Project 1 February 12, 2017

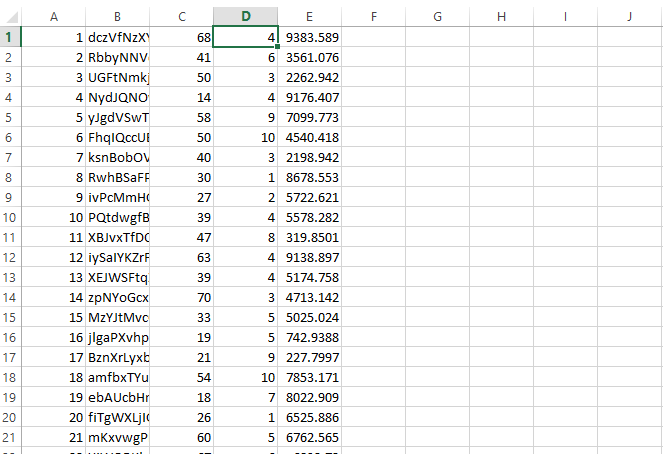
Big Data Management Spring 2017

Team 14 Wenlei Cao and Caitlin Kuhlman

1. The file Q1/DataGenerator.java generates two data files: customers.csv and transactions.csv. Below are screenshots of the resulting data.



**Customers Table**



**Transactions Table**

**3 Mapreduce Jobs.**

We have a single package Q3 which contains code for all the MapReduce queries. All the jobs require command line arguments to run secifying input and output paths. This is the usage for each job:

3.1 hadoop jar <JARFILE> Q3.Query1 <PATH TO CUSTOMERS> <OUTPUT PATH>

3.2 hadoop jar <JARFILE> Q3.Query2 <PATH TO TRANSACTIONS> <OUTPUT PATH>

3.3 hadoop jar <JARFILE> Q3.Query3 <PATH TO CUSTOMERS> <PATH TO TRANSACTIONS> <OUTPUT PATH>

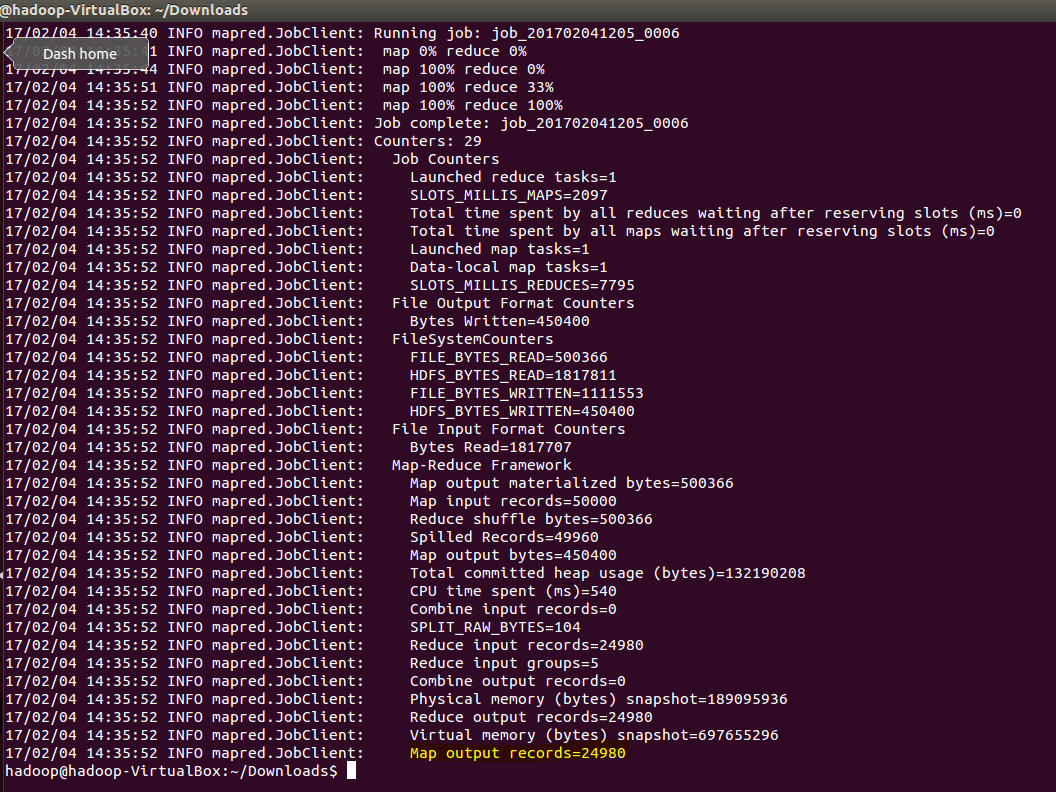
3.4 hadoop jar <JARFILE> Q3.Query4 <PATH TO CUSTOMERS> <PATH TO TRANSACTIONS> <OUTPUT PATH>

3.5 hadoop jar <JARFILE> Q3.query5.Query5 <PATH TO CUSTOMERS> <PATH TO TRANSACTIONS> <OUTPUT PATH>

3.1

In Q3/Query1.java a single job reports the customers whose CountryCode is between 2 and 6 (inclusive). This is completed using only a map phase and no reducer. The output is the same format as in the customers file. Below is a screenshot of the output form this hadoop job. In our generated file there were 24,980 records out of 50,000 customer with country code between 2-6





3.2

Q3/Query2.java reports for every customer, the number of transactions that customer did and the total sum of these transactions. This is done using a custom datatype called TransCountWritable.java which implements the WritableComparable interface.

The TransCountWritable object holds the transaction data for each customer. In the mapper, for each transaction, the output key is the customer id and the output value is a new TransCountWritable object which holds the totalTrans value for this one transaction.

A combiner is used to aggregate the count and total of transactions for one customer. It is a separate class from the reducer which reads in one customer id and a list of associated TransCountWritable objects. The counts and values of the transactions are aggregated and output in a single TransCountWritable object.

The same aggregation is performed in the reducer, but the output is a NullWritable key (empty) and a Text value with the new record for this customer. The output format is : CustomerID, NumTransactions, TotalSum.

Below is a sample screenshot of the output.

File: /tmp/customers.csv/part-r-OOOOO 
. Vtmp/customers.csv 
Goto • 
Go back to dir listin 
Advanced view/download options 
View Next chunk 
1, 92, 43034.4 
2, 79, 37820 203 
3, 112, 55001 484 
4, 112, 59234. 992 
s, 100, 50139. z-_ 
6, 108, 54330. 246 
7, 95, 46040. 715 
8, 111, 54854 305 
g, 107, 52010. 086 
10, 85, 40245. og 
11, 95, 43882.57 
12, 93, 46818.8 
13, 104, 54540. 523 
14, 110, 55666. 082 
15, 103, 55025.117 
16, 107, 53380. 207 
17, 95, 48307 496 
18, 105, 52289.18 
19, 101, 48325. 56 
20, 108, 60290 023 
21, 100, 49793. 27 
22, 98, 52933. 707 
23, 95, 062 
24, 85, 44320. 137 
25, 100, 48944. 043 
26, 95, 47549. 402 

3.3

For Query 3 we completed 2 alternative implementations. Q3/Query3.java also uses the TransCountWritable object to complete the join. MultipleInputs are used to read from customers and transactions files and separate mappers handle each. Both mappers output the key CustomerId, and the results are joined in the reducer.

An alternative implementation is found in package Q3.query3\_w. The driver file is query3.java. It follows the logic below:

Select

c.customerID,

c.name,

c.salary,

Count(t.transID),

Sum(t.transtotal),

Min (t.transNumItems)

From customers c

Join transactions t on c.ID = t.custID



Here is a sample output:

File: /tmp/output23/part-r-00000 
Goto : 
tmp/output23 
Go back to dir listin 
Advanced view/download options 
View Next chunk 
, dczVfNzxyoHi, 9383.589, 92, 43034. 40145599999, ı 
RbbyNNVcxptfzG, 3561 0757, 79, 37820. 20530999999, 
, UGFtNmkjNnceTzsHoJşo, 2262.9424, 112, 55001 48619899999, 
2, 
, NydJQN0vQ,Nyqcy, 9176.407, 112, 59234. 99032599999, 
, yJgdvswTXqqroLQZz, 7099.7734, 100, 50139.110992000016, 
Fhq1QccU8cMC, 4540.4175, 108, 54330. 246476, ı 
, ksn80b0VNnKyHDJ, 2198.9424, 95, 46040. 71284, ı 
Rwh8SaFPoMbGNYKwXRC, 8678. 553, 111, 54854. 30506099999, 
, ivPcMmHGjfRittE, 5722.621, 107, 52010. 08273800001, 
PQtdwgf8LkKEG, 5578.2817, 85, 40245. 088126, ı 
10, 
XBJvxTfDQidezj, 319.8501, 95, 43882. 569547000014, 
iySa1YKZrFLcNcRO, 9138.897, 93, 46818. 797419, 
12, 
XEJRSFtqXtfM, 5174.7583, 104, 54540. 527356000035, 
13, 
zpNYoGcxhcNNtU0yk, 4713.1416, 110, 55666. 08054399998, 
14, 
MzYJtMvcCHednnv, 5025.0244, 103, 55025.1168995, ı 
15, 
16, 
8znxrLyxbJN1pxFR-x, 227. 79965, 95, 48307. 497372999984, 
17, 
amfbxTYuUf1gMoxxqk, 7853.1714, 105, 52289.175811500005, 
18, 
ebAUcbHmrDpvvKk, 8022.909, 101, 48325. 55510699999, 
19, 
fiTgRXLj1G, 6525.8857, 108, 60290. 02492500001, 
20, 
mKxvwgPMybagEv1BeF, 6762.5645, 100, 49793. 27267899997, 
21, 
XjRQ8K1nNYbD, 6832.72, 98, 52933. 70579499998, 
22, 
vtovoBbUdFhJ1, 7743.1235, 95, 50793. 06434499999, 
23, 
oseRhgopbaTDKEZ, 6055.5767, 85, 44320.136718, ı 
24, 
LpNS,NkaPzMYTG, 9987 636, 100, 48944. 04193099998, 
25, 
KaLDtssssazomkLkos, 2374.6895, 95, 47549. 40226599998, 
26, 

Result file



3.4 Also contains 2 alternate implementations. In Q3.Query4.java a broadcast join is used to complete the task in a single MapReduce job. Customer info is read in by each Map task and put into a hashtable in a setup function. An object called a Trans4Writable is used to hold intermediate data.

In the package Q3.query4 the same task is implemented with 2 MapReduce jobs. The implementation follows the logic below:

Output format: CountryCode, NumberOfCustomers, MinTransTotal, MaxTransTotal

Select

c.countrycode,

Count(c.customerID),

min(t.transtotal),

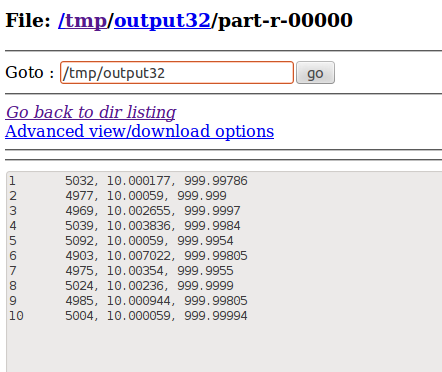
max(t.transtotal)

From customers c

Join transactions t on c.ID = t.custID

Group by c.countrycode





Result file



3.5

The implementation for query 5 is found in package Q3.query5

Logic

Select

c.customername,

count (t.transactioniD)

From customers c

Join transactions t on c.ID = t.custID

Group by c.customername

having count(t.transactioniD) >100 (5000000/50000)



File: /tmp/output28/part-r-OOOOO 
Vtmp/output28 
Goto : 
Go back to dir listin 
Advanced view/download options 
View Next chunk 
UGFtNmkjNnceTzsHoJS0, 
NydJQN0vQAyqcy, 
112 
Fhq1QccUBcMC, 108 
RwhBSaFPoMbG,WKwMC, 
ivPcMmHGjfRittE, 107 
XEJRSFtqXtfM, 104 
112 
zpWoGcxhcNNtUOyk, 110 
MzYJtMvcCHednnv , 
103 
jlgapxvhpE1, 107 
amfbxTYuUflgMoXxqk, 105 
ebAUcbHm rDpVvKk, 101 
fiTgRXLj1G, 108 
ZHUQvs1Qqqa, 112 
IzeXVDUKiss, 
aLpOhZuxpokseJsfK, 106 
eFnJufRijUNISUHKJ, 108 
zUKfUFYXSrEtnn, 116 
nTR-XPhTXtbtJ, 108 
mjBPDdCbMvcqA0pj, 107 
RZeGaaedHTn1EDSFUoV, 
ipTAGbLMudPZ, 120 
OhhFPooajENW.'NGqQNP, 103 
khvwJSNtRRKxwMz, 107 
"GitLNgTDnGcfFO, 
HuaNepAaqbecX, 
105 
RicSXhk1RX, 109 



4.1

/\*Write an Apache Pig query that reports the customer names that have the least number of transactions\*/



See the customer\_id and transaction number, min 51

(16523, 68) 
(21530 , 68) 
(15239,67) 
(14043,67) 
(38564, 67) 
(45607 , 67) 
(45095, 66) 
(44646 , 64) 
(47126, 63) 
(47871, 63) 
(29002 , 59) 
(24608, 51) 
2017-02-05 [main] ERROR org.apache. 
pig. tools.grunt.Grunt - 
ERROR 10 
00: Error during parsing. Encountered " <PATH> 
"D3-ORDERED DESC 
at line 2, c 
lumn 6. 
Was expecting one of: 
< EOF> 

Maxium 145

(17506,140) 
(28883 , 140) 
(37328 , 140) 
(5707 , 141) 
(2839 , 142) 
(18398, 143) 
(6346 , 145) 

 Retrieve the lowest customer name

228 [main] INFO org.apache . pig. backend. hadoop. 
2017-02-05 12:09:14, 
ne . util . MapRedUttl 
- Total input paths to process 
(ZQubROFZQmuo,51) 
grunt> 

File: /tmp/PIG_Q1.csv/part-r-OOOOO 
Goto : Vtmp/PlG Qi.csv 
Go back to dir listin 
Advanced view/download options 
ZQubRoFzQmJ',No , 51 



4.2

/\*

Write an Apache Pig query that join Customers and Transactions using Broadcast (replicated) join. The query reports for each customer the following info:

CustomerID, Name, Salary, NumOf Transactions, TotalSum, MinItem

number of transaction total number of transaction,

totalsum is sum of transtotal

minum number of item in transaction

\*/



Code using 'replicated'

Result file

(49988 , vejHDNwmtweTEQYDcn , 9671.999 , 77 , 42907.84966182709 , 1) 
(49989 , sww0PhcqguhzxjE1 , 5999.9507 , 95 , 48367.03768348694, 1) 
(49990, 
axNcrqLRdqzjGFeHzw , 6904.114 , 99 , 50608.32879924774, 1) 
(49991 , DygyjQFXYVEALLdOKWZ , 4405.227 , 109 , 52607.33077812195 , 1) 
(49992 , AEWBhrAYqpHKZeE , 5267.631 , 106 , 55242.20034599304, 1) 
(49993, 
kxnwzmhsmwttB , 1989.1237 , 114 , 53473.36391830444 , 1) 
(49994 , 
eabMIYSHsKubXYgzvv , 5709.443 , 117 , 59660.40106201172, 1) 
(49995 , otwRJtw1nAmZKMf1 , 6963.8047 , 115 , 63341.01471328735 , 1) 
(49996, 
ABcgcnwNxgLVADPNL , 5244.646 , 87 , 39914.57150268555 , 1) 
(49997, 
tTFe1CKzUFY0j0NYh , 284.10858 , 100 , 50321.8217048645 , 1) 
(49998, 
cVkDzzfSQSBNtnck , 1424.8522 , 90 , 40804.66194152832 , 1) 
(49999, 
dbuqysJoudmR , 9311.197 , 100 , 50218.591426849365 , 1) 
( 50000 , twtcoxZk1YBQheyp , 2496.2346 , 101 ,47782.20666885376 , 1) 
grunt> 

 Show CustomerID, Name, Salary, NumOf Transactions, TotalSum, MinItem

File: /tmp/PIG_Q2.csv/part-r-OOOOO 
. Vtmp/PlG Q2.csv 
Goto • 
Go back to dir listin 
Advanced view/download options 
View Next chunk 
1, dczVfNzxyoHi , 9383 589, 92, 43034 40139389038, 
2, RbbyNNVcxptfzG, 3561 0757 , 79, 37820 205266952515, 
3, UGFtNmkjNnCeTzsHoJSO, 2262 9424, 112, 55001 48632621765, 
4, NydJQN0vQAyqcy , 9176 407 , 112, 59234 990156173706, 
6, Fhq1QccUBcMC, 4540 4175, 108, 54330 246509552, 
7, ksn80b0VNnKyHDJ , 2198 9424, 95, 46040 71289539337 , 1 
, Rwh8SaFPoMbGNYKwMC, 8678 553, 111 , 54854 305086135864, 
g, ivPcMmHGj ttE, 5722 621 , 107 , 52010 08278656006, 
10, PQtdwgf8LkKEG, 5578 2817, 85, 40245 088148117065, 1 
11 , X8JvxTfDQidezj , 319 8501 , 95, 43882 56953620911 , 1 
12, iySa1YKZrFLcNcRO, 9138 897 , 93, 46818 79751968384, 
13, XEJRSFtqXtfM, 5174 7583, 104, 54540 527338027954, 
14, zpWoGcxhcNNtUOyk , 4713 1416, 110, 55666 08066558838, 
15, MzYJtMvcCHednnv , 5025 0244, 103, 55025 11685466766, 
16, j , 742 9388, 107 , 53380 211141586304, 
17, aznxrLyxbJN1pxFR-x, 227 79965, 95, 48307 497371673584, 
18, amfbxTYuUf1gMoxxqk , 7853 1714, 105, 52289 17593193054, 
19, ebAUcbHm rDpvvKk , 8022 gog, 101 , 48325 555203437805, 
20, fiTgRXLj1G, 6525 8857 , 108, 60290 024965286255, 
21 , mKxvwgPMybagEv1BeF, 6762 5645, 100, 49793 27274131775, 
22, XjRQBK1nNYbD, 6832 72, 98, 52933 705921173096, 1 
23, , 7743 1235, 95, 50793 06426048279, 
24, oseRhgopbaTDKEZ, 6055 5767 , 85, 44320 136865615845, 
25, LpNSAkaPzMYTG, 9987 636, 100, 48944 0419960022, 
26, KaLDtSSssaZQmkLkos , 2374 6895, 95, 47549 40232658386, 

Result file



4.3

/\*

Write an Apache Pig query that reports the Country Codes having number of customers greater than 5000 or less than 2000

\*/



 Country code and customer number

adoop-VirtualBox: — 
30b DAG: 
2017-02-05 14:20:29 , 405 [main] INFO org.apache.pig.backend.hadoop . executionengt 
ne .mapReduceLayer .MapReduceLauncher - 
Success ! 
2017-02-05 14:20:29 , 406 [main] INFO org.apache.ptg.data.SchemaTupleBackend - 
y [pig. schematuple] was not set... 
will not generate code. 
2017-02-05 14:20:29 , 409 [main] INFO org.apache.hadoop.mapreduce.lib.input.Filel 
nputFormat - Total input paths to process 
1 
2017-02-05 14:20: 29,409 [main] INFO org.apache.pig.backend.hadoop.executtonengt 
ne.uttl.MapRedUttl - Total input paths to process 
(1, 5032) 
(2,4977) 
(3,4969) 
(4, 5039) 
(5, 5092) 
(6,4903) 
(7,4975) 
(8 , 5024) 
(9,4985) 
(10, 5004) 
grunt> 

Dump show

"e .mapReduceLayer .MapReduceLauncher - 
Success ! 
'017-02-05 14:22:19, 320 [main] INFO org.apache.ptg.data.SchemaTupleBackend 
[pig. schematuple] was not set. 
will not generate code. 
'017-02-05 14:22:19, 326 [main] INFO org.apache.hadoop.mapreduce.lib.input.F 
IputFormat - Total input paths to process 
.1 
'017-02-05 14:22:19, 326 [main] INFO org.apache.pig.backend.hadoop.executton 
le.uttl.MapRedUttl - Total input paths to process 
a, 5032) 
g, 5039) 
, 5024) 
, 5004) 
Jrunt> 

Filter >5000 or < 2000

File: /tmp/PIG_Q3.csv/part-r-OOOOO 
Goto : 
/tmp/PlG Q3.csv 
go 
Go back to dir listin 
Advanced view/download options 
1, 5032 
4, 5039 
s, 5092 
8, 5024 
10, 5004 

Result file

