POPU

Date 01/02/2021

Topic

Page No. 12

Experiment-4

Q. Transformations

1. Remove data file from Databricks file system.

→ dbutils.fs.rm("FileStore/tables/data.txt")

2. FlatMap.

→
```
data = sc.textFile("FileStore/tables/mydata.txt")
rdd = data.flatMap(lambda x: x.split(" "))
rdd.collect()
```

3. Filter

→
```
ls = [1,2,3,4,5]
data = sc.parallelize(ls)
f = data.filter(lambda x: x!=5)
f.collect()
```

4. Distinct

→
```
data = sc.textFile("FileStore/tables/mydata.txt")
rdd = data.flatMap(lambda x: x.split(" ")).distinct()
rdd.collect()
```

5. MapValues

→
```
rdd = sc.parallelize([(1,2),(2,4),(3,5),(4,7),(5,8),
                      (6,23),(7,45)])
rdd.mapValues(lambda x: x+1).collect()
```

Teacher's Signature .....

6. flatMapValues

→
```
data = sc.textFile("FileStore/tables/user_address.txt")
a = data.map(lambda x: x.split("\t"))
b = a.map(lambda x: (x[0], x[1]))
c = b.flatMapValues(lambda x: x.split(","))
c.collect()
```

7. Key

→
```
data = sc.textFile("FileStore/tables/spark.txt")
a = data.flatMap(lambda x: x.split(" "))
b = a.map(lambda x: (x,1)).keys()
b.collect()
```

8. Values

→
```
data = sc.textFile("FileStore/tables/spark.txt")
a = data.flatMap(lambda x: x.split(" "))
b = a.map(lambda x: (x,1)).values()
b.collect()
```

9. sortByKeys

→
```
data = sc.textFile("FileStore/tables/spark.txt")
a = data.flatMap(lambda x: x.split(" "))
b = a.map(lambda x: (x,1)).sortByKey()
b.collect()
```

10 groupByKey

→
```
data = sc.textFile("FileStore/tables/spark.txt")
a = data.flatMap(lambda x: x.split(" "))
b = a.map(lambda x: (x, 1)).groupByKey()
b.collect()
```

11 reduceByKey

→
```
data = sc.textFile("FileStore/tables/spark.txt")
a = data.flatMap(lambda a: x.split(" "))
b = a.map(lambda a: (x,1)).reduceByKey(lambda V1,V2:
                                       V1+V2)
b.collect()
```

12 Joins on RDD's (Inner joins between two RDDs)

→
```
rdd = sc.textFile("FileStore/tables/spark.txt").flatMap(
        lambda x: x.split(" ")).map(lambda
        x: (x,1))
rdd1 = sc.textFile("FileStore/tables/hadoop.txt").flatMap(
        lambda x: x.split(" ")).map(lambda a: (x, 1))
rdd2 = rdd.join(rdd1)
rdd2.collect()
```

13 Joins on files

→
```
rdd = sc.textFile("FileStore/tables/log.txt").map(lambda
        l: l.split(" ").map(lambda K: (K[3], K[0]))
rdd1 = sc.textFile("FileStore/tables/userdetails.csv").map(
        lambda l: l.split(",").map(lambda K: (K[0], K[1]))
rdd2 = rdd.join(rdd1)
rdd2.collect()
```

Teacher's Signature

Date - 01/02/2021

Outputs:

1 Remove data file
Out[1]: True

2 FlatMap.
Out[2]: ['1','2','3', ... 'a','b','c','d','d','b','s','d','t']

3 Filter.
Out[3]: [1,2,3,4]

4 Distinct.
Out[4]: ['1','2','3', ...'a','b','5','6','7','w','t']

5 MapValues
Out[5]: [(1,3),(2,5),(3,6),(4,8),(5,9),(6,24),(7,46)]

6 FlatMapValues
Out[6]: [('usr001',"PMB201"),
('usr001','228, Park aves'),
('usr001','New York'),
- - - - - -
- - - - - -
('usr089','New York'),
('usr089','NY')]

Teacher's Signature .....

7 Key
Out[7]: ['spark',
'run',
'programs',
- - - - - - -
- - - - -
- - - -
'applications',
'quickly']

8 values
Out[8]: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

9 sortByKey
Out[9]: [('100x', 1),
('lazy', 1),
- - -
- - -
- - -
('use', 1)
('write', 1)]

10 groupByKey
Out[27]: [('spark', <pyspark.resultiterable.ResultIterable
at 0x1fe8c512450>),
- - - - - - - - - - - - -
- - - - - - - - - - - - - -
('applications', <pyspark.resultiterable.ResultIterable
at 0x1fe8c51273903>)]

Teacher's Signature ....

11 reduceByKey

▶(a) Spark Jobs

Out[11]: [('spark', 1),
('run', 1),
('programs', 2),
('in', 1),
('easy', 1),
.
.
.
('write', 1),
('applications', 1)]

12 Inner join b/w two RDDs

Out[5]: [('to', (1, 1)), ('and', (1, 1)), ('data', (1, 2))]

13 Join on files

Out[13]: [('7352', ('64.242.88.11', 'steven')),
('6291', ('64.242.88.10', 'Mark'))]