Coding Challenge

1. Extract, Transform, Load in Pyspark.

ETL (Extract, Transform, Load) is a process used to prepare and integrate data for analysis and storage. In PySpark, ETL is performed on large datasets by using Spark's parallel processing capabilities. Here we can make use of clusters to perform the activity.

Extract involves retrieving data from various sources such as databases, APIs, or flat files. PySpark's built-in connectors and libraries, such as spark.read, support various file formats like CSV, JSON and Parquet for seamless extraction.

Transform is the core step where raw data is cleaned, enriched, and reshaped to meet business requirements. PySpark's DataFrame API and SQL capabilities enable transformations such as filtering, aggregations, joins, and column operations. The distributed architecture ensures transformations are fast and scalable, even for massive datasets.

Load involves storing the processed data into a target system, such as a database or a data warehouse, using PySpark's write functions. This completes the data pipeline, ensuring prepared data is readily available for analysis.

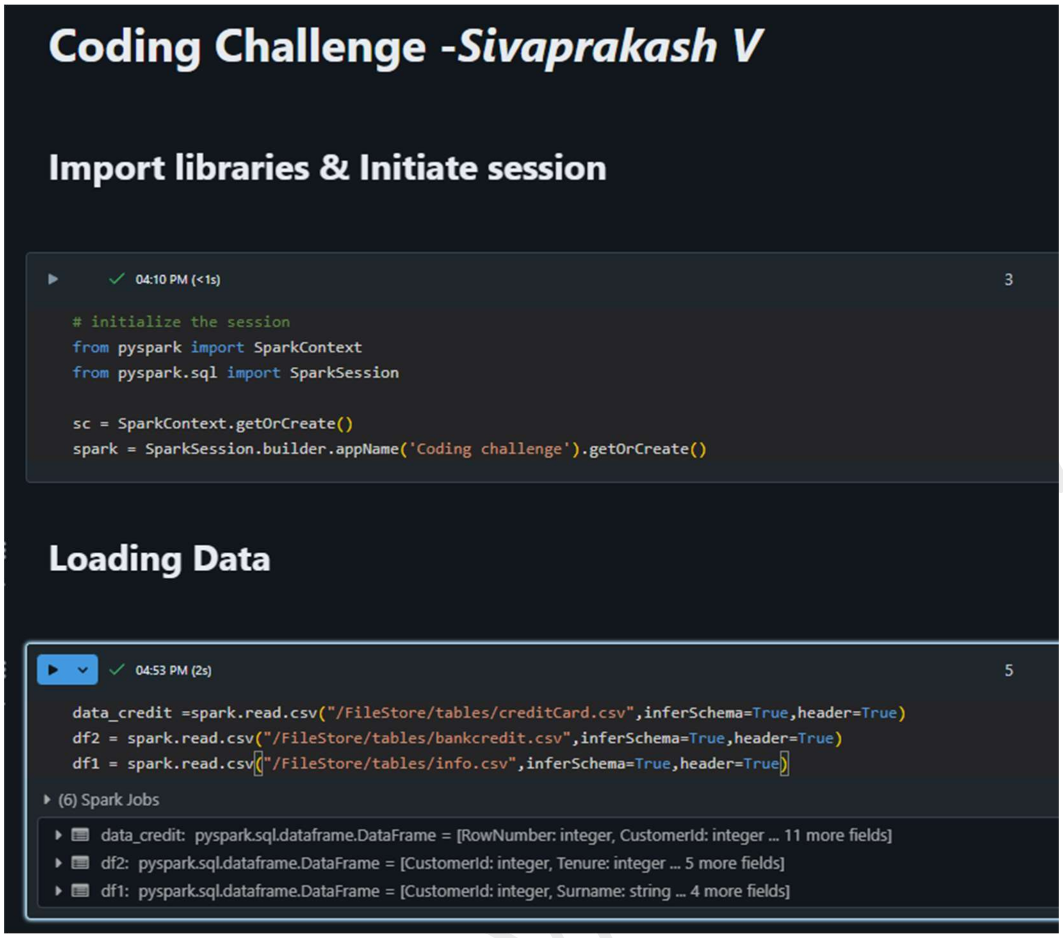
2. Queries & Solutions on Credit Dataset.

Read me:

For every query first I have written the Spark SQL code followed by PySpark code. question will be written in markdown.

1. First solution is on Spark SQl. For performing this initially I have created temporary view table. Then only we can perform SQL operations.
2. Second solution is on pyspark. I have performed these operations on the created dataframe itself.

For joins I have taken the credit data and keeping CustomerID column as common column I have splitted the dataset into 2 dataframes namely df1 and df2. The schema for the same can be viewed in page number 5 [under 4. JOINS division]. Similar to above to perform SQL operations in it I have created views for the 2 dataframes namely table1 and table2. Then performed joins operations in it.



Coding Challenge -*Sivaprakash V*

Import libraries & Initiate session

04:10 PM (<1s) 3

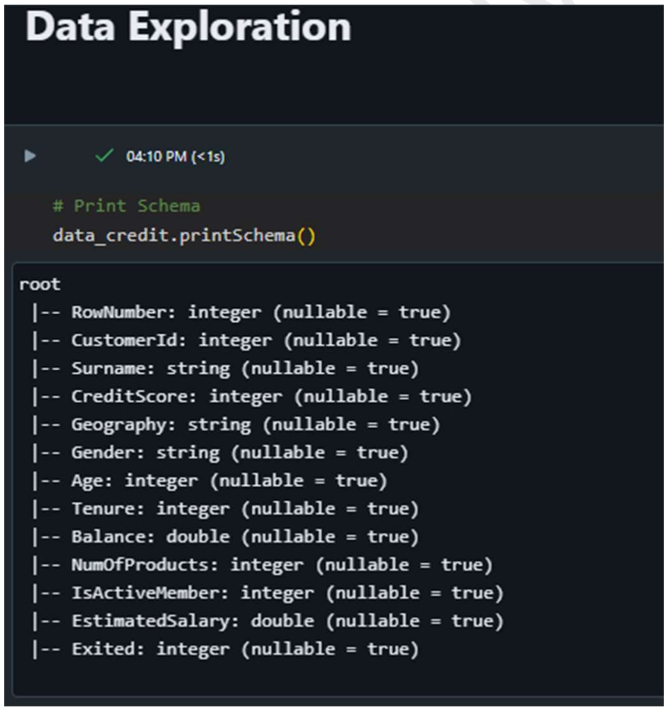
```
# initialize the session
from pyspark import SparkContext
from pyspark.sql import SparkSession

sc = SparkContext.getOrCreate()
spark = SparkSession.builder.appName('Coding challenge').getOrCreate()
```

Loading Data

04:53 PM (2s) 5

```
data_credit =spark.read.csv("/FileStore/tables/creditCard.csv",inferSchema=True,header=True)
df2 = spark.read.csv("/FileStore/tables/bankcredit.csv",inferSchema=True,header=True)
df1 = spark.read.csv("/FileStore/tables/info.csv",inferSchema=True,header=True)
```

```
(6) Spark Jobs
data_credit: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]
df2: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Tenure: integer ... 5 more fields]
df1: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 4 more fields]
```

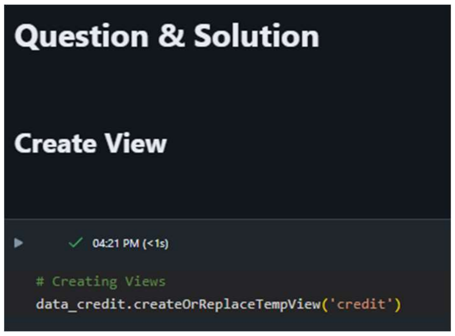


Data Exploration

04:10 PM (<1s)

```
# Print Schema
data_credit.printSchema()
```

```
root
 |-- RowNumber: integer (nullable = true)
 |-- CustomerId: integer (nullable = true)
 |-- Surname: string (nullable = true)
 |-- CreditScore: integer (nullable = true)
 |-- Geography: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Tenure: integer (nullable = true)
 |-- Balance: double (nullable = true)
 |-- NumOfProducts: integer (nullable = true)
 |-- IsActiveMember: integer (nullable = true)
 |-- EstimatedSalary: double (nullable = true)
 |-- Exited: integer (nullable = true)
```

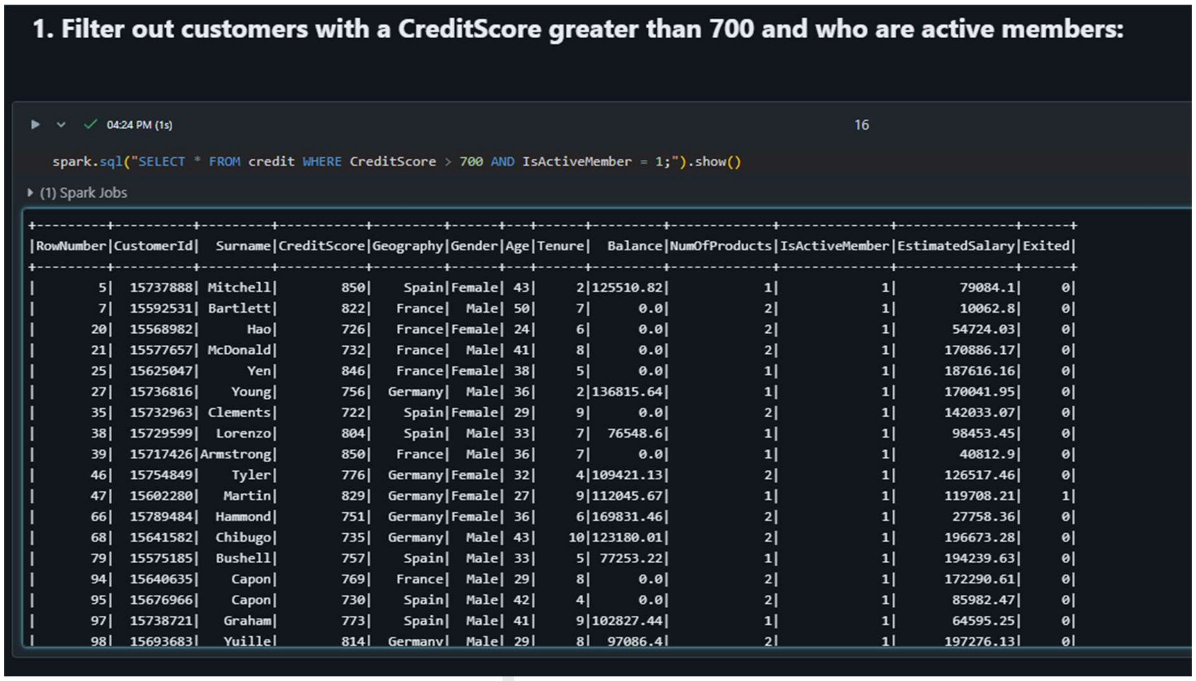


Question & Solution

Create View

04:21 PM (<1s)

```
# Creating Views
data_credit.createOrReplaceTempView('credit')
```



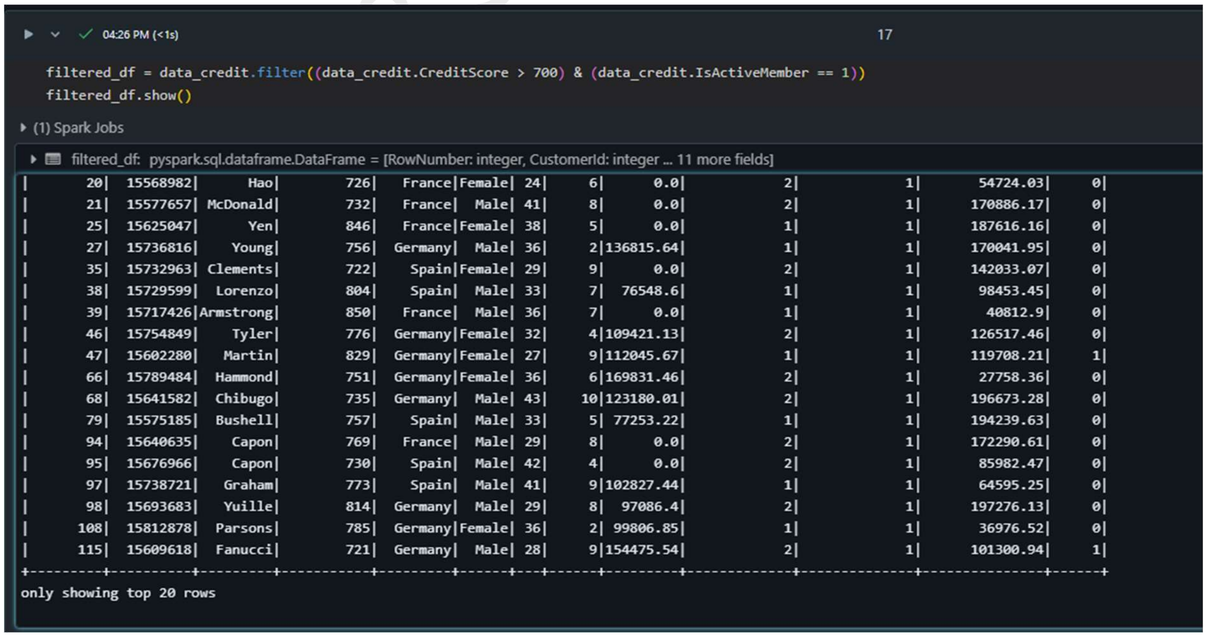
1. Filter out customers with a CreditScore greater than 700 and who are active members:

04:24 PM (1s) 16

```
spark.sql("SELECT * FROM credit WHERE CreditScore > 700 AND IsActiveMember = 1;").show()
```

(1) Spark Jobs

```
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
|        7|  15592531| Bartlett|        822|   France|  Male| 50|     7|      0.0|            2|             1|        10062.8|     0|
|       20|  15568982|      Hao|        726|   France|Female| 24|     6|      0.0|            2|             1|       54724.03|     0|
|       21|  15577657| McDonald|        732|   France|  Male| 41|     8|      0.0|            2|             1|      170886.17|     0|
|       25|  15625047|      Yen|        846|   France|Female| 38|     5|      0.0|            1|             1|      187616.16|     0|
|       27|  15736816|    Young|        756|  Germany|  Male| 36|     2|136815.64|            1|             1|      170041.95|     0|
|       35|  15732963| Clements|        722|    Spain|Female| 29|     9|      0.0|            2|             1|      142033.07|     0|
|       38|  15729599|  Lorenzo|        804|    Spain|  Male| 33|     7|  76548.6|            1|             1|       98453.45|     0|
|       39|  15717426|Armstrong|        850|   France|  Male| 36|     7|      0.0|            1|             1|        40812.9|     0|
|       46|  15754849|    Tyler|        776|  Germany|Female| 32|     4|109421.13|            2|             1|      126517.46|     0|
|       47|  15602280|   Martin|        829|  Germany|Female| 27|     9|112045.67|            1|             1|      119708.21|     1|
|       66|  15789484|  Hammond|        751|  Germany|Female| 36|     6|169831.46|            2|             1|       27758.36|     0|
|       68|  15641582|  Chibugo|        735|  Germany|  Male| 43|    10|123180.01|            2|             1|      196673.28|     0|
|       79|  15575185|  Bushell|        757|    Spain|  Male| 33|     5| 77253.22|            1|             1|      194239.63|     0|
|       94|  15640635|    Capon|        769|   France|  Male| 29|     8|      0.0|            2|             1|      172290.61|     0|
|       95|  15676966|    Capon|        730|    Spain|  Male| 42|     4|      0.0|            2|             1|       85982.47|     0|
|       97|  15738721|   Graham|        773|    Spain|  Male| 41|     9|102827.44|            1|             1|       64595.25|     0|
|       98|  15693683|   Yuille|        814|  Germany|  Male| 29|     8|  97086.4|            2|             1|      197276.13|     0|
```



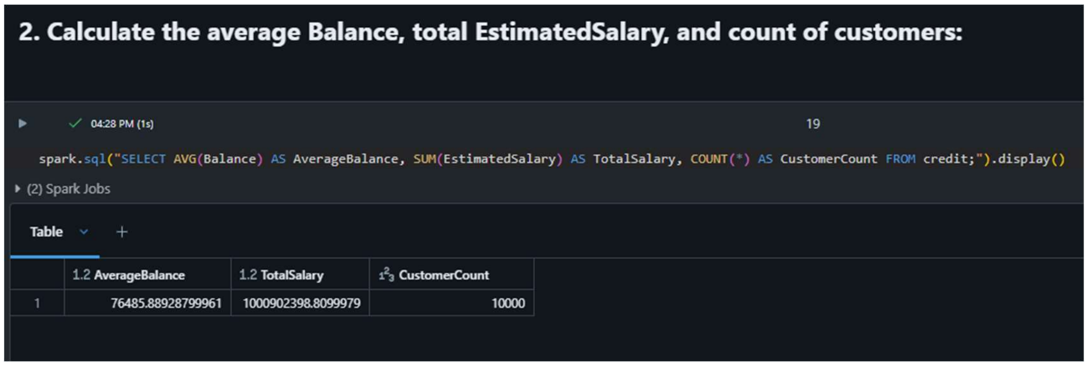
04:26 PM (<1s) 17

```
filtered_df = data_credit.filter((data_credit.CreditScore > 700) & (data_credit.IsActiveMember == 1))
filtered_df.show()
```

(1) Spark Jobs

filtered_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

```
|       20|  15568982|      Hao|        726|   France|Female| 24|     6|      0.0|            2|             1|       54724.03|     0|
|       21|  15577657| McDonald|        732|   France|  Male| 41|     8|      0.0|            2|             1|      170886.17|     0|
|       25|  15625047|      Yen|        846|   France|Female| 38|     5|      0.0|            1|             1|      187616.16|     0|
|       27|  15736816|    Young|        756|  Germany|  Male| 36|     2|136815.64|            1|             1|      170041.95|     0|
|       35|  15732963| Clements|        722|    Spain|Female| 29|     9|      0.0|            2|             1|      142033.07|     0|
|       38|  15729599|  Lorenzo|        804|    Spain|  Male| 33|     7|  76548.6|            1|             1|       98453.45|     0|
|       39|  15717426|Armstrong|        850|   France|  Male| 36|     7|      0.0|            1|             1|        40812.9|     0|
|       46|  15754849|    Tyler|        776|  Germany|Female| 32|     4|109421.13|            2|             1|      126517.46|     0|
|       47|  15602280|   Martin|        829|  Germany|Female| 27|     9|112045.67|            1|             1|      119708.21|     1|
|       66|  15789484|  Hammond|        751|  Germany|Female| 36|     6|169831.46|            2|             1|       27758.36|     0|
|       68|  15641582|  Chibugo|        735|  Germany|  Male| 43|    10|123180.01|            2|             1|      196673.28|     0|
|       79|  15575185|  Bushell|        757|    Spain|  Male| 33|     5| 77253.22|            1|             1|      194239.63|     0|
|       94|  15640635|    Capon|        769|   France|  Male| 29|     8|      0.0|            2|             1|      172290.61|     0|
|       95|  15676966|    Capon|        730|    Spain|  Male| 42|     4|      0.0|            2|             1|       85982.47|     0|
|       97|  15738721|   Graham|        773|    Spain|  Male| 41|     9|102827.44|            1|             1|       64595.25|     0|
|       98|  15693683|   Yuille|        814|  Germany|  Male| 29|     8|  97086.4|            2|             1|      197276.13|     0|
|      108|  15812878|  Parsons|        785|  Germany|Female| 36|     2| 99806.85|            1|             1|       36976.52|     0|
|      115|  15609618|  Fanucci|        721|  Germany|  Male| 28|     9|154475.54|            2|             1|      101300.94|     1|
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
only showing top 20 rows
```



2. Calculate the average Balance, total EstimatedSalary, and count of customers:

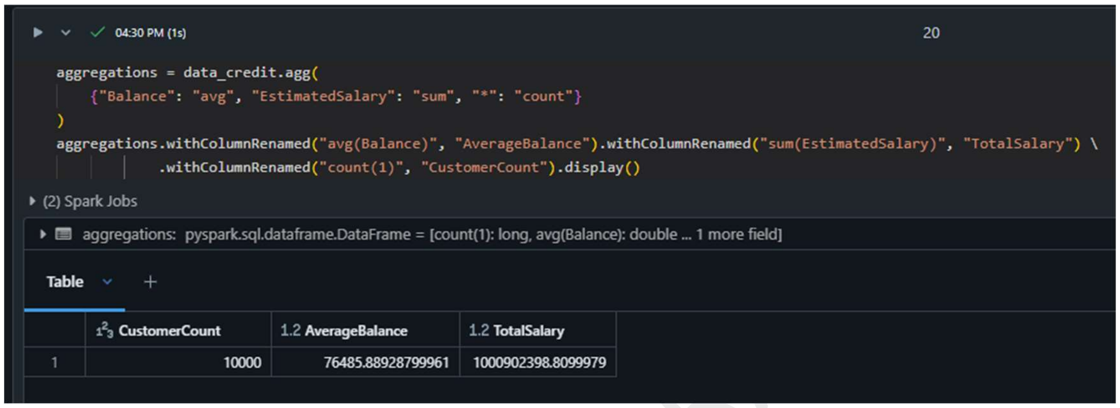
04:28 PM (1s) 19

```
spark.sql("SELECT AVG(Balance) AS AverageBalance, SUM(EstimatedSalary) AS TotalSalary, COUNT(*) AS CustomerCount FROM credit;").display()
```

(2) Spark Jobs

Table

	1.2 AverageBalance	1.2 TotalSalary	CustomerCount
1	76485.88928799961	1000902398.8099979	10000



04:30 PM (1s) 20

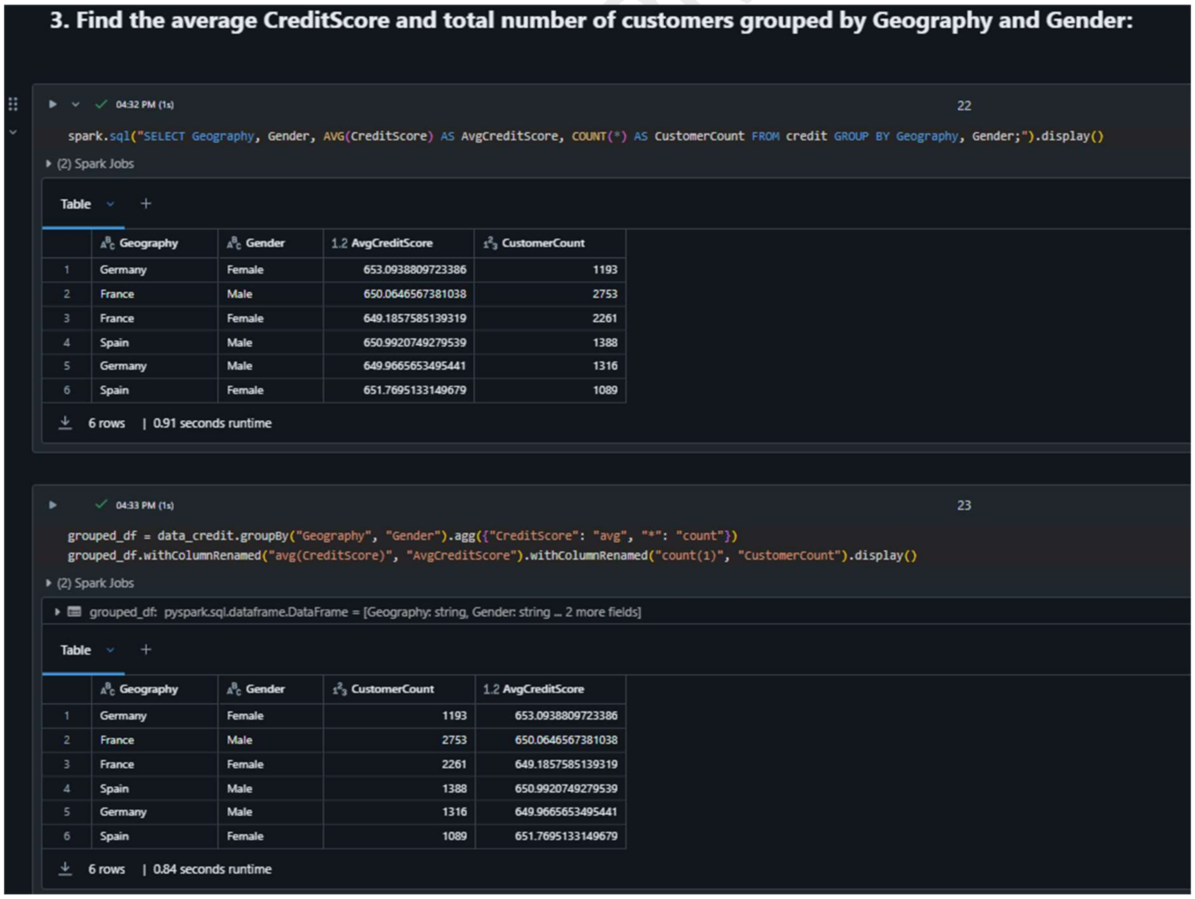
```
aggregations = data_credit.agg(
    {"Balance": "avg", "EstimatedSalary": "sum", "*": "count"}
)
aggregations.withColumnRenamed("avg(Balance)", "AverageBalance").withColumnRenamed("sum(EstimatedSalary)", "TotalSalary") \
            .withColumnRenamed("count(1)", "CustomerCount").display()
```

(2) Spark Jobs

aggregations: pyspark.sql.dataframe.DataFrame = [count(1): long, avg(Balance): double ... 1 more field]

Table

	CustomerCount	1.2 AverageBalance	1.2 TotalSalary
1	10000	76485.88928799961	1000902398.8099979



3. Find the average CreditScore and total number of customers grouped by Geography and Gender:

04:32 PM (1s) 22

```
spark.sql("SELECT Geography, Gender, AVG(CreditScore) AS AvgCreditScore, COUNT(*) AS CustomerCount FROM credit GROUP BY Geography, Gender;").display()
```

(2) Spark Jobs

Table

	Geography	Gender	1.2 AvgCreditScore	CustomerCount
1	Germany	Female	653.0938809723386	1193
2	France	Male	650.0646567381038	2753
3	France	Female	649.1857585139319	2261
4	Spain	Male	650.9920749279539	1388
5	Germany	Male	649.9665653495441	1316
6	Spain	Female	651.7695133149679	1089

6 rows | 0.91 seconds runtime

04:33 PM (1s) 23

```
grouped_df = data_credit.groupBy("Geography", "Gender").agg({"CreditScore": "avg", "*": "count"})
grouped_df.withColumnRenamed("avg(CreditScore)", "AvgCreditScore").withColumnRenamed("count(1)", "CustomerCount").display()
```

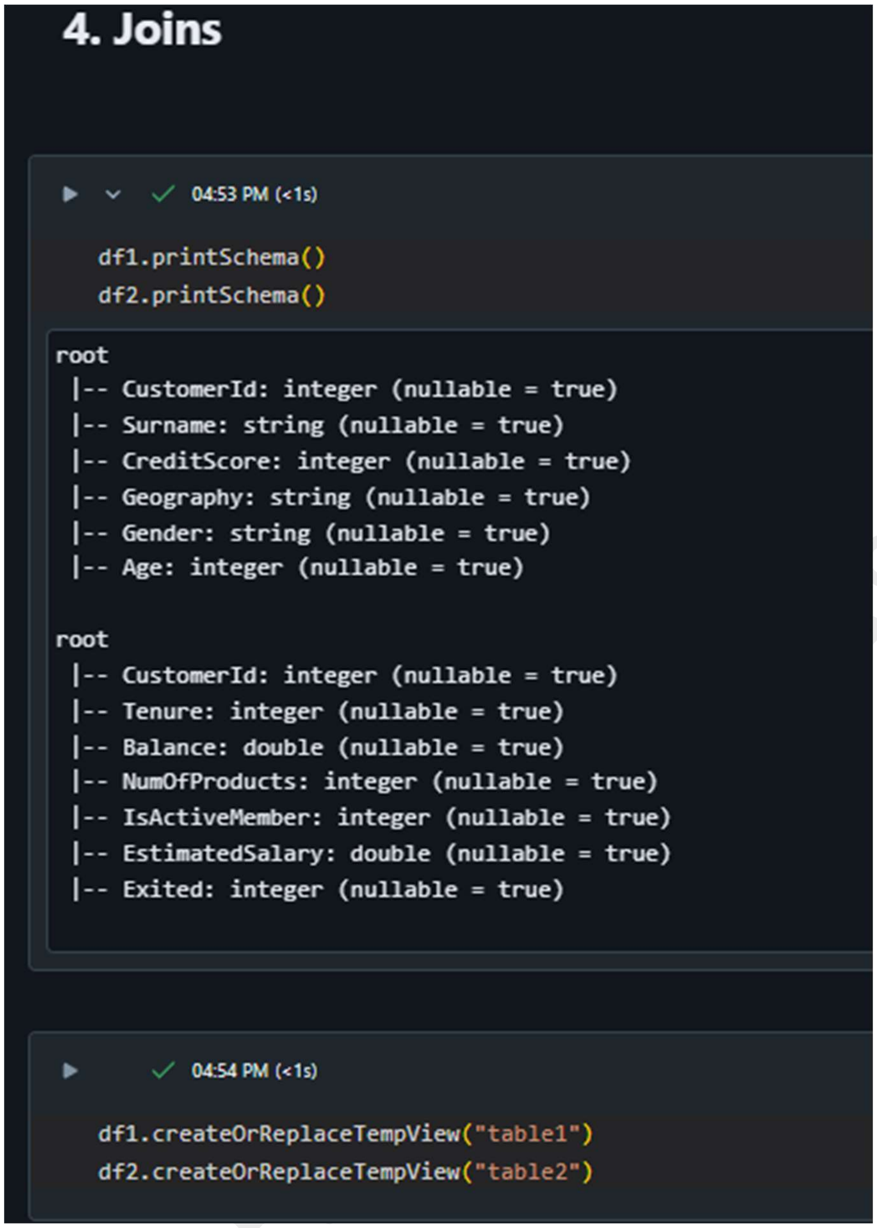
(2) Spark Jobs

grouped_df: pyspark.sql.dataframe.DataFrame = [Geography: string, Gender: string ... 2 more fields]

Table

	Geography	Gender	CustomerCount	1.2 AvgCreditScore
1	Germany	Female	1193	653.0938809723386
2	France	Male	2753	650.0646567381038
3	France	Female	2261	649.1857585139319
4	Spain	Male	1388	650.9920749279539
5	Germany	Male	1316	649.9665653495441
6	Spain	Female	1089	651.7695133149679

6 rows | 0.84 seconds runtime



4. Joins

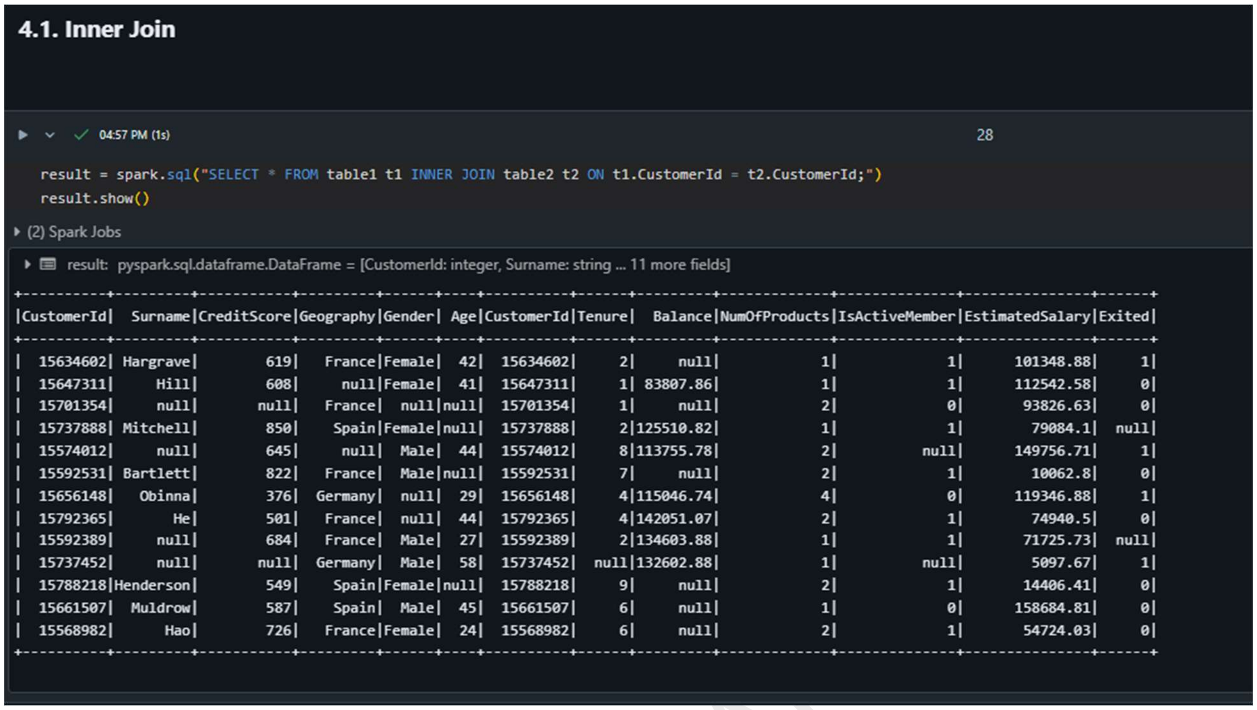
04:53 PM (<1s)

```
df1.printSchema()
df2.printSchema()
```

```
root
 |-- CustomerId: integer (nullable = true)
 |-- Surname: string (nullable = true)
 |-- CreditScore: integer (nullable = true)
 |-- Geography: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Age: integer (nullable = true)

root
 |-- CustomerId: integer (nullable = true)
 |-- Tenure: integer (nullable = true)
 |-- Balance: double (nullable = true)
 |-- NumOfProducts: integer (nullable = true)
 |-- IsActiveMember: integer (nullable = true)
 |-- EstimatedSalary: double (nullable = true)
 |-- Exited: integer (nullable = true)
```

04:54 PM (<1s)

```
df1.createOrReplaceTempView("table1")
df2.createOrReplaceTempView("table2")
```



4.1. Inner Join

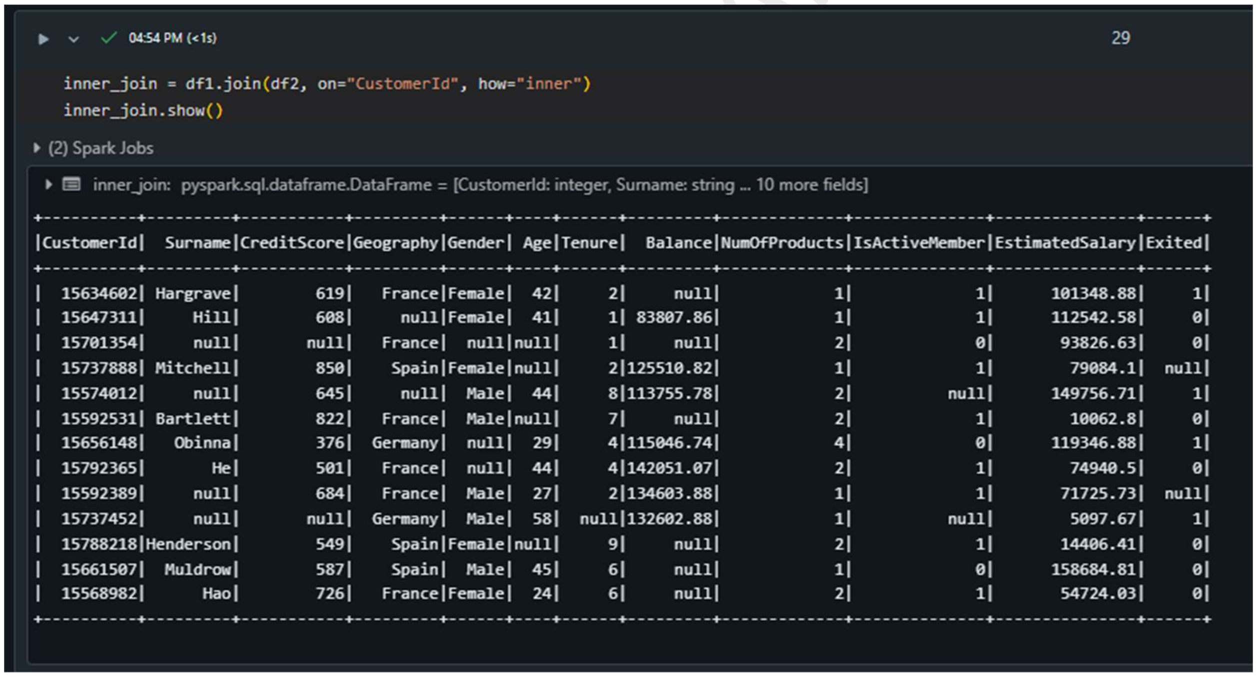
04:57 PM (1s) 28

```
result = spark.sql("SELECT * FROM table1 t1 INNER JOIN table2 t2 ON t1.CustomerId = t2.CustomerId;")
result.show()
```

(2) Spark Jobs

result: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 11 more fields]

```
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|CustomerId|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
|  15634602| Hargrave|        619|   France|Female|  42|  15634602|     2|     null|            1|             1|      101348.88|     1|
|  15647311|     Hill|        608|     null|Female|  41|  15647311|     1| 83807.86|            1|             1|      112542.58|     0|
|  15701354|     null|       null|   France|  null|null|  15701354|     1|     null|            2|             0|       93826.63|     0|
|  15737888| Mitchell|        850|    Spain|Female|null|  15737888|     2|125510.82|            1|             1|        79084.1|  null|
|  15574012|     null|        645|     null|  Male|  44|  15574012|     8|113755.78|            2|          null|      149756.71|     1|
|  15592531| Bartlett|        822|   France|  Male|null|  15592531|     7|     null|            2|             1|        10062.8|     0|
|  15656148|   Obinna|        376|  Germany|  null|  29|  15656148|     4|115046.74|            4|             0|      119346.88|     1|
|  15792365|       He|        501|   France|  null|  44|  15792365|     4|142051.07|            2|             1|        74940.5|     0|
|  15592389|     null|        684|   France|  Male|  27|  15592389|     2|134603.88|            1|             1|       71725.73|  null|
|  15737452|     null|       null|  Germany|  Male|  58|  15737452|  null|132602.88|            1|          null|        5097.67|     1|
|  15788218|Henderson|        549|    Spain|Female|null|  15788218|     9|     null|            2|             1|       14406.41|     0|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|  15661507|     6|     null|            1|             0|      158684.81|     0|
|  15568982|      Hao|        726|   France|Female|  24|  15568982|     6|     null|            2|             1|       54724.03|     0|
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
```



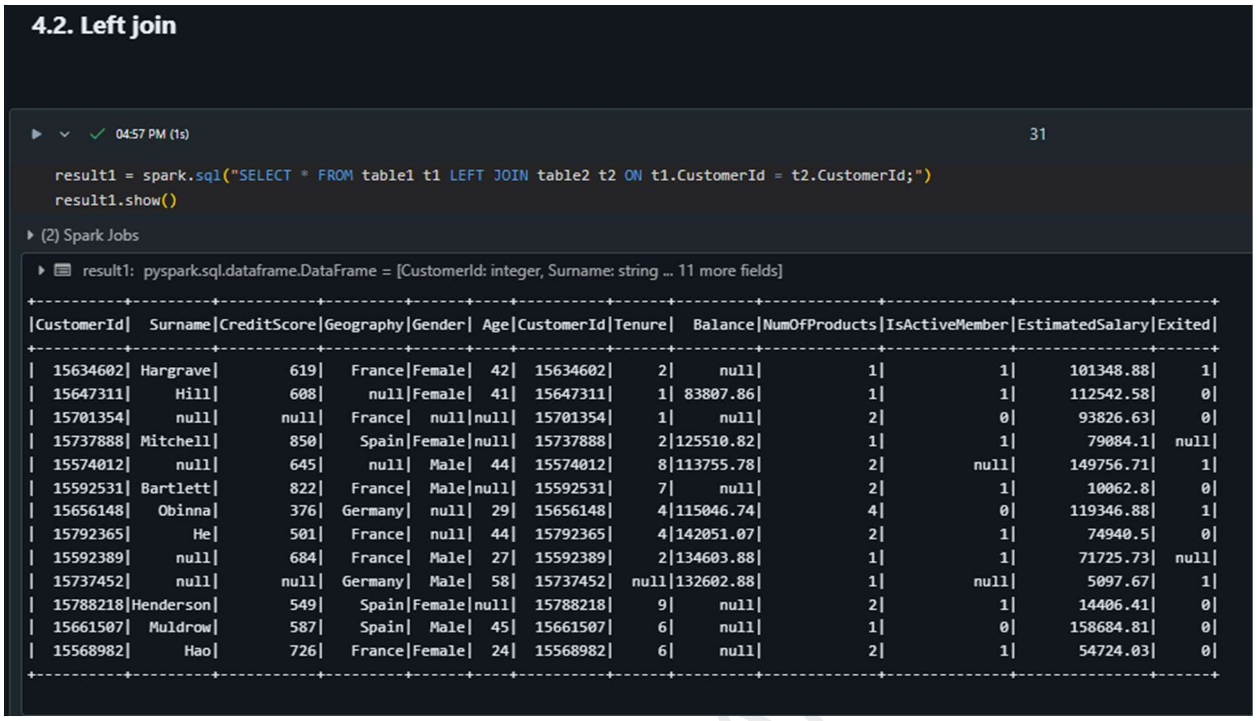
04:54 PM (<1s) 29

```
inner_join = df1.join(df2, on="CustomerId", how="inner")
inner_join.show()
```

(2) Spark Jobs

inner_join: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 10 more fields]

```
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|  15634602| Hargrave|        619|   France|Female|  42|     2|     null|            1|             1|      101348.88|     1|
|  15647311|     Hill|        608|     null|Female|  41|     1| 83807.86|            1|             1|      112542.58|     0|
|  15701354|     null|       null|   France|  null|null|     1|     null|            2|             0|       93826.63|     0|
|  15737888| Mitchell|        850|    Spain|Female|null|     2|125510.82|            1|             1|        79084.1|  null|
|  15574012|     null|        645|     null|  Male|  44|     8|113755.78|            2|          null|      149756.71|     1|
|  15592531| Bartlett|        822|   France|  Male|null|     7|     null|            2|             1|        10062.8|     0|
|  15656148|   Obinna|        376|  Germany|  null|  29|     4|115046.74|            4|             0|      119346.88|     1|
|  15792365|       He|        501|   France|  null|  44|     4|142051.07|            2|             1|        74940.5|     0|
|  15592389|     null|        684|   France|  Male|  27|     2|134603.88|            1|             1|       71725.73|  null|
|  15737452|     null|       null|  Germany|  Male|  58|  null|132602.88|            1|          null|        5097.67|     1|
|  15788218|Henderson|        549|    Spain|Female|null|     9|     null|            2|             1|       14406.41|     0|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|     6|     null|            1|             0|      158684.81|     0|
|  15568982|      Hao|        726|   France|Female|  24|     6|     null|            2|             1|       54724.03|     0|
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
```



4.2. Left join

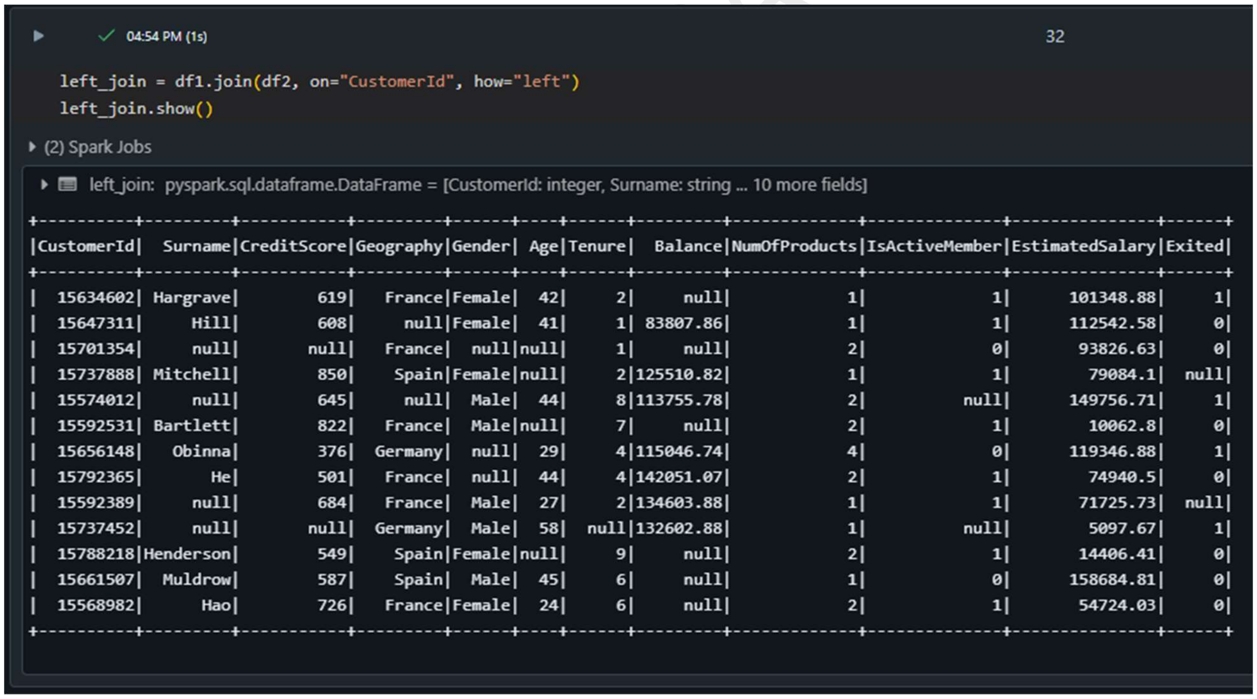
04:57 PM (1s) 31

```
result1 = spark.sql("SELECT * FROM table1 t1 LEFT JOIN table2 t2 ON t1.CustomerId = t2.CustomerId;")
result1.show()
```

(2) Spark Jobs

result1: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 11 more fields]

```
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|CustomerId|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
|  15634602| Hargrave|        619|   France|Female|  42|  15634602|     2|     null|            1|             1|      101348.88|     1|
|  15647311|     Hill|        608|     null|Female|  41|  15647311|     1| 83807.86|            1|             1|      112542.58|     0|
|  15701354|     null|       null|   France|  null|null|  15701354|     1|     null|            2|             0|       93826.63|     0|
|  15737888| Mitchell|        850|    Spain|Female|null|  15737888|     2|125510.82|            1|             1|        79084.1|  null|
|  15574012|     null|        645|     null|  Male|  44|  15574012|     8|113755.78|            2|          null|      149756.71|     1|
|  15592531| Bartlett|        822|   France|  Male|null|  15592531|     7|     null|            2|             1|        10062.8|     0|
|  15656148|   Obinna|        376|  Germany|  null|  29|  15656148|     4|115046.74|            4|             0|      119346.88|     1|
|  15792365|       He|        501|   France|  null|  44|  15792365|     4|142051.07|            2|             1|        74940.5|     0|
|  15592389|     null|        684|   France|  Male|  27|  15592389|     2|134603.88|            1|             1|       71725.73|  null|
|  15737452|     null|       null|  Germany|  Male|  58|  15737452|  null|132602.88|            1|          null|        5097.67|     1|
|  15788218|Henderson|        549|    Spain|Female|null|  15788218|     9|     null|            2|             1|       14406.41|     0|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|  15661507|     6|     null|            1|             0|      158684.81|     0|
|  15568982|      Hao|        726|   France|Female|  24|  15568982|     6|     null|            2|             1|       54724.03|     0|
+----------+---------+-----------+---------+------+----+----------+------+---------+-------------+--------------+---------------+------+
```



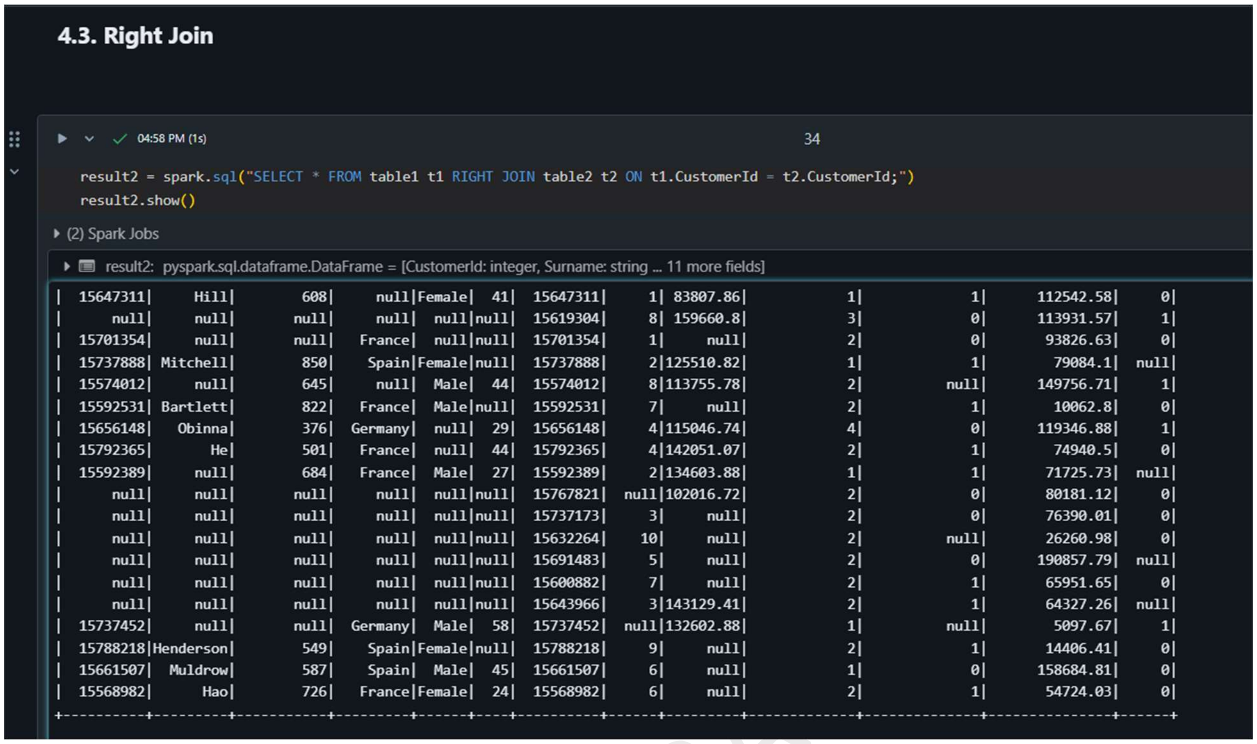
04:54 PM (1s) 32

```
left_join = df1.join(df2, on="CustomerId", how="left")
left_join.show()
```

(2) Spark Jobs

left_join: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 10 more fields]

```
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|  15634602| Hargrave|        619|   France|Female|  42|     2|     null|            1|             1|      101348.88|     1|
|  15647311|     Hill|        608|     null|Female|  41|     1| 83807.86|            1|             1|      112542.58|     0|
|  15701354|     null|       null|   France|  null|null|     1|     null|            2|             0|       93826.63|     0|
|  15737888| Mitchell|        850|    Spain|Female|null|     2|125510.82|            1|             1|        79084.1|  null|
|  15574012|     null|        645|     null|  Male|  44|     8|113755.78|            2|          null|      149756.71|     1|
|  15592531| Bartlett|        822|   France|  Male|null|     7|     null|            2|             1|        10062.8|     0|
|  15656148|   Obinna|        376|  Germany|  null|  29|     4|115046.74|            4|             0|      119346.88|     1|
|  15792365|       He|        501|   France|  null|  44|     4|142051.07|            2|             1|        74940.5|     0|
|  15592389|     null|        684|   France|  Male|  27|     2|134603.88|            1|             1|       71725.73|  null|
|  15737452|     null|       null|  Germany|  Male|  58|  null|132602.88|            1|          null|        5097.67|     1|
|  15788218|Henderson|        549|    Spain|Female|null|     9|     null|            2|             1|       14406.41|     0|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|     6|     null|            1|             0|      158684.81|     0|
|  15568982|      Hao|        726|   France|Female|  24|     6|     null|            2|             1|       54724.03|     0|
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
```



4.3. Right Join

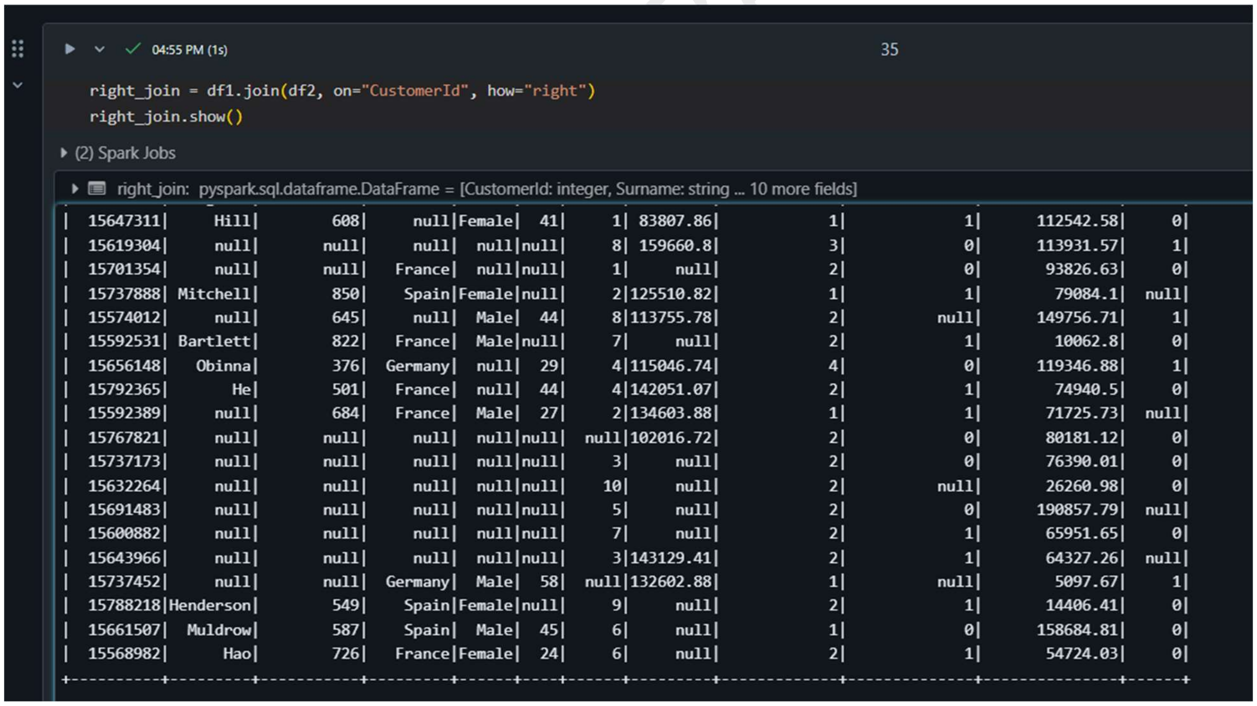
04:58 PM (1s) 34

```
result2 = spark.sql("SELECT * FROM table1 t1 RIGHT JOIN table2 t2 ON t1.CustomerId = t2.CustomerId;")
result2.show()
```

(2) Spark Jobs

result2: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 11 more fields]

```
| 15647311|     Hill|       608|     null|Female|  41| 15647311|    1| 83807.86|            1|            1|     112542.58|     0|
|     null|     null|      null|     null|  null|null| 15619304|    8| 159660.8|            3|            0|     113931.57|     1|
| 15701354|     null|      null|   France|  null|null| 15701354|    1|     null|            2|            0|      93826.63|     0|
| 15737888| Mitchell|       850|    Spain|Female|null| 15737888|    2|125510.82|            1|            1|       79084.1|  null|
| 15574012|     null|       645|     null|  Male|  44| 15574012|    8|113755.78|            2|         null|     149756.71|     1|
| 15592531| Bartlett|       822|   France|  Male|null| 15592531|    7|     null|            2|            1|       10062.8|     0|
| 15656148|   Obinna|       376|  Germany|  null|  29| 15656148|    4|115046.74|            4|            0|     119346.88|     1|
| 15792365|       He|       501|   France|  null|  44| 15792365|    4|142051.07|            2|            1|       74940.5|     0|
| 15592389|     null|       684|   France|  Male|  27| 15592389|    2|134603.88|            1|            1|      71725.73|  null|
|     null|     null|      null|     null|  null|null| 15767821| null|102016.72|            2|            0|      80181.12|     0|
|     null|     null|      null|     null|  null|null| 15737173|    3|     null|            2|            0|      76390.01|     0|
|     null|     null|      null|     null|  null|null| 15632264|   10|     null|            2|         null|      26260.98|     0|
|     null|     null|      null|     null|  null|null| 15691483|    5|     null|            2|            0|     190857.79|  null|
|     null|     null|      null|     null|  null|null| 15600882|    7|     null|            2|            1|      65951.65|     0|
|     null|     null|      null|     null|  null|null| 15643966|    3|143129.41|            2|            1|      64327.26|  null|
| 15737452|     null|      null|  Germany|  Male|  58| 15737452| null|132602.88|            1|         null|       5097.67|     1|
| 15788218|Henderson|       549|    Spain|Female|null| 15788218|    9|     null|            2|            1|      14406.41|     0|
| 15661507|  Muldrow|       587|    Spain|  Male|  45| 15661507|    6|     null|            1|            0|     158684.81|     0|
| 15568982|      Hao|       726|   France|Female|  24| 15568982|    6|     null|            2|            1|      54724.03|     0|
+---------+---------+----------+---------+------+----+---------+-----+---------+-------------+-------------+--------------+------+
```



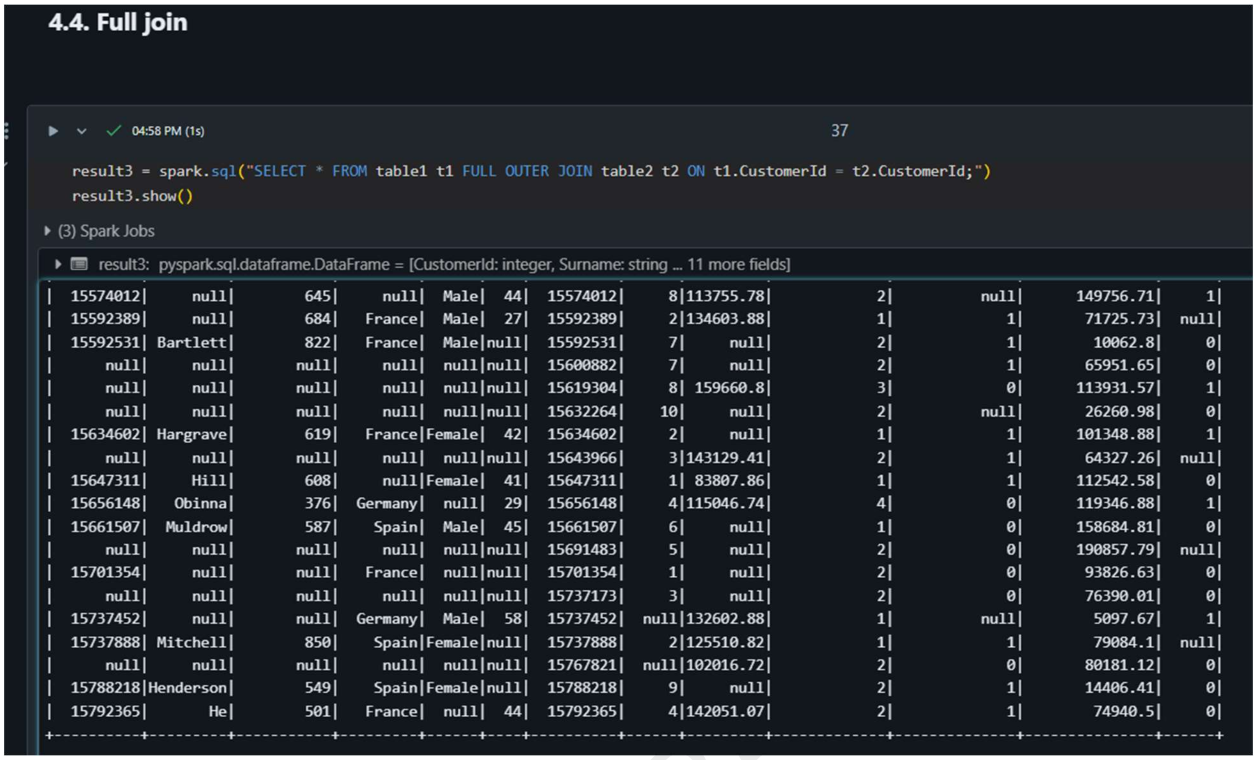
04:55 PM (1s) 35

```
right_join = df1.join(df2, on="CustomerId", how="right")
right_join.show()
```

(2) Spark Jobs

right_join: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 10 more fields]

```
| 15647311|     Hill|       608|     null|Female|  41|    1| 83807.86|            1|            1|     112542.58|     0|
| 15619304|     null|      null|     null|  null|null|    8| 159660.8|            3|            0|     113931.57|     1|
| 15701354|     null|      null|   France|  null|null|    1|     null|            2|            0|      93826.63|     0|
| 15737888| Mitchell|       850|    Spain|Female|null|    2|125510.82|            1|            1|       79084.1|  null|
| 15574012|     null|       645|     null|  Male|  44|    8|113755.78|            2|         null|     149756.71|     1|
| 15592531| Bartlett|       822|   France|  Male|null|    7|     null|            2|            1|       10062.8|     0|
| 15656148|   Obinna|       376|  Germany|  null|  29|    4|115046.74|            4|            0|     119346.88|     1|
| 15792365|       He|       501|   France|  null|  44|    4|142051.07|            2|            1|       74940.5|     0|
| 15592389|     null|       684|   France|  Male|  27|    2|134603.88|            1|            1|      71725.73|  null|
| 15767821|     null|      null|     null|  null|null| null|102016.72|            2|            0|      80181.12|     0|
| 15737173|     null|      null|     null|  null|null|    3|     null|            2|            0|      76390.01|     0|
| 15632264|     null|      null|     null|  null|null|   10|     null|            2|         null|      26260.98|     0|
| 15691483|     null|      null|     null|  null|null|    5|     null|            2|            0|     190857.79|  null|
| 15600882|     null|      null|     null|  null|null|    7|     null|            2|            1|      65951.65|     0|
| 15643966|     null|      null|     null|  null|null|    3|143129.41|            2|            1|      64327.26|  null|
| 15737452|     null|      null|  Germany|  Male|  58| null|132602.88|            1|         null|       5097.67|     1|
| 15788218|Henderson|       549|    Spain|Female|null|    9|     null|            2|            1|      14406.41|     0|
| 15661507|  Muldrow|       587|    Spain|  Male|  45|    6|     null|            1|            0|     158684.81|     0|
| 15568982|      Hao|       726|   France|Female|  24|    6|     null|            2|            1|      54724.03|     0|
+---------+---------+----------+---------+------+----+-----+---------+-------------+-------------+--------------+------+
```



4.4. Full join

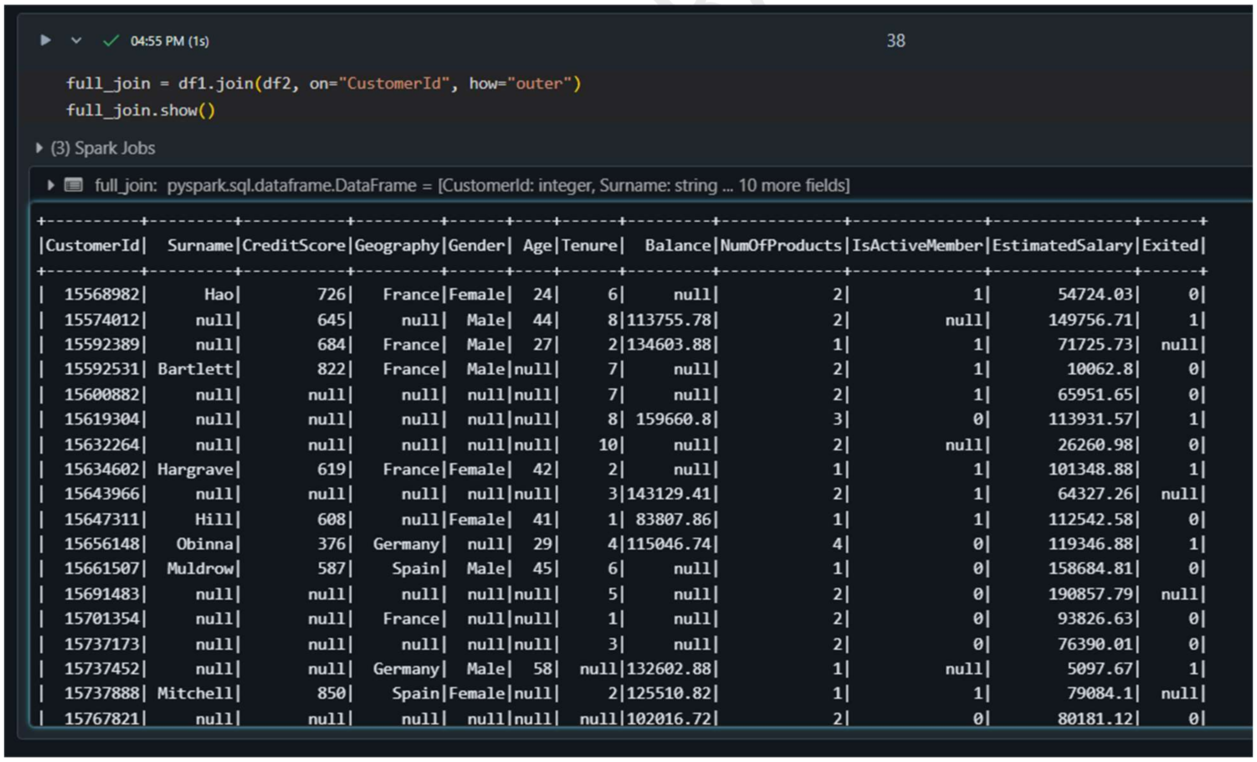
04:58 PM (1s) 37

```
result3 = spark.sql("SELECT * FROM table1 t1 FULL OUTER JOIN table2 t2 ON t1.CustomerId = t2.CustomerId;")
result3.show()
```

(3) Spark Jobs

result3: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 11 more fields]

```
| 15574012|     null|      645|     null|  Male|  44| 15574012|     8|113755.78|            2|           null|      149756.71|     1|
| 15592389|     null|      684|   France|  Male|  27| 15592389|     2|134603.88|            1|              1|       71725.73|  null|
| 15592531| Bartlett|      822|   France|  Male|null| 15592531|     7|     null|            2|              1|        10062.8|     0|
|     null|     null|     null|     null|  null|null| 15600882|     7|     null|            2|              1|       65951.65|     0|
|     null|     null|     null|     null|  null|null| 15619304|     8| 159660.8|            3|              0|      113931.57|     1|
|     null|     null|     null|     null|  null|null| 15632264|    10|     null|            2|           null|       26260.98|     0|
| 15634602| Hargrave|      619|   France|Female|  42| 15634602|     2|     null|            1|              1|      101348.88|     1|
|     null|     null|     null|     null|  null|null| 15643966|     3|143129.41|            2|              1|       64327.26|  null|
| 15647311|     Hill|      608|     null|Female|  41| 15647311|     1| 83807.86|            1|              1|      112542.58|     0|
| 15656148|   Obinna|      376|  Germany|  null|  29| 15656148|     4|115046.74|            4|              0|      119346.88|     1|
| 15661507|  Muldrow|      587|    Spain|  Male|  45| 15661507|     6|     null|            1|              0|      158684.81|     0|
|     null|     null|     null|     null|  null|null| 15691483|     5|     null|            2|              0|      190857.79|  null|
| 15701354|     null|     null|   France|  null|null| 15701354|     1|     null|            2|              0|       93826.63|     0|
|     null|     null|     null|     null|  null|null| 15737173|     3|     null|            2|              0|       76390.01|     0|
| 15737452|     null|     null|  Germany|  Male|  58| 15737452|  null|132602.88|            1|           null|        5097.67|     1|
| 15737888| Mitchell|      850|    Spain|Female|null| 15737888|     2|125510.82|            1|              1|        79084.1|  null|
|     null|     null|     null|     null|  null|null| 15767821|  null|102016.72|            2|              0|       80181.12|     0|
| 15788218|Henderson|      549|    Spain|Female|null| 15788218|     9|     null|            2|              1|       14406.41|     0|
| 15792365|       He|      501|   France|  null|  44| 15792365|     4|142051.07|            2|              1|        74940.5|     0|
+---------+---------+---------+---------+------+----+---------+------+---------+-------------+---------------+---------------+------+
```



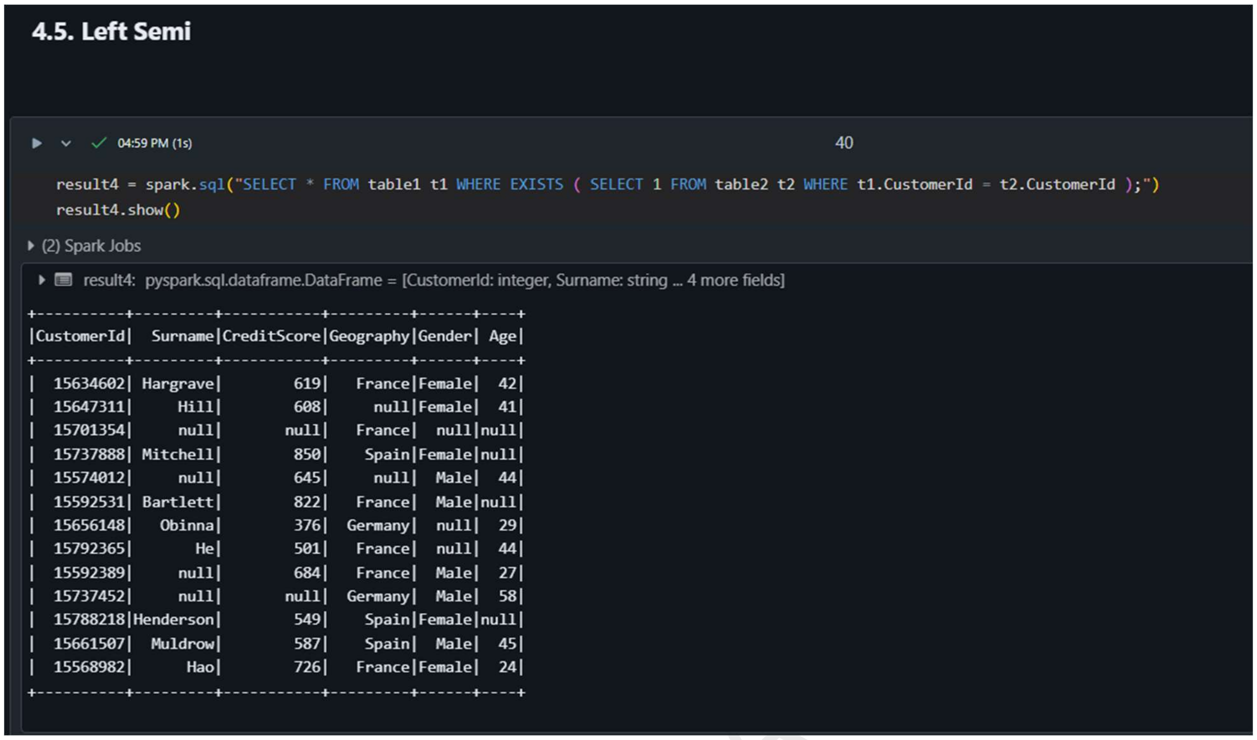
04:55 PM (1s) 38

```
full_join = df1.join(df2, on="CustomerId", how="outer")
full_join.show()
```

(3) Spark Jobs

full_join: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 10 more fields]

```
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+----------+---------+-----------+---------+------+----+------+---------+-------------+--------------+---------------+------+
|  15568982|      Hao|        726|   France|Female|  24|     6|     null|            2|             1|       54724.03|     0|
|  15574012|     null|        645|     null|  Male|  44|     8|113755.78|            2|          null|      149756.71|     1|
|  15592389|     null|        684|   France|  Male|  27|     2|134603.88|            1|             1|       71725.73|  null|
|  15592531| Bartlett|        822|   France|  Male|null|     7|     null|            2|             1|        10062.8|     0|
|  15600882|     null|       null|     null|  null|null|     7|     null|            2|             1|       65951.65|     0|
|  15619304|     null|       null|     null|  null|null|     8| 159660.8|            3|             0|      113931.57|     1|
|  15632264|     null|       null|     null|  null|null|    10|     null|            2|          null|       26260.98|     0|
|  15634602| Hargrave|        619|   France|Female|  42|     2|     null|            1|             1|      101348.88|     1|
|  15643966|     null|       null|     null|  null|null|     3|143129.41|            2|             1|       64327.26|  null|
|  15647311|     Hill|        608|     null|Female|  41|     1| 83807.86|            1|             1|      112542.58|     0|
|  15656148|   Obinna|        376|  Germany|  null|  29|     4|115046.74|            4|             0|      119346.88|     1|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|     6|     null|            1|             0|      158684.81|     0|
|  15691483|     null|       null|     null|  null|null|     5|     null|            2|             0|      190857.79|  null|
|  15701354|     null|       null|   France|  null|null|     1|     null|            2|             0|       93826.63|     0|
|  15737173|     null|       null|     null|  null|null|     3|     null|            2|             0|       76390.01|     0|
|  15737452|     null|       null|  Germany|  Male|  58|  null|132602.88|            1|          null|        5097.67|     1|
|  15737888| Mitchell|        850|    Spain|Female|null|     2|125510.82|            1|             1|        79084.1|  null|
|  15767821|     null|       null|     null|  null|null|  null|102016.72|            2|             0|       80181.12|     0|
```



4.5. Left Semi

04:59 PM (1s) 40

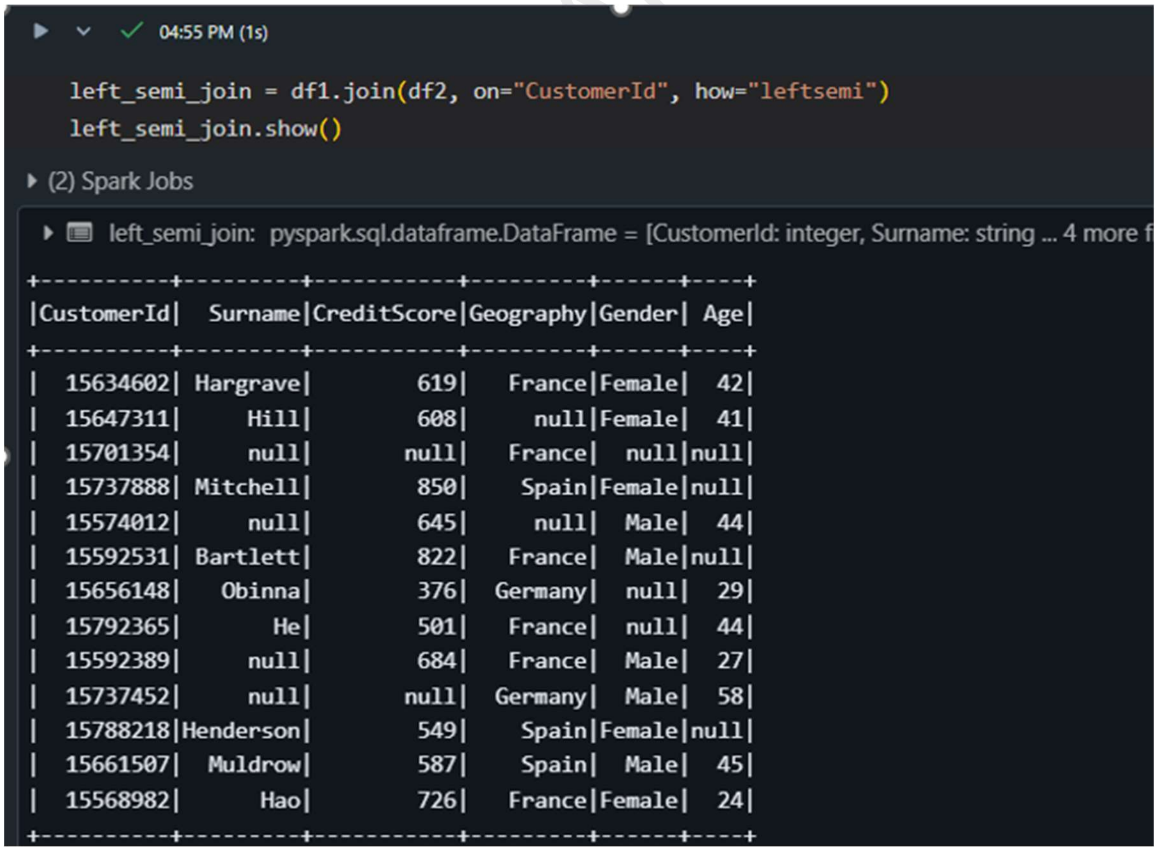
```
result4 = spark.sql("SELECT * FROM table1 t1 WHERE EXISTS ( SELECT 1 FROM table2 t2 WHERE t1.CustomerId = t2.CustomerId );")
result4.show()
```

(2) Spark Jobs

result4: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 4 more fields]

```
+----------+---------+-----------+---------+------+----+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|
+----------+---------+-----------+---------+------+----+
|  15634602| Hargrave|        619|   France|Female|  42|
|  15647311|     Hill|        608|     null|Female|  41|
|  15701354|     null|       null|   France|  null|null|
|  15737888| Mitchell|        850|    Spain|Female|null|
|  15574012|     null|        645|     null|  Male|  44|
|  15592531| Bartlett|        822|   France|  Male|null|
|  15656148|   Obinna|        376|  Germany|  null|  29|
|  15792365|       He|        501|   France|  null|  44|
|  15592389|     null|        684|   France|  Male|  27|
|  15737452|     null|       null|  Germany|  Male|  58|
|  15788218|Henderson|        549|    Spain|Female|null|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|
|  15568982|      Hao|        726|   France|Female|  24|
+----------+---------+-----------+---------+------+----+
```

Left Semi join is not available in SQL but can be performed by using the sub queries concept.



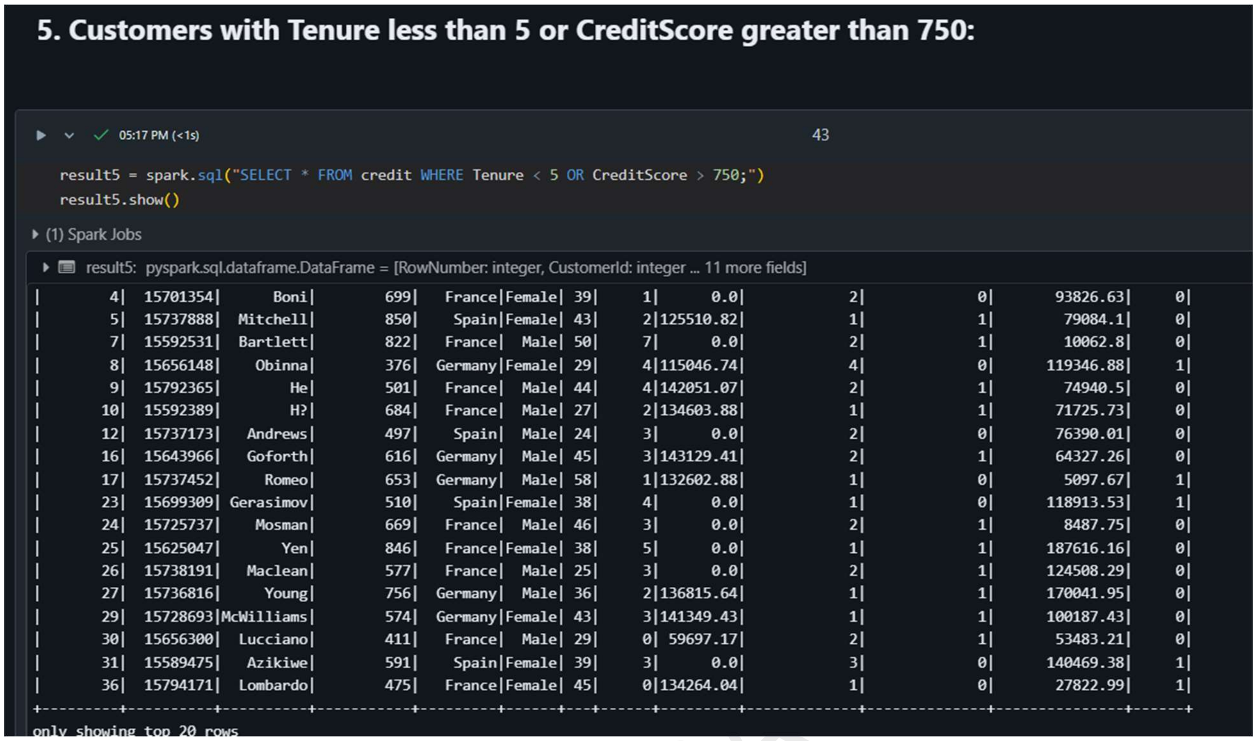
04:55 PM (1s)

```
left_semi_join = df1.join(df2, on="CustomerId", how="leftsemi")
left_semi_join.show()
```

(2) Spark Jobs

left_semi_join: pyspark.sql.dataframe.DataFrame = [CustomerId: integer, Surname: string ... 4 more f

```
+----------+---------+-----------+---------+------+----+
|CustomerId|  Surname|CreditScore|Geography|Gender| Age|
+----------+---------+-----------+---------+------+----+
|  15634602| Hargrave|        619|   France|Female|  42|
|  15647311|     Hill|        608|     null|Female|  41|
|  15701354|     null|       null|   France|  null|null|
|  15737888| Mitchell|        850|    Spain|Female|null|
|  15574012|     null|        645|     null|  Male|  44|
|  15592531| Bartlett|        822|   France|  Male|null|
|  15656148|   Obinna|        376|  Germany|  null|  29|
|  15792365|       He|        501|   France|  null|  44|
|  15592389|     null|        684|   France|  Male|  27|
|  15737452|     null|       null|  Germany|  Male|  58|
|  15788218|Henderson|        549|    Spain|Female|null|
|  15661507|  Muldrow|        587|    Spain|  Male|  45|
|  15568982|      Hao|        726|   France|Female|  24|
+----------+---------+-----------+---------+------+----+
```



5. Customers with Tenure less than 5 or CreditScore greater than 750:

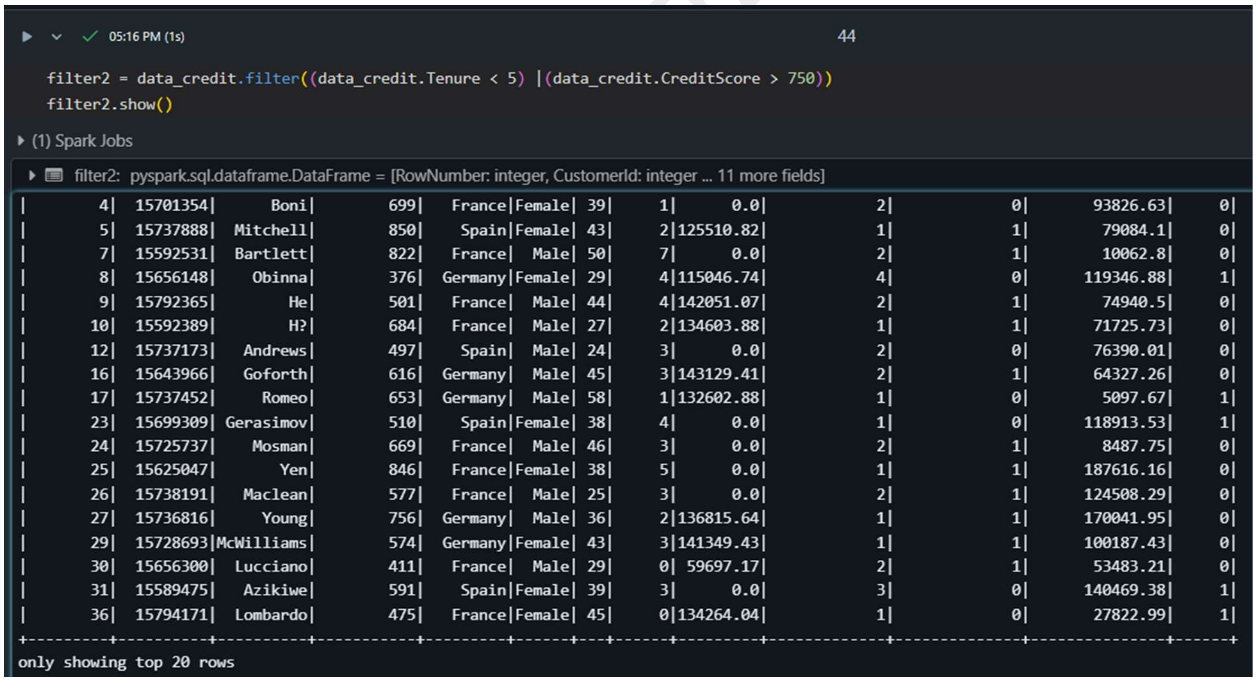
05:17 PM (<1s) 43

```
result5 = spark.sql("SELECT * FROM credit WHERE Tenure < 5 OR CreditScore > 750;")
result5.show()
```

(1) Spark Jobs

result5: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

```
|        4|  15701354|      Boni|       699|    France|Female| 39|     1|       0.0|            2|             0|      93826.63|     0|
|        5|  15737888|  Mitchell|       850|     Spain|Female| 43|     2| 125510.82|            1|             1|       79084.1|     0|
|        7|  15592531|  Bartlett|       822|    France|  Male| 50|     7|       0.0|            2|             1|       10062.8|     0|
|        8|  15656148|    Obinna|       376|   Germany|Female| 29|     4| 115046.74|            4|             0|     119346.88|     1|
|        9|  15792365|        He|       501|    France|  Male| 44|     4| 142051.07|            2|             1|       74940.5|     0|
|       10|  15592389|        H?|       684|    France|  Male| 27|     2| 134603.88|            1|             1|      71725.73|     0|
|       12|  15737173|   Andrews|       497|     Spain|  Male| 24|     3|       0.0|            2|             0|      76390.01|     0|
|       16|  15643966|   Goforth|       616|   Germany|  Male| 45|     3| 143129.41|            2|             1|      64327.26|     0|
|       17|  15737452|     Romeo|       653|   Germany|  Male| 58|     1| 132602.88|            1|             0|       5097.67|     1|
|       23|  15699309|  Gerasimov|      510|     Spain|Female| 38|     4|       0.0|            1|             0|     118913.53|     1|
|       24|  15725737|    Mosman|       669|    France|  Male| 46|     3|       0.0|            2|             1|       8487.75|     0|
|       25|  15625047|       Yen|       846|    France|Female| 38|     5|       0.0|            1|             1|     187616.16|     0|
|       26|  15738191|   Maclean|       577|    France|  Male| 25|     3|       0.0|            2|             1|     124508.29|     0|
|       27|  15736816|     Young|       756|   Germany|  Male| 36|     2| 136815.64|            1|             1|     170041.95|     0|
|       29|  15728693|McWilliams|       574|   Germany|Female| 43|     3| 141349.43|            1|             1|     100187.43|     0|
|       30|  15656300|  Lucciano|       411|    France|  Male| 29|     0|  59697.17|            2|             1|      53483.21|     0|
|       31|  15589475|   Azikiwe|       591|     Spain|Female| 39|     3|       0.0|            3|             0|     140469.38|     1|
|       36|  15794171|  Lombardo|       475|    France|Female| 45|     0| 134264.04|            1|             0|      27822.99|     1|
+---------+----------+----------+----------+----------+------+---+------+----------+-------------+--------------+--------------+------+
only showing top 20 rows
```



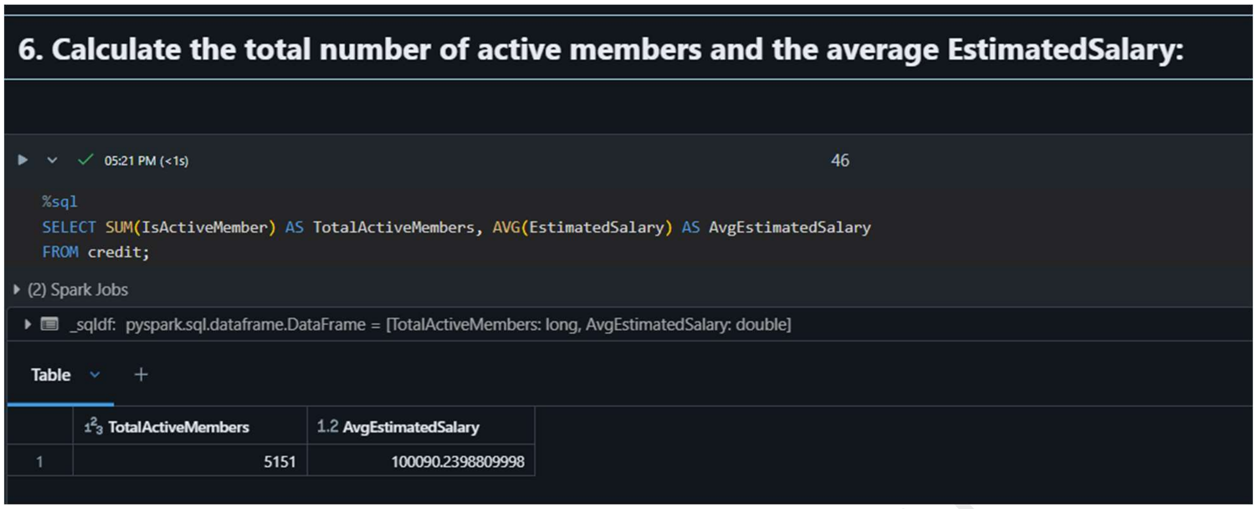
05:16 PM (1s) 44

```
filter2 = data_credit.filter((data_credit.Tenure < 5) |(data_credit.CreditScore > 750))
filter2.show()
```

(1) Spark Jobs

filter2: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

```
|        4|  15701354|      Boni|       699|    France|Female| 39|     1|       0.0|            2|             0|      93826.63|     0|
|        5|  15737888|  Mitchell|       850|     Spain|Female| 43|     2| 125510.82|            1|             1|       79084.1|     0|
|        7|  15592531|  Bartlett|       822|    France|  Male| 50|     7|       0.0|            2|             1|       10062.8|     0|
|        8|  15656148|    Obinna|       376|   Germany|Female| 29|     4| 115046.74|            4|             0|     119346.88|     1|
|        9|  15792365|        He|       501|    France|  Male| 44|     4| 142051.07|            2|             1|       74940.5|     0|
|       10|  15592389|        H?|       684|    France|  Male| 27|     2| 134603.88|            1|             1|      71725.73|     0|
|       12|  15737173|   Andrews|       497|     Spain|  Male| 24|     3|       0.0|            2|             0|      76390.01|     0|
|       16|  15643966|   Goforth|       616|   Germany|  Male| 45|     3| 143129.41|            2|             1|      64327.26|     0|
|       17|  15737452|     Romeo|       653|   Germany|  Male| 58|     1| 132602.88|            1|             0|       5097.67|     1|
|       23|  15699309|  Gerasimov|      510|     Spain|Female| 38|     4|       0.0|            1|             0|     118913.53|     1|
|       24|  15725737|    Mosman|       669|    France|  Male| 46|     3|       0.0|            2|             1|       8487.75|     0|
|       25|  15625047|       Yen|       846|    France|Female| 38|     5|       0.0|            1|             1|     187616.16|     0|
|       26|  15738191|   Maclean|       577|    France|  Male| 25|     3|       0.0|            2|             1|     124508.29|     0|
|       27|  15736816|     Young|       756|   Germany|  Male| 36|     2| 136815.64|            1|             1|     170041.95|     0|
|       29|  15728693|McWilliams|       574|   Germany|Female| 43|     3| 141349.43|            1|             1|     100187.43|     0|
|       30|  15656300|  Lucciano|       411|    France|  Male| 29|     0|  59697.17|            2|             1|      53483.21|     0|
|       31|  15589475|   Azikiwe|       591|     Spain|Female| 39|     3|       0.0|            3|             0|     140469.38|     1|
|       36|  15794171|  Lombardo|       475|    France|Female| 45|     0| 134264.04|            1|             0|      27822.99|     1|
+---------+----------+----------+----------+----------+------+---+------+----------+-------------+--------------+--------------+------+
only showing top 20 rows
```



6. Calculate the total number of active members and the average EstimatedSalary:

```
05:21 PM (<1s)                                        46

%sql
SELECT SUM(IsActiveMember) AS TotalActiveMembers, AVG(EstimatedSalary) AS AvgEstimatedSalary
FROM credit;
```

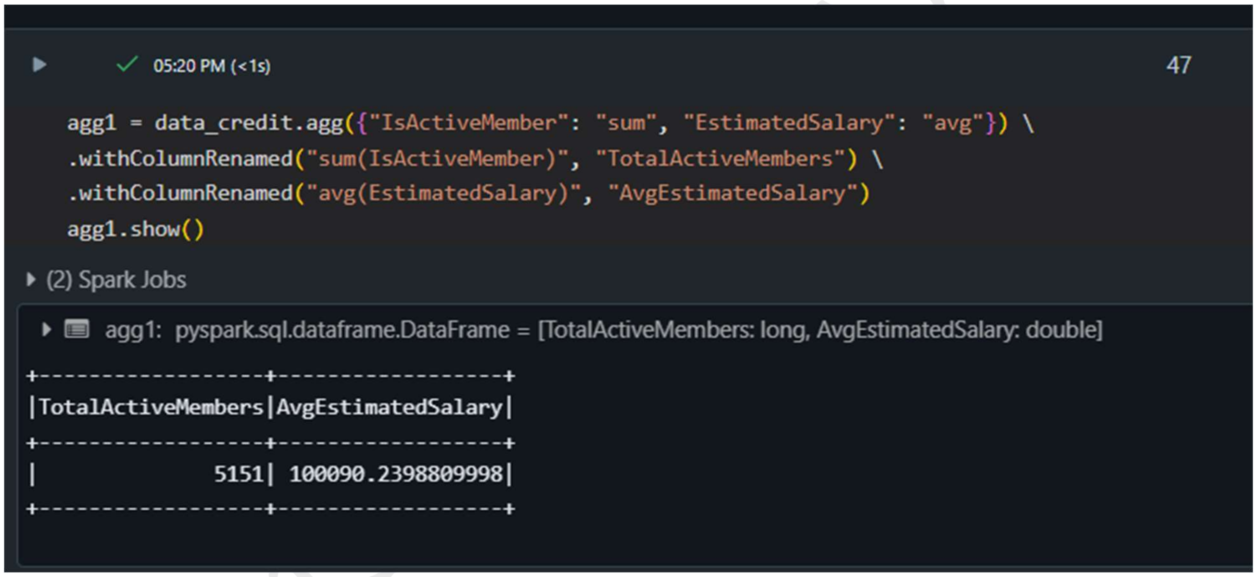
(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [TotalActiveMembers: long, AvgEstimatedSalary: double]

Table

	TotalActiveMembers	AvgEstimatedSalary
1	5151	100090.2398809998

The above one done using SQL magic function in pyspark – Databricks platform.



```
05:20 PM (<1s)                                        47

agg1 = data_credit.agg({"IsActiveMember": "sum", "EstimatedSalary": "avg"}) \
.withColumnRenamed("sum(IsActiveMember)", "TotalActiveMembers") \
.withColumnRenamed("avg(EstimatedSalary)", "AvgEstimatedSalary")
agg1.show()
```

(2) Spark Jobs

agg1: pyspark.sql.dataframe.DataFrame = [TotalActiveMembers: long, AvgEstimatedSalary: double]

```
+------------------+------------------+
|TotalActiveMembers|AvgEstimatedSalary|
+------------------+------------------+
|              5151| 100090.2398809998|
+------------------+------------------+
```

Thank you

Sivaprakash V