In [57]:

```
import pyspark

```

In [58]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, explode, col
```

In [59]:

```
spark = SparkSession.builder \
    .appName("WordCountExample") \
    .getOrCreate()
spark
```

Out[59]:

SparkSession - in-memory

SparkContext

[Spark UI (http://192.168.0.11:4040)](http://192.168.0.11:4040)

Version

`v3.2.1`

Master

`local[*]`

AppName

`WordCountExample`

In [ ]:

```

```

In [54]:

```
text_file_path = "/Users/myyntiimac/Desktop/Churn_Modelling.csv"

# Read the text file into a DataFrame
df = spark.read.csv(text_file_path)
```

In [55]:

```
df.show()
```

```
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
|      _c0|       _c1|      _c2|        _c3|      _c4|   _c5|_c6|   _c
7|      _c8|          _c9|     _c10|          _c11|           _c12|  _
c13|
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenur
e|  Balance|NumOfProducts|HasCrCard|IsActiveMember|EstimatedSalary|Exi
ted|
|        1|  15634602| Hargrave|        619|   France|Female| 42|
2|        0|            1|        1|             1|      101348.88|
1|
|        2|  15647311|     Hill|        608|    Spain|Female| 41|
1| 83807.86|            1|        0|             1|      112542.58|
0|
|        3|  15619304|     Onio|        502|   France|Female| 42|
8| 159660.8|            3|        1|             0|      113931.57|
1|
|        4|  15701354|     Boni|        699|   France|Female| 39|
1|        0|            2|        0|             0|       93826.63|
0|
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|
2|125510.82|            1|        1|             1|        79084.1|
0|
|        6|  15574012|      Chu|        645|    Spain|  Male| 44|
8|113755.78|            2|        1|             0|      149756.71|
1|
|        7|  15592531| Bartlett|        822|   France|  Male| 50|
7|        0|            2|        1|             1|        10062.8|
0|
|        8|  15656148|   Obinna|        376|  Germany|Female| 29|
4|115046.74|            4|        1|             0|      119346.88|
1|
|        9|  15792365|       He|        501|   France|  Male| 44|
4|142051.07|            2|        0|             1|        74940.5|
0|
|       10|  15592389|       H?|        684|   France|  Male| 27|
2|134603.88|            1|        1|             1|       71725.73|
0|
|       11|  15767821|   Bearce|        528|   France|  Male| 31|
6|102016.72|            2|        0|             0|       80181.12|
0|
|       12|  15737173|  Andrews|        497|    Spain|  Male| 24|
3|        0|            2|        1|             0|       76390.01|
0|
|       13|  15632264|      Kay|        476|   France|Female| 34|    1
0|        0|            2|        1|             0|       26260.98|
0|
|       14|  15691483|     Chin|        549|   France|Female| 25|
5|        0|            2|        0|             0|      190857.79|
0|
|       15|  15600882|    Scott|        635|    Spain|Female| 35|
7|        0|            2|        1|             1|       65951.65|
0|
|       16|  15643966|  Goforth|        616|  Germany|  Male| 45|
3|143129.41|            2|        0|             1|       64327.26|
0|
|       17|  15737452|    Romeo|        653|  Germany|  Male| 58|
1|132602.88|            1|        1|             0|        5097.67|
1|
|        18|  15788218|Henderson|       549|    Spain|Female| 24|
9|        0|            2|        1|             1|       14406.41|
0|
|        19|  15661507|  Muldrow|       587|    Spain|  Male| 45|
6|        0|            1|        0|             0|      158684.81|
0|
+----------+----------+---------+----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
only showing top 20 rows
```

In [49]:

```
df1=spark.read.csv(text_file_path,header=True,inferSchema=True)
```

In [50]:

```
df1.show()
```

```
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenur
e|  Balance|NumOfProducts|HasCrCard|IsActiveMember|EstimatedSalary|Exi
ted|
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
|        1|  15634602| Hargrave|        619|   France|Female| 42|
2|      0.0|            1|        1|             1|      101348.88|
1|
|        2|  15647311|     Hill|        608|    Spain|Female| 41|
1| 83807.86|            1|        0|             1|      112542.58|
0|
|        3|  15619304|     Onio|        502|   France|Female| 42|
8| 159660.8|            3|        1|             0|      113931.57|
1|
|        4|  15701354|     Boni|        699|   France|Female| 39|
1|      0.0|            2|        0|             0|       93826.63|
0|
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|
2|125510.82|            1|        1|             1|        79084.1|
0|
|        6|  15574012|      Chu|        645|    Spain|  Male| 44|
8|113755.78|            2|        1|             0|      149756.71|
1|
|        7|  15592531| Bartlett|        822|   France|  Male| 50|
7|      0.0|            2|        1|             1|        10062.8|
0|
|        8|  15656148|   Obinna|        376|  Germany|Female| 29|
4|115046.74|            4|        1|             0|      119346.88|
1|
|        9|  15792365|       He|        501|   France|  Male| 44|
4|142051.07|            2|        0|             1|        74940.5|
0|
|       10|  15592389|       H?|        684|   France|  Male| 27|
2|134603.88|            1|        1|             1|       71725.73|
0|
|       11|  15767821|   Bearce|        528|   France|  Male| 31|
6|102016.72|            2|        0|             0|       80181.12|
0|
|       12|  15737173|  Andrews|        497|    Spain|  Male| 24|
3|      0.0|            2|        1|             0|       76390.01|
0|
|       13|  15632264|      Kay|        476|   France|Female| 34|     1
0|      0.0|            2|        1|             0|       26260.98|
0|
|       14|  15691483|     Chin|        549|   France|Female| 25|
5|      0.0|            2|        0|             0|      190857.79|
0|
|       15|  15600882|    Scott|        635|    Spain|Female| 35|
7|      0.0|            2|        1|             1|       65951.65|
0|
|       16|  15643966|  Goforth|        616|  Germany|  Male| 45|
3|143129.41|            2|        0|             1|       64327.26|
0|
|       17|  15737452|    Romeo|        653|  Germany|  Male| 58|
1|132602.88|            1|        1|             0|        5097.67|
1|
|       18|  15788218|Henderson|        549|    Spain|Female| 24|
9|      0.0|            2|        1|             1|       14406.41|
0|
|       19|  15661507|  Muldrow|        587|    Spain|  Male| 45|
6|      0.0|            1|        0|             0|      158684.81|
0|
|       20|  15568982|      Hao|        726|   France|Female| 24|
6|      0.0|            2|        1|             1|       54724.03|
0|
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+---------+--------------+---------------+---
---+
only showing top 20 rows
```

In [34]:

```
df1.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)
 |-- HasCrCard: integer (nullable = true)
 |-- IsActiveMember: integer (nullable = true)
 |-- EstimatedSalary: double (nullable = true)
 |-- Exited: integer (nullable = true)
```

In [35]:

```
type(df1)
```

Out[35]:

```
pyspark.sql.dataframe.DataFrame
```

In [36]:

```
df1.head(5)
```

Out[36]:

```
[Row(RowNumber=1, CustomerId=15634602, Surname='Hargrave', CreditScore
=619, Geography='France', Gender='Female', Age=42, Tenure=2, Balance=
0.0, NumOfProducts=1, HasCrCard=1, IsActiveMember=1, EstimatedSalary=1
01348.88, Exited=1),
 Row(RowNumber=2, CustomerId=15647311, Surname='Hill', CreditScore=60
8, Geography='Spain', Gender='Female', Age=41, Tenure=1, Balance=8380
7.86, NumOfProducts=1, HasCrCard=0, IsActiveMember=1, EstimatedSalary=
112542.58, Exited=0),
 Row(RowNumber=3, CustomerId=15619304, Surname='Onio', CreditScore=50
2, Geography='France', Gender='Female', Age=42, Tenure=8, Balance=1596
60.8, NumOfProducts=3, HasCrCard=1, IsActiveMember=0, EstimatedSalary=
113931.57, Exited=1),
 Row(RowNumber=4, CustomerId=15701354, Surname='Boni', CreditScore=69
9, Geography='France', Gender='Female', Age=39, Tenure=1, Balance=0.0,
NumOfProducts=2, HasCrCard=0, IsActiveMember=0, EstimatedSalary=93826.
63, Exited=0),
 Row(RowNumber=5, CustomerId=15737888, Surname='Mitchell', CreditScore
=850, Geography='Spain', Gender='Female', Age=43, Tenure=2, Balance=12
5510.82, NumOfProducts=1, HasCrCard=1, IsActiveMember=1, EstimatedSala
ry=79084.1, Exited=0)]
```

In [37]:

```
df1.head(10)
```

Out[37]:

```
[Row(RowNumber=1, CustomerId=15634602, Surname='Hargrave', CreditScore
=619, Geography='France', Gender='Female', Age=42, Tenure=2, Balance=
0.0, NumOfProducts=1, HasCrCard=1, IsActiveMember=1, EstimatedSalary=1
01348.88, Exited=1),
 Row(RowNumber=2, CustomerId=15647311, Surname='Hill', CreditScore=60
8, Geography='Spain', Gender='Female', Age=41, Tenure=1, Balance=8380
7.86, NumOfProducts=1, HasCrCard=0, IsActiveMember=1, EstimatedSalary=
112542.58, Exited=0),
 Row(RowNumber=3, CustomerId=15619304, Surname='Onio', CreditScore=50
2, Geography='France', Gender='Female', Age=42, Tenure=8, Balance=1596
60.8, NumOfProducts=3, HasCrCard=1, IsActiveMember=0, EstimatedSalary=
113931.57, Exited=1),
 Row(RowNumber=4, CustomerId=15701354, Surname='Boni', CreditScore=69
9, Geography='France', Gender='Female', Age=39, Tenure=1, Balance=0.0,
NumOfProducts=2, HasCrCard=0, IsActiveMember=0, EstimatedSalary=93826.
63, Exited=0),
 Row(RowNumber=5, CustomerId=15737888, Surname='Mitchell', CreditScore
=850, Geography='Spain', Gender='Female', Age=43, Tenure=2, Balance=12
5510.82, NumOfProducts=1, HasCrCard=1, IsActiveMember=1, EstimatedSala
ry=79084.1, Exited=0),
 Row(RowNumber=6, CustomerId=15574012, Surname='Chu', CreditScore=645,
Geography='Spain', Gender='Male', Age=44, Tenure=8, Balance=113755.78,
NumOfProducts=2, HasCrCard=1, IsActiveMember=0, EstimatedSalary=14975
6.71, Exited=1),
 Row(RowNumber=7, CustomerId=15592531, Surname='Bartlett', CreditScore
=822, Geography='France', Gender='Male', Age=50, Tenure=7, Balance=0.
0, NumOfProducts=2, HasCrCard=1, IsActiveMember=1, EstimatedSalary=100
62.8, Exited=0),
 Row(RowNumber=8, CustomerId=15656148, Surname='Obinna', CreditScore=3
76, Geography='Germany', Gender='Female', Age=29, Tenure=4, Balance=11
5046.74, NumOfProducts=4, HasCrCard=1, IsActiveMember=0, EstimatedSala
ry=119346.88, Exited=1),
 Row(RowNumber=9, CustomerId=15792365, Surname='He', CreditScore=501,
Geography='France', Gender='Male', Age=44, Tenure=4, Balance=142051.0
7, NumOfProducts=2, HasCrCard=0, IsActiveMember=1, EstimatedSalary=749
40.5, Exited=0),
 Row(RowNumber=10, CustomerId=15592389, Surname='H?', CreditScore=684,
Geography='France', Gender='Male', Age=27, Tenure=2, Balance=134603.8
8, NumOfProducts=1, HasCrCard=1, IsActiveMember=1, EstimatedSalary=717
25.73, Exited=0)]
```

In [38]:

```python
column_name = 'Surname'

# Select the specific attribute (column)
selected_column = df1.select(column_name)

# Show the result
selected_column.show()

```

```
+---------+
|  Surname|
+---------+
| Hargrave|
|     Hill|
|     Onio|
|     Boni|
| Mitchell|
|      Chu|
| Bartlett|
|   Obinna|
|       He|
|       H?|
|   Bearce|
|  Andrews|
|      Kay|
|     Chin|
|    Scott|
|  Goforth|
|    Romeo|
|Henderson|
|  Muldrow|
|      Hao|
+---------+
only showing top 20 rows
```

In [39]:

```
column_name1 = 'Surname'

column_name2 = 'Geography'

# Select the specific attribute (column)
selected_column = df1.select(column_name1,column_name2 )

# Show the result
selected_column.show()
```

```
+---------+---------+
|  Surname|Geography|
+---------+---------+
| Hargrave|   France|
|     Hill|    Spain|
|     Onio|   France|
|     Boni|   France|
| Mitchell|    Spain|
|      Chu|    Spain|
| Bartlett|   France|
|   Obinna|  Germany|
|       He|   France|
|       H?|   France|
|   Bearce|   France|
|  Andrews|    Spain|
|      Kay|   France|
|     Chin|   France|
|    Scott|    Spain|
|  Goforth|  Germany|
|    Romeo|  Germany|
|Henderson|    Spain|
|  Muldrow|    Spain|
|      Hao|   France|
+---------+---------+
only showing top 20 rows
```

df1.describe().show()

In [45]:

```
df1.describe().show()
```

```
+-------+------------------+-----------------+-------+----------------
-+---------+------+------------------+------------------+-------------
----+------------------+-------------------+-------------------+------
-----------+-------------------+
|summary|         RowNumber|       CustomerId|Surname|      CreditScor
e|Geography|Gender|               Age|            Tenure|          Bal
ance|     NumOfProducts|          HasCrCard|     IsActiveMember|  Esti
matedSalary|             Exited|
+-------+------------------+-----------------+-------+----------------
-+---------+------+------------------+------------------+-------------
----+------------------+-------------------+-------------------+------
-----------+-------------------+
|  count|             10000|            10000|  10000|            1000
0|    10000| 10000|             10000|             10000|            1
0000|             10000|              10000|              10000|
10000|              10000|
|   mean|            5000.5|  1.56909405694E7|   null|         650.528
8|     null|  null|           38.9218|            5.0128|76485.8892879
9961|            1.5302|             0.7055|             0.5151|10009
0.2398809998|             0.2037|
| stddev|2886.8956799071675|71936.18612274907|   null|96.6532987361303
5|     null|  null|10.487806451704587|2.8921743770496837|62397.4052023
8599|0.5816543579989917|0.45584046447513327|0.49979692845891815|57510.
49281769821|0.40276858399486065|
|    min|                 1|         15565701|  Abazu|              35
0|   France|Female|                18|                 0|
0.0|                 1|                  0|                  0|
11.58|                  0|
|    max|             10000|         15815690| Zuyeva|              85
0|    Spain|  Male|                92|                10|        25089
8.09|                 4|                  1|                  1|
199992.48|                  1|
+-------+------------------+-----------------+-------+----------------
-+---------+------+------------------+------------------+-------------
----+------------------+-------------------+-------------------+------
-----------+-------------------+
```

In [51]:

```
# You can replace 'old_column_name' and 'new_column_name' with the actual names
old_column_name = 'HasCrCard'
new_column_name = 'crcard'

# Rename the column
df1 = df1.withColumnRenamed(old_column_name, new_column_name)

# Show the DataFrame to verify the new column name
df1.show()
```

```
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+------+--------------+---------------+------
+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenur
e|  Balance|NumOfProducts|crcard|IsActiveMember|EstimatedSalary|Exited
|
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+------+--------------+---------------+------
+
|        1|  15634602| Hargrave|        619|   France|Female| 42|     
2|      0.0|            1|     1|             1|      101348.88|     1
|
|        2|  15647311|     Hill|        608|    Spain|Female| 41|     
1| 83807.86|            1|     0|             1|      112542.58|     0
|
|        3|  15619304|     Onio|        502|   France|Female| 42|     
8| 159660.8|            3|     1|             0|      113931.57|     1
|
|        4|  15701354|     Boni|        699|   France|Female| 39|     
1|      0.0|            2|     0|             0|       93826.63|     0
|
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|     
2|125510.82|            1|     1|             1|        79084.1|     0
|
|        6|  15574012|      Chu|        645|    Spain|  Male| 44|     
8|113755.78|            2|     1|             0|      149756.71|     1
|
|        7|  15592531| Bartlett|        822|   France|  Male| 50|     
7|      0.0|            2|     1|             1|        10062.8|     0
|
|        8|  15656148|   Obinna|        376|  Germany|Female| 29|     
4|115046.74|            4|     1|             0|      119346.88|     1
|
|        9|  15792365|       He|        501|   France|  Male| 44|     
4|142051.07|            2|     0|             1|        74940.5|     0
|
|       10|  15592389|       H?|        684|   France|  Male| 27|     
2|134603.88|            1|     1|             1|       71725.73|     0
|
|       11|  15767821|   Bearce|        528|   France|  Male| 31|     
6|102016.72|            2|     0|             0|       80181.12|     0
|
|       12|  15737173|  Andrews|        497|    Spain|  Male| 24|     
3|      0.0|            2|     1|             0|       76390.01|     0
|
|       13|  15632264|      Kay|        476|   France|Female| 34|    1
0|      0.0|            2|     1|             0|       26260.98|     0
|
|       14|  15691483|     Chin|        549|   France|Female| 25|     
5|      0.0|            2|     0|             0|      190857.79|     0
|
|       15|  15600882|    Scott|        635|    Spain|Female| 35|     
7|      0.0|            2|     1|             1|       65951.65|     0
|
|       16|  15643966|  Goforth|        616|  Germany|  Male| 45|     
3|143129.41|            2|     0|             1|       64327.26|     0
|
|       17|  15737452|    Romeo|        653|  Germany|  Male| 58|     
1|132602.88|            1|     1|             0|        5097.67|     1
|
|       18|  15788218|Henderson|        549|    Spain|Female| 24|
9|      0.0|            2|    1|             1|       14406.41|     0
|
|       19|  15661507|  Muldrow|         587|    Spain|  Male| 45|
6|      0.0|            1|    0|             0|      158684.81|     0
|
|       20|  15568982|      Hao|         726|   France|Female| 24|
6|      0.0|            2|    1|             1|       54724.03|     0
|
+---------+----------+---------+-----------+---------+------+---+-----
-+---------+-------------+------+--------------+---------------+------
+
only showing top 20 rows
```

In [ ]:

```
 
```