```
In [57]:
```

```
1 import pyspark
2
```

In [58]:

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

In [59]:

```
1 spark = SparkSession.builder \
2     .appName("WordCountExample") \
3     .getOrCreate()
4     spark
```

Out[59]:

SparkSession - in-memory

SparkContext

Spark UI (http://192.168.0.11:4040)

Version

v3.2.1

Master

local[*]

AppName

WordCountExample

In []:

1

In [54]:

```
1 text_file_path = "/Users/myyntiimac/Desktop/Churn_Modelling.csv"
2 
3 # Read the text file into a DataFrame
4 df = spark.read.csv(text_file_path)
```

In [55]:

1 df.show()

```
_c0|
                  _c1|
                            _c2|
                                                          _c5|_c6|
                                        _c3|
                      c9
                               _c10
7
        _c8
   |RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenur
    Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exi
ted
             15634602 | Hargrave |
                                        619
                                                France | Female | 42 |
2 |
          0 |
                        1 |
                                   1 |
                                                  1 |
                                                          101348.88
1 |
             15647311
                                        608
                                                 Spain|Female| 41|
         2
                           Hill
   83807.86
1
                        1 |
                                   0 |
                                                          112542.58
0 |
         3 |
             15619304
                           Onio
                                        502
                                                France | Female | 42 |
                        3 |
   159660.8
                                                          113931.57
                                   1 |
8 |
1 |
             15701354
                                        699
                                                France|Female| 39|
         4
                           Boni
          0 |
                        2
                                   0 |
1 |
                                                           93826.63
0 |
             15737888 | Mitchell |
                                                 Spain|Female| 43|
                                        850
2 | 125510.82 |
                                                            79084.1
                        1
                                   1 |
                                                  1 |
0
             15574012
                                                 Spain|
                            Chu
                                        645
                                                         Male 44
8 | 113755.78 |
                        2
                                   1 |
                                                  0 |
                                                          149756.71
1 |
             15592531 | Bartlett|
         7 |
                                        822
                                                France
                                                         Male | 50 |
7 |
          0
                                   1 |
                                                  1 |
                                                            10062.8
0
         8
             15656148
                         Obinna
                                        376
                                               Germany | Female | 29 |
4 | 115046.74 |
                                                          119346.88
                                                  0 |
                                   1 |
1
         9 |
             15792365
                                         501
                                                France
                                                         Male | 44|
                             Не
                        2 |
4 | 142051.07 |
                                   0 |
                                                  1 |
                                                            74940.5
             15592389
                             H?
                                         684
                                                France
                                                         Male | 27|
        10
2 | 134603.88 |
                        1
                                   1 |
                                                  1 |
                                                           71725.73
0 |
        11
             15767821
                         Bearce
                                         528
                                                France
                                                         Male | 31|
6 | 102016.72 |
                                   0 |
                        2
                                                  0 |
                                                           80181.12
0
                        Andrews
                                         497
                                                         Male | 24 |
        12
             15737173
                                                 Spain
                                  1 |
                                                           76390.01
3
          0
                        2
                                                  0 |
0
                                                France | Female | 34 |
        13
             15632264
                            Kay
                                        476
                        2
                                                           26260.98
0
          0
                                   1 |
0
             15691483
                                                France|Female| 25|
        14
                           Chin
                                         549
                                                          190857.79
          0 |
                        2
                                   0 |
5 |
0 |
             15600882
                                                 Spain|Female| 35|
        15
                          Scott
                                        635
7
          0 |
                        2 |
                                   1 |
                                                  1 |
                                                           65951.65
0 |
             15643966
                        Goforth
                                               Germany
        16
                                        616
                                                        Male 45
3 | 143129.41 |
                        2
                                  0 |
                                                  1 |
                                                           64327.26
0 |
                                        653
                                               Germany | Male | 58 |
        17
             15737452
                          Romeo
```

```
1|132602.88|
                     1 |
                              1 |
                                           0 |
                                                    5097.67
1 |
           15788218 | Henderson |
                                   549
                                          Spain|Female| 24|
       18|
                              1 |
9 |
        0 |
                     2 |
                                           1 |
                                                   14406.41
0 |
                     Muldrow
       19|
           15661507
                                   587
                                          Spain | Male | 45 |
                              0 |
                                                  158684.81
        0 |
                     1 |
                                           0 |
6
0 |
  _____+
only showing top 20 rows
```

In [49]:

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

In [50]:

1 df1.show()

|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenur Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exi __________ 15634602 | Hargrave | 619 France | Female | 42 | 1 2 0.0 1 1 | 1 | 101348.88 1 | 2 15647311 Hill 608 Spain|Female| 41| 83807.86 1 | 0 | 1 | 112542.58 1 | 0 15619304 502 France | Female | 42 | 3 | Onio 8 159660.8 3 | 1 | 113931.57 1 4 | 15701354 Boni 699 France | Female | 39 | 2 | 0 | 93826.63 0.0 1 | 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 149756.71 1 | 0 | 1 15592531 | Bartlett| 7 822 France Male 50 7 0.0 2 1 | 1 | 10062.8 0 | 376 Germany | Female | 29 | 15656148 Obinna 8 4 | 115046.74 | 4 | 1 | 119346.88 0 | 1 | 9 | 15792365 Не 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 15767821 Bearce 528 France Male 31 11 6 | 102016.72 | 0 | 0 | 80181.12 0 | 12 15737173 Andrews 497 Spain| Male | 24 | 3 0.0 2 1 | 0 | 76390.01 0 France|Female| 34| 13 15632264 Kay 476 2 | 1 | 0 0.0 0 | 26260.98 0 France | Female | 25 | 14 15691483 Chin 549 2 | 190857.79 5 0.0 0 | 0 15600882 Spain|Female| 35| 15 Scott 635 65951.65 0.0 1 | 7 | 2 1 | 0 | 15643966 Goforth 616 Germany | Male | 45| 16 3 | 143129.41 | 2 0 | 64327.26 1 | 0 15737452 Romeo 653 Germany| Male | 58 17 1 | 132602.88 | 1 1 | 5097.67 0 | 1 | 15788218 | Henderson | Spain|Female| 24| 18 549

```
27/07/2023, 19:12
                               Spark exploring - Jupyter Notebook
 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 |
                   2
                           1 |
                                             54724.03
 6
       0.0
                                      1 |
 0 |
    _+____+__
 only showing top 20 rows
```

In [34]:

```
1 dfl.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]:

```
1 type(df1)
```

Out[35]:

pyspark.sql.dataframe.DataFrame

In [36]:

1 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, EstimatedSalary=79084.1, Exited=0)]

In [37]:

1 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, EstimatedSalary=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, EstimatedSalary=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]:

```
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
       Не
       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 = dfl.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
       Не
            France
       H?
            France
   Bearce
           France
  Andrews
             Spain
      Kay
            France
            France
     Chin
    Scott
             Spain
  Goforth|
           Germany
    Romeo
           Germany
Henderson
             Spain
  Muldrow
             Spain
      Hao
            France
+----+
only showing top 20 rows
```

df1.describe().show()

In [45]:

1 df1.describe().show()

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

				eography Gender er EstimatedSala	
 ++-	+	+	+	' 	++
+ 1 2 0.0		Hargrave 1 1	619	France Female 1 101348	
 2 1 83807.86		Hill 1 0	608	Spain Female	
 3 8 159660.8		Onio 3 1	502	France Female	
 4 1 0.0		Boni 2 0	699	France Female	
 5 2 125510.82	15737888	Mitchell 1 1	850	Spain Female	
 6 8 113755.78	•	Chu 2 1	645	Spain Male 0 149756	
 7 7 0.0		Bartlett 2 1	822	France Male	
 8 4 115046.74		•	376	Germany Female	
 9 4 142051.07	15792365	He 2 0	501	France Male 1 74940	
 10 2 134603.88	15592389	H? 1 1	684	France Male 1 71725	27 .73 0
 11 6 102016.72			528	France Male 0 80181	
 12 3 0.0		Andrews 2 1	497	Spain Male 0 76390	
 13 0 0.0	15632264	Kay 2 1	476	France Female	
 14 5 0.0	15691483	Chin 2 0	549	France Female 0 190857	
 15 7 0.0			635	Spain Female	
 16 3 143129.41	•	•	616	Germany Male 1 64327	
 17 1 132602.88			653	Germany Male 0 5097	
18	15788218 1	Henderson	549	Spain Female	24

In []:

only showing top 20 rows

1