BDA_Lab3

Name :- Kunal Sanjay Patil

PRN :- 20190802025

```
!pip install pyspark

     Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.2
     Requirement already satisfied: py4j==0.10.9.3 in /usr/local/lib/python3.7/dist-packag
```

```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
```

```
collect_rdd = sc.parallelize([5,6,7,8,9])
print(collect_rdd.collect())

     [5, 6, 7, 8, 9]
```

```
count_rdd = sc.parallelize([3,6,0,3,1,8])
print(count_rdd.count())

     6
```

```
take_rdd = sc.parallelize([4,7,8,9,1,5])
print(take_rdd.take(3))

     [4, 7, 8]
```

```
reduce_rdd = sc.parallelize([6,5,4,8])
print(reduce_rdd.reduce(lambda x, y : x + y))

     23
```

```
save_rdd = sc.parallelize([1,2,3,4,5,6])
save_rdd.saveAsTextFile('KP1.txt')
```

```
my_rdd = sc.parallelize([7,8,9,6])
print(my_rdd.map(lambda x: x+ 10).collect())

     [17, 18, 19, 16]
```

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

     ['This', 'is', 'PySpark', 'RDD', 'Transformations']
```

```
filter_rdd = sc.parallelize([8,9,7,4,2,3])
print(filter_rdd.filter(lambda x: x%2 == 0).collect())
```

```
[8, 4, 2]
```

```
marks_rdd = sc.parallelize([('Mahi', 25), ('Harsh', 26), ('Shreya', 22), ('Neel', 29), ('R
print(marks_rdd.reduceByKey(lambda x, y: x + y).collect())
```

```
[('Shreya', 50), ('Rohan', 44), ('Rahul', 23), ('Mahi', 25), ('Harsh', 26), ('Neel',
```

```
marks_rdd = sc.parallelize([('Mmahi', 25), ('Harsh', 26), ('Shreya', 22), ('Neel', 29), ('
dict_rdd = marks_rdd.groupByKey().collect()
for key, value in dict_rdd:
    print(key, list(value))
```

```
Mmahi [25]
Shreya [22, 28]
Rohan [22, 22]
Rahul [23]
Harsh [26]
Neel [29]
Swati [19]
Abhay [26]
```

```
marks_rdd = sc.parallelize([('Mahi', 25), ('Harsh', 26), ('Shreya', 22), ('Neel', 29), ('R
print(marks_rdd.sortByKey('ascending').collect())
```

```
[('Abhay', 26), ('Harsh', 26), ('Mahi', 25), ('Neel', 29), ('Rahul', 23), ('Rohan',
```