COMP 543, Tools and Models for Data Science

Chengyin Liu, cl93

Assignment #3

#3.1 Task 1

Write a MapReduce program that checks all of the files and computes the total "net ingredient cost" of prescription items dispensed for each PERIOD in the data set (total pounds and pence from the NIC field).

The result that my last MapReduce job wrote out:

```
INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=3536
                FILE: Number of bytes written=55968791
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=38236
                HDFS: Number of bytes written=403
                HDFS: Number of read operations=680
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=32
                S3: Number of bytes read=20943208677
                S3: Number of bytes written=0
                S3: Number of read operations=0
                S3: Number of large read operations=0
                S3: Number of write operations=0
        Job Counters
                Killed map tasks=2
                Killed reduce tasks=1
                Launched map tasks=316
                Launched reduce tasks=16
                Data-local map tasks=316
                Total time spent by all maps in occupied slots (ms)=250754895
                Total time spent by all reduces in occupied slots (ms)=132539400
                Total time spent by all map tasks (ms)=5572331
                Total time spent by all reduce tasks (ms)=1472660
                Total vcore-milliseconds taken by all map tasks=5572331
                Total vcore-milliseconds taken by all reduce tasks=1472660
                Total megabyte-milliseconds taken by all map tasks=8024156640
```

```
        Total megabyte-milliseconds taken by all reduce tasks=4241260800
    Map-Reduce Framework
        Map input records=150653175
        Map output records=150653160
        Map output bytes=2259797400
        Map output materialized bytes=86268
        Input split bytes=38236
        Combine input records=150653160
        Combine output records=316
        Reduce input groups=15
        Reduce shuffle bytes=86268
        Reduce input records=316
        Reduce output records=15
        Spilled Records=632
        Shuffled Maps =5056
        Failed Shuffles=0
        Merged Map outputs=5056
        GC time elapsed (ms)=136024
        CPU time spent (ms)=4616270
        Physical memory (bytes) snapshot=250763554816
        Virtual memory (bytes) snapshot=1098758893568
        Total committed heap usage (bytes)=234859528192
    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=20943208677
    File Output Format Counters
        Bytes Written=403
-----------------------------------------------
```

The result I got:

```
Period, Total net ingredient cost
----------------------------------------
201607      7.345053891600001E8
201608      7.337382873699934E8
201609      7.619078214700094E8
201610      7.533392691600046E8
201611      7.659713188500162E8
201612      7.750924622800076E8
201701      7.241018585499976E8
```

201702 6.75422474089994E8
201703 7.77765728080004E8
201704 6.785792095900078E8
201705 7.517331963300071E8
201706 7.821127311700062E8
201707 7.610782987999952E8
201708 7.419382048000088E8
201709 7.549630222200073E8

#3.2 Task 2

Write a MapReduce program that computes the 5 practices that issued the prescriptions with the highest total net ingredient cost in the data set.

The result that my last MapReduce job wrote out:

INFO mapreduce.Job: Counters: 50
File System Counters
FILE: Number of bytes read=2058
FILE: Number of bytes written=5395663
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=264783
HDFS: Number of bytes written=131
HDFS: Number of read operations=96
HDFS: Number of large read operations=0
HDFS: Number of write operations=32
Job Counters
Killed map tasks=1
Launched map tasks=16
Launched reduce tasks=16
Data-local map tasks=16
Total time spent by all maps in occupied slots (ms)=5492565
Total time spent by all reduces in occupied slots (ms)=6119730
Total time spent by all map tasks (ms)=122057
Total time spent by all reduce tasks (ms)=67997
Total vcore-milliseconds taken by all map tasks=122057
Total vcore-milliseconds taken by all reduce tasks=67997
Total megabyte-milliseconds taken by all map tasks=175762080
Total megabyte-milliseconds taken by all reduce tasks=195831360
Map-Reduce Framework
Map input records=10724
Map output records=80

```
        Map output bytes=2329
        Map output materialized bytes=6265
        Input split bytes=2384
        Combine input records=0
        Combine output records=0
        Reduce input groups=1
        Reduce shuffle bytes=6265
        Reduce input records=80
        Reduce output records=5
        Spilled Records=160
        Shuffled Maps =256
        Failed Shuffles=0
        Merged Map outputs=256
        GC time elapsed (ms)=5128
        CPU time spent (ms)=31790
        Physical memory (bytes) snapshot=10761256960
        Virtual memory (bytes) snapshot=123187638272
        Total committed heap usage (bytes)=10250354688
    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=262399
    File Output Format Counters
        Bytes Written=131
```

The result I got:

Practice, Total net ingredient cost

M85063	1.3157490440000126E7
Y01008	1.131368996000013E7
B82005	1.031304486000015E7
J82155	9059811.700000098
K83002	8703067.79000007
