Akilesh K

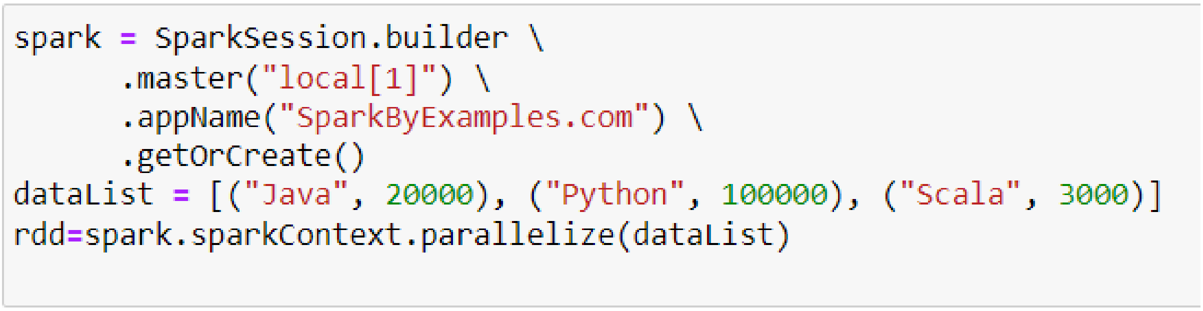
k.akilesh123@gmail.com

Data engineering - Batch 1

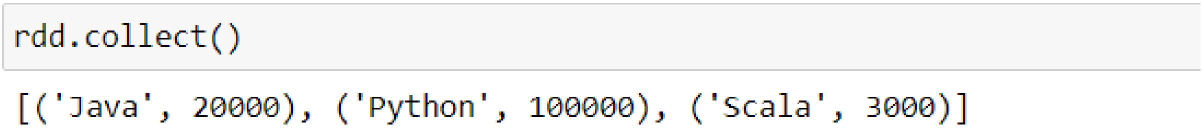
Date: 06-02-24

DAY 13 - PYSPARK -Action, Transformation

Parallellize dataframe



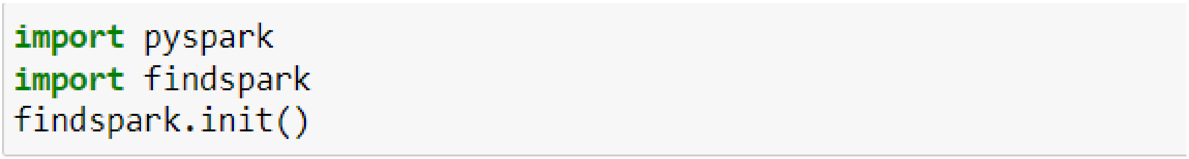
```
spark = SparkSession.builder \
      .master("local[1]") \
      .appName("SparkByExamples.com") \
      .getOrCreate()
dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
rdd=spark.sparkContext.parallelize(dataList)
```



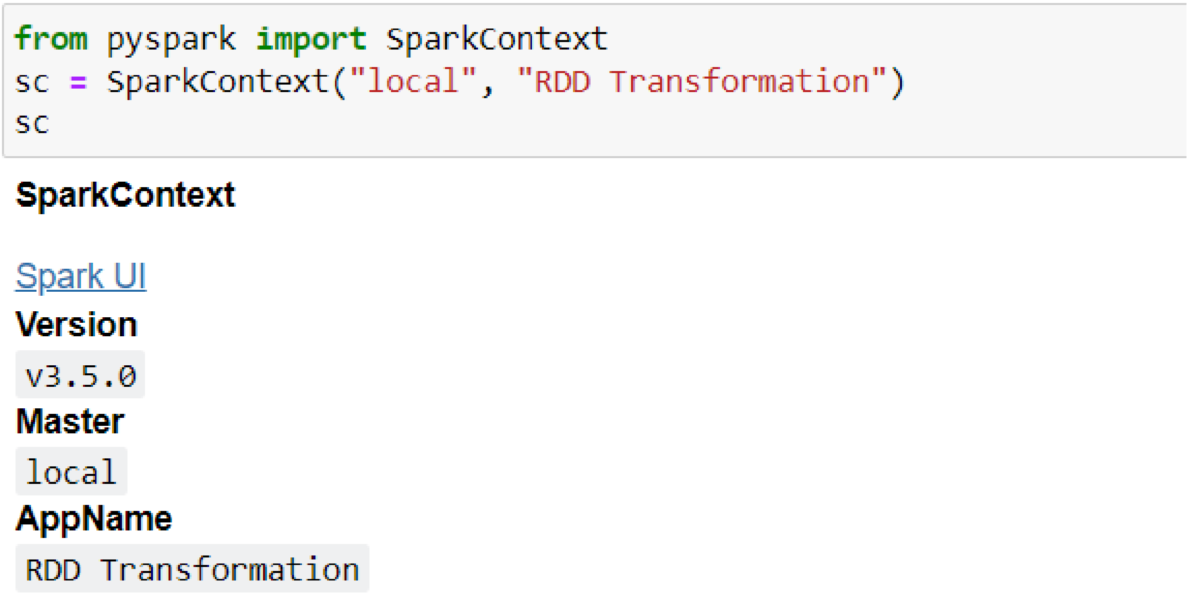
```
rdd.collect()
```

```
[('Java', 20000), ('Python', 100000), ('Scala', 3000)]
```

Import findspark



```
import pyspark
import findspark
findspark.init()
```



```
from pyspark import SparkContext
sc = SparkContext("local", "RDD Transformation")
sc
```

SparkContext

Spark UI

Version

`v3.5.0`

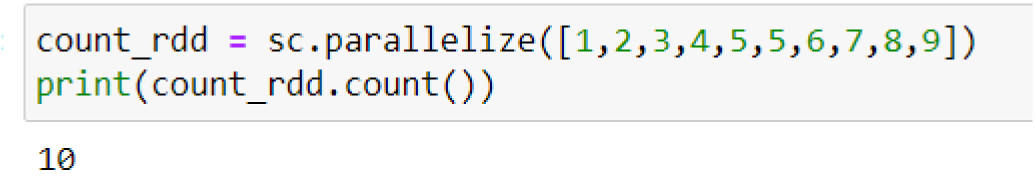
Master

`local`

AppName

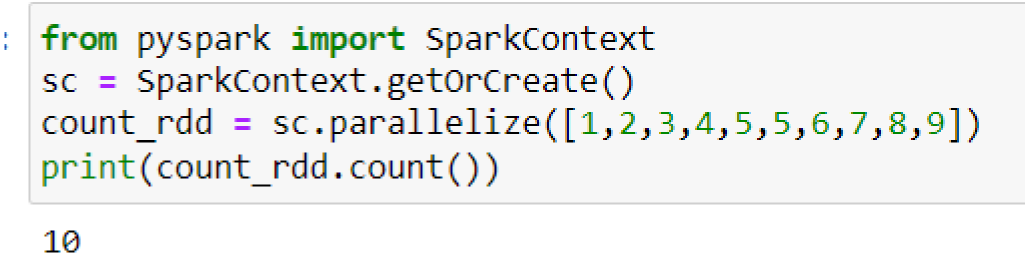
`RDD Transformation`

Count



```
count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
print(count_rdd.count())
```

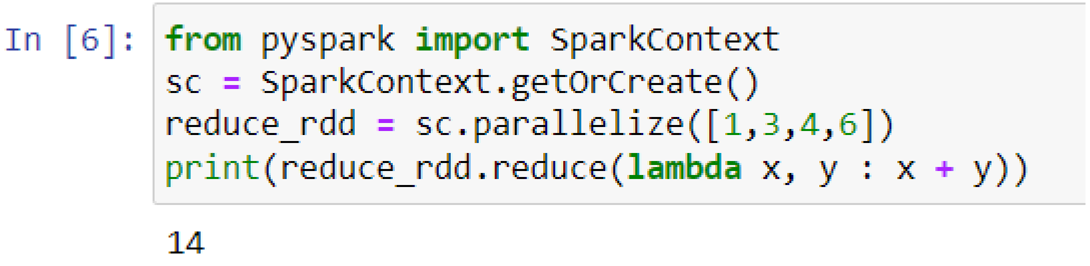
```
10
```



```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
print(count_rdd.count())
```

```
10
```

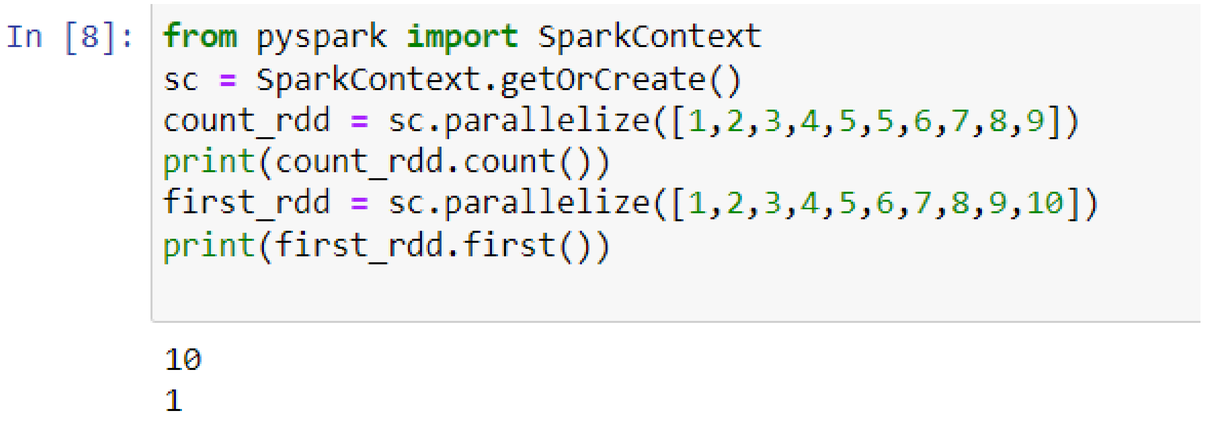
Lambda



```
In [6]: from pyspark import SparkContext
        sc = SparkContext.getOrCreate()
        reduce_rdd = sc.parallelize([1,3,4,6])
        print(reduce_rdd.reduce(lambda x, y : x + y))
```

```
14
```

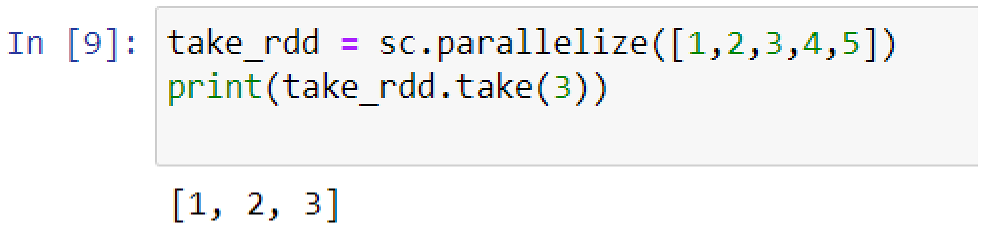
First()



```
In [8]: from pyspark import SparkContext
        sc = SparkContext.getOrCreate()
        count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
        print(count_rdd.count())
        first_rdd = sc.parallelize([1,2,3,4,5,6,7,8,9,10])
        print(first_rdd.first())
```

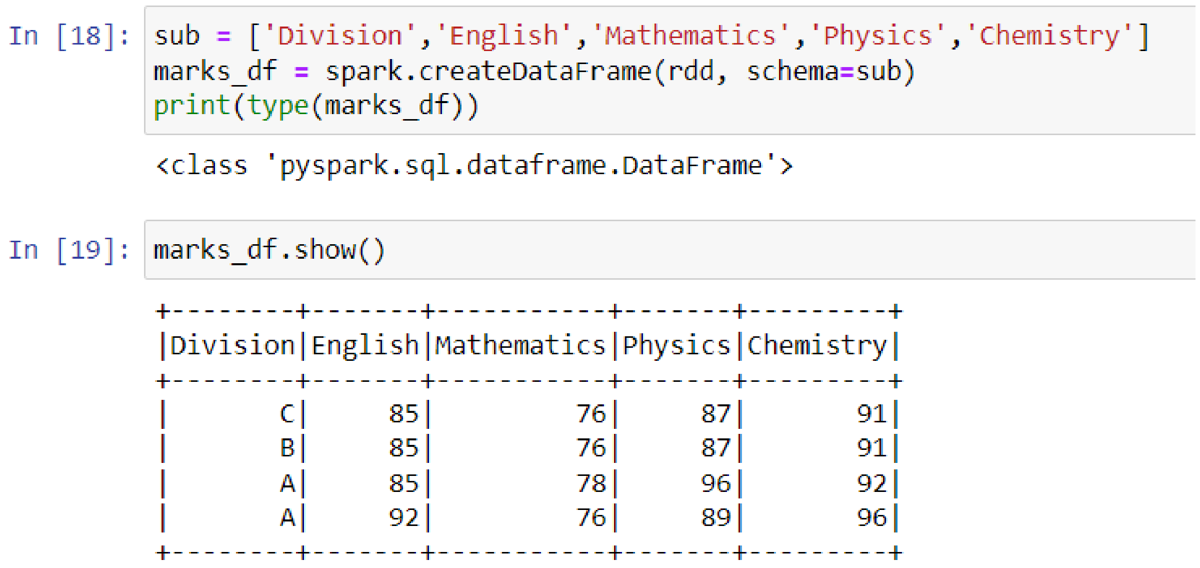
```
10
1
```

Take



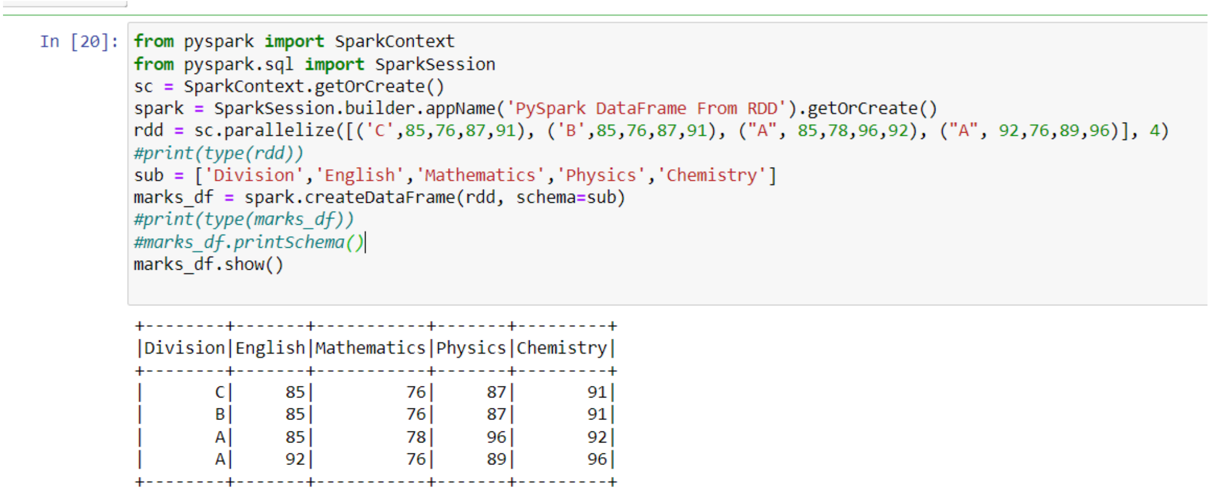
```
In [9]: take_rdd = sc.parallelize([1,2,3,4,5])
        print(take_rdd.take(3))

[1, 2, 3]
```

Set up dataframe



```
In [18]: sub = ['Division','English','Mathematics','Physics','Chemistry']
         marks_df = spark.createDataFrame(rdd, schema=sub)
         print(type(marks_df))

<class 'pyspark.sql.dataframe.DataFrame'>
```

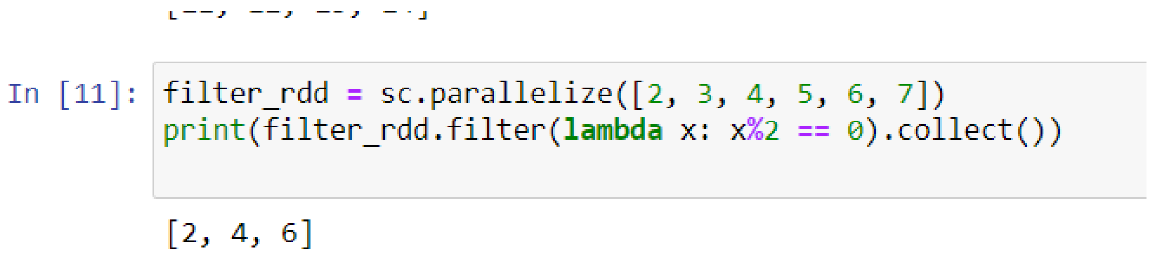
```
In [19]: marks_df.show()

+--------+-------+-----------+-------+---------+
|Division|English|Mathematics|Physics|Chemistry|
+--------+-------+-----------+-------+---------+
|       C|     85|         76|     87|       91|
|       B|     85|         76|     87|       91|
|       A|     85|         78|     96|       92|
|       A|     92|         76|     89|       96|
+--------+-------+-----------+-------+---------+
```



```
In [20]: from pyspark import SparkContext
         from pyspark.sql import SparkSession
         sc = SparkContext.getOrCreate()
         spark = SparkSession.builder.appName('PySpark DataFrame From RDD').getOrCreate()
         rdd = sc.parallelize([('C',85,76,87,91), ('B',85,76,87,91), ("A", 85,78,96,92), ("A", 92,76,89,96)], 4)
         #print(type(rdd))
         sub = ['Division','English','Mathematics','Physics','Chemistry']
         marks_df = spark.createDataFrame(rdd, schema=sub)
         #print(type(marks_df))
         #marks_df.printSchema()
         marks_df.show()

+--------+-------+-----------+-------+---------+
|Division|English|Mathematics|Physics|Chemistry|
+--------+-------+-----------+-------+---------+
|       C|     85|         76|     87|       91|
|       B|     85|         76|     87|       91|
|       A|     85|         78|     96|       92|
|       A|     92|         76|     89|       96|
+--------+-------+-----------+-------+---------+
```

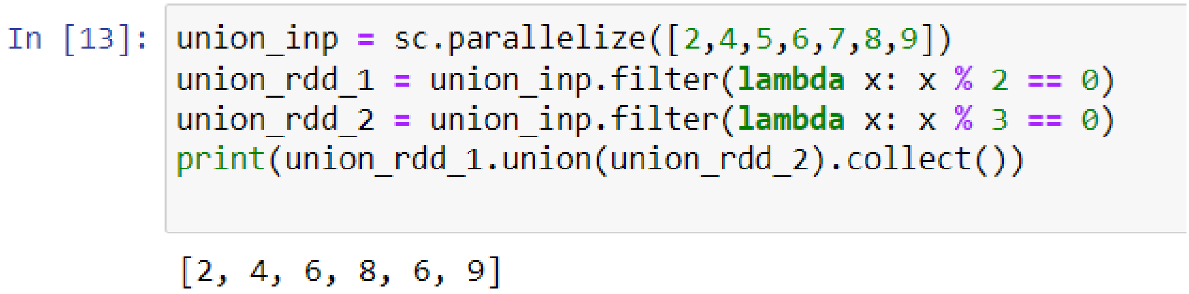
Transformations

filter



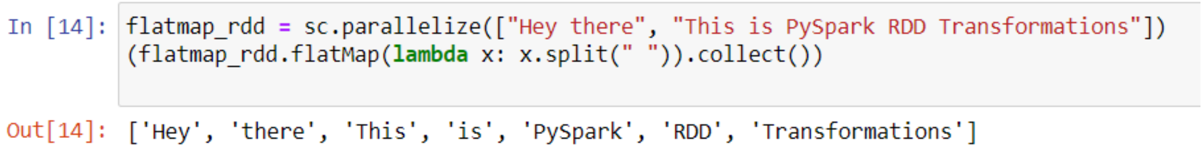
```
In [11]: filter_rdd = sc.parallelize([2, 3, 4, 5, 6, 7])
         print(filter_rdd.filter(lambda x: x%2 == 0).collect())

         [2, 4, 6]
```

Union



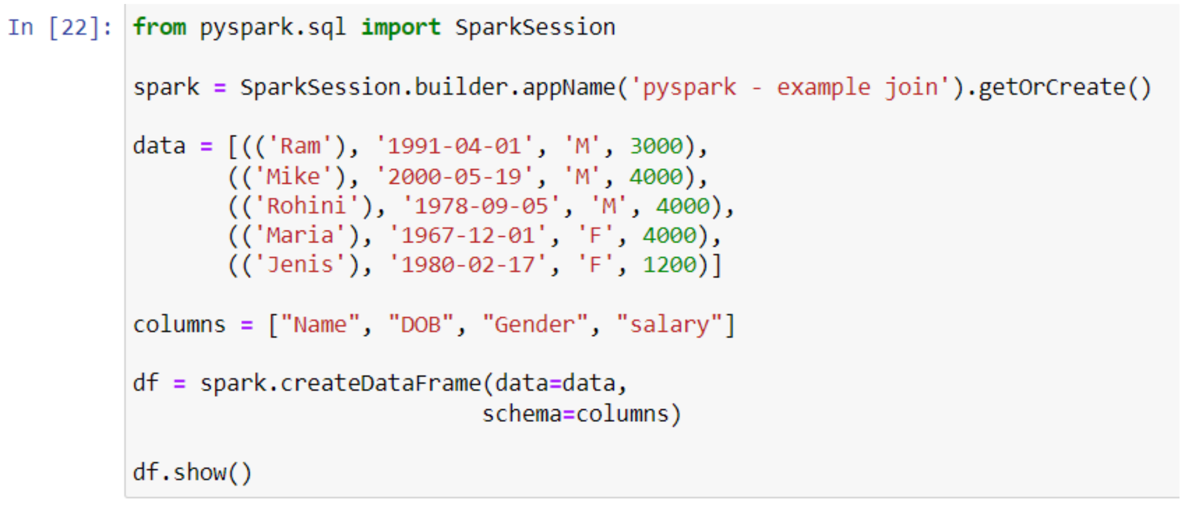
```
In [13]: union_inp = sc.parallelize([2,4,5,6,7,8,9])
         union_rdd_1 = union_inp.filter(lambda x: x % 2 == 0)
         union_rdd_2 = union_inp.filter(lambda x: x % 3 == 0)
         print(union_rdd_1.union(union_rdd_2).collect())

         [2, 4, 6, 8, 6, 9]
```

Flatmap



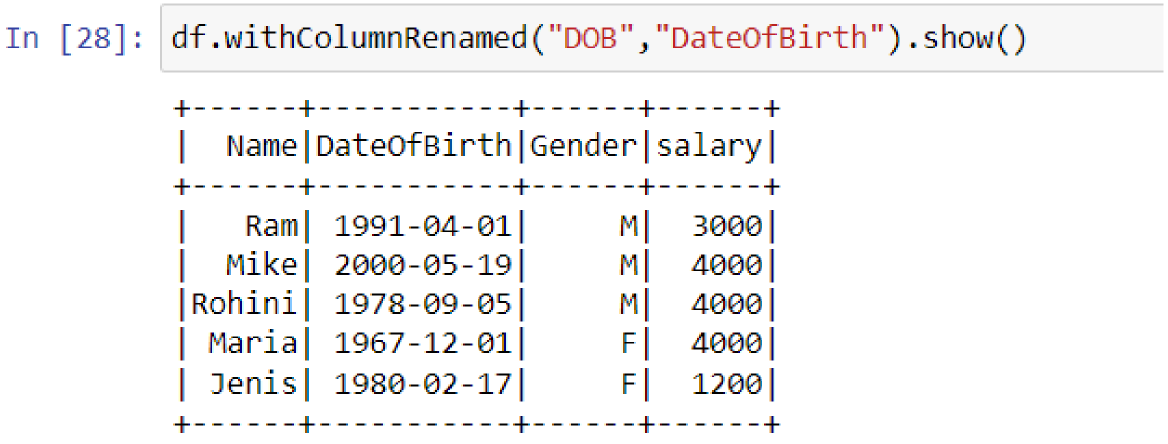
```
In [14]: flatmap_rdd = sc.parallelize(["Hey there", "This is PySpark RDD Transformations"])
         (flatmap_rdd.flatMap(lambda x: x.split(" ")).collect())

Out[14]: ['Hey', 'there', 'This', 'is', 'PySpark', 'RDD', 'Transformations']
```

Creating dataframe



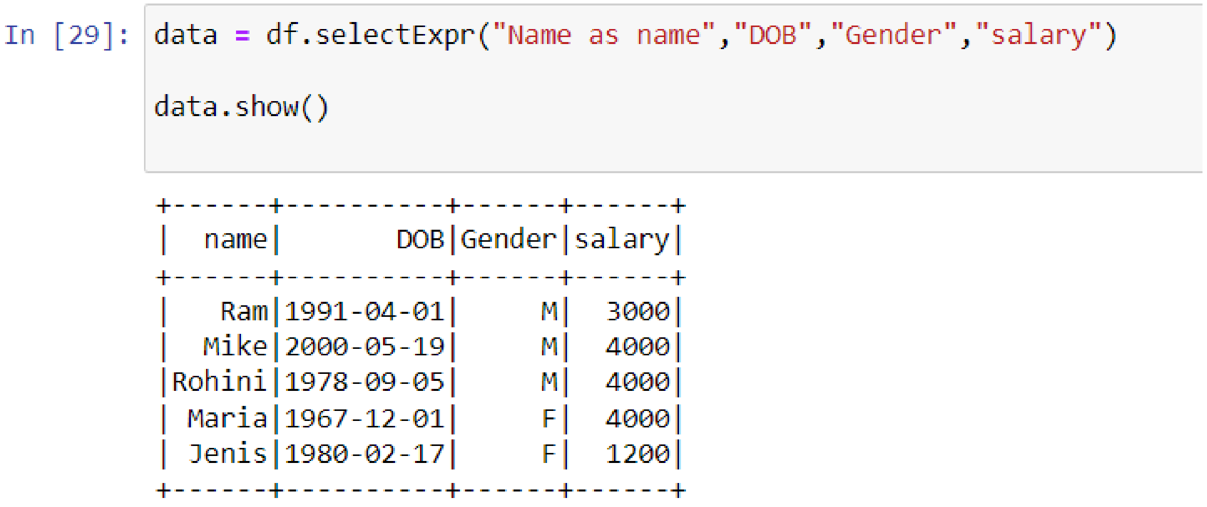
```
In [22]: from pyspark.sql import SparkSession

         spark = SparkSession.builder.appName('pyspark - example join').getOrCreate()

         data = [(('Ram'), '1991-04-01', 'M', 3000),
                (('Mike'), '2000-05-19', 'M', 4000),
                (('Rohini'), '1978-09-05', 'M', 4000),
                (('Maria'), '1967-12-01', 'F', 4000),
                (('Jenis'), '1980-02-17', 'F', 1200)]

         columns = ["Name", "DOB", "Gender", "salary"]

         df = spark.createDataFrame(data=data,
                                    schema=columns)

         df.show()
```

Rename column



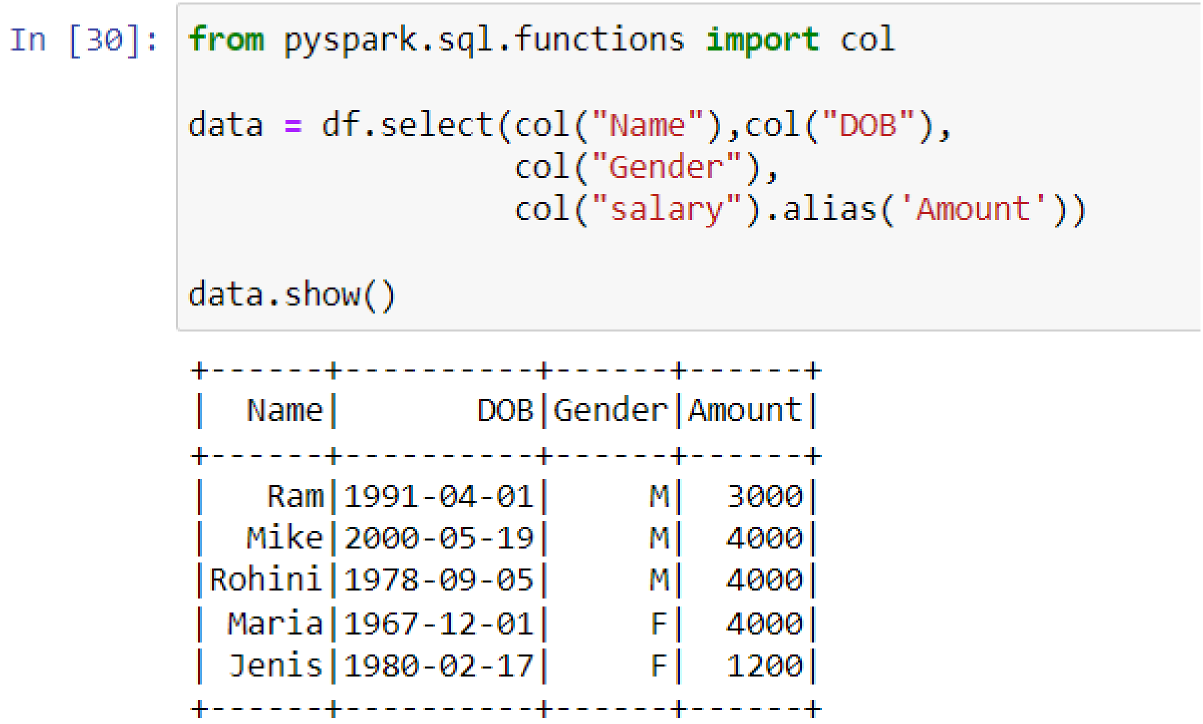
```
In [28]: df.withColumnRenamed("DOB","DateOfBirth").show()

         +------+-----------+------+------+
         |  Name|DateOfBirth|Gender|salary|
         +------+-----------+------+------+
         |   Ram| 1991-04-01|     M|  3000|
         |  Mike| 2000-05-19|     M|  4000|
         |Rohini| 1978-09-05|     M|  4000|
         | Maria| 1967-12-01|     F|  4000|
         | Jenis| 1980-02-17|     F|  1200|
         +------+-----------+------+------+
```

Rename using expression



```
In [29]: data = df.selectExpr("Name as name","DOB","Gender","salary")

         data.show()

         +------+----------+------+------+
         |  name|       DOB|Gender|salary|
         +------+----------+------+------+
         |   Ram|1991-04-01|     M|  3000|
         |  Mike|2000-05-19|     M|  4000|
         |Rohini|1978-09-05|     M|  4000|
         | Maria|1967-12-01|     F|  4000|
         | Jenis|1980-02-17|     F|  1200|
         +------+----------+------+------+
```

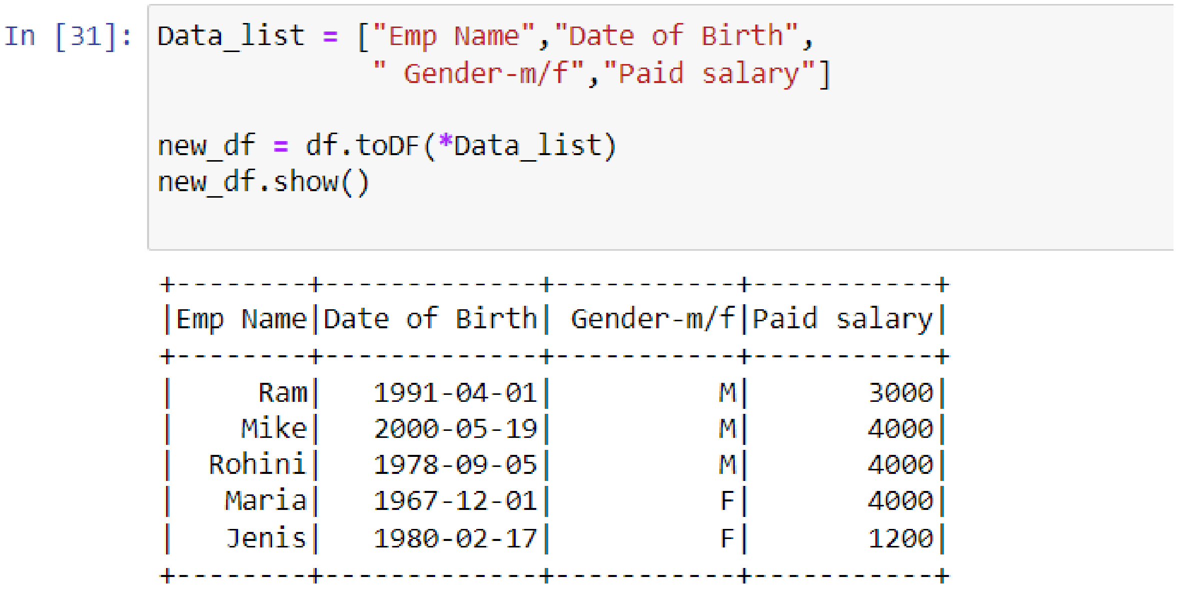
Rename multiple column



```
In [30]: from pyspark.sql.functions import col

         data = df.select(col("Name"),col("DOB"),
                          col("Gender"),
                          col("salary").alias('Amount'))

         data.show()
```

```
+------+----------+------+------+
|  Name|       DOB|Gender|Amount|
+------+----------+------+------+
|   Ram|1991-04-01|     M|  3000|
|  Mike|2000-05-19|     M|  4000|
|Rohini|1978-09-05|     M|  4000|
| Maria|1967-12-01|     F|  4000|
| Jenis|1980-02-17|     F|  1200|
+------+----------+------+------+
```

To DF



```
In [31]: Data_list = ["Emp Name","Date of Birth",
                      " Gender-m/f","Paid salary"]

         new_df = df.toDF(*Data_list)
         new_df.show()
```

```
+--------+-------------+-----------+-----------+
|Emp Name|Date of Birth| Gender-m/f|Paid salary|
+--------+-------------+-----------+-----------+
|     Ram|   1991-04-01|          M|       3000|
|    Mike|   2000-05-19|          M|       4000|
|  Rohini|   1978-09-05|          M|       4000|
|   Maria|   1967-12-01|          F|       4000|
|   Jenis|   1980-02-17|          F|       1200|
+--------+-------------+-----------+-----------+
```