Hadoop Streaming - Join

myMapper.py

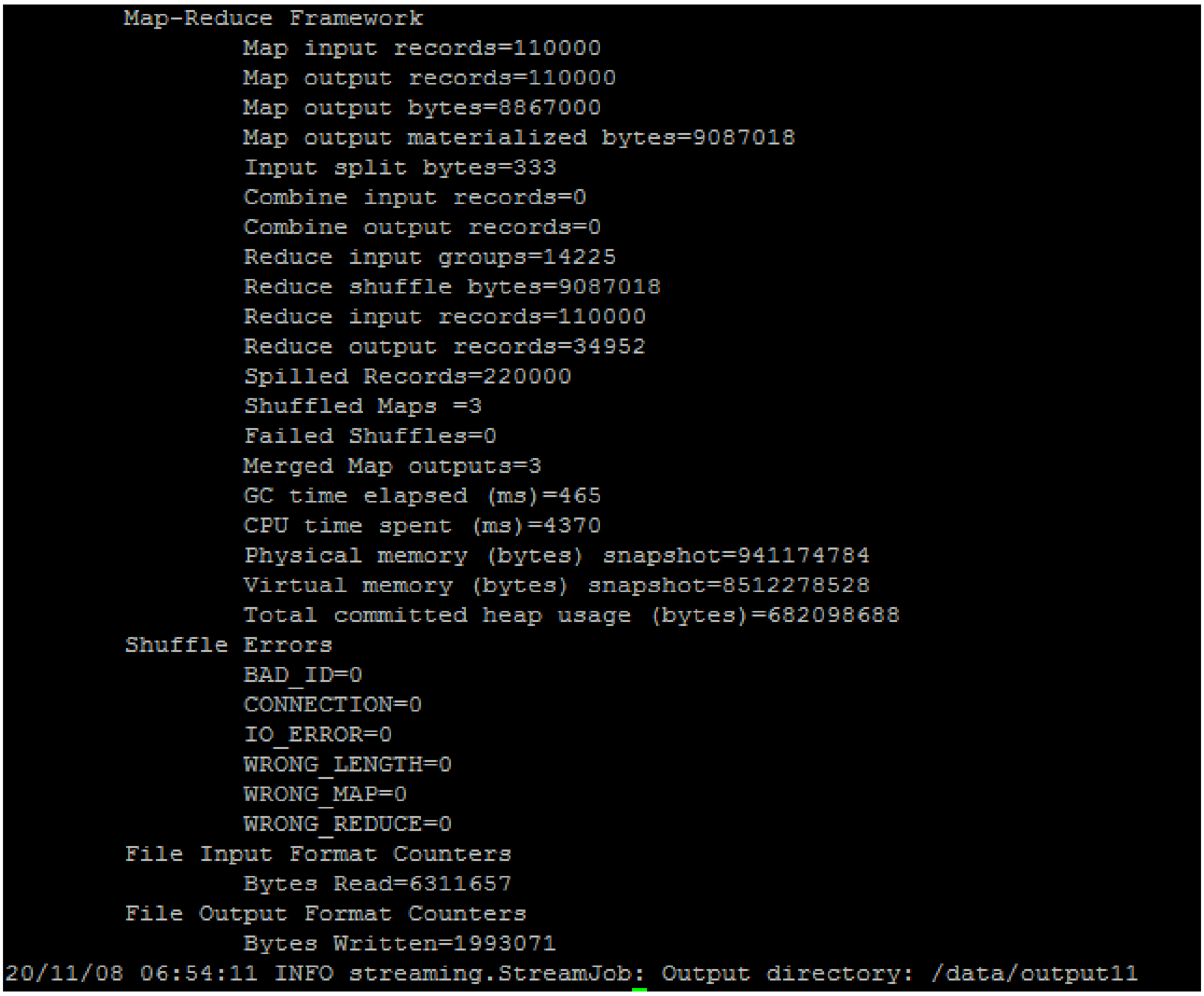
```
#!/usr/bin/python
import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    split = line.split('|')
    key = None
    if split[0].startswith('EMP'): #Mapper1, Employer
    key = split[1],split[2]
        print "%s\t%s\t%s\t%s"% (key,split[1],split[2],split[3],'Employer')
    else:                                   #Mapper2, Customer
        key = split[1],split[2]
    print "%s\t%s\t%s\t%s"% (key,split[1],split[2],split[3],'Customer')
```

myReducer.py

```
#!/usr/bin/python

import sys
import itertools

currentKey = None
valsEmployer = []
valsCustomer = []
employervalue = None
customervalue = None
phonevalue = None
addressvalue = None
fname = None
lname = None

# input comes from STDIN
for line in sys.stdin:

    split = line.strip().split('\t')    # 'Q11 \t Val1 \t Val2 \t Val3'
    key = split[0]
    line_value = '\t'.join(split[1:])

    if currentKey == key:  # Same key
        if line_value.endswith('Employer'):
        employervalue = line_value.strip().split('\t')
            fname = employervalue[0]
            lname = employervalue[1]
            phonevalue = employervalue[2]
            valsEmployer.append([fname, lname, phonevalue])
        if line_value.endswith('Customer'):
            customervalue = line_value.strip().split('\t')
            addressvalue = customervalue[2]
            valsCustomer.append(addressvalue)
    else:
        if currentKey:
            lenEmployer = len(valsEmployer)
            lenCustomer = len(valsCustomer)
            if (lenEmployer*lenCustomer > 0):
                for i in valsEmployer:
                    for j in valsCustomer:
                        print '%s\t%s' %('\t'.join(i),j)
        currentKey = key
        valsEmployer = []
        valsCustomer = []
        fname = None
        lname  = None
        phonevalue = None
        addressvalue = None
        if line_value.endswith('Employer'):
        employervalue = line_value.strip().split('\t')
            fname = employervalue[0]
            lname = employervalue[1]
            phonevalue = employervalue[2]
            valsEmployer.append([fname, lname, phonevalue])
        elif line_value.endswith('Customer'):
        customervalue = line_value.strip().split('\t')
            addressvalue = customervalue[2]
            valsCustomer.append(addressvalue)
lenEmployer = len(valsEmployer)
lenCustomer = len(valsCustomer)
if (lenEmployer*lenCustomer > 0):
    for i in valsEmployer:
        for j in valsCustomer:
            print '%s\t%s' %('\t'.join(i),j)
```

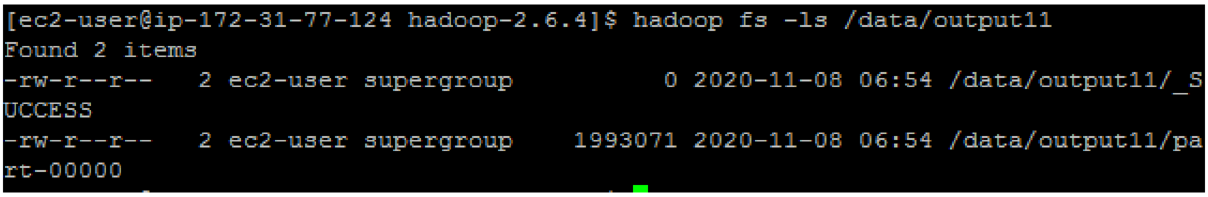
```
hadoop jar hadoop-streaming-2.6.4.jar -input /data/joinEmployerCustomer/ -mapper
../myMapper.py -file ../myMapper.py -reducer ../myReducer.py -file ../myReducer.py -output
/data/output11
```

Output:



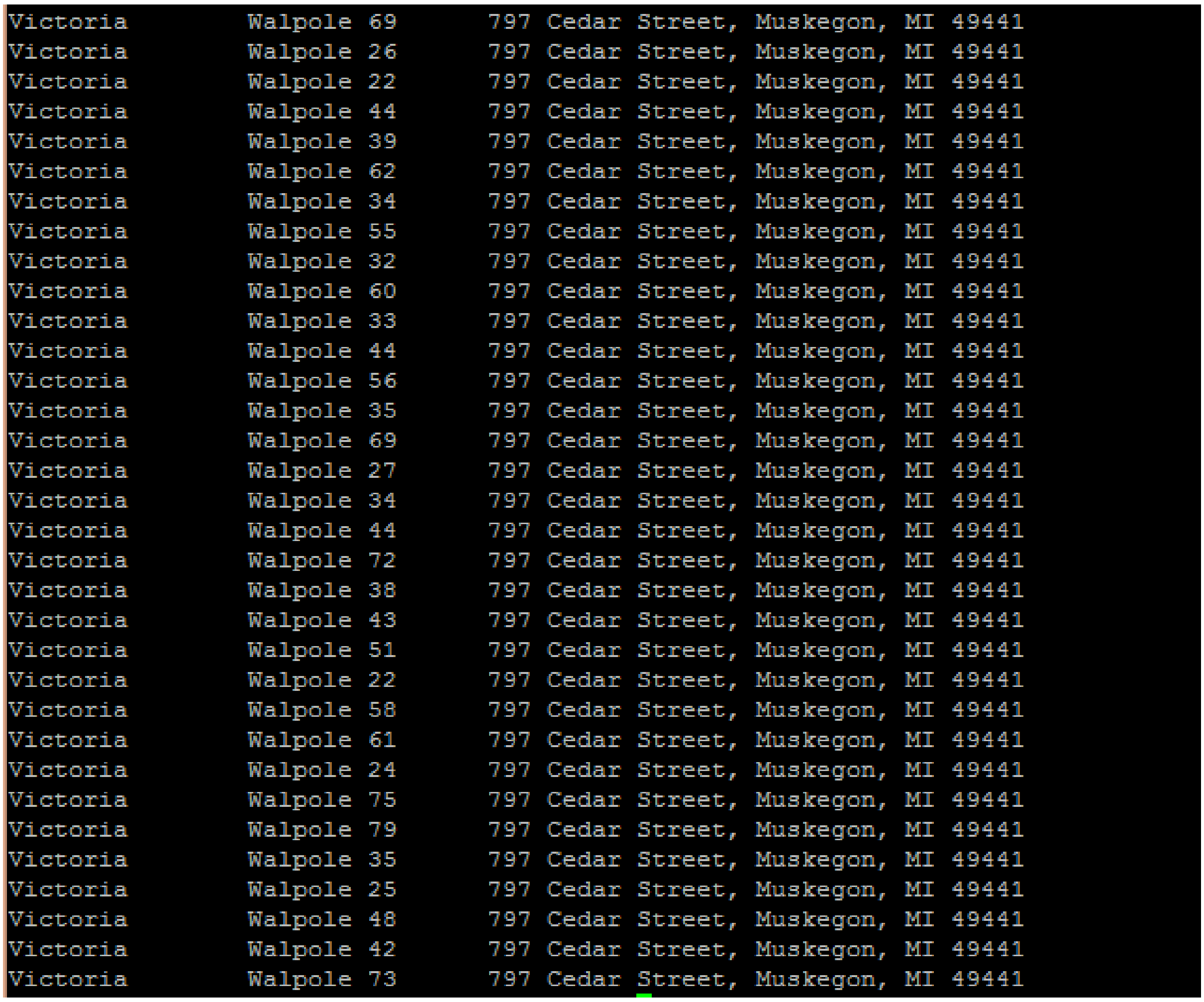
```
        Map-Reduce Framework
                Map input records=110000
                Map output records=110000
                Map output bytes=8867000
                Map output materialized bytes=9087018
                Input split bytes=333
                Combine input records=0
                Combine output records=0
                Reduce input groups=14225
                Reduce shuffle bytes=9087018
                Reduce input records=110000
                Reduce output records=34952
                Spilled Records=220000
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=465
                CPU time spent (ms)=4370
                Physical memory (bytes) snapshot=941174784
                Virtual memory (bytes) snapshot=8512278528
                Total committed heap usage (bytes)=682098688
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=6311657
        File Output Format Counters
                Bytes Written=1993071
20/11/08 06:54:11 INFO streaming.StreamJob: Output directory: /data/output11
```

hadoop fs -ls /data/output11

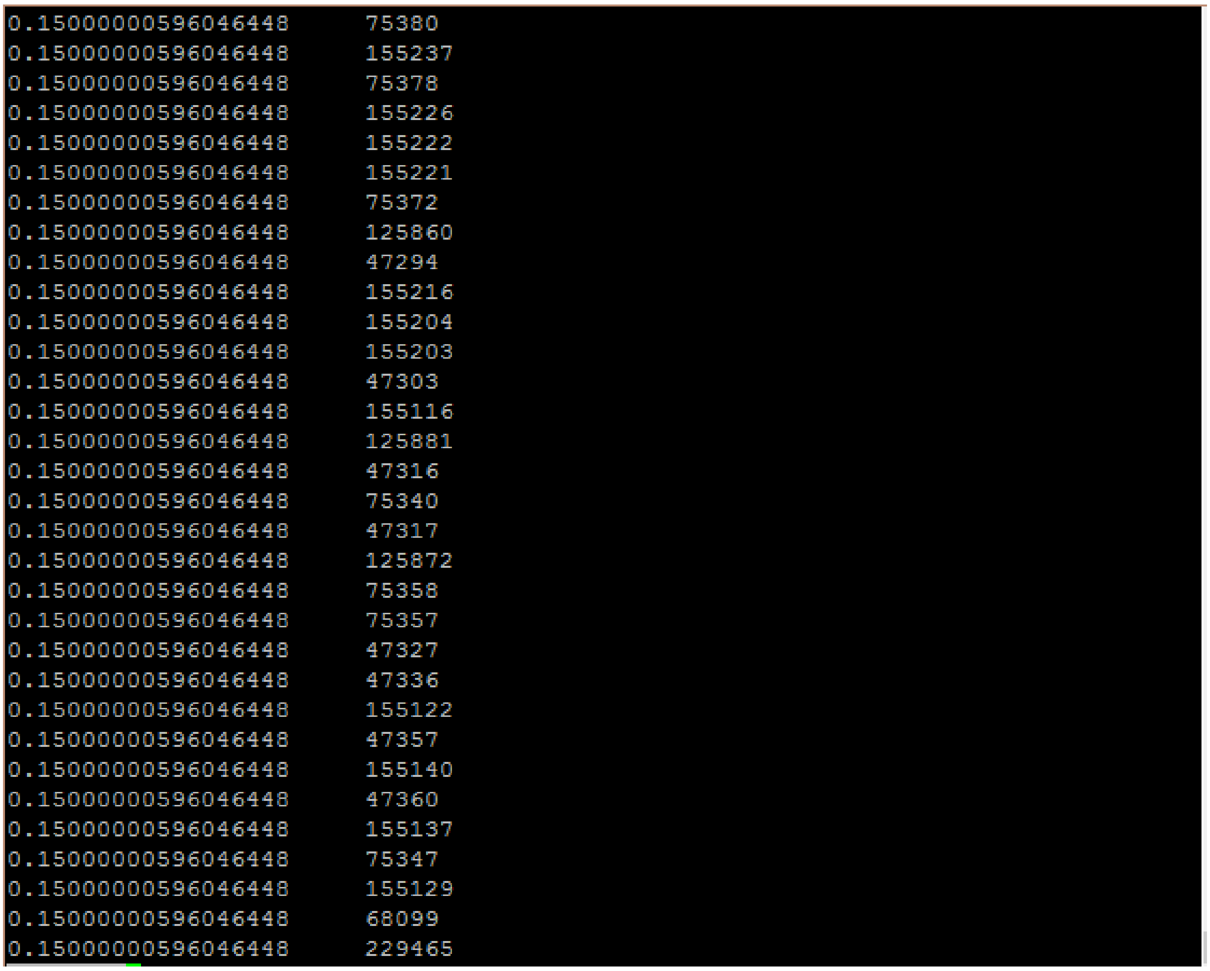


```
[ec2-user@ip-172-31-77-124 hadoop-2.6.4]$ hadoop fs -ls /data/output11
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-08 06:54 /data/output11/_S
UCCESS
-rw-r--r--   2 ec2-user supergroup    1993071 2020-11-08 06:54 /data/output11/pa
rt-00000
```

hadoop fs -cat /data/output11/part-00000



```
Victoria        Walpole 69      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 26      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 22      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 44      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 39      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 62      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 34      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 55      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 32      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 60      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 33      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 44      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 56      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 35      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 69      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 27      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 34      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 44      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 72      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 38      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 43      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 51      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 22      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 58      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 61      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 24      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 75      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 79      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 35      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 25      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 48      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 42      797 Cedar Street, Muskegon, MI 49441
Victoria        Walpole 73      797 Cedar Street, Muskegon, MI 49441
```



```
0.15000000596046448     75380
0.15000000596046448     155237
0.15000000596046448     75378
0.15000000596046448     155226
0.15000000596046448     155222
0.15000000596046448     155221
0.15000000596046448     75372
0.15000000596046448     125860
0.15000000596046448     47294
0.15000000596046448     155216
0.15000000596046448     155204
0.15000000596046448     155203
0.15000000596046448     47303
0.15000000596046448     155116
0.15000000596046448     125881
0.15000000596046448     47316
0.15000000596046448     75340
0.15000000596046448     47317
0.15000000596046448     125872
0.15000000596046448     75358
0.15000000596046448     75357
0.15000000596046448     47327
0.15000000596046448     47336
0.15000000596046448     155122
0.15000000596046448     47357
0.15000000596046448     155140
0.15000000596046448     47360
0.15000000596046448     155137
0.15000000596046448     75347
0.15000000596046448     155129
0.15000000596046448     68099
0.15000000596046448     229465
```