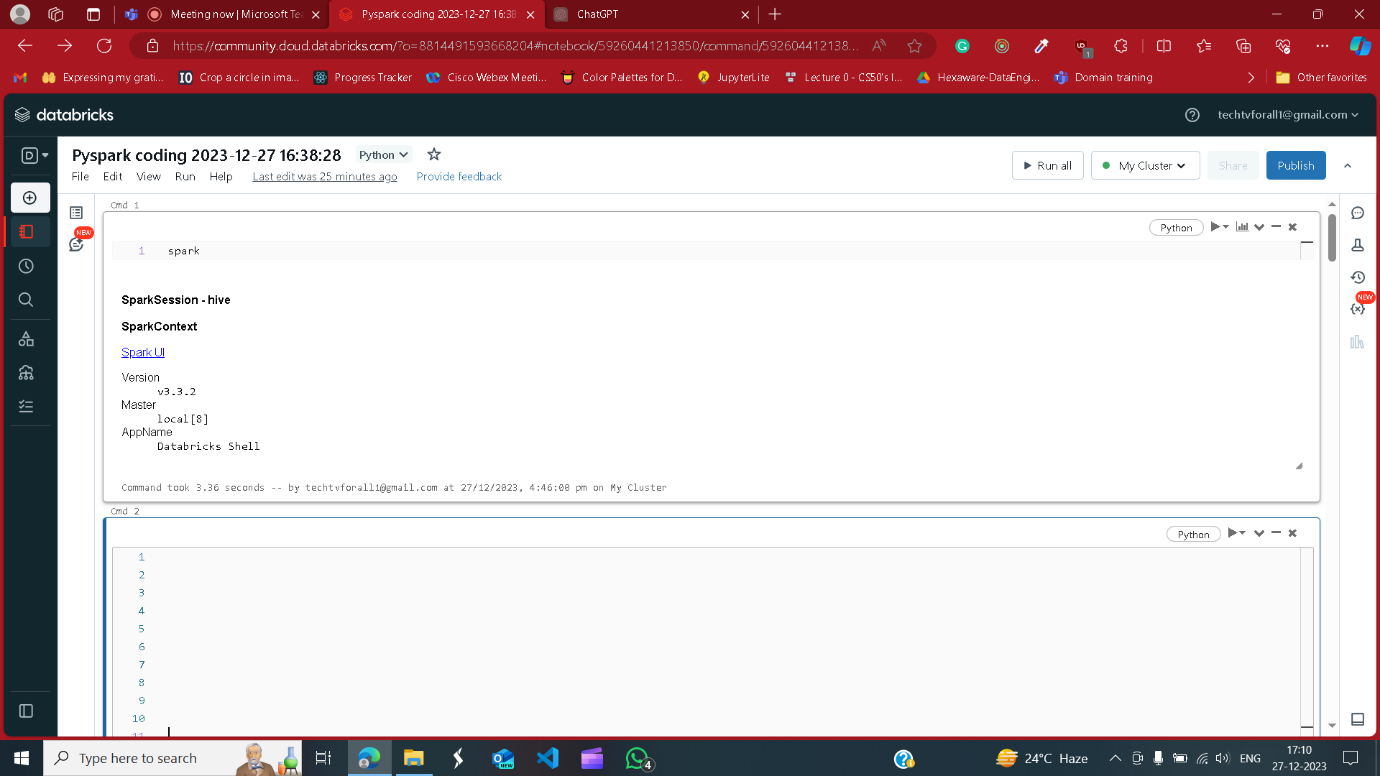
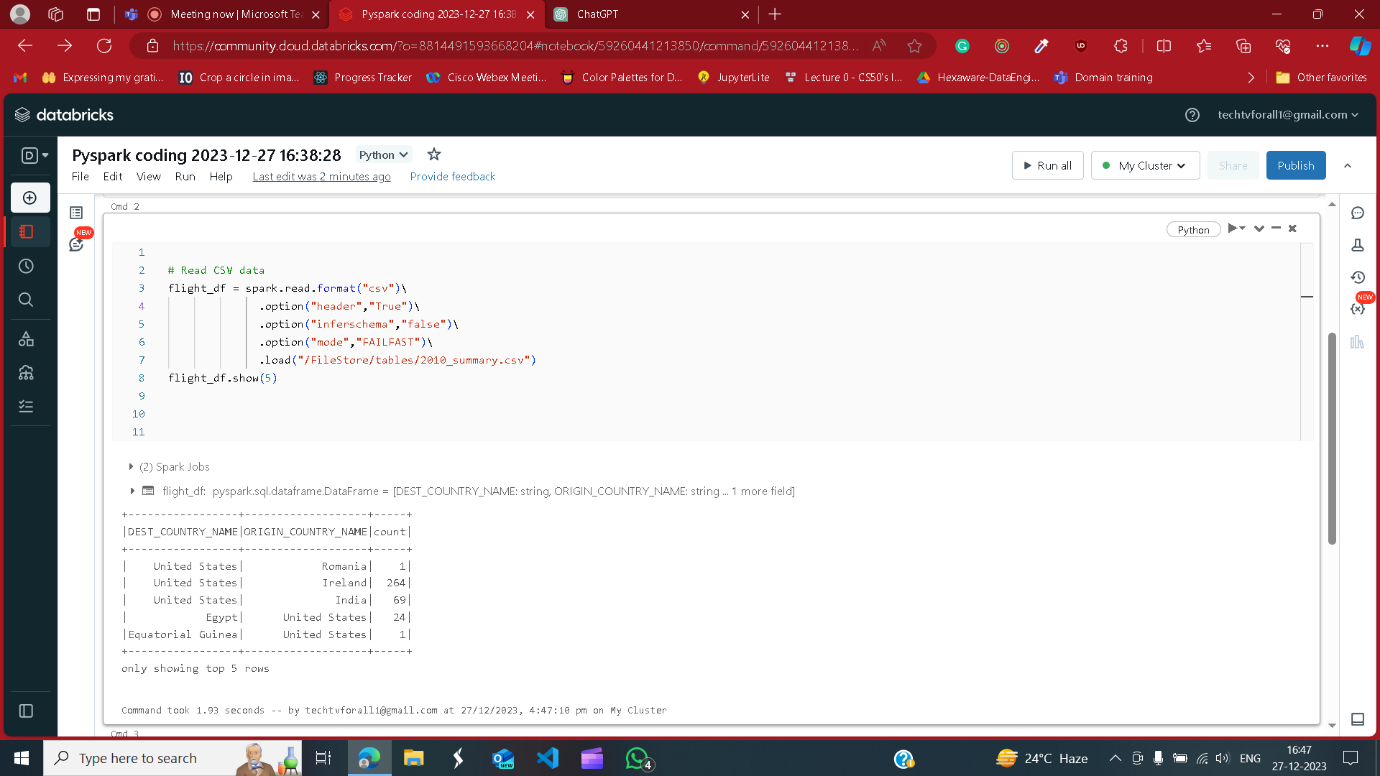
PYSPARK CODING ASSESSMENT 27-12-2023

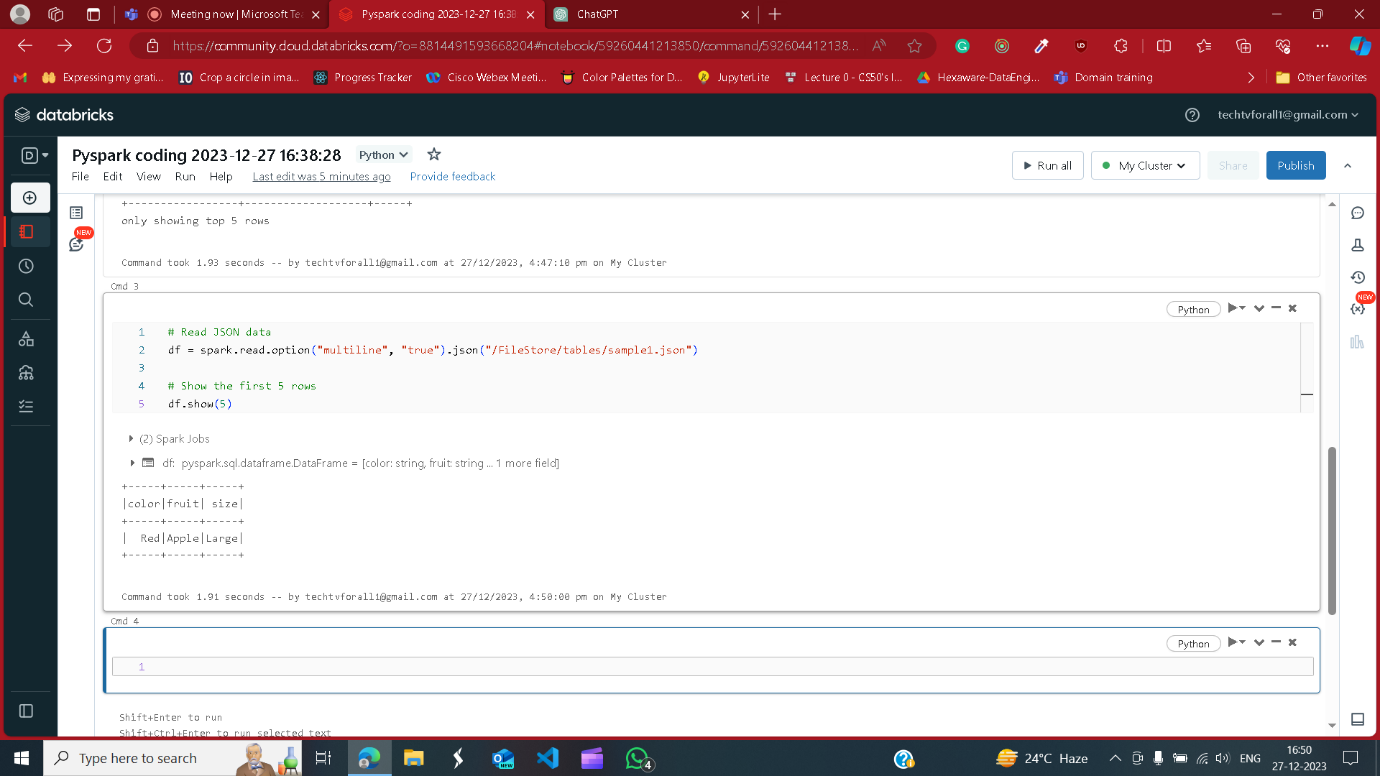
**1.Implement Processing JSON and CSV data with PySpark**

**Started spark session**



Reading CSV file

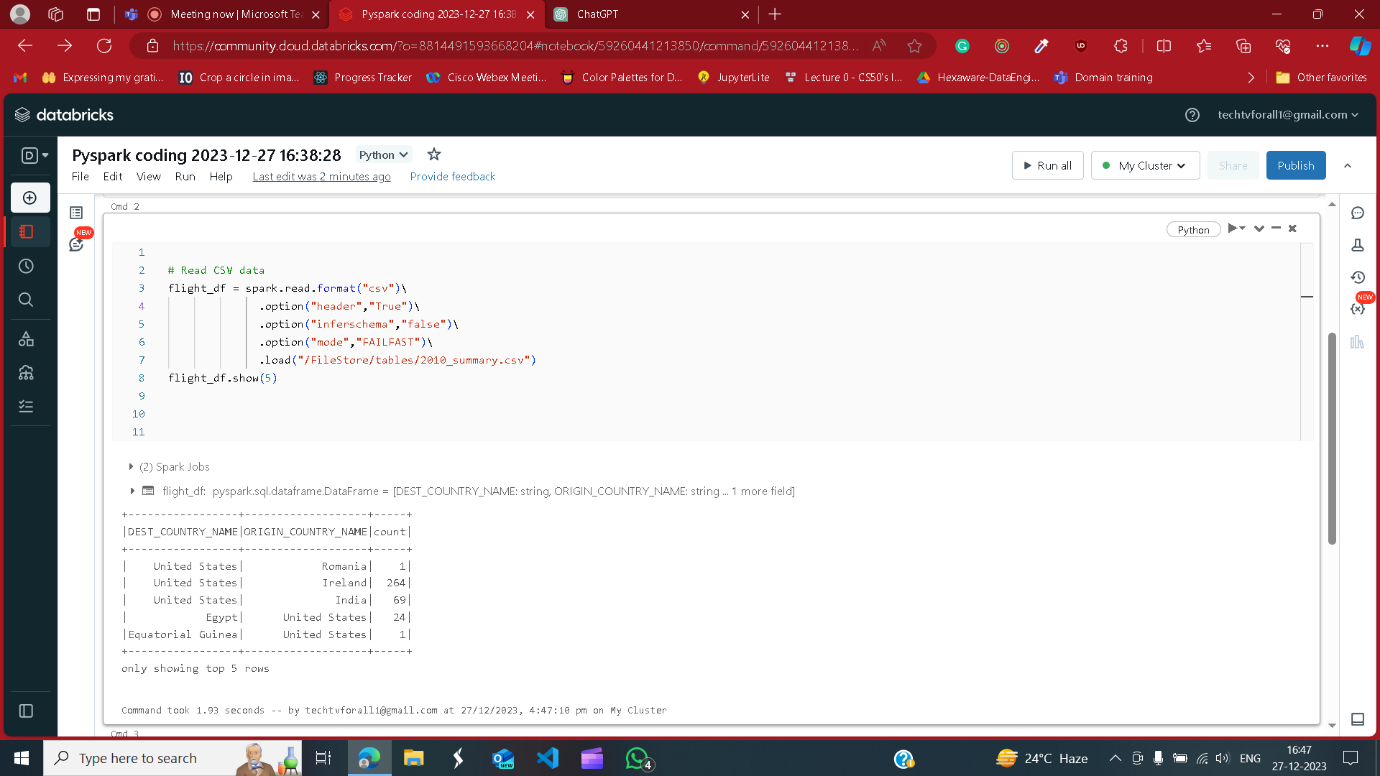
Reading JSON file

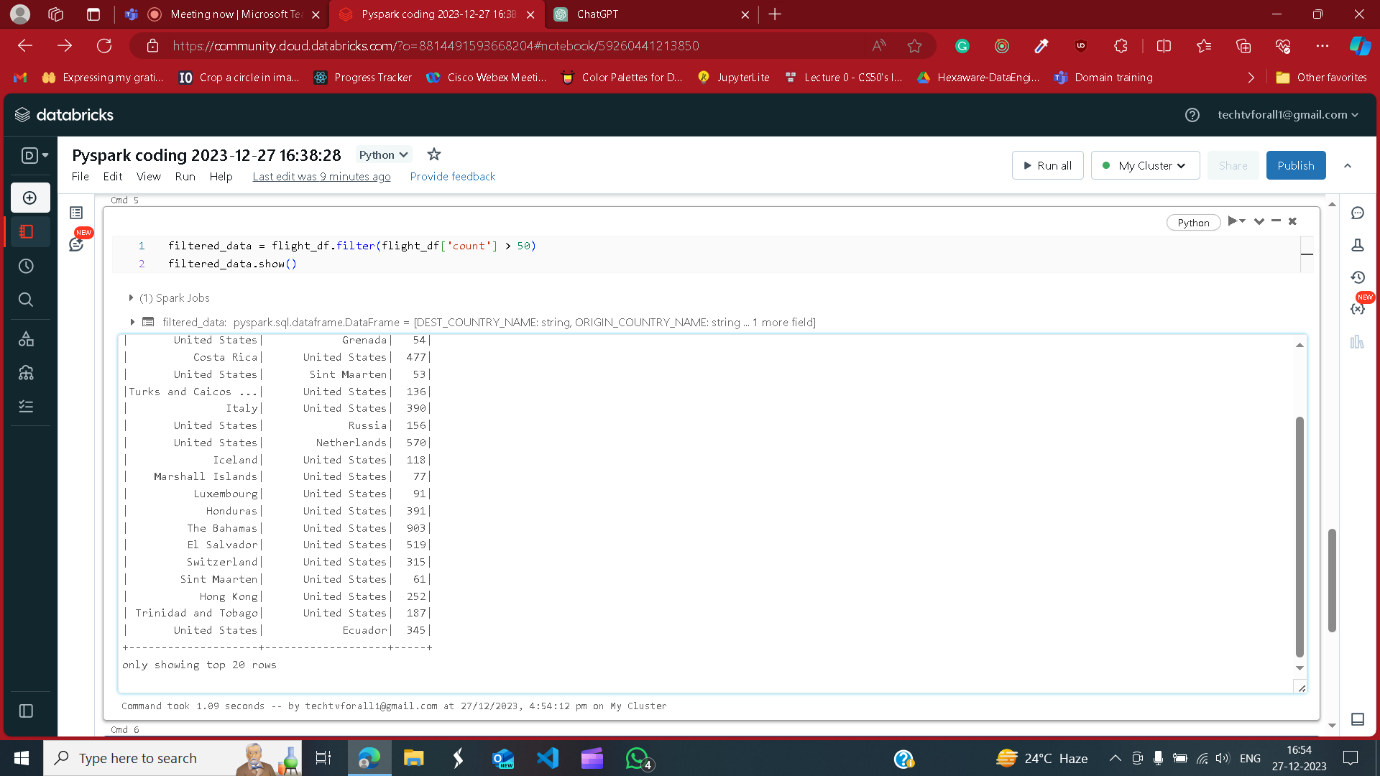


**2.Explain ETL (Extract, Transform, Load) with PySpark**

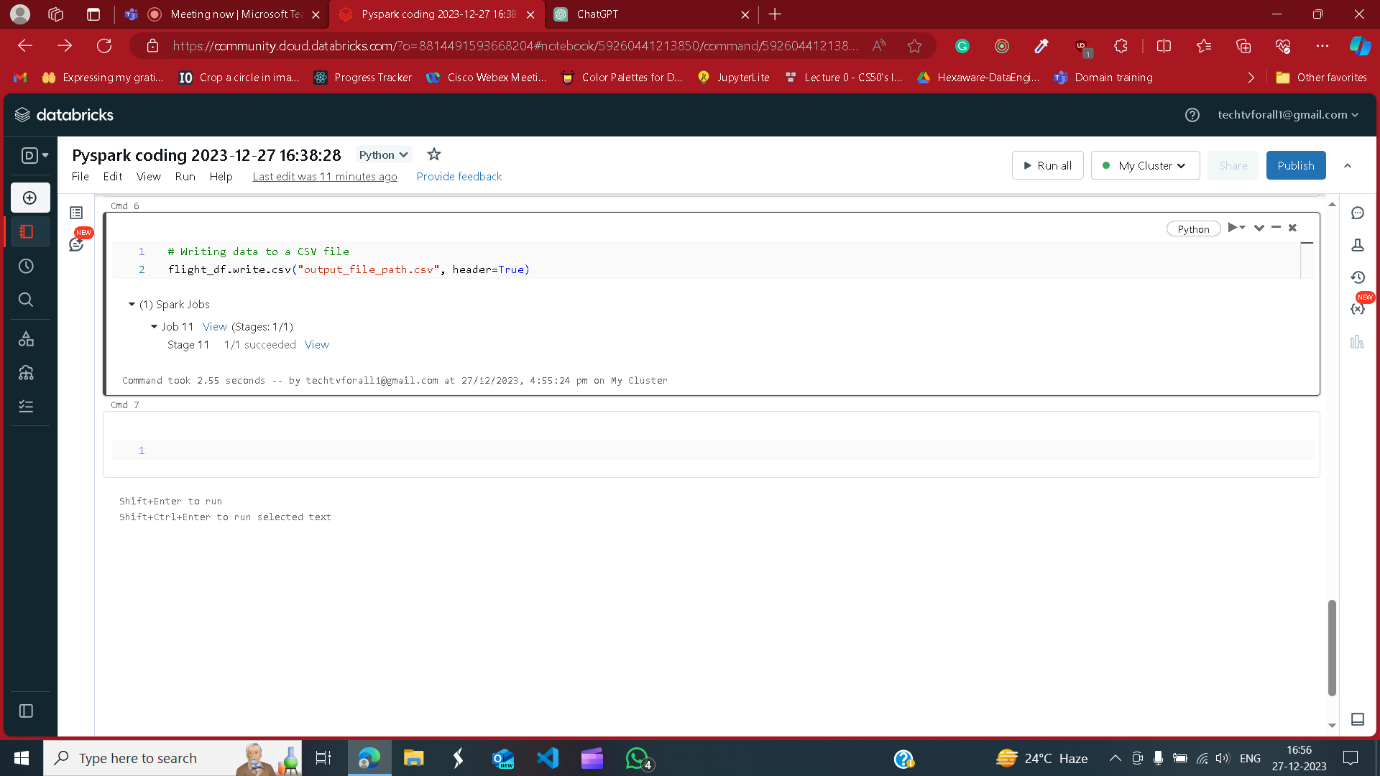
ETL (Extract, Transform, Load) is a data integration process that involves extracting data from various sources, transforming and loading to different databases.

Extracting – reading data from a file



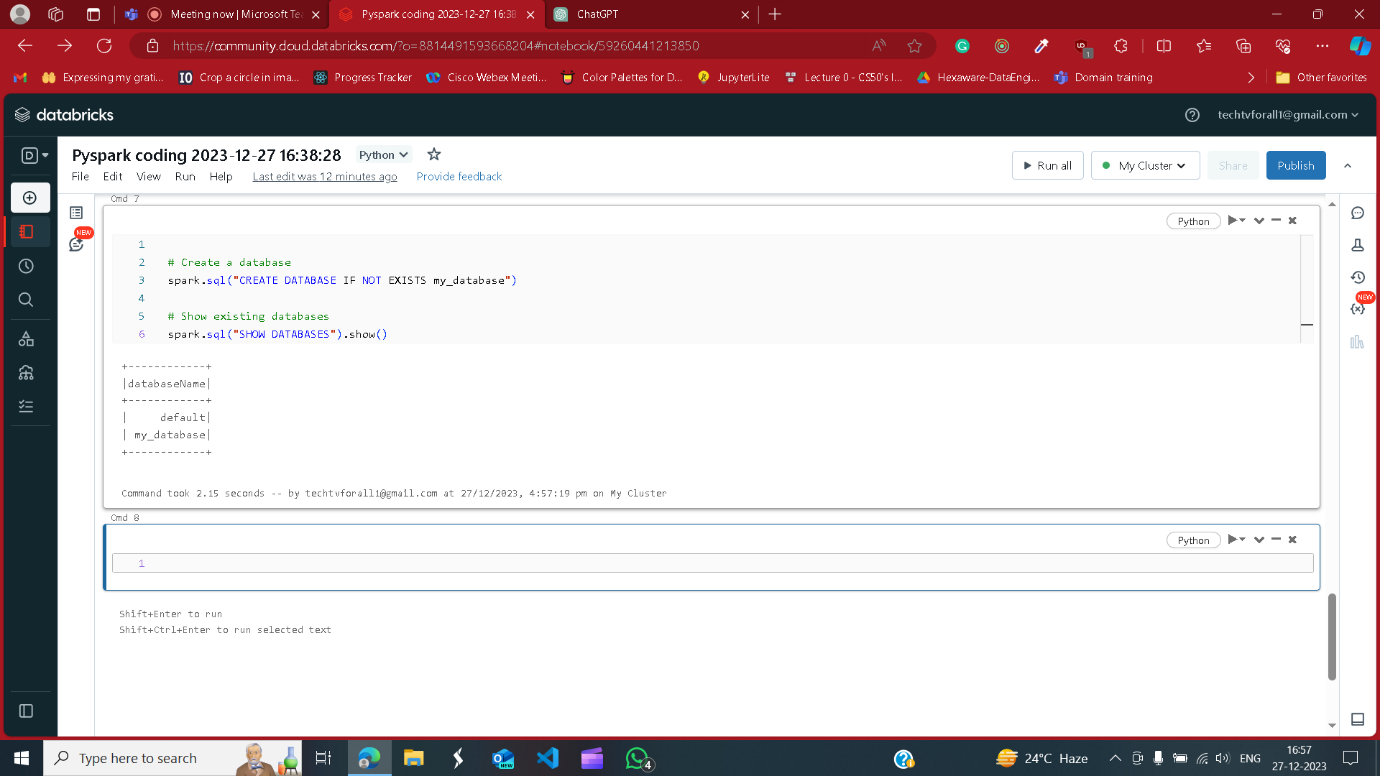
Transforming

**Loading** here creating a new file

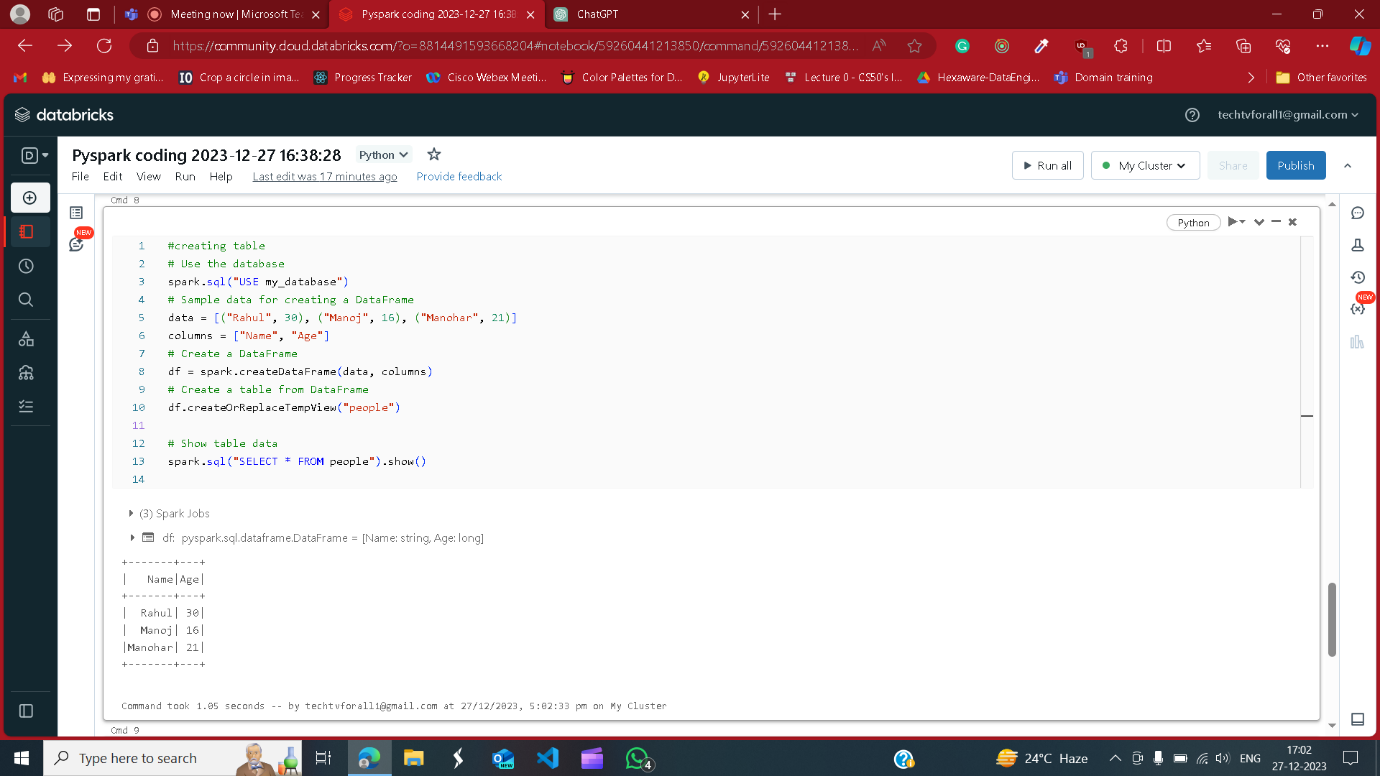


**3.Using Spark SQL - Creating databases, tables**

**Creating database**

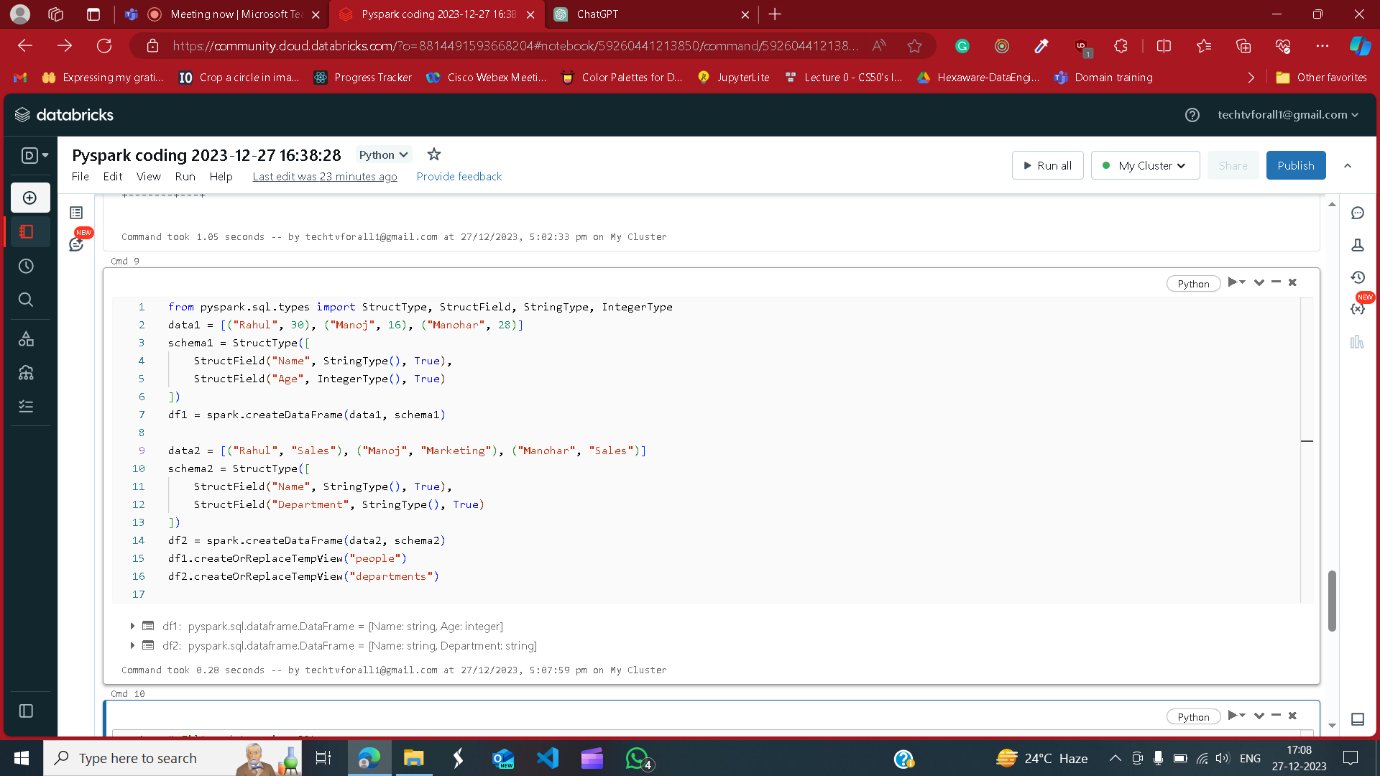


Creating table

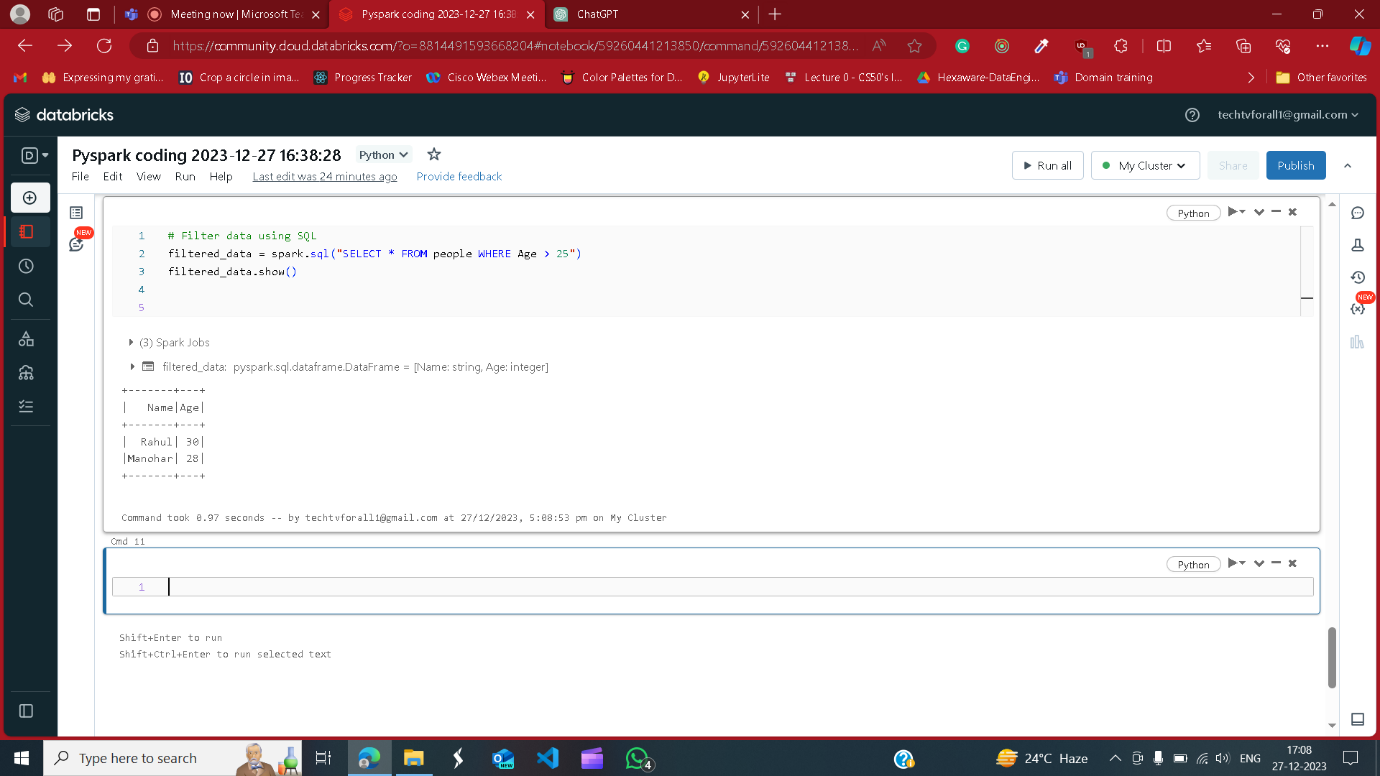


**4.Using Spark SQL - Transformations such as Filter, Join, Simple Aggregations, GroupBy.**

**Creating dataframes and table**

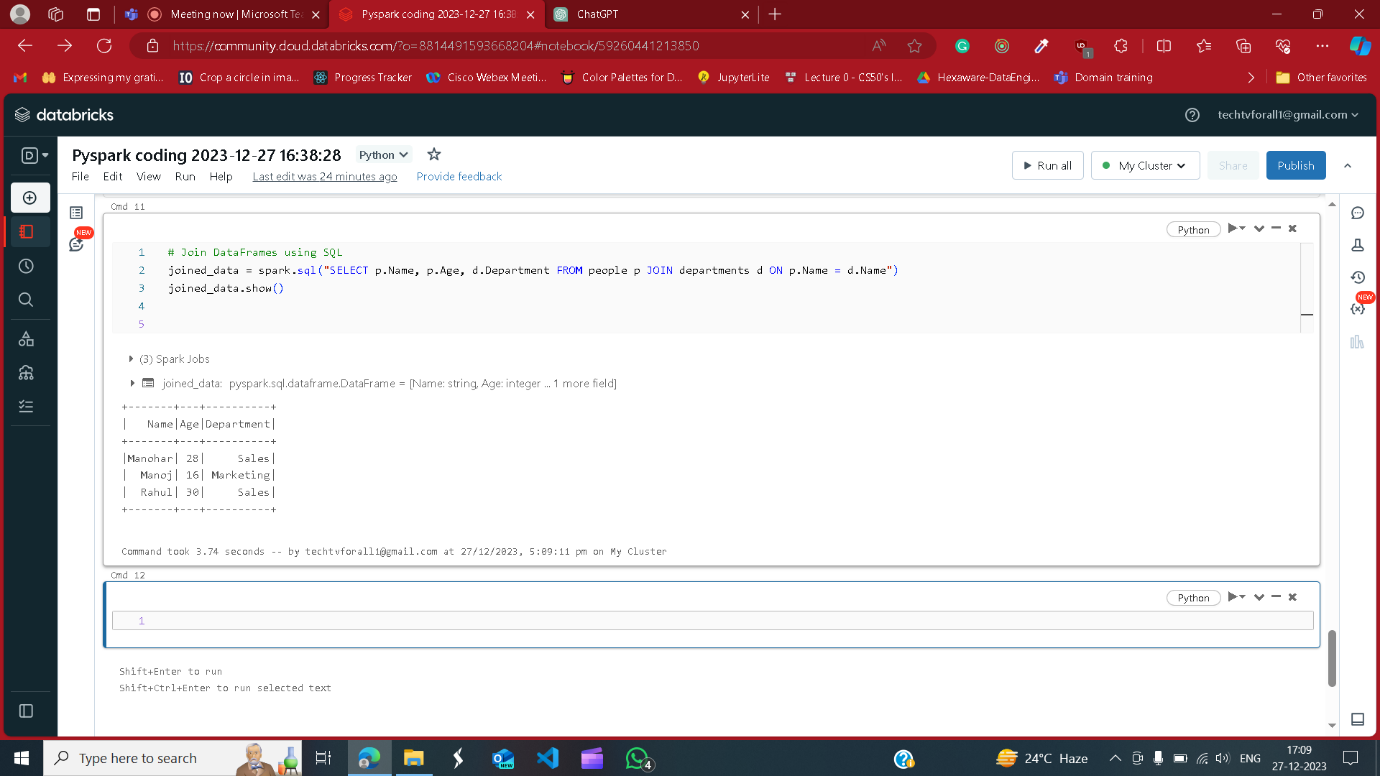


Filter

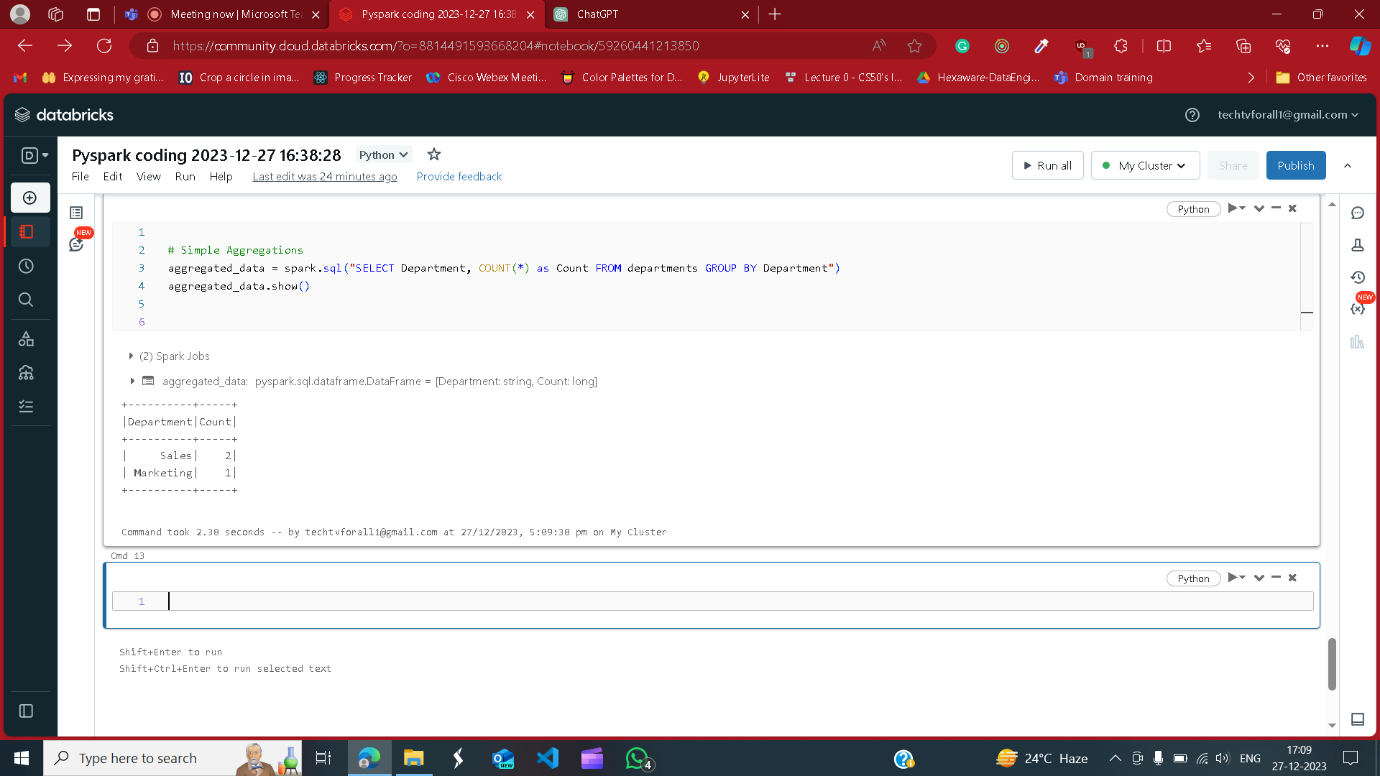


Join

It joins two or more table based on related column



Simple Aggregation



Group by and aggregation

Group by group the same related row or same kind of row

