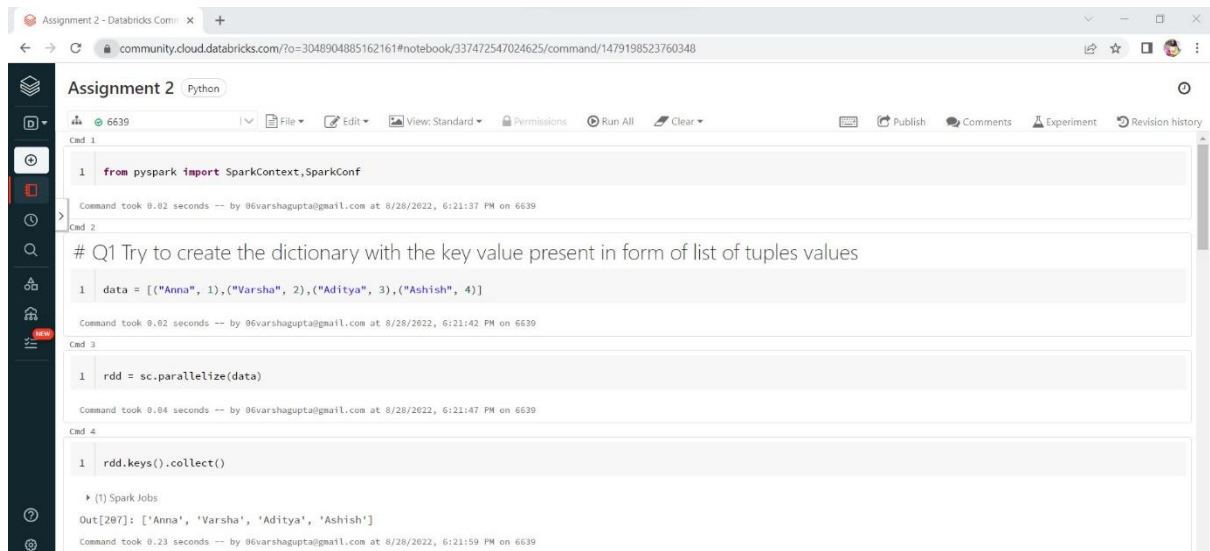


Task-19

Name: Varsha Gupta

Referral id: DIRSS3129

Q1. Try to create the dictionary with the key value present in form of list of tuples values



The screenshot shows a Databricks notebook titled "Assignment 2" in Python. It contains four commands (Cmd 1 to Cmd 4) and their outputs. Cmd 1 imports SparkContext and SparkConf. Cmd 2 defines a list of tuples 'data'. Cmd 3 creates an RDD from 'data'. Cmd 4 collects the keys of the RDD, resulting in the output: Out[287]: ['Anna', 'Varsha', 'Aditya', 'Ashish'].

```
1 from pyspark import SparkContext, SparkConf

Command took 0.02 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 6:21:37 PM on 6639

Cmd 2
# Q1 Try to create the dictionary with the key value present in form of list of tuples values

1 data = [("Anna", 1), ("Varsha", 2), ("Aditya", 3), ("Ashish", 4)]

Command took 0.02 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 6:21:42 PM on 6639

Cmd 3
1 rdd = sc.parallelize(data)

Command took 0.04 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 6:21:47 PM on 6639

Cmd 4
1 rdd.keys().collect()

Out[287]: ['Anna', 'Varsha', 'Aditya', 'Ashish']

Command took 0.23 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 6:21:59 PM on 6639
```

Q2. reduceByKey & groupByKey - cost taking operation



The screenshot shows a Databricks notebook with two commands (Cmd 5 and Cmd 6) and their outputs. Cmd 5 collects the values of the RDD from the previous command, resulting in the output: Out[288]: [1, 2, 3, 4]. Cmd 6 defines a list 'List' for Q2. Cmd 7 creates an RDD from 'List'.

```
1 rdd.values().collect()

Out[288]: [1, 2, 3, 4]

Command took 0.22 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 6:22:08 PM on 6639

Cmd 6
# Q2 reduceByKey & groupByKey - cost taking operation

1 List = [('alex', 2), ('alex', 4), ('alex', 8), ('jane', 3), ('jane', 7), ('rafa', 1), ('rafa', 3), ('rafa', 5), ('rafa', 6), ('clint', 9)]

Command took 0.02 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 7:55:14 PM on 6639

Cmd 7
1 rdd1=spark.sparkContext.parallelize(List)

Command took 0.04 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 7:55:20 PM on 6639
```

+

📄

🕒

🔍

📊

🔗

NEW

⚙️

👤

📁

```

1 rdd2=rdd1.reduceByKey(lambda a,b: a+b)
2 for element in rdd2.collect():
3     print(element)

```

▶ (1) Spark Jobs

```

('clint', 9)
('alex', 14)
('jane', 10)
('rafa', 15)

```

Command took 0.85 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 7:55:26 PM on 6639

Cmd 9

```

1 new = rdd1.groupByKey()
2 new.collect()

```

▶ (1) Spark Jobs

```

Out[223]: [('clint', <pyspark.resultiterable.ResultIterable at 0x7f1cc82f3040>),
('alex', <pyspark.resultiterable.ResultIterable at 0x7f1cc82fe730>),
('jane', <pyspark.resultiterable.ResultIterable at 0x7f1cc82fe760>),
('rafa', <pyspark.resultiterable.ResultIterable at 0x7f1cc82fe7f0>)]

```

Command took 0.37 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 8:02:39 PM on 6639

Cmd 10

```

1 rdd1.groupByKey().mapValues(len).collect()

```

▶ (1) Spark Jobs

```

Out[224]: [('clint', 1), ('alex', 3), ('jane', 2), ('rafa', 4)]

```

Command took 0.39 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 8:02:55 PM on 6639

Q3. For every age try to find out how many friends are present but using countByValue operation

🔍

📄

🕒

🔍

📊

🔗

NEW

⚙️

👤

📁

Cmd 11

```

# Q3 For every age try to find out how many friends are present but using countByValue operation
1 data_rdd = sc.textFile("dbfs:/FileStore/learning/FriendsData.csv")

```

Command took 0.06 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 8:12:33 PM on 6639

Cmd 12

```

1 def transformData(lines):
2     fields = lines.split(",")
3     pid = fields[0]
4     p_name = fields[1]
5     p_age = fields[2]
6     p_friends = fields[3]
7
8     return (p_age,p_friends)

```

Command took 0.03 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 9:15:05 PM on 6639

Cmd 13

```
> 1 gen = data_rdd.map(lambda lines: lines.split(',')[2])  
2 gen.countByValue()
```

▶ (1) Spark Jobs

```
Out[302]: defaultdict(int,  
    {'50': 17,  
     '26': 25,  
     '55': 13,  
     '40': 17,  
     '68': 22,  
     '59': 9,  
     '37': 9,  
     '54': 13,  
     '38': 15,  
     '53': 7,  
     '56': 6,  
     '43': 7,  
     '36': 10,  
     '22': 7,  
     '35': 8,  
     '45': 13,  
     '60': 7,  
     '67': 16,  
     '19': 11,  
     '30': 11,  
     '51': 7,  
     '25': 11,  
     '21': 8,  
     '42': 6,  
     '49': 6,  
     '48': 10,  
     '39': 7,  
     '32': 11,  
     '58': 11,  
     '64': 12,  
     '31': 8,  
     '52': 11,  
     '24': 5,  
     '20': 5,  
     '62': 13,  
     '41': 9,  
     '44': 12,  
     '69': 10,  
     '65': 5,  
     '61': 9,  
     '28': 10,  
     '66': 9,  
     '46': 13,  
     '29': 12,  
     '18': 8,  
     '47': 9,  
     '34': 6,  
     '63': 4,  
     '23': 10})
```

Command took 0.27 seconds -- by 06varshagupta@gmail.com at 8/28/2022, 10:18:36 PM on 6639