|  |
| --- |
| **Assignment#3** |
|  |  |
|  | **Bank Marketing Campaign Data Analysis with PySpark SQL** |
|  |  |
|  | **Create PySpark Application - bank-marketing-analysis.py. Perform below operations.** |
|  | **a) Load Bank Marketing Campaign Data from csv file** |
|  | **b) Get AgeGroup wise SubscriptionCount** |
|  | **c) Write the output in parquet file format** |
|  | **d) Load the data from parquet file written above** |
|  | **e) Show the data** |
|  | **f) Filter AgeGroup with SubcriptionCount > 2000 and write into Avro file format** |
|  | **g) Load the data from avro file written above** |
|  | **h) Show the data** |
|  |  |
|  |  |

**import pyspark**

**from pyspark.sql import SparkSession**

**spark = SparkSession.builder.appName('pyspark-examples').getOrCreate()**

**df=spark.read.options(delimiter=";",header=True, inferSchema=True).csv("/home/tech/futurense\_hadoop-pyspark/labs/dataset/bankmarket/bankmarketdata.csv")**

**df.registerTempTable("bankmarketing")**

**out=spark.sql("select (case when age<13 then 'Kids' when age<20 then 'Teenagers' when age<31 then 'Youngsters' when age<50 then 'Middleagers 'else 'Seniors' end) as ag,count(\*) sub\_cnt from bankmarketing where y='yes' group by ag order by sub\_cnt desc ")**

**out.select('ag','sub\_cnt').write.format('parquet').save('/home/tech/spark/bankmarketing/new\_parquett')**

**parq\_df=spark.read.format('parquet').load('/home/tech/spark/bankmarketing/new\_parquett')**

**parq\_df.show()**

**parq\_df.registerTempTable("bankmarketing1")**

**out1=spark.sql("select \* from bankmarketing1 where sub\_cnt>2000 ")**

**out1.show()**

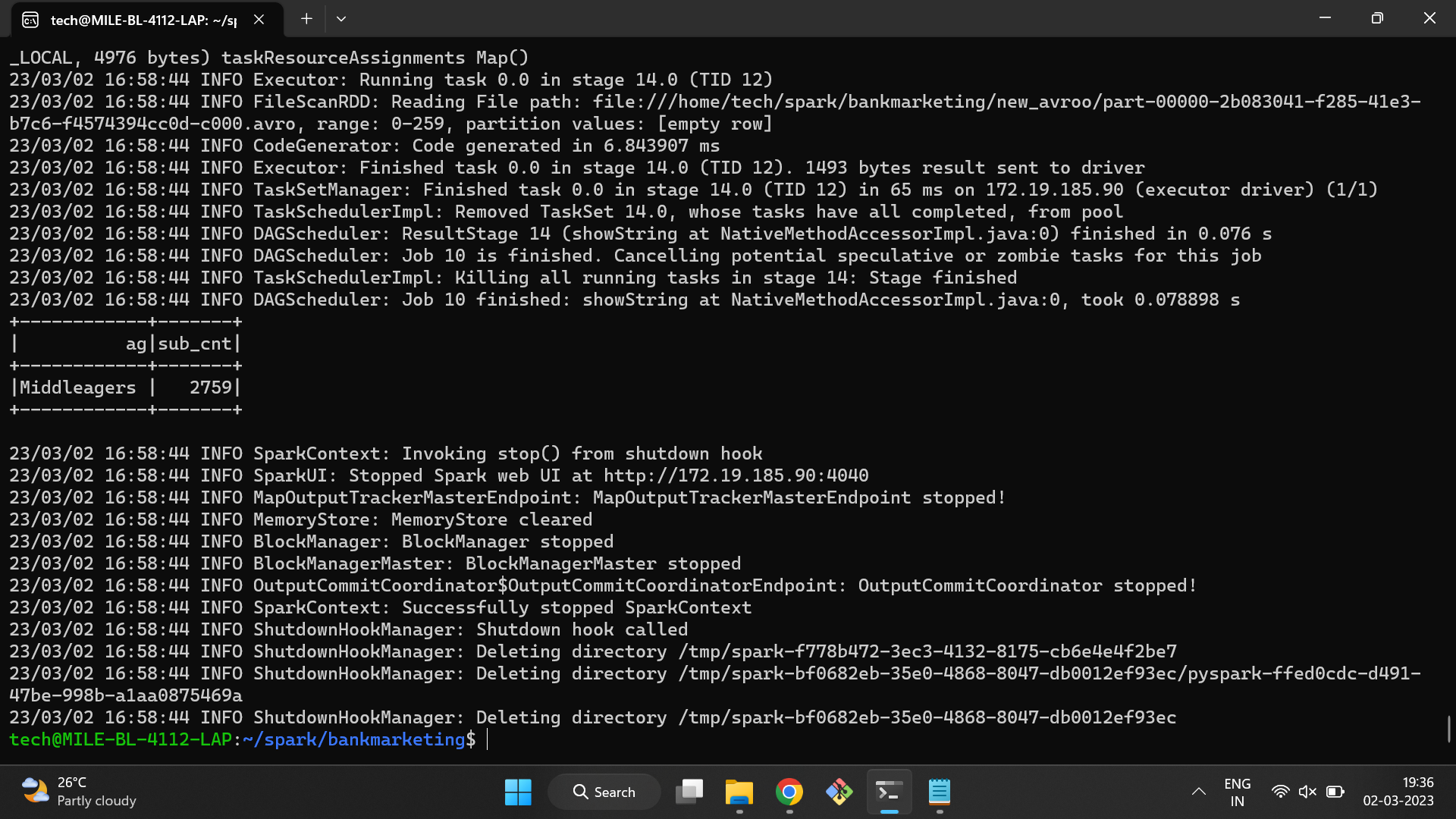
**out1.select('ag','sub\_cnt').write.format("avro").save('/home/tech/spark/bankmarketing/new\_avroo')**

**avro\_df=spark.read.format("avro").load('/home/tech/spark/bankmarketing/new\_avroo')**

**avro\_df.show()**

**1] spark-submit bank-marketing-analysis.py => Runs in local mode**

spark-submit --packages org.apache.spark:spark-avro\_2.12:3.3.2 ~/bank-marketing-analysis.py



**2] Run in cluster mode**

spark-submit --master spark://MILE-BL-4112-LAP.localdomain:7077 --packages org.apache.spark:spark-avro\_2.12:3.3.2 ~/bank-marketing-analysis.py

