What is PySpark?

- PySpark is the Python API for Apache Spark, an open-source distributed computing framework.
- It is used for big data processing, offering support for data analysis, machine learning, and stream processing.
- PySpark provides high-level abstractions like Resilient Distributed Datasets (RDDs), DataFrames, and SQL APIs.

Initiating a Spark Session

- A Spark session provides an entry point for interacting with Spark functionality.
- To create a Spark session in PySpark:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('Pyspark first program').getOrCreate()
```

  - appName: Sets the name of the application.
  - Use .getOrCreate() to reuse an existing session if available.

Spark Context

- The Spark Context (sc) is the core abstraction in PySpark, responsible for managing the connection to the Spark cluster.
- It can be accessed from a Spark session:

  sc = spark.sparkContext

Creating an RDD in PySpark

- What is an RDD?
  - RDD (Resilient Distributed Dataset) is the fundamental data structure in PySpark, providing fault tolerance and parallel operations. There are two ways to create dataframe

1. **From a Collection:**

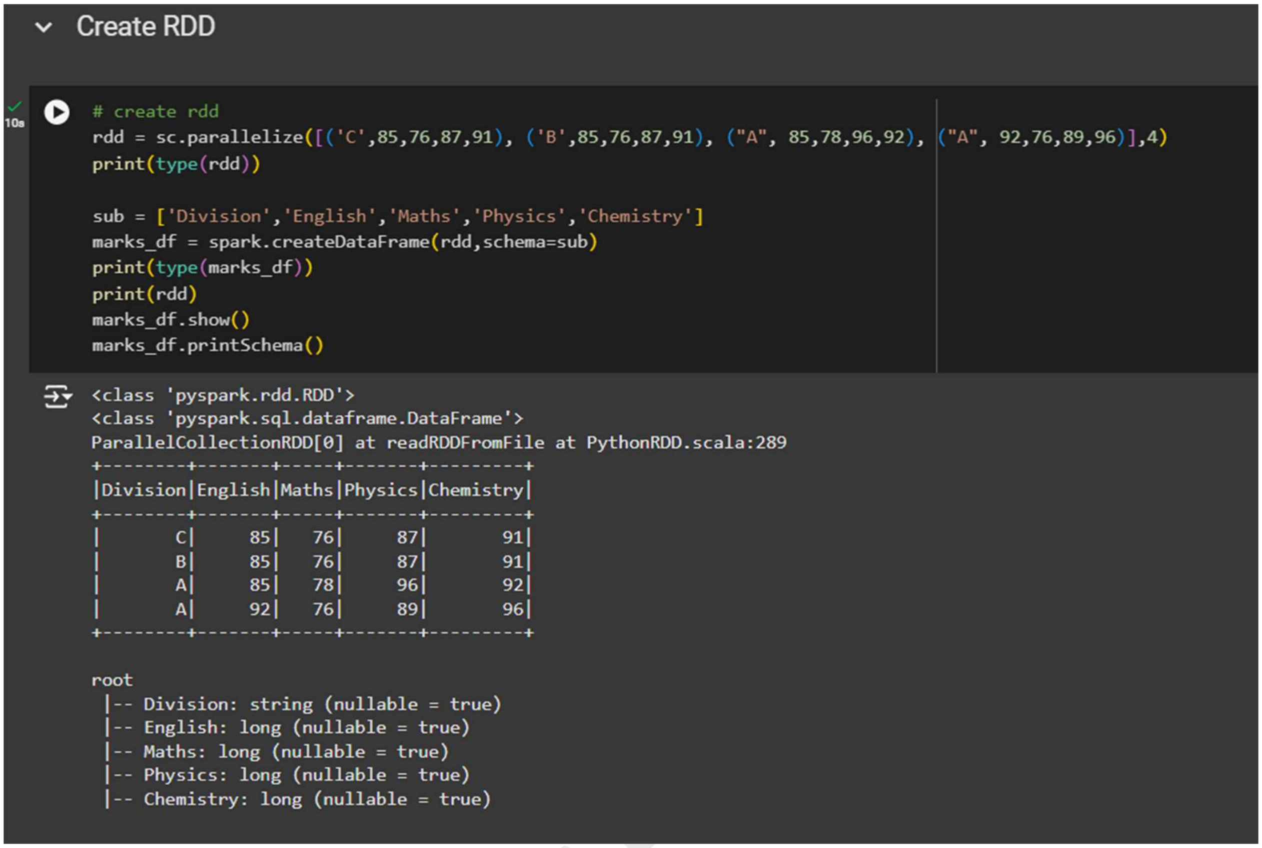
data = [1, 2, 3, 4, 5]

rdd = sc.parallelize(data)

2. **From an External File:**

rdd = sc.textFile("path/to/file.txt")

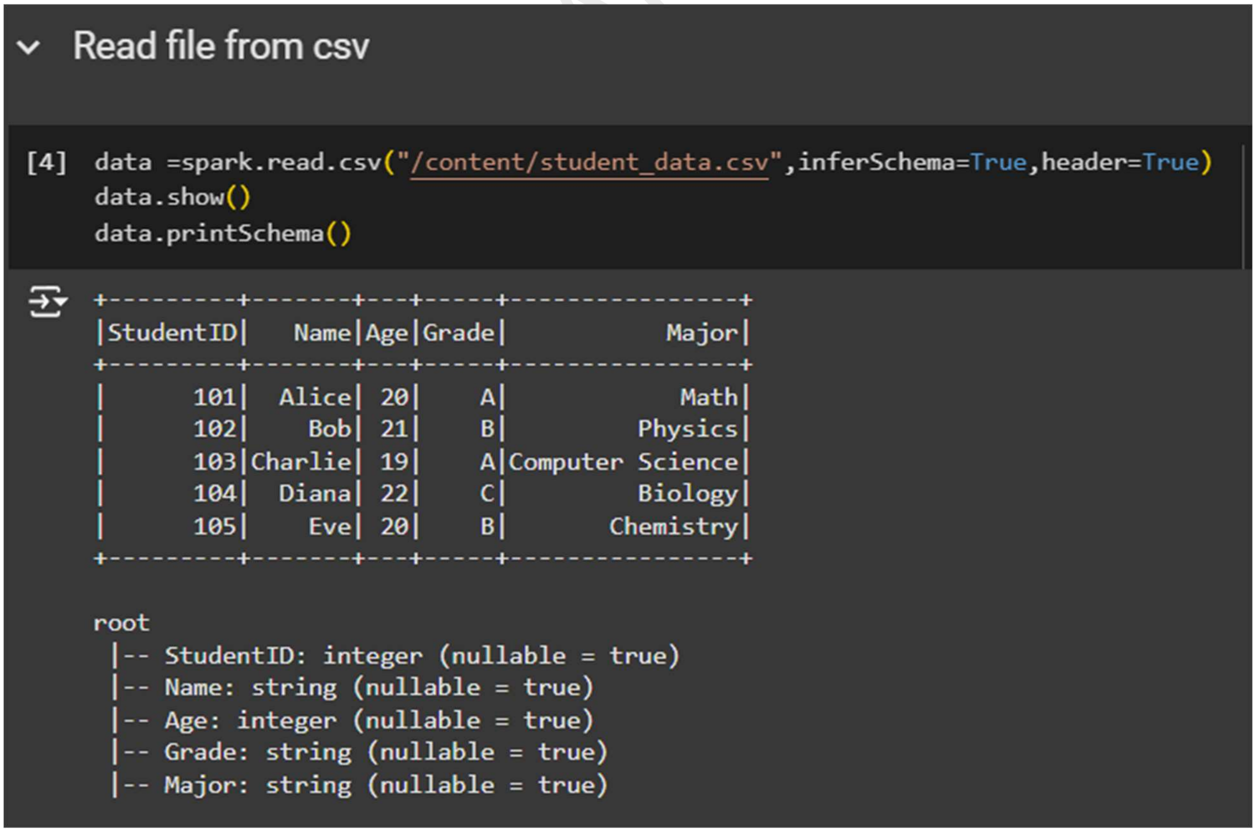
Basic RDD Operations: Transformations: return a new RDD. Actions: return results.



Create RDD

```
# create 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)
print(type(rdd))

sub = ['Division','English','Maths','Physics','Chemistry']
marks_df = spark.createDataFrame(rdd,schema=sub)
print(type(marks_df))
print(rdd)
marks_df.show()
marks_df.printSchema()
```

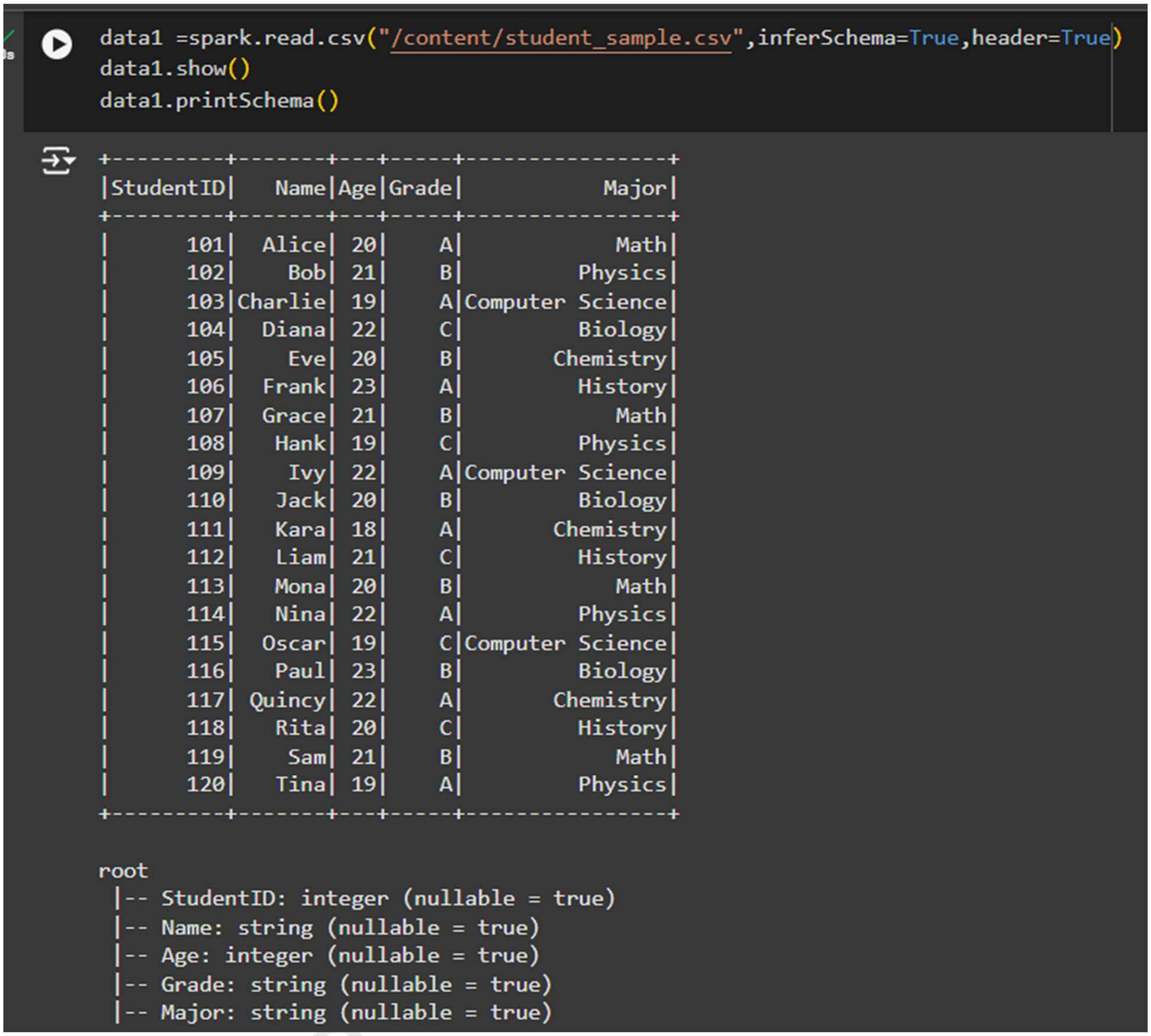
```
<class 'pyspark.rdd.RDD'>
<class 'pyspark.sql.dataframe.DataFrame'>
ParallelCollectionRDD[0] at readRDDFromFile at PythonRDD.scala:289
+--------+-------+-----+-------+---------+
|Division|English|Maths|Physics|Chemistry|
+--------+-------+-----+-------+---------+
|       C|     85|   76|     87|       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)
 |-- Maths: long (nullable = true)
 |-- Physics: long (nullable = true)
 |-- Chemistry: long (nullable = true)
```



Read file from csv

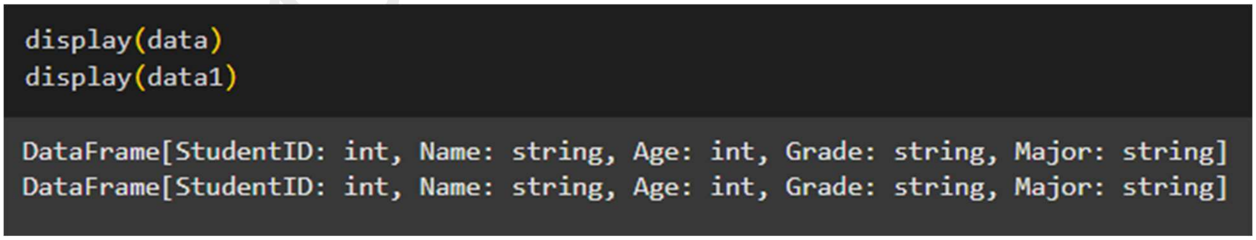
```
[4] data =spark.read.csv("/content/student_data.csv",inferSchema=True,header=True)
    data.show()
    data.printSchema()
```

```
+---------+-------+---+-----+----------------+
|StudentID|   Name|Age|Grade|           Major|
+---------+-------+---+-----+----------------+
|      101|  Alice| 20|    A|            Math|
|      102|    Bob| 21|    B|         Physics|
|      103|Charlie| 19|    A|Computer Science|
|      104|  Diana| 22|    C|         Biology|
|      105|    Eve| 20|    B|       Chemistry|
+---------+-------+---+-----+----------------+

root
 |-- StudentID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Grade: string (nullable = true)
 |-- Major: string (nullable = true)
```



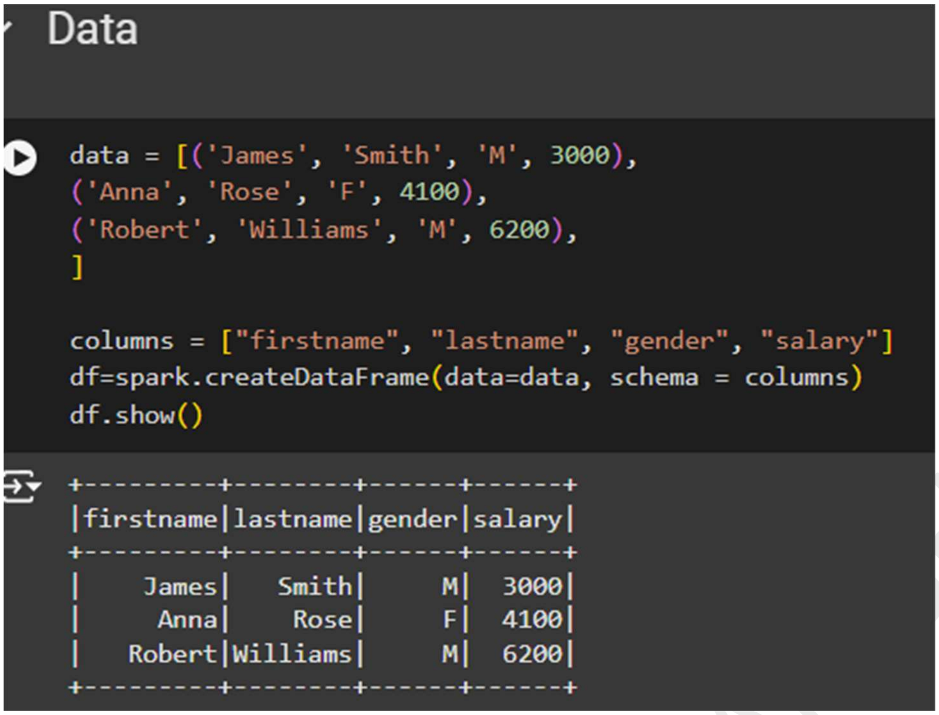
```
data1 =spark.read.csv("/content/student_sample.csv",inferSchema=True,header=True)
data1.show()
data1.printSchema()
```

```
+---------+-------+---+-----+----------------+
|StudentID|   Name|Age|Grade|           Major|
+---------+-------+---+-----+----------------+
|      101|  Alice| 20|    A|            Math|
|      102|    Bob| 21|    B|         Physics|
|      103|Charlie| 19|    A|Computer Science|
|      104|  Diana| 22|    C|         Biology|
|      105|    Eve| 20|    B|       Chemistry|
|      106|  Frank| 23|    A|         History|
|      107|  Grace| 21|    B|            Math|
|      108|   Hank| 19|    C|         Physics|
|      109|    Ivy| 22|    A|Computer Science|
|      110|   Jack| 20|    B|         Biology|
|      111|   Kara| 18|    A|       Chemistry|
|      112|   Liam| 21|    C|         History|
|      113|   Mona| 20|    B|            Math|
|      114|   Nina| 22|    A|         Physics|
|      115|  Oscar| 19|    C|Computer Science|
|      116|   Paul| 23|    B|         Biology|
|      117| Quincy| 22|    A|       Chemistry|
|      118|   Rita| 20|    C|         History|
|      119|    Sam| 21|    B|            Math|
|      120|   Tina| 19|    A|         Physics|
+---------+-------+---+-----+----------------+

root
 |-- StudentID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Grade: string (nullable = true)
 |-- Major: string (nullable = true)
```



```
display(data)
display(data1)
```

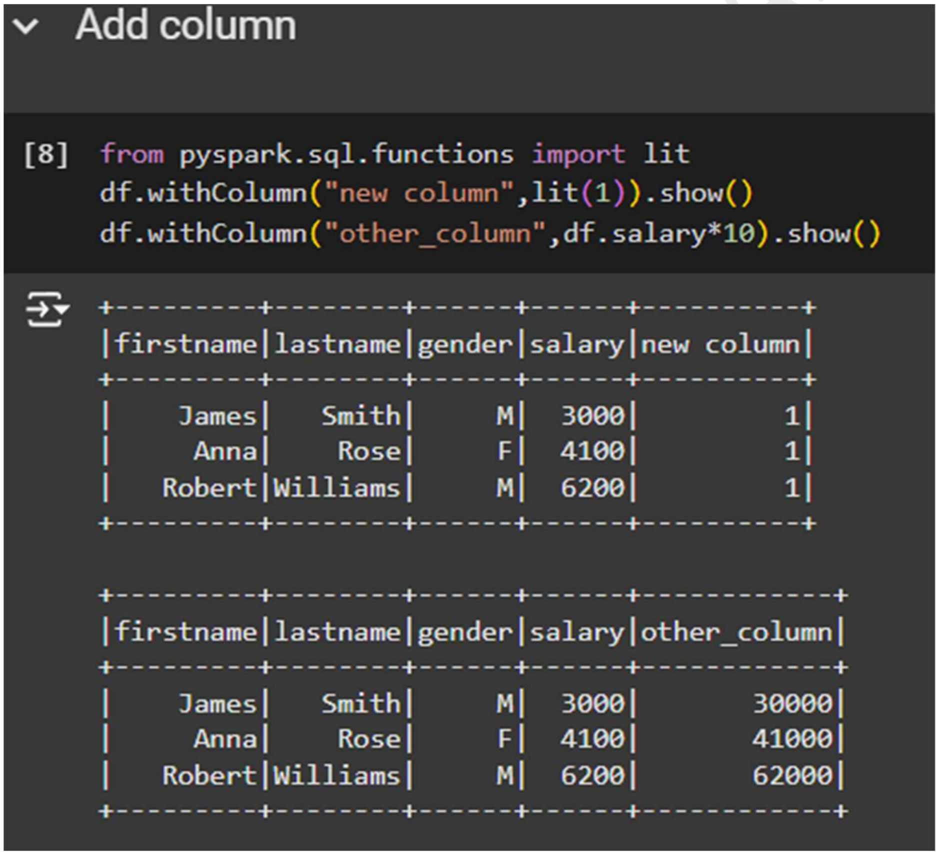
```
DataFrame[StudentID: int, Name: string, Age: int, Grade: string, Major: string]
DataFrame[StudentID: int, Name: string, Age: int, Grade: string, Major: string]
```



Data

```
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()
```

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



Add column

```
[8] from pyspark.sql.functions import lit
    df.withColumn("new column",lit(1)).show()
    df.withColumn("other_column",df.salary*10).show()
```

```
+---------+--------+------+------+----------+
|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|
|     Anna|    Rose|     F|  4100|       41000|
|   Robert|Williams|     M|  6200|       62000|
+---------+--------+------+------+------------+
```