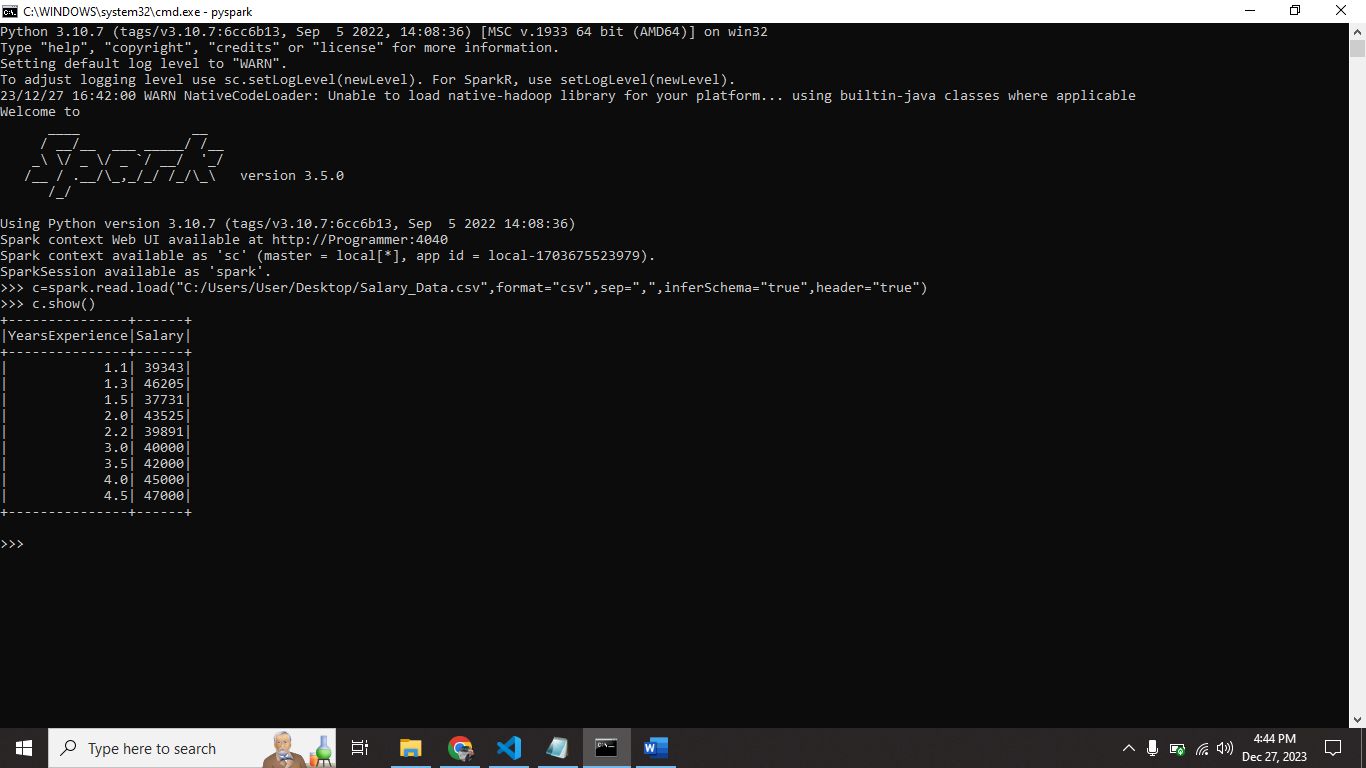
**Enayathulla A**

**Pyspark Assessment**

**27/12/2023**

**Q1. Implement Processing CSV data with PySpark**

I read my local CSV (Salaty\_Data) file in cmd using Pyspark Commands.



**Q2.Explain ETL:**

ETL🡪 Extract, Transform, Load.

Extract🡺 It retrieve data from various sources.

Transform 🡺 Cleaning, aggregate and manipulation the data.

Load🡺 Processed data store into various data sources.

Step1) we want to get raw data from nodes that process is consider as extract.

Step2) we process the data in using format or readable format that is called transform.

Step3) The cooked data stored into file format or else some data sources for analytics purpose.

**All the pyspark commands are run in databricks community cloud**

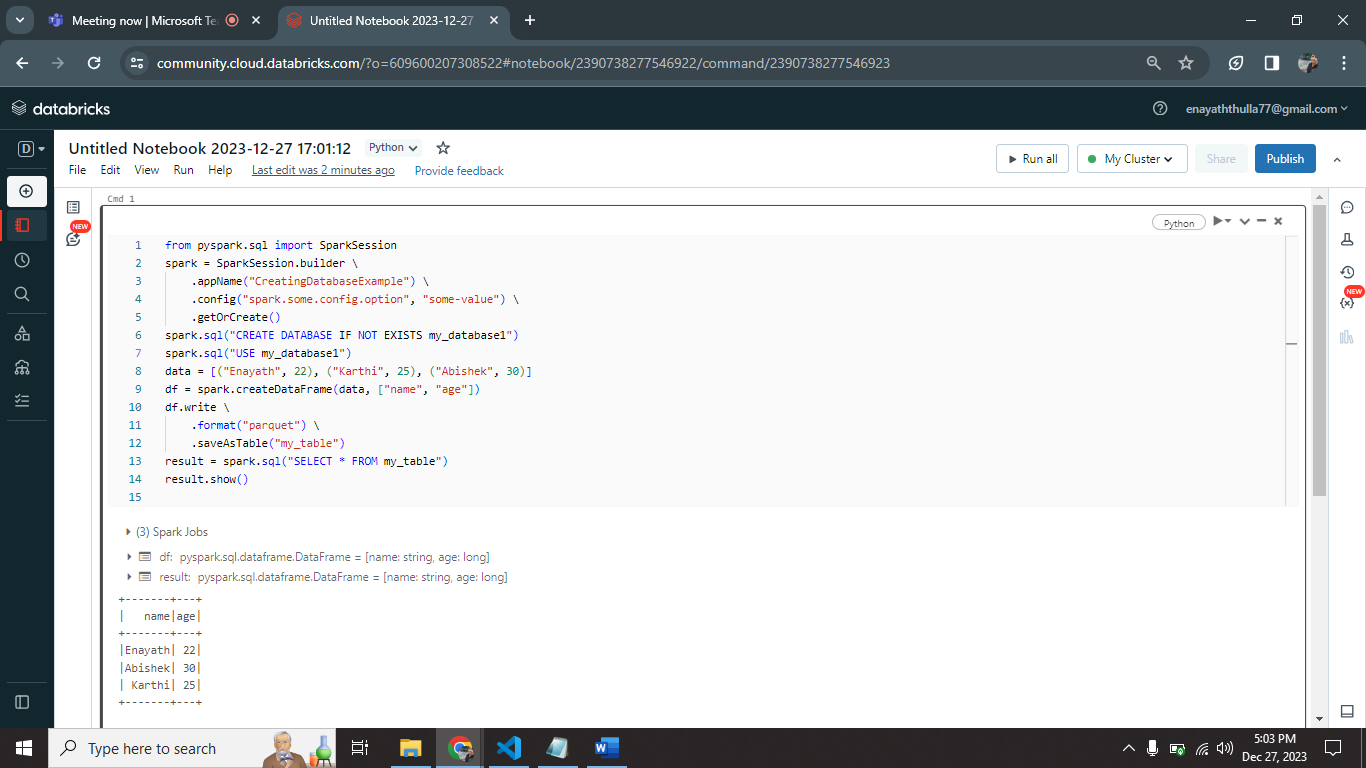
**Q3. Using Spark SQL - Creating databases, tables**

I created the database(my\_database1) using Pyspark.

I created the table (my\_table) using Pyspark.

In result variable I stored SQL command.

Using result.show() I displayed the table.



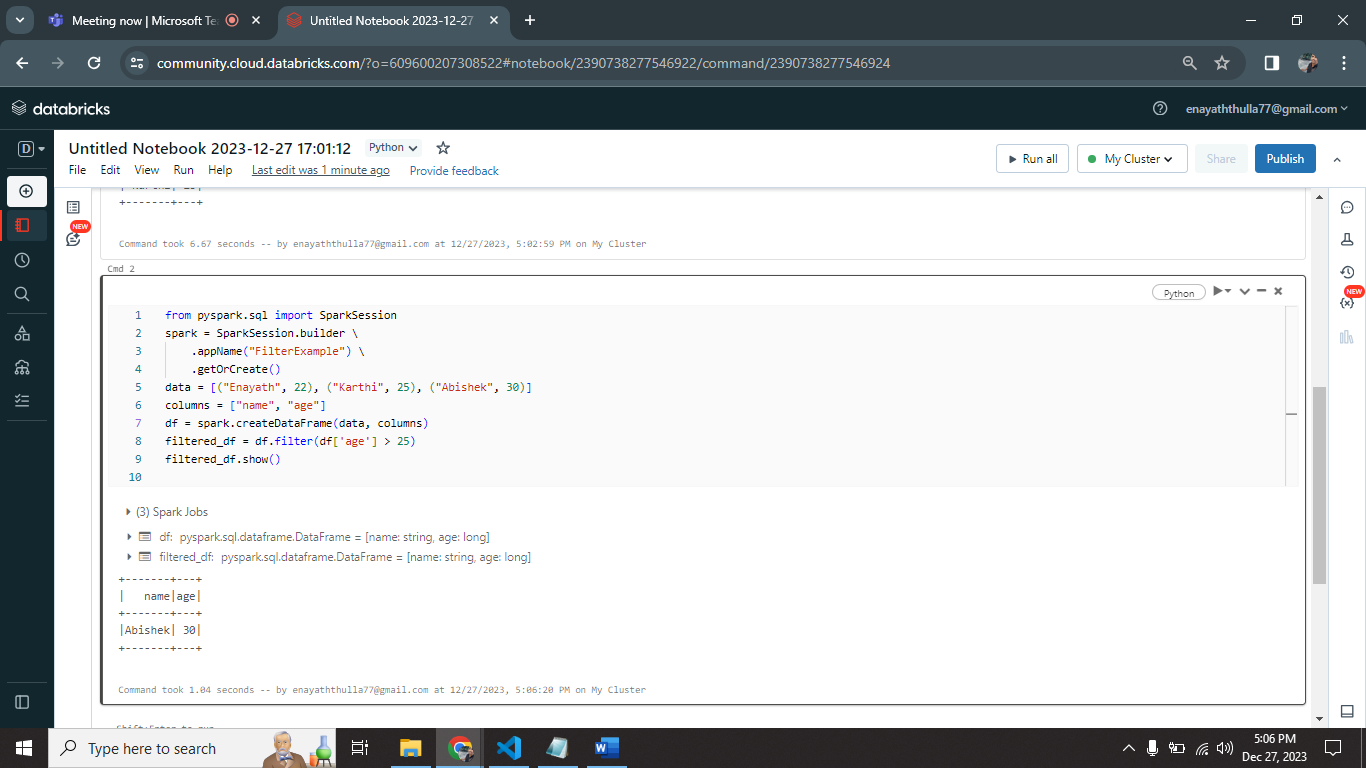
**Q4. Using Spark SQL - Transformations such as Filter, Join, Simple Aggregations, GroupBy.**

**Filter:**

I created the data frame (df) with some row values.

I filtered table where age is greater than 25 using “filtered\_df = df.filter(df['age'] > 25” this command.

I displayed filtered table using “filtered\_df.show()” command.

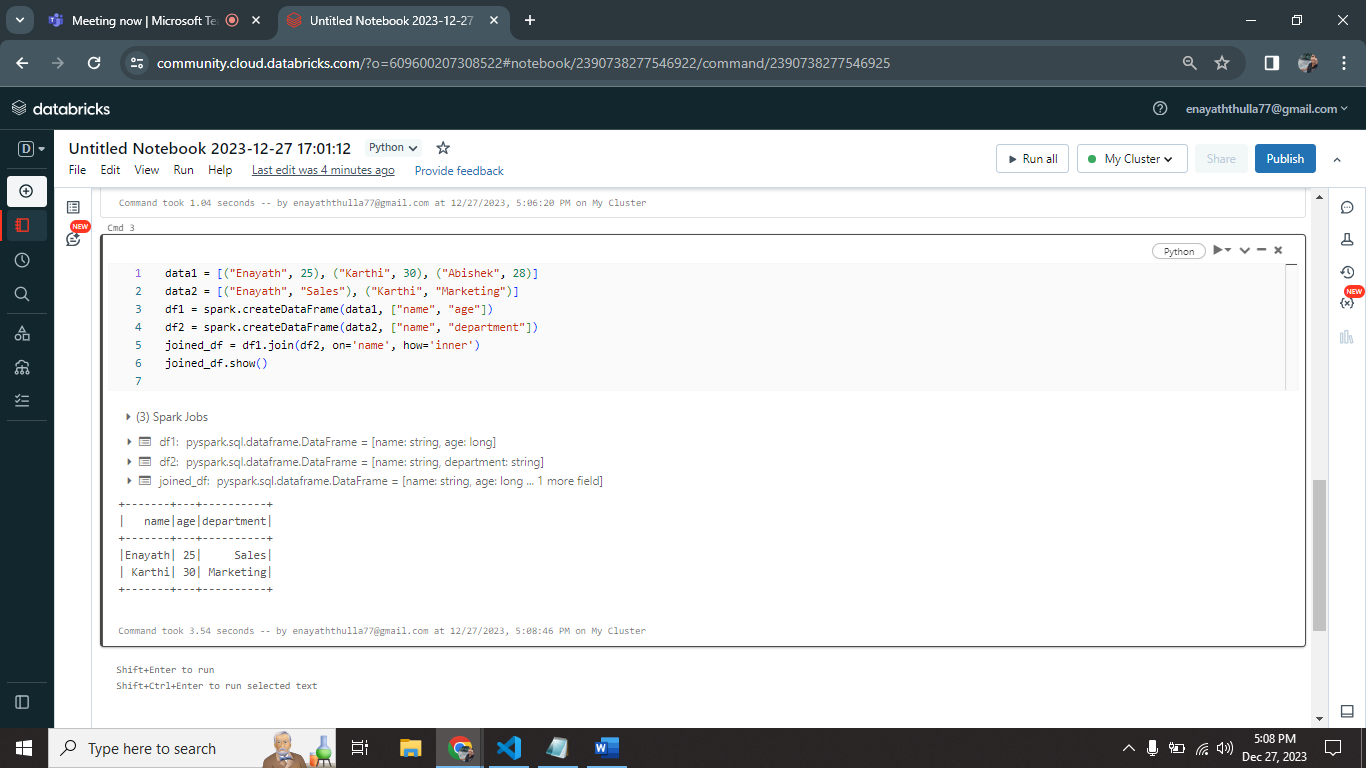


**Join:**

I created df1 and df2 dataframes with the values of data1, data2.

I joined two different dataframes using “joined\_df = df1.join(df2, on='name', how='inner')” this command.

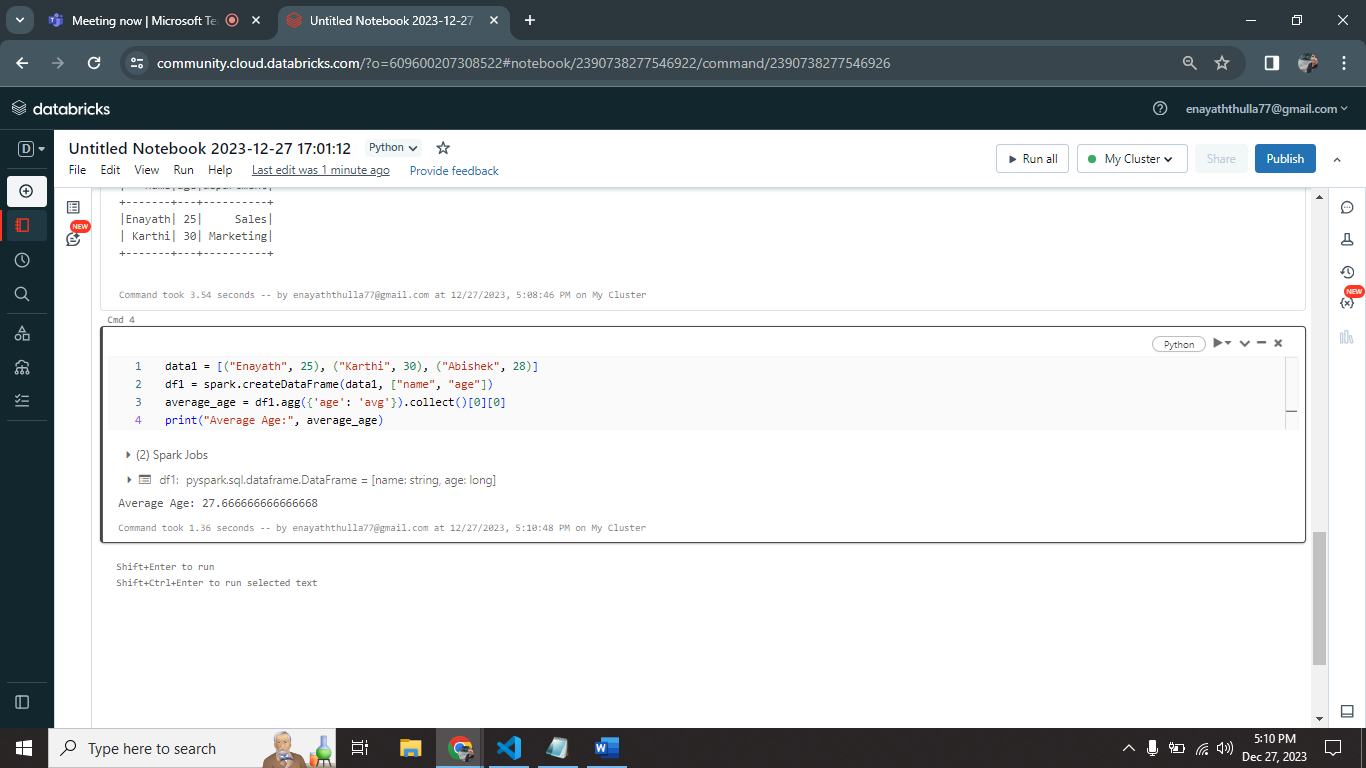
I displayed joined table output using “joined\_df.show()”.



**Aggregation:**

I created df1 dataframe for performing one simple aggregation function.

It display the average age of age column.



**Group by:**

I created df1 dataframe with the help of data1 values.

I group by this table using “grouped\_df = df1.groupBy('age').count()” this command.

It display values group by age.

