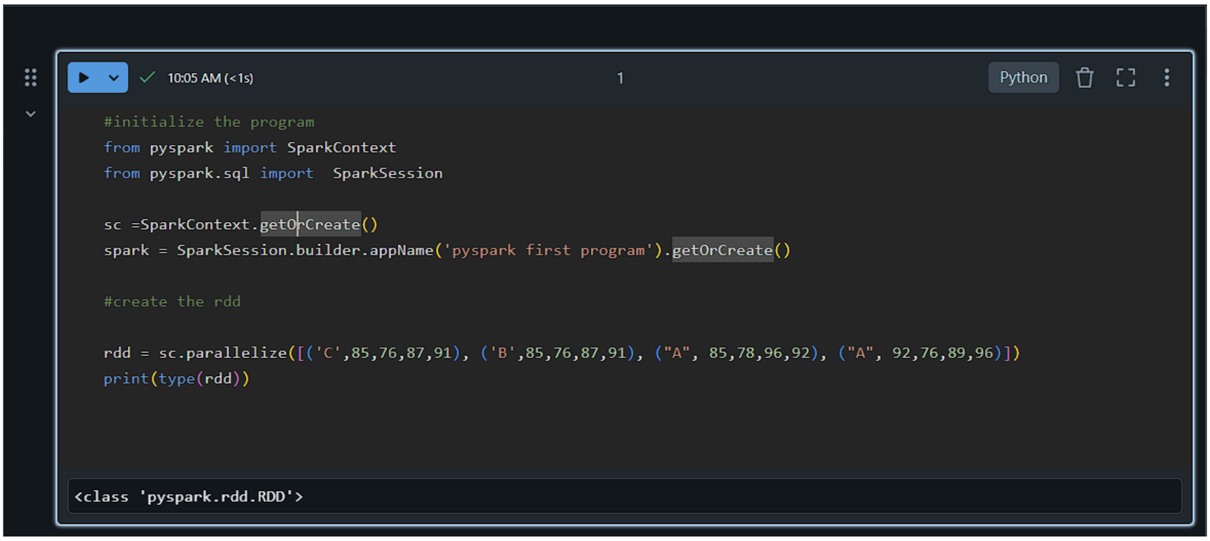
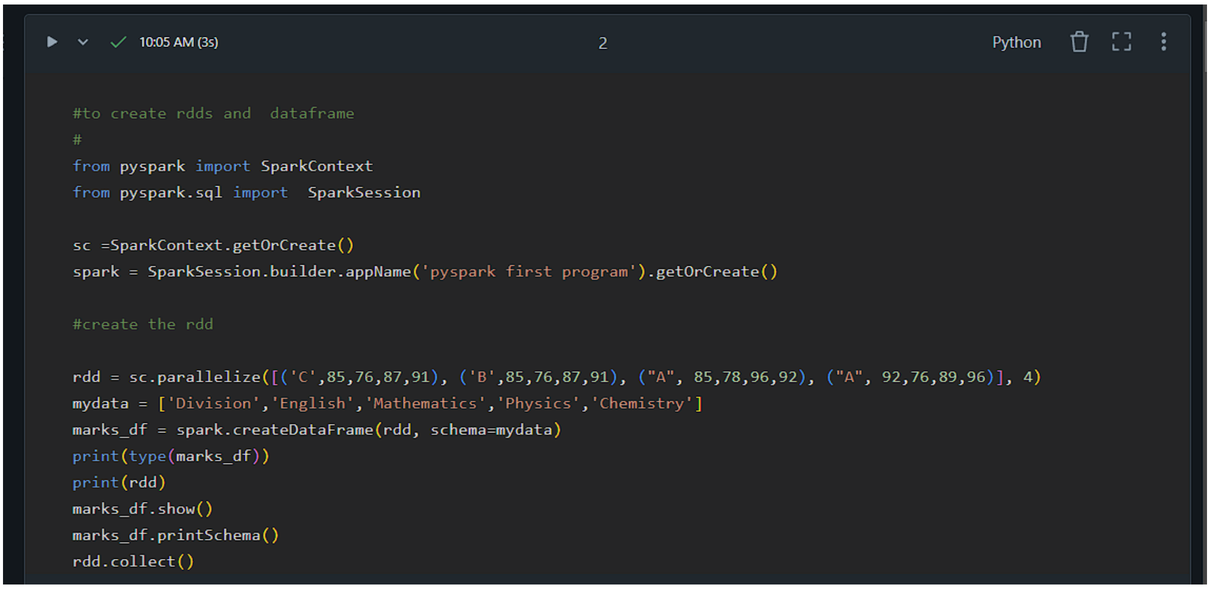
PySpark Hands-On Day-2

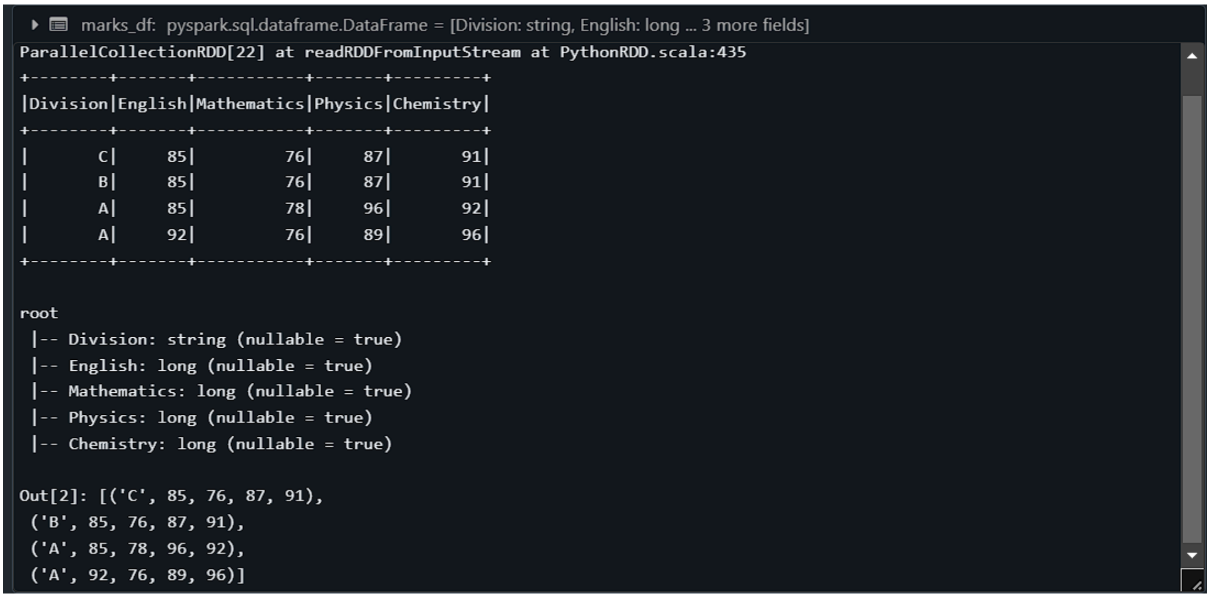


```
#initialize the program
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()

#create the rdd

rdd = sc.parallelize([('C',85,76,87,91), ('B',85,76,87,91), ("A", 85,78,96,92), ("A", 92,76,89,96)])
print(type(rdd))
```

```
<class 'pyspark.rdd.RDD'>
```

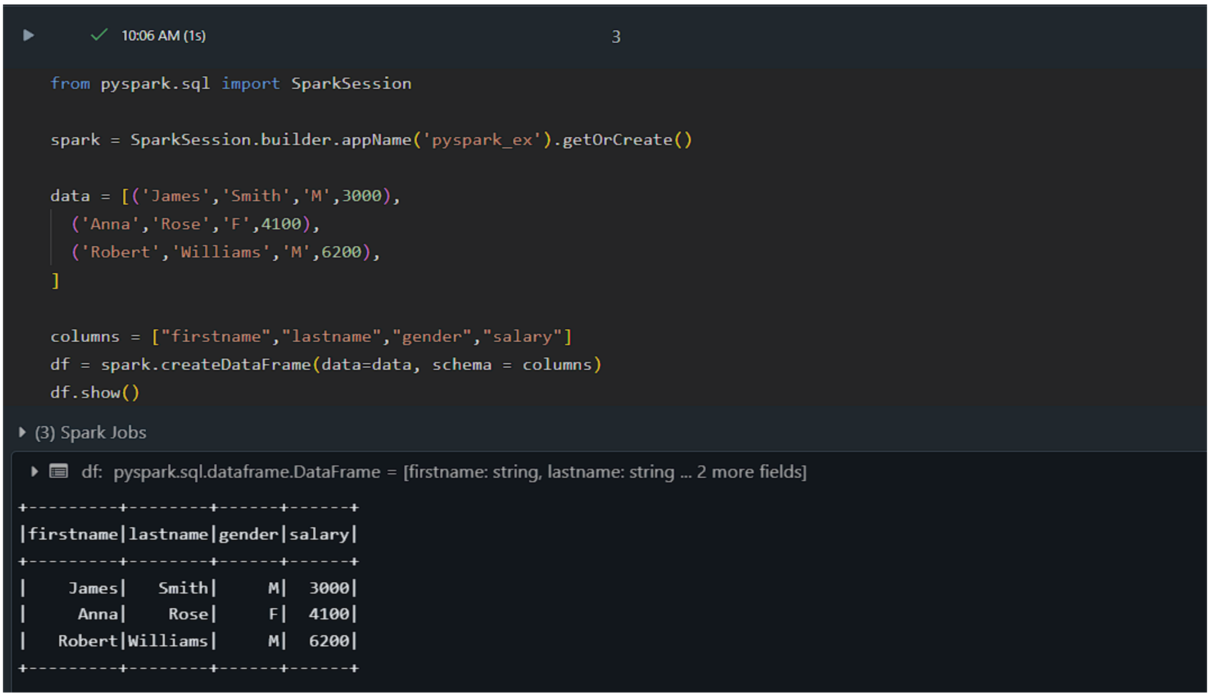


```
#to create rdds and  dataframe
#
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()

#create the rdd

rdd = sc.parallelize([('C',85,76,87,91), ('B',85,76,87,91), ("A", 85,78,96,92), ("A", 92,76,89,96)], 4)
mydata = ['Division','English','Mathematics','Physics','Chemistry']
marks_df = spark.createDataFrame(rdd, schema=mydata)
print(type(marks_df))
print(rdd)
marks_df.show()
marks_df.printSchema()
rdd.collect()
```



marks_df: pyspark.sql.dataframe.DataFrame = [Division: string, English: long ... 3 more fields]

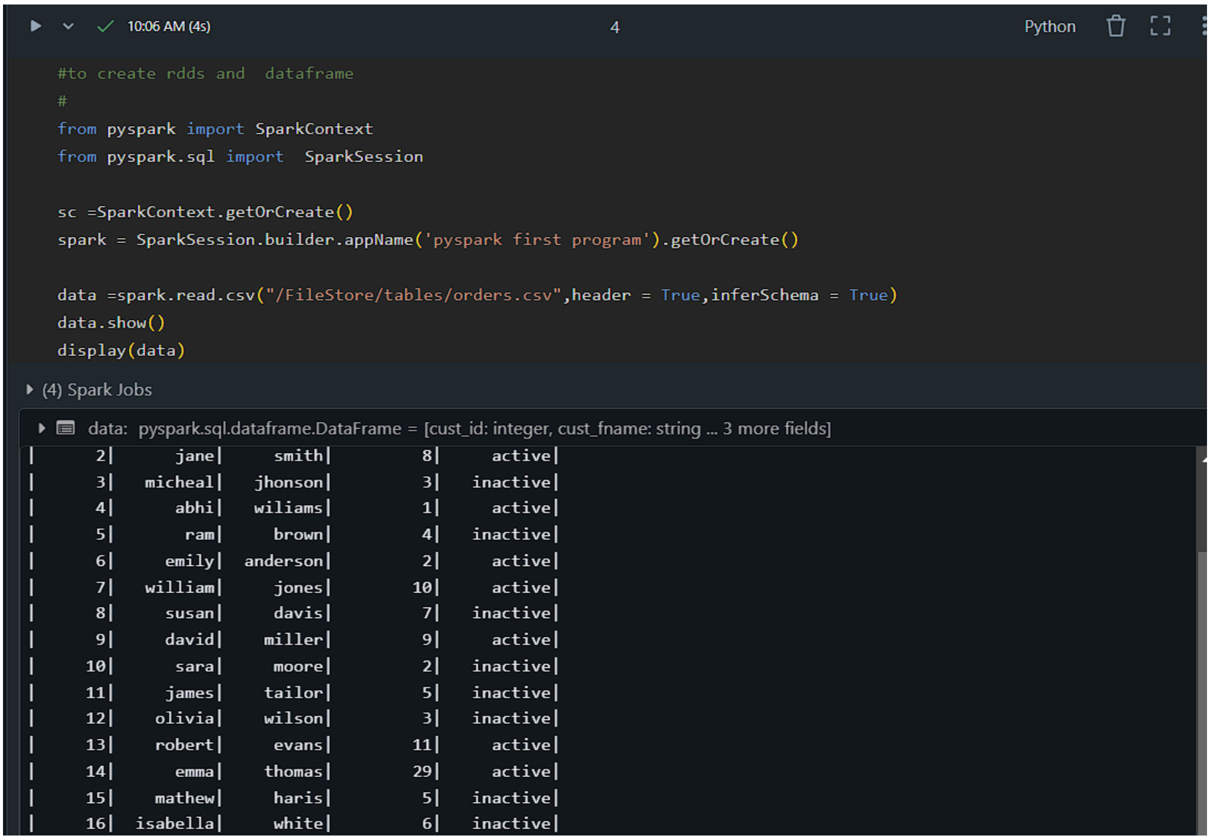
```
ParallelCollectionRDD[22] at readRDDFromInputStream at PythonRDD.scala:435
+--------+-------+-----------+-------+---------+
|Division|English|Mathematics|Physics|Chemistry|
+--------+-------+-----------+-------+---------+
|       C|     85|         76|     87|       91|
|       B|     85|         76|     87|       91|
|       A|     85|         78|     96|       92|
|       A|     92|         76|     89|       96|
+--------+-------+-----------+-------+---------+

root
 |-- Division: string (nullable = true)
 |-- English: long (nullable = true)
 |-- Mathematics: long (nullable = true)
 |-- Physics: long (nullable = true)
 |-- Chemistry: long (nullable = true)

Out[2]: [('C', 85, 76, 87, 91),
 ('B', 85, 76, 87, 91),
 ('A', 85, 78, 96, 92),
 ('A', 92, 76, 89, 96)]
```



```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('pyspark_ex').getOrCreate()

data = [('James','Smith','M',3000),
  ('Anna','Rose','F',4100),
  ('Robert','Williams','M',6200),
]

columns = ["firstname","lastname","gender","salary"]
df = spark.createDataFrame(data=data, schema = columns)
df.show()
```

df: pyspark.sql.dataframe.DataFrame = [firstname: string, lastname: string ... 2 more fields]

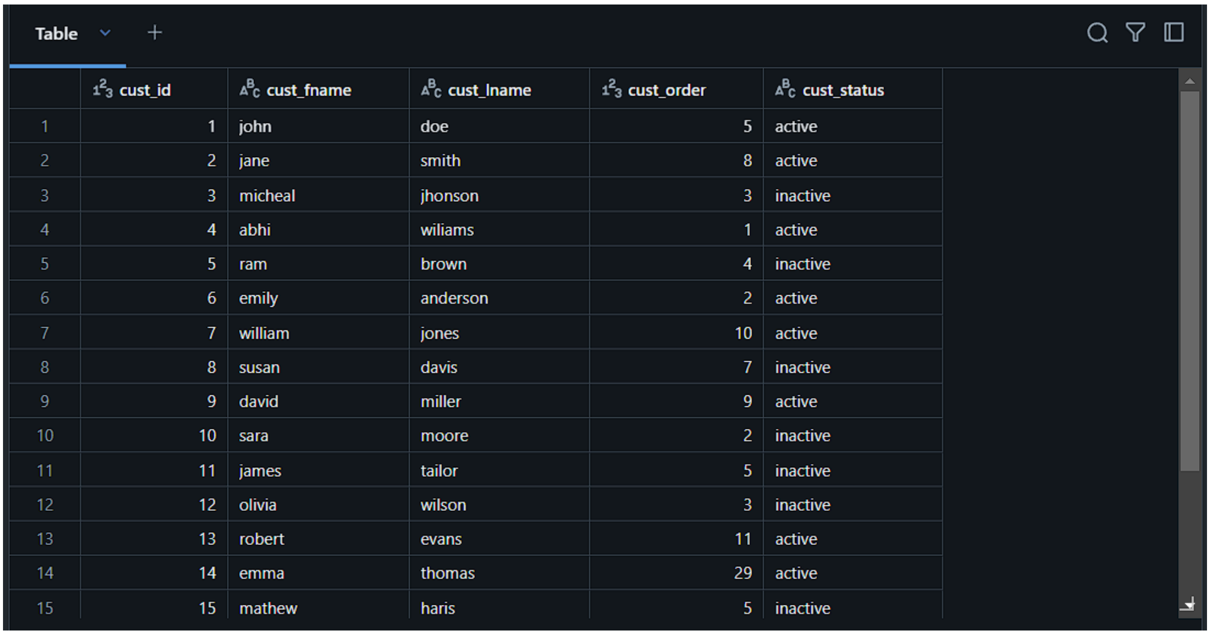
```
+---------+--------+------+------+
|firstname|lastname|gender|salary|
+---------+--------+------+------+
|    James|   Smith|     M|  3000|
|     Anna|    Rose|     F|  4100|
|   Robert|Williams|     M|  6200|
+---------+--------+------+------+
```



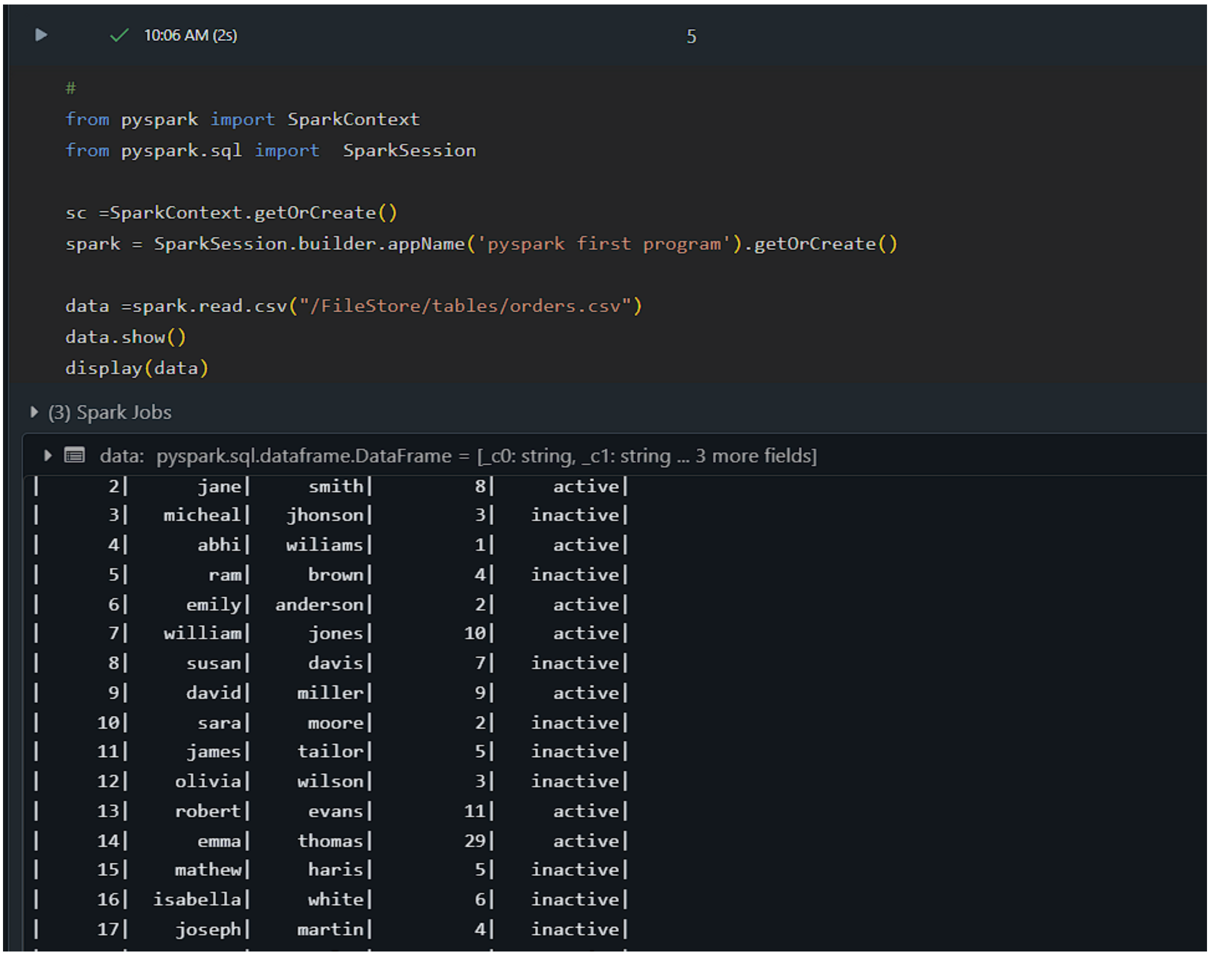
```
#to create rdds and  dataframe
#
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()

data =spark.read.csv("/FileStore/tables/orders.csv",header = True,inferSchema = True)
data.show()
display(data)
```

data: pyspark.sql.dataframe.DataFrame = [cust_id: integer, cust_fname: string ... 3 more fields]

```
|      2|      jane|     smith|          8|  active|
|      3|   micheal|   jhonson|          3|inactive|
|      4|      abhi|   wiliams|          1|  active|
|      5|       ram|     brown|          4|inactive|
|      6|     emily|  anderson|          2|  active|
|      7|   william|     jones|         10|  active|
|      8|     susan|     davis|          7|inactive|
|      9|     david|    miller|          9|  active|
|     10|      sara|     moore|          2|inactive|
|     11|     james|    tailor|          5|inactive|
|     12|    olivia|    wilson|          3|inactive|
|     13|    robert|     evans|         11|  active|
|     14|      emma|    thomas|         29|  active|
|     15|    mathew|     haris|          5|inactive|
|     16|  isabella|     white|          6|inactive|
```



	cust_id	cust_fname	cust_lname	cust_order	cust_status
1	1	john	doe	5	active
2	2	jane	smith	8	active
3	3	micheal	jhonson	3	inactive
4	4	abhi	wiliams	1	active
5	5	ram	brown	4	inactive
6	6	emily	anderson	2	active
7	7	william	jones	10	active
8	8	susan	davis	7	inactive
9	9	david	miller	9	active
10	10	sara	moore	2	inactive
11	11	james	tailor	5	inactive
12	12	olivia	wilson	3	inactive
13	13	robert	evans	11	active
14	14	emma	thomas	29	active
15	15	mathew	haris	5	inactive

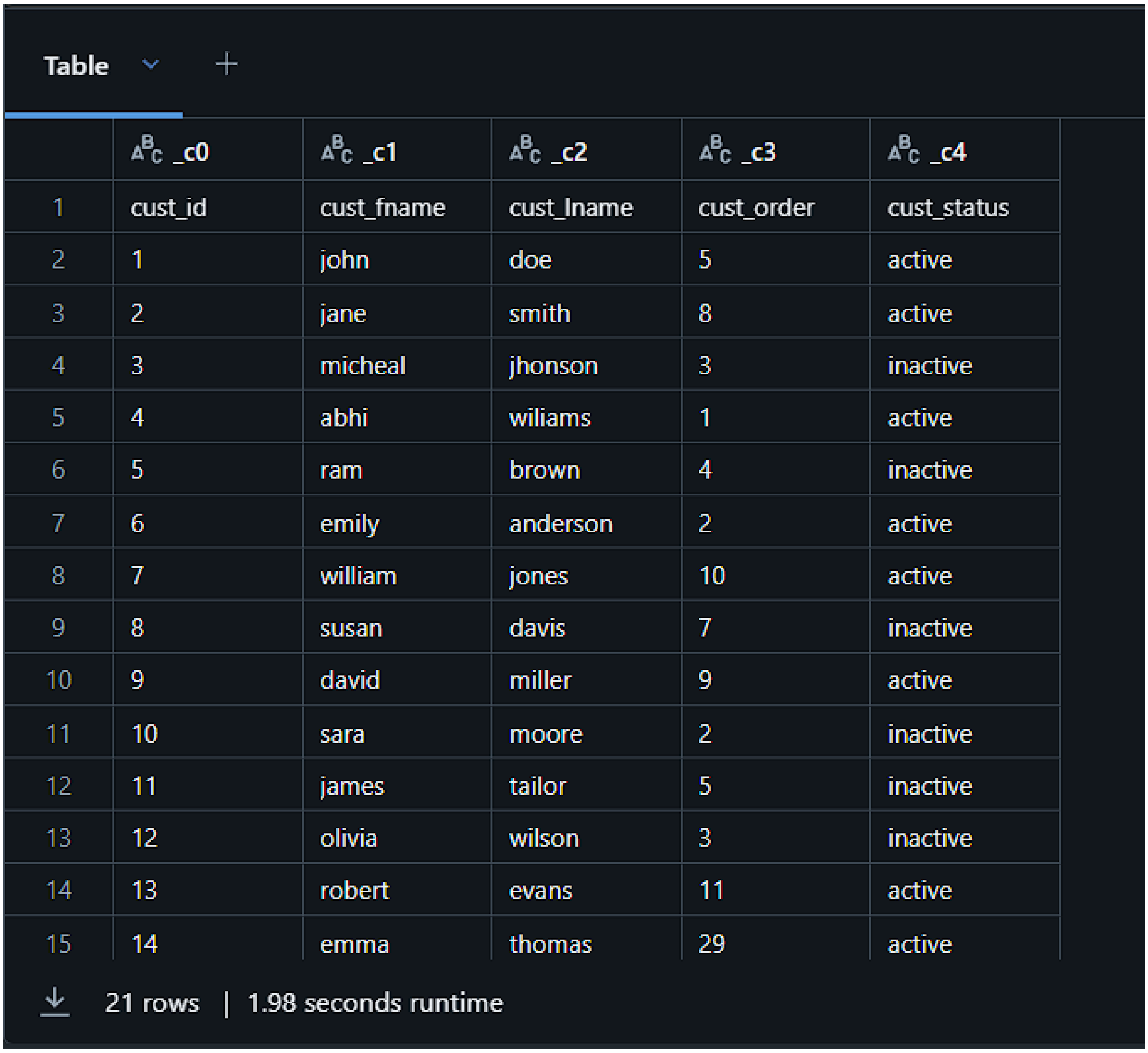


```
#
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()

data =spark.read.csv("/FileStore/tables/orders.csv")
data.show()
display(data)
```

(3) Spark Jobs

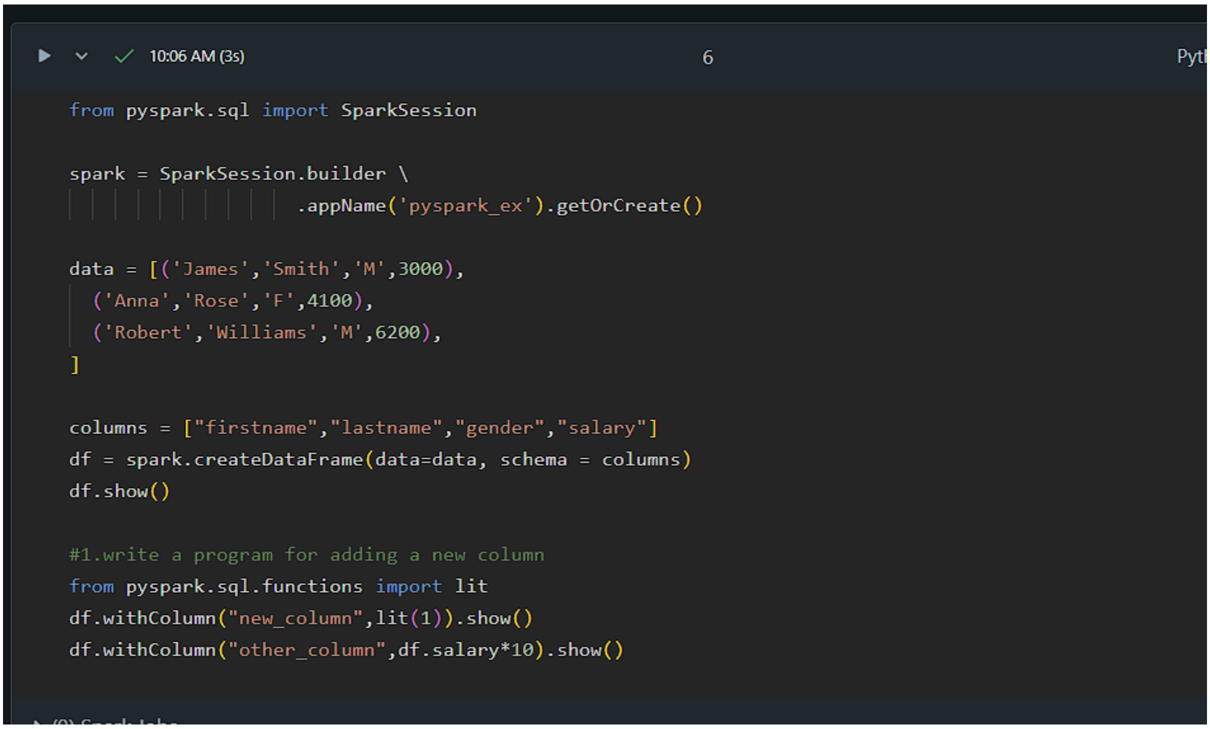
data: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 3 more fields]

```
|      2|    jane|    smith|         8|    active|
|      3| micheal|  jhonson|         3|  inactive|
|      4|    abhi|  wiliams|         1|    active|
|      5|     ram|    brown|         4|  inactive|
|      6|   emily| anderson|         2|    active|
|      7| william|    jones|        10|    active|
|      8|   susan|    davis|         7|  inactive|
|      9|   david|   miller|         9|    active|
|     10|    sara|    moore|         2|  inactive|
|     11|   james|   tailor|         5|  inactive|
|     12|  olivia|   wilson|         3|  inactive|
|     13|  robert|    evans|        11|    active|
|     14|    emma|   thomas|        29|    active|
|     15|  mathew|    haris|         5|  inactive|
|     16|isabella|    white|         6|  inactive|
|     17|  joseph|   martin|         4|  inactive|
```

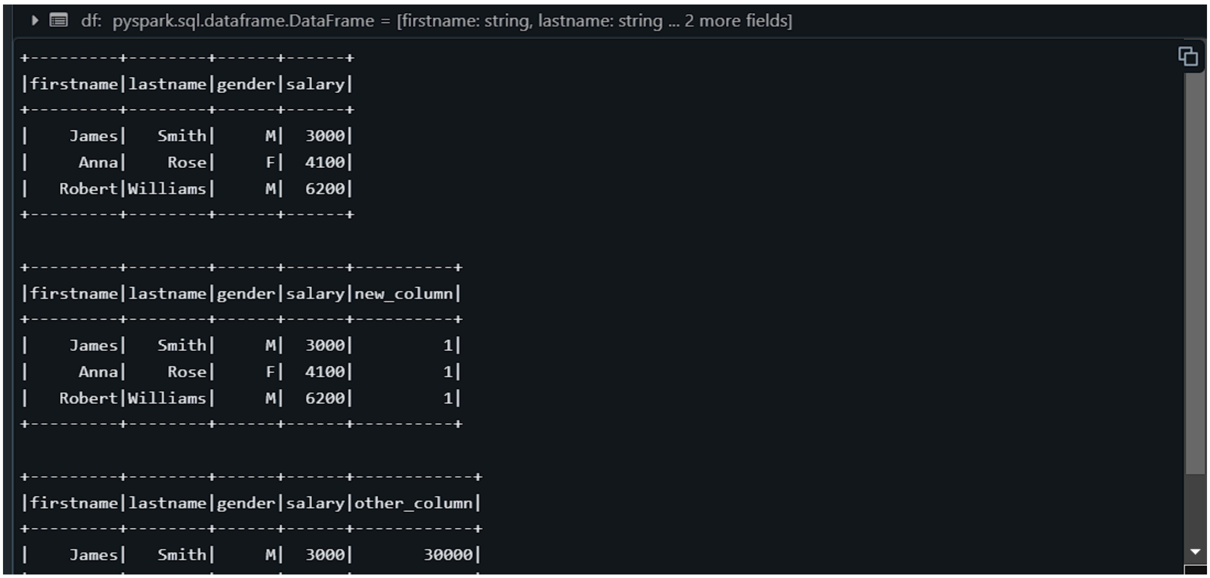


	_c0	_c1	_c2	_c3	_c4
1	cust_id	cust_fname	cust_lname	cust_order	cust_status
2	1	john	doe	5	active
3	2	jane	smith	8	active
4	3	micheal	jhonson	3	inactive
5	4	abhi	wiliams	1	active
6	5	ram	brown	4	inactive
7	6	emily	anderson	2	active
8	7	william	jones	10	active
9	8	susan	davis	7	inactive
10	9	david	miller	9	active
11	10	sara	moore	2	inactive
12	11	james	tailor	5	inactive
13	12	olivia	wilson	3	inactive
14	13	robert	evans	11	active
15	14	emma	thomas	29	active

21 rows | 1.98 seconds runtime

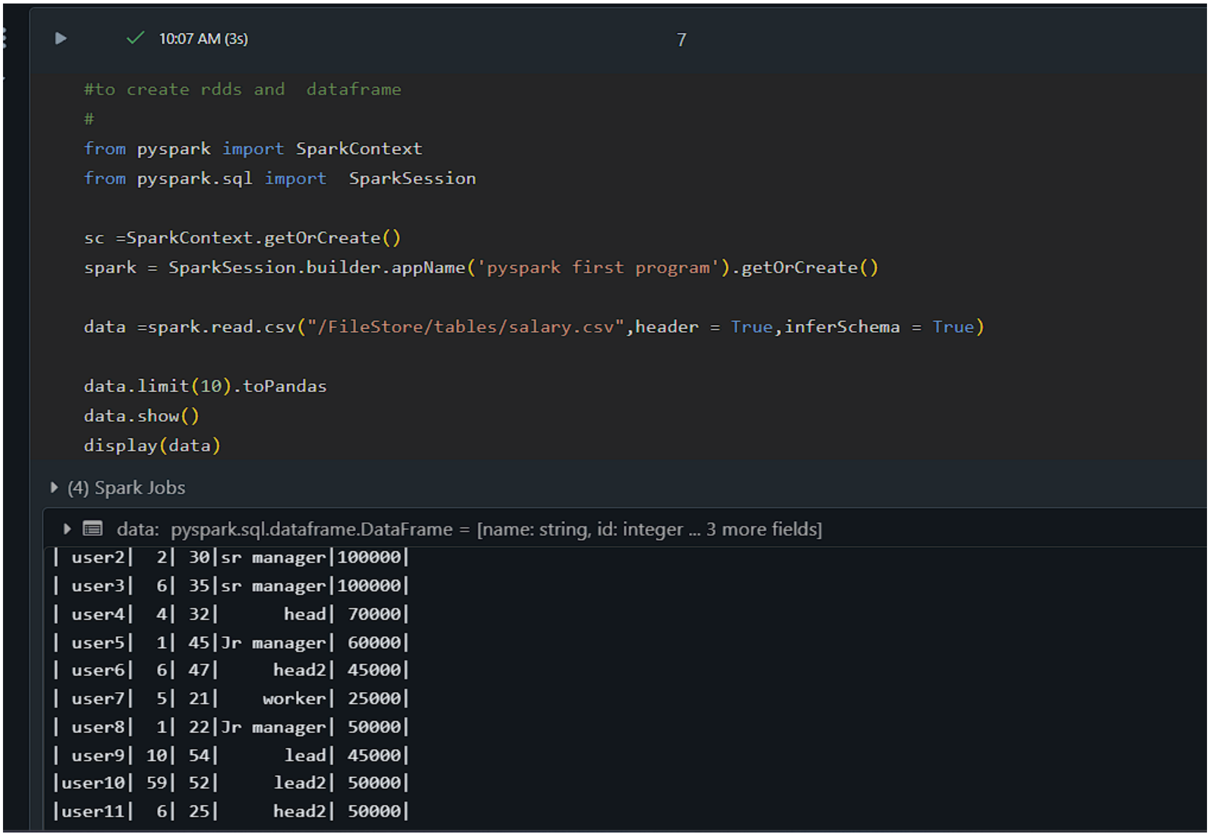


```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
                    .appName('pyspark_ex').getOrCreate()

data = [('James','Smith','M',3000),
  ('Anna','Rose','F',4100),
  ('Robert','Williams','M',6200),
]

columns = ["firstname","lastname","gender","salary"]
df = spark.createDataFrame(data=data, schema = columns)
df.show()

#1.write a program for adding a new column
from pyspark.sql.functions import lit
df.withColumn("new_column",lit(1)).show()
df.withColumn("other_column",df.salary*10).show()
```



df: pyspark.sql.dataframe.DataFrame = [firstname: string, lastname: string ... 2 more fields]

```
+---------+--------+------+------+
|firstname|lastname|gender|salary|
+---------+--------+------+------+
|    James|   Smith|     M|  3000|
|     Anna|    Rose|     F|  4100|
|   Robert|Williams|     M|  6200|
+---------+--------+------+------+

+---------+--------+------+------+----------+
|firstname|lastname|gender|salary|new_column|
+---------+--------+------+------+----------+
|    James|   Smith|     M|  3000|         1|
|     Anna|    Rose|     F|  4100|         1|
|   Robert|Williams|     M|  6200|         1|
+---------+--------+------+------+----------+

+---------+--------+------+------+------------+
|firstname|lastname|gender|salary|other_column|
+---------+--------+------+------+------------+
|    James|   Smith|     M|  3000|       30000|
```



10:07 AM (3s)

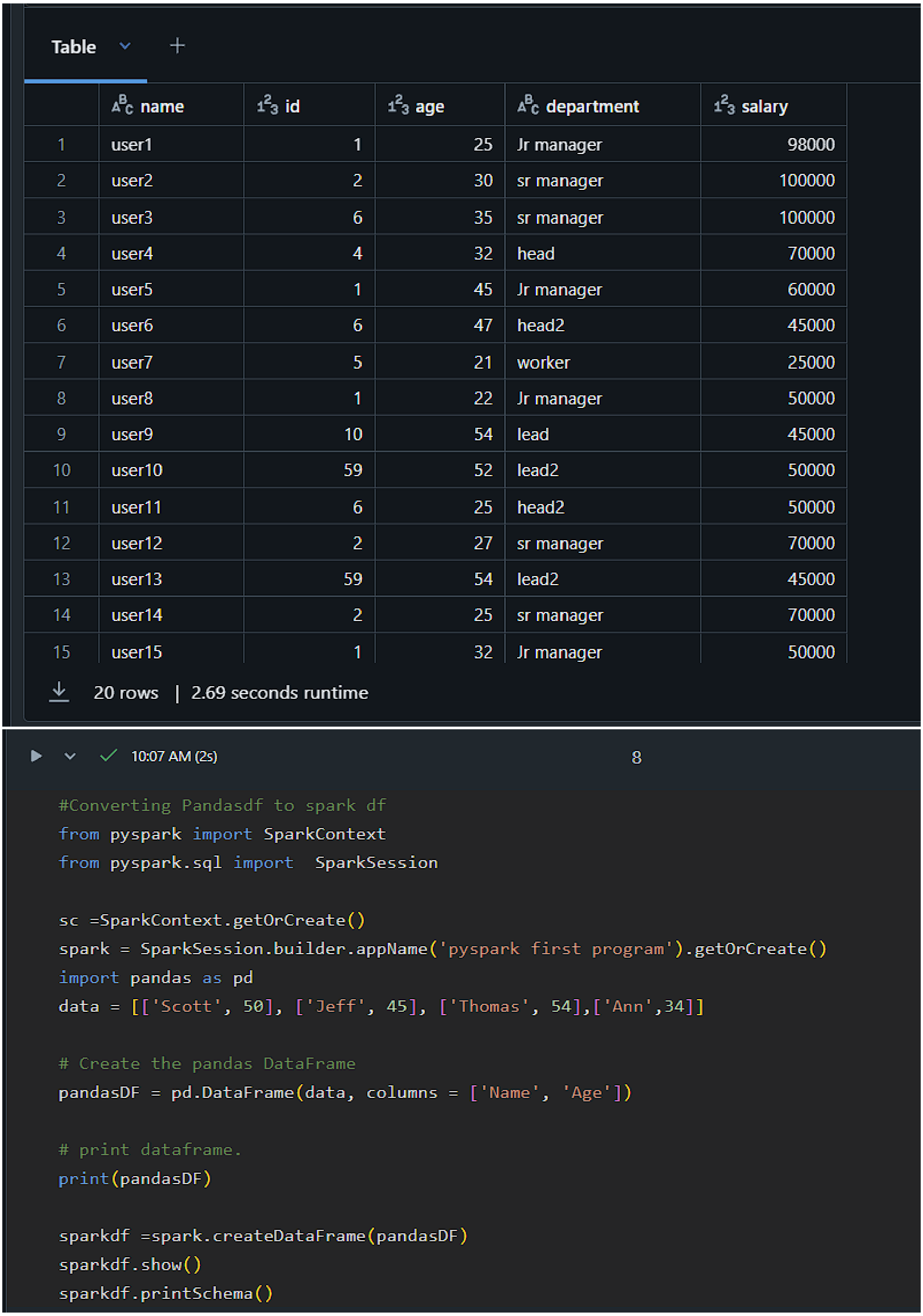
7

```
#to create rdds and  dataframe
#
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()

data =spark.read.csv("/FileStore/tables/salary.csv",header = True,inferSchema = True)

data.limit(10).toPandas
data.show()
display(data)
```

(4) Spark Jobs

data: pyspark.sql.dataframe.DataFrame = [name: string, id: integer ... 3 more fields]

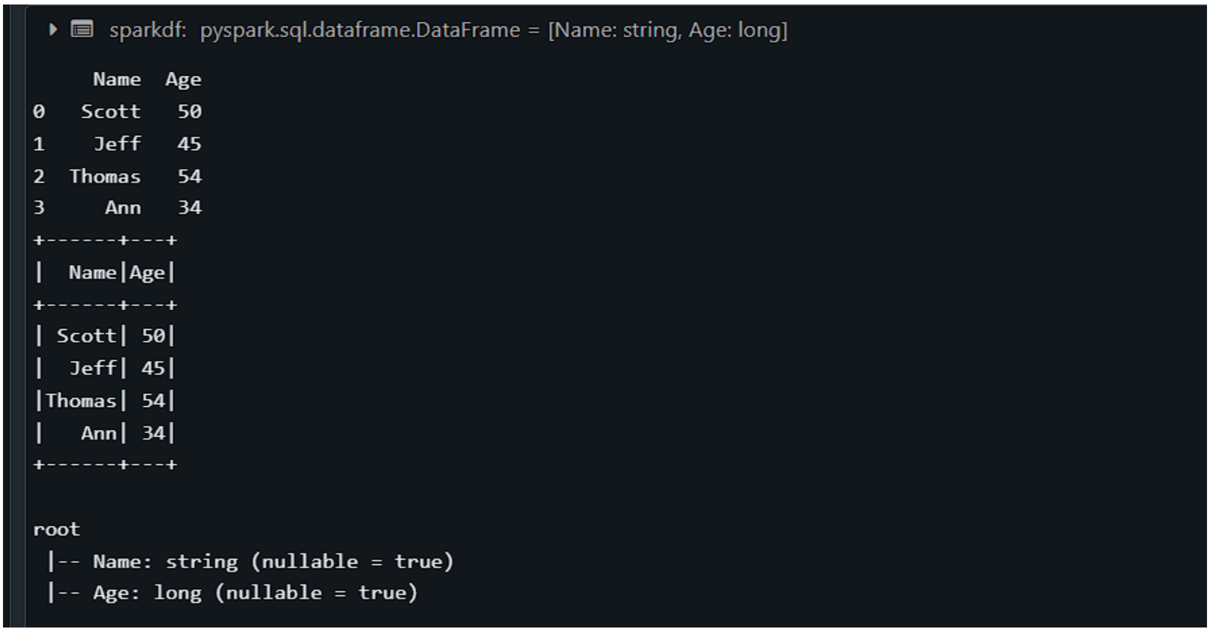
```
| user2|  2| 30|sr manager|100000|
| user3|  6| 35|sr manager|100000|
| user4|  4| 32|      head| 70000|
| user5|  1| 45|Jr manager| 60000|
| user6|  6| 47|     head2| 45000|
| user7|  5| 21|    worker| 25000|
| user8|  1| 22|Jr manager| 50000|
| user9| 10| 54|      lead| 45000|
|user10| 59| 52|     lead2| 50000|
|user11|  6| 25|     head2| 50000|
```



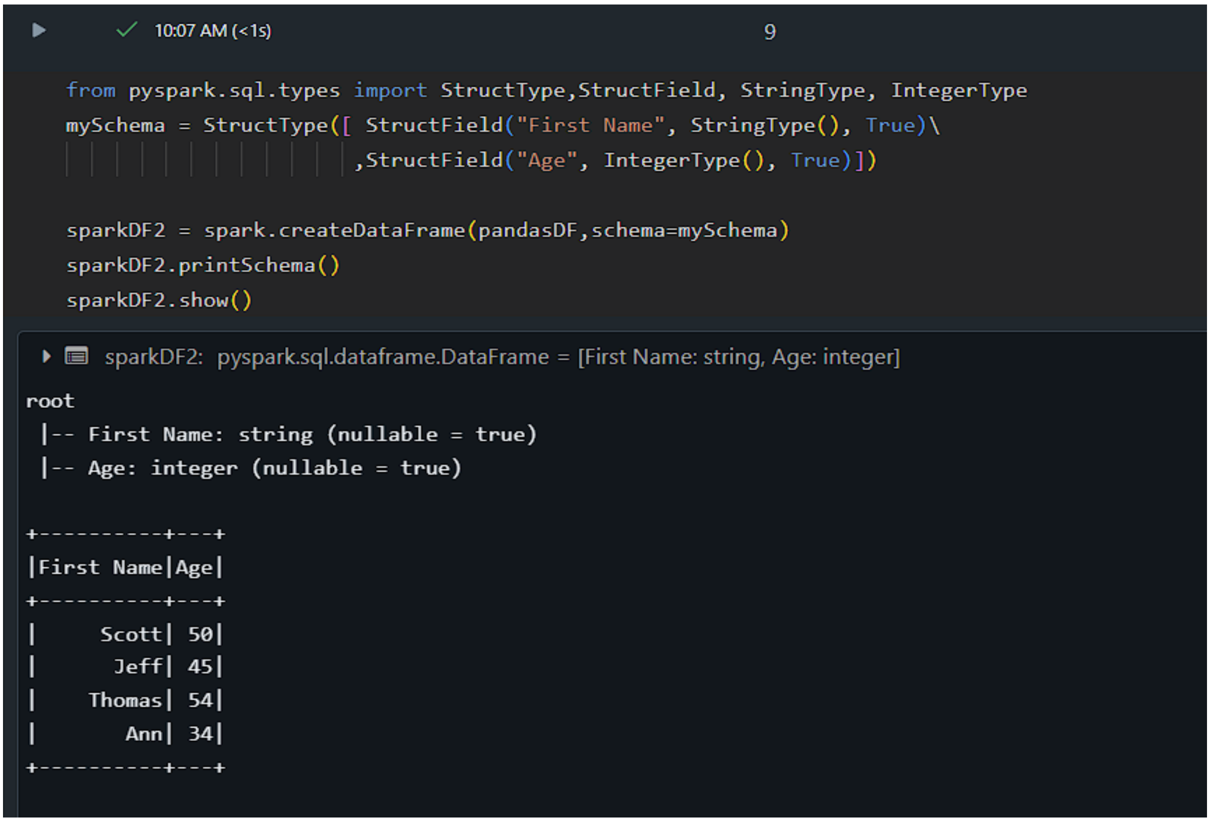
	name	id	age	department	salary
1	user1	1	25	Jr manager	98000
2	user2	2	30	sr manager	100000
3	user3	6	35	sr manager	100000
4	user4	4	32	head	70000
5	user5	1	45	Jr manager	60000
6	user6	6	47	head2	45000
7	user7	5	21	worker	25000
8	user8	1	22	Jr manager	50000
9	user9	10	54	lead	45000
10	user10	59	52	lead2	50000
11	user11	6	25	head2	50000
12	user12	2	27	sr manager	70000
13	user13	59	54	lead2	45000
14	user14	2	25	sr manager	70000
15	user15	1	32	Jr manager	50000

20 rows | 2.69 seconds runtime

```
#Converting Pandasdf to spark df
from pyspark import SparkContext
from pyspark.sql import  SparkSession

sc =SparkContext.getOrCreate()
spark = SparkSession.builder.appName('pyspark first program').getOrCreate()
import pandas as pd
data = [['Scott', 50], ['Jeff', 45], ['Thomas', 54],['Ann',34]]

# Create the pandas DataFrame
pandasDF = pd.DataFrame(data, columns = ['Name', 'Age'])

# print dataframe.
print(pandasDF)

sparkdf =spark.createDataFrame(pandasDF)
sparkdf.show()
sparkdf.printSchema()
```

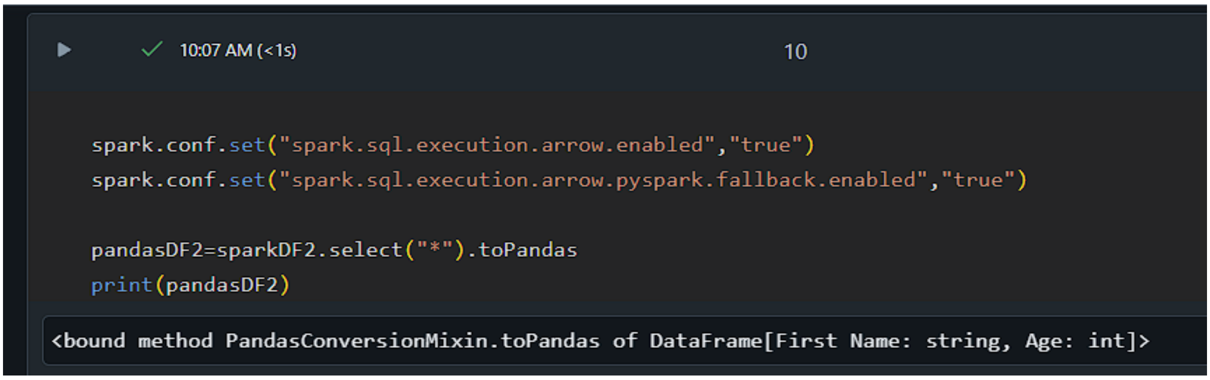


```
sparkdf:  pyspark.sql.dataframe.DataFrame = [Name: string, Age: long]

     Name  Age
0   Scott   50
1    Jeff   45
2  Thomas   54
3     Ann   34
+------+---+
|  Name|Age|
+------+---+
| Scott| 50|
|  Jeff| 45|
|Thomas| 54|
|   Ann| 34|
+------+---+


root
 |-- Name: string (nullable = true)
 |-- Age: long (nullable = true)
```



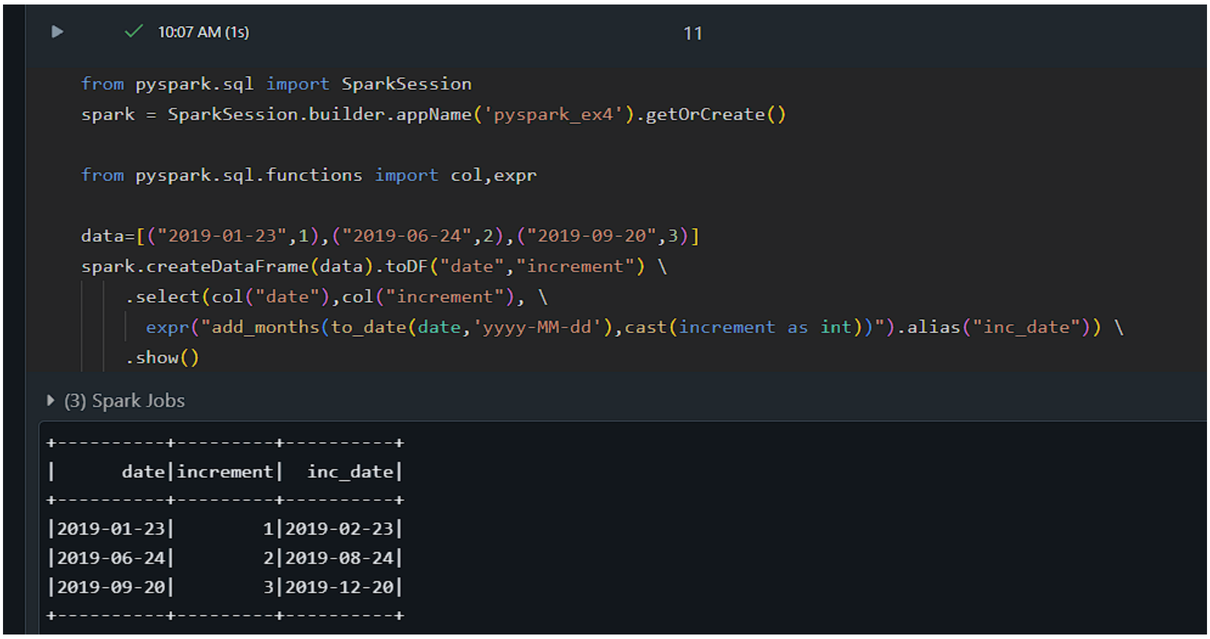
```
from pyspark.sql.types import StructType,StructField, StringType, IntegerType
mySchema = StructType([ StructField("First Name", StringType(), True)\
                        ,StructField("Age", IntegerType(), True)])

sparkDF2 = spark.createDataFrame(pandasDF,schema=mySchema)
sparkDF2.printSchema()
sparkDF2.show()
```

```
sparkDF2:  pyspark.sql.dataframe.DataFrame = [First Name: string, Age: integer]
root
 |-- First Name: string (nullable = true)
 |-- Age: integer (nullable = true)

+----------+---+
|First Name|Age|
+----------+---+
|     Scott| 50|
|      Jeff| 45|
|    Thomas| 54|
|       Ann| 34|
+----------+---+
```



```
spark.conf.set("spark.sql.execution.arrow.enabled","true")
spark.conf.set("spark.sql.execution.arrow.pyspark.fallback.enabled","true")

pandasDF2=sparkDF2.select("*").toPandas
print(pandasDF2)
```

```
<bound method PandasConversionMixin.toPandas of DataFrame[First Name: string, Age: int]>
```



```python
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('pyspark_ex4').getOrCreate()

from pyspark.sql.functions import col,expr

data=[("2019-01-23",1),("2019-06-24",2),("2019-09-20",3)]
spark.createDataFrame(data).toDF("date","increment") \
    .select(col("date"),col("increment"), \
      expr("add_months(to_date(date,'yyyy-MM-dd'),cast(increment as int))").alias("inc_date")) \
    .show()
```

```
+----------+---------+----------+
|      date|increment|  inc_date|
+----------+---------+----------+
|2019-01-23|        1|2019-02-23|
|2019-06-24|        2|2019-08-24|
|2019-09-20|        3|2019-12-20|
+----------+---------+----------+
```