TRANSFORMATIONS RDD

14 March 2024 16:14

TRANSFORMATIONS:

- Map trans working with strings and numbers
- Flat map trans working with strings and numbers - (word count)
- Filter trans - filter via lambda expression
- Group by , sort by
- Group by key , sort by key, reduce by key
- Distinct, count, count by value

Cmd 1

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

Command took 2.12 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:22:49 PM on Sumanthz

Cmd 2

```
data1 = [1,2,3,4,5,6,7,8,9,10]
data2 = ['Sai','Sumanth']

rdd1 = gcp.parallelize(data1)
rdd2 = gcp.parallelize(data2)
```

Command took 1.35 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:22:49 PM on Sumanthz

Cmd 3

```
#Map transformaton working with numbers
newrdd1 = rdd1.map(lambda x: int(x)+100)
newrdd1.collect()
```

▸ (1) Spark Jobs

```
Out[3]: [101, 102, 103, 104, 105, 106, 107, 108, 109, 110]
```

Command took 6.18 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:22:49 PM on Sumanthz

Cmd 4

```
#Map transformaton working with Strings
newrdd2 = rdd2.map(lambda x: x+' '+x)
newrdd2.collect()
```

▸ (1) Spark Jobs

```
Out[4]: ['Sai Sai', 'Sumanth Sumanth']
```

Command took 0.76 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:22:49 PM on Sumanthz

```
#Flat Map transformaton working with numbers
def add_mul(x):
    return [x+2,x*2]
newrdd3 = rdd1.flatMap(add_mul)
newrdd3.collect()
```

▸ (1) Spark Jobs

```
Out[6]: [3, 2, 4, 4, 5, 6, 6, 8, 7, 10, 8, 12, 9, 14, 10, 16, 11, 18, 12, 20]
```

Command took 0.84 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:25:18 PM on Sumanthz

Cmd 6

```
newrdd3 = rdd1.flatMap(lambda x: x+10) # 'int' object is not iterable'
newrdd3.collect()
```

▸ (1) Spark Jobs

```
⊞org.apache.spark.SparkException: Job aborted due to stage failure: Task 4 in sta
6.0 (TID 52) (ip-10-172-252-227.us-west-2.compute.internal executor driver): org.ap
not iterable'. Full traceback below:
```

Command took 1.10 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:26:03 PM on Sumanthz

Cmd 7

```
newrdd3 = rdd1.flatMap(lambda x: [x+10])
newrdd3.collect()
```

▸ (1) Spark Jobs

```
Out[8]: [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
```

Command took 0.78 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:25:50 PM on Sumanthz

Cmd 8

```
#Flat Map transformaton working with strings
def strings(x):
    return [x,x,x]
newrdd4 = rdd2.flatMap(strings)
newrdd4.collect()
```

▸ (1) Spark Jobs

```
Out[10]: ['Sai', 'Sai', 'Sai', 'Sumanth', 'Sumanth', 'Sumanth']
```

Cmd 9

```
#Filter transformations (based on true or false value)
def is_odd(x):
    return x % 2 !=0
newrdd5 = rdd1.filter(is_odd)
newrdd5.collect()
```

▸ (1) Spark Jobs

```
Out[11]: [1, 3, 5, 7, 9]
```

Command took 0.32 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:10 PM on Sumanthz

Cmd 10

```
newrdd6 = rdd1.filter(lambda x: x % 2 ==0)   #is Even
newrdd6.collect()
```

▸ (1) Spark Jobs

```
Out[12]: [2, 4, 6, 8, 10]
```

Command took 0.39 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:13 PM on Sumanthz

Cmd 11

```
# group by
data3 = [[2,'d'],[3,'e'],[3,'f'],[4,'0'],[1,'a'],[2,'b'],[3,'c']]
rdd3 = gcp.parallelize(data3)

grouped_data=rdd3.groupBy(lambda x: x[0])
for k,v in grouped_data.collect():
    print('group',k,':',list(v))
```

▸ (1) Spark Jobs

```
group 1 : [[1, 'a']]
group 2 : [[2, 'd'], [2, 'b']]
group 3 : [[3, 'e'], [3, 'f'], [3, 'c']]
group 4 : [[4, '0']]
```

Command took 2.56 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:15 PM on Sumanthz

Cmd 12

```
# group by Key
data3 = [[2,'d'],[3,'e'],[3,'f'],[4,'0'],[1,'a'],[2,'b'],[3,'c']]
rdd3 = gcp.parallelize(data3)

grouped_data=rdd3.groupByKey()
for k,v in grouped_data.collect():
    print('group',k,':',list(v))
```

▸ (1) Spark Jobs

```
group 1 : ['a']
group 2 : ['d', 'b']
group 3 : ['e', 'f', 'c']
group 4 : ['0']
```

Command took 1.05 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:18 PM on Sumanthz

Cmd 13

```
# sort by with numbers
data4 = [5,88,34,465,1,0,88.5,22/7,3.14]
rdd4 = gcp.parallelize(data4)

sorted_data1=rdd4.sortBy(lambda x: x)
sorted_data1.collect()
```

▸ (3) Spark Jobs

```
Out[25]: [0, 1, 3.14, 3.142857142857143, 5, 34, 88, 88.5, 465]
```

Command took 1.22 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:38:39 PM on Sumanthz

Cmd 14

```
# sort by with strings
data5 = ['z','g','a','w','1','1082748']
rdd5 = gcp.parallelize(data5)

sorted_data2=rdd5.sortBy(lambda x: x)
sorted_data2.collect()
```

▸ (3) Spark Jobs

```
Out[22]: ['1', '1082748', 'a', 'g', 'w', 'z']
```

Command took 1.02 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:37:29 PM on Sumanthz

Cmd 15

```
# Distinct
data6 = [1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,'1',2,2,44,4,55,55,6,'z','g','a','w','a','w']
rdd6 = gcp.parallelize(data6)

distinct_data=rdd6.distinct()
distinct_data.collect()
```

▸ (1) Spark Jobs

```
Out[17]: ['g', 1, 'z', 'a', '1', 2, 44, 4, 'w', 6, 55]
```

Command took 0.77 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:26 PM on Sumanthz

Cmd 16

```
#count
distinct_data.count()
```

▸ (1) Spark Jobs

```
Out[18]: 11
```

Command took 0.35 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:28:29 PM on Sumanthz

Cmd 17

```
#count by value
count_by_value = rdd6.countByValue()
rdd6.countByValue()

#for k,v in count_by_value.items():
    #print(type(k),k,':',v)
```

▸ (2) Spark Jobs

```
Out[20]: defaultdict(int,
             {1: 17,
              '1': 1,
              2: 2,
              44: 1,
              4: 1,
              55: 2,
              6: 1,
              'z': 1,
              'g': 1,
              'a': 2,
              'w': 2})
```

Command took 0.96 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 4:34:33 PM on Sumanthz

Cmd 18

```
data3=[(1,2),(2,4),(1,6),(2,8),(3,10),('a',5),('b',7),('a',2)]
rdd7=gcp.parallelize(data3)
newrdd7=rdd7.reduceByKey(lambda x,y: x+y)
newrdd7.collect()
```

▸ (1) Spark Jobs

```
Out[31]: [(1, 8), ('a', 7), (2, 12), (3, 10), ('b', 7)]
```

Command took 1.13 seconds -- by saisumanth2002pss@gmail.com at 3/18/2024, 5:34:43 PM on Sumanthz