Project 1 Report

Jiexuan Sun
Congyang Wang

Warning:

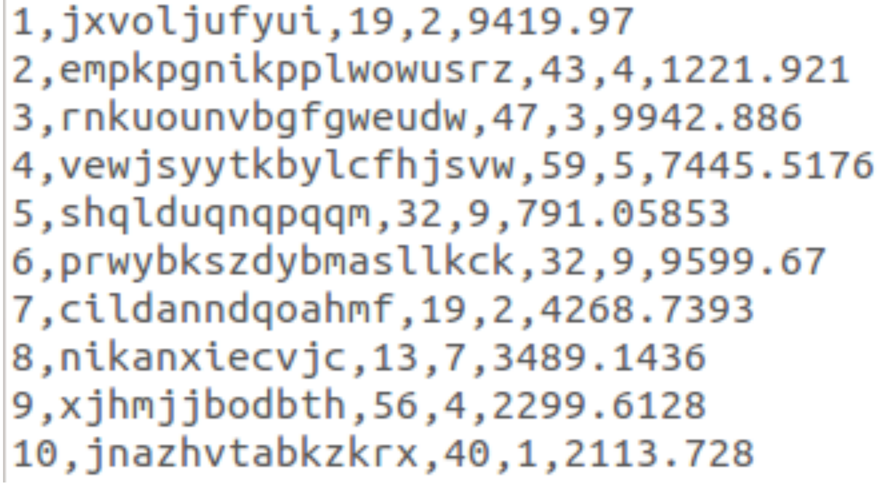
The path for mapreduce java code is input/1/customers and input/1/transactions

The input path for Apache Pig is input/1/customers.txt and input/1/transactions.txt
The output path for Apache Pig is output/Pig#_out.txt (# is the number of question)

For other details, please refer the source code.

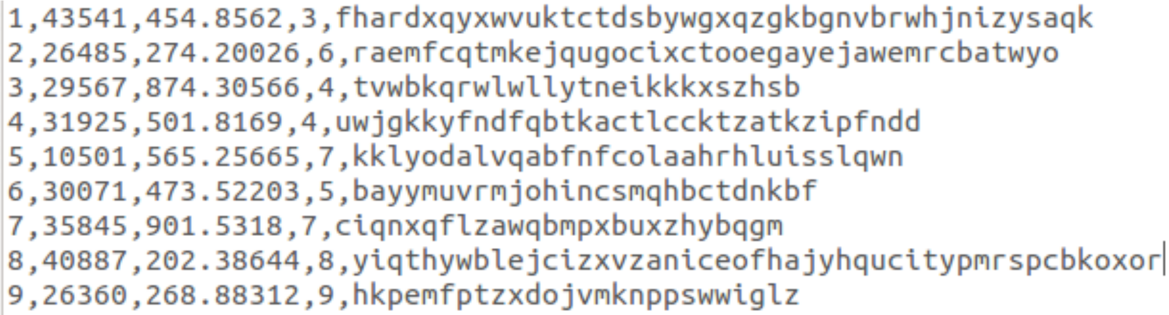
1.Creating Dataset

Sample Result (Customers):



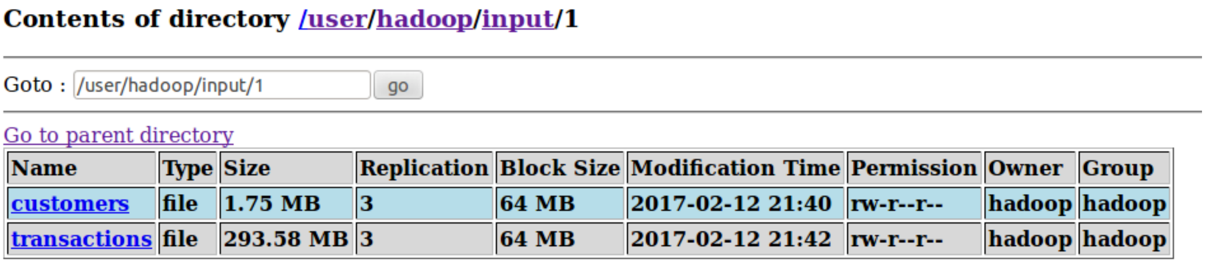
```
1,jxvoljufyui,19,2,9419.97
2,empkpgnikpplwowusrz,43,4,1221.921
3,rnkuounvbgfgweudw,47,3,9942.886
4,vewjsyytkbylcfhjsvw,59,5,7445.5176
5,shqlduqnqpqqm,32,9,791.05853
6,prwybkszdybmasllkck,32,9,9599.67
7,cildanndqoahmf,19,2,4268.7393
8,nikanxiecvjc,13,7,3489.1436
9,xjhmjjbodbth,56,4,2299.6128
10,jnazhvtabkzkrx,40,1,2113.728
```

Sample Result (Transactions):



```
1,43541,454.8562,3,fhardxqyxwvuktctdsbywgxqzgkbgnvbrwhjnizysaqk
2,26485,274.20026,6,raemfcqtmkejqugocixctooegayejawemrcbatwyo
3,29567,874.30566,4,tvwbkqrwlwllytneikkkxszhsb
4,31925,501.8169,4,uwjgkkyfndfqbtkactlccktzatkzipfndd
5,10501,565.25665,7,kklyodalvqabfnfcolaahrhluisslqwn
6,30071,473.52203,5,bayymuvrmjohincsmqhbctdnkbf
7,35845,901.5318,7,ciqnxqflzawqbmpxbuxzhybqgm
8,40887,202.38644,8,yiqthywblejcizxvzaniceofhajyhqucitypmrspcbkoxor
9,26360,268.88312,9,hkpemfptzxdojvmknppswwiglz
```

2.Uploading Data into Hadoop



Contents of directory /user/hadoop/input/1

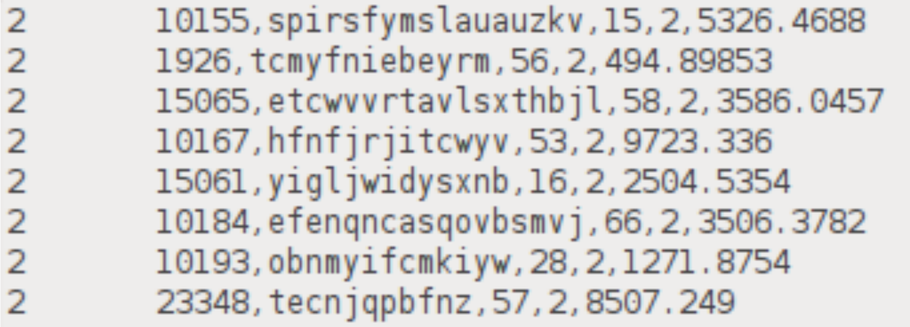
Goto : /user/hadoop/input/1 go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
customers	file	1.75 MB	3	64 MB	2017-02-12 21:40	rw-r--r--	hadoop	hadoop
transactions	file	293.58 MB	3	64 MB	2017-02-12 21:42	rw-r--r--	hadoop	hadoop

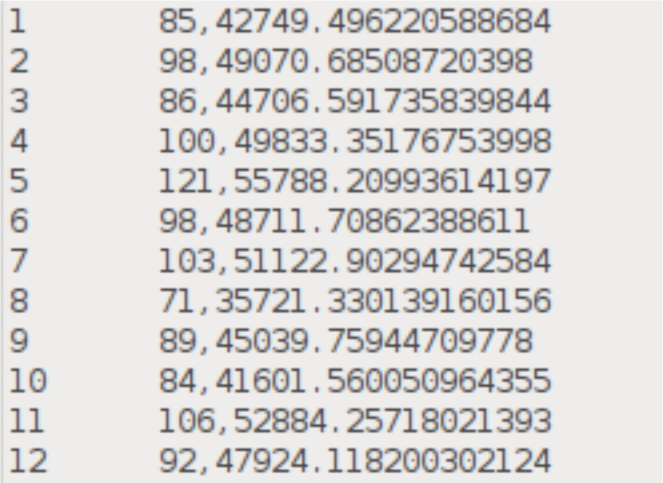
3.Writing MapReduce Jobs

1)Sample Result:



```
2	10155,spirsfymslauauzkv,15,2,5326.4688
2	1926,tcmyfniebeyrm,56,2,494.89853
2	15065,etcwvvrtavlsxthbjl,58,2,3586.0457
2	10167,hfnfjrjitcwyv,53,2,9723.336
2	15061,yigljwidysxnb,16,2,2504.5354
2	10184,efenqncasqovbsmvj,66,2,3506.3782
2	10193,obnmyifcmkiyw,28,2,1271.8754
2	23348,tecnjqpbfnz,57,2,8507.249
```

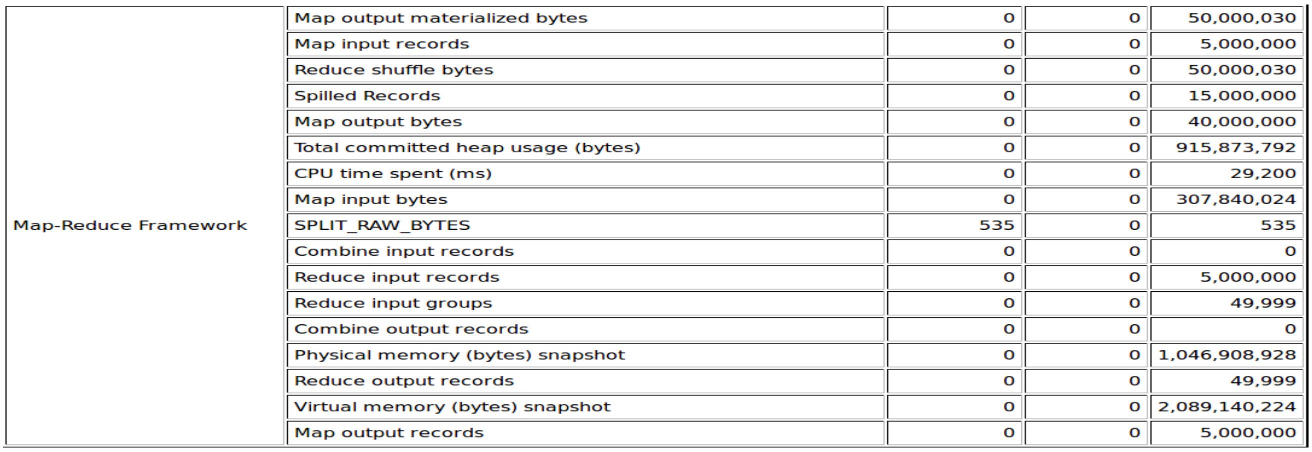
2)Sample Result:



```
1	85,42749.496220588684
2	98,49070.68508720398
3	86,44706.591735839844
4	100,49833.35176753998
5	121,55788.20993614197
6	98,48711.70862388611
7	103,51122.90294742584
8	71,35721.330139160156
9	89,45039.75944709778
10	84,41601.560050964355
11	106,52884.25718021393
12	92,47924.118200302124
```

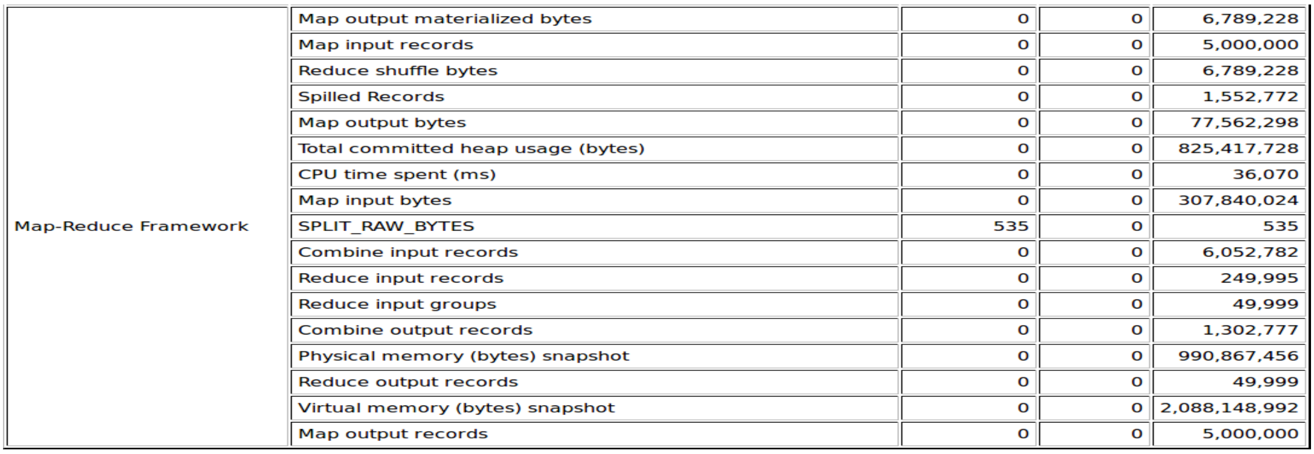
Run Time Analysis between with and without combiner:

Without Combiner:



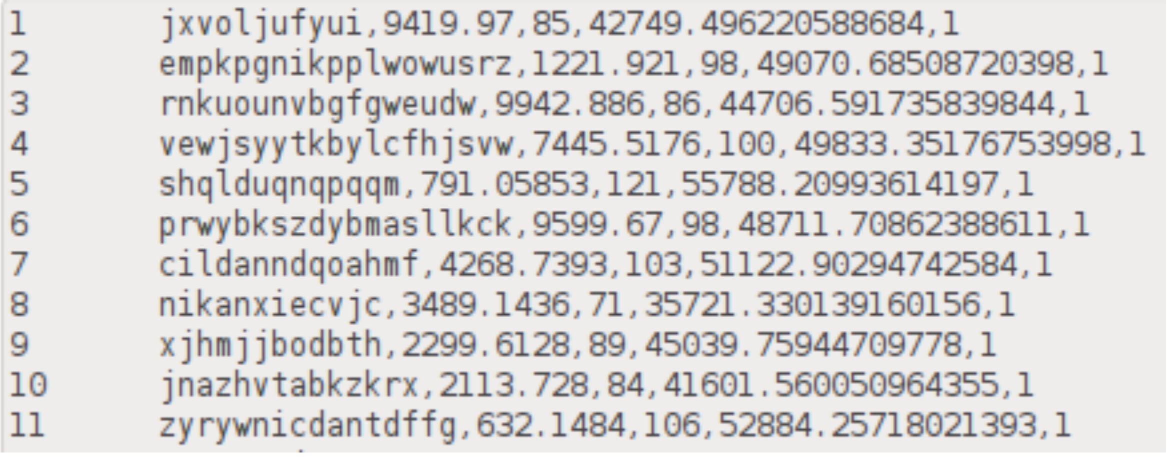
Map-Reduce Framework	Map output materialized bytes	0	0	50,000,030
	Map input records	0	0	5,000,000
	Reduce shuffle bytes	0	0	50,000,030
	Spilled Records	0	0	15,000,000
	Map output bytes	0	0	40,000,000
	Total committed heap usage (bytes)	0	0	915,873,792
	CPU time spent (ms)	0	0	29,200
	Map input bytes	0	0	307,840,024
	SPLIT_RAW_BYTES	535	0	535
	Combine input records	0	0	0
	Reduce input records	0	0	5,000,000
	Reduce input groups	0	0	49,999
	Combine output records	0	0	0
	Physical memory (bytes) snapshot	0	0	1,046,908,928
	Reduce output records	0	0	49,999
	Virtual memory (bytes) snapshot	0	0	2,089,140,224
	Map output records	0	0	5,000,000

With Combiner:



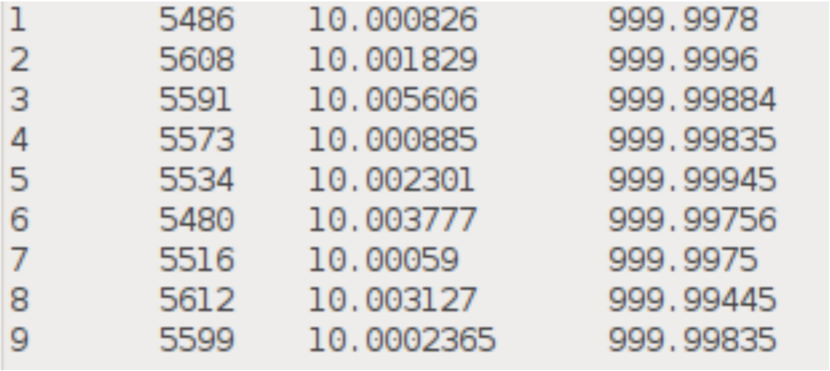
Map-Reduce Framework	Map output materialized bytes	0	0	6,789,228
	Map input records	0	0	5,000,000
	Reduce shuffle bytes	0	0	6,789,228
	Spilled Records	0	0	1,552,772
	Map output bytes	0	0	77,562,298
	Total committed heap usage (bytes)	0	0	825,417,728
	CPU time spent (ms)	0	0	36,070
	Map input bytes	0	0	307,840,024
	SPLIT_RAW_BYTES	535	0	535
	Combine input records	0	0	6,052,782
	Reduce input records	0	0	249,995
	Reduce input groups	0	0	49,999
	Combine output records	0	0	1,302,777
	Physical memory (bytes) snapshot	0	0	990,867,456
	Reduce output records	0	0	49,999
	Virtual memory (bytes) snapshot	0	0	2,088,148,992
	Map output records	0	0	5,000,000

3)Sample Result:



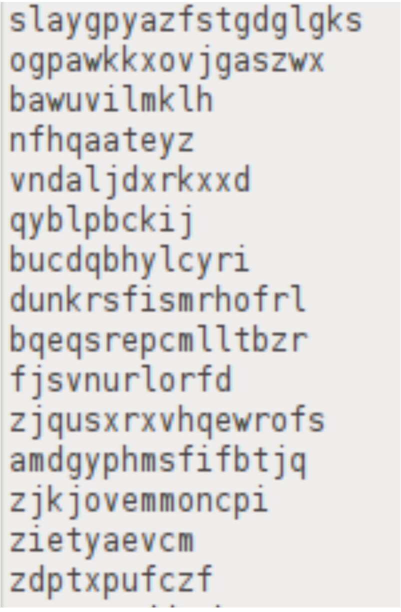
```
1       jxvoljufyui,9419.97,85,42749.496220588684,1
2       empkpgnikpplwowusrz,1221.921,98,49070.68508720398,1
3       rnkuounvbgfgweudw,9942.886,86,44706.591735839844,1
4       vewjsyytkbylcfhjsvw,7445.5176,100,49833.35176753998,1
5       shqlduqnqpqqm,791.05853,121,55788.20993614197,1
6       prwybkszdybmasllkck,9599.67,98,48711.70862388611,1
7       cildanndqoahmf,4268.7393,103,51122.90294742584,1
8       nikanxiecvjc,3489.1436,71,35721.330139160156,1
9       xjhmjjbodbth,2299.6128,89,45039.75944709778,1
10      jnazhvtabkzkrx,2113.728,84,41601.560050964355,1
11      zyrywnicdantdffg,632.1484,106,52884.25718021393,1
```

4)Sample Result:



```
1       5486    10.000826       999.9978
2       5608    10.001829       999.9996
3       5591    10.005606       999.99884
4       5573    10.000885       999.99835
5       5534    10.002301       999.99945
6       5480    10.003777       999.99756
7       5516    10.00059        999.9975
8       5612    10.003127       999.99445
9       5599    10.0002365      999.99835
```

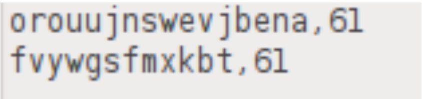
5)Sample Result:



```
slaygpyazfstgdglgks
ogpawkkxovjgaszwx
bawuvilmklh
nfhqaateyz
vndaljdxrkxxd
qyblpbckij
bucdqbhylcyri
dunkrsfismrhofrl
bqeqsrepcmlltbzr
fjsvnurlorfd
zjqusxrxvhqewrofs
amdgyphmsfifbtjq
zjkjovemmoncpi
zietyaevcm
zdptxpufczf
```

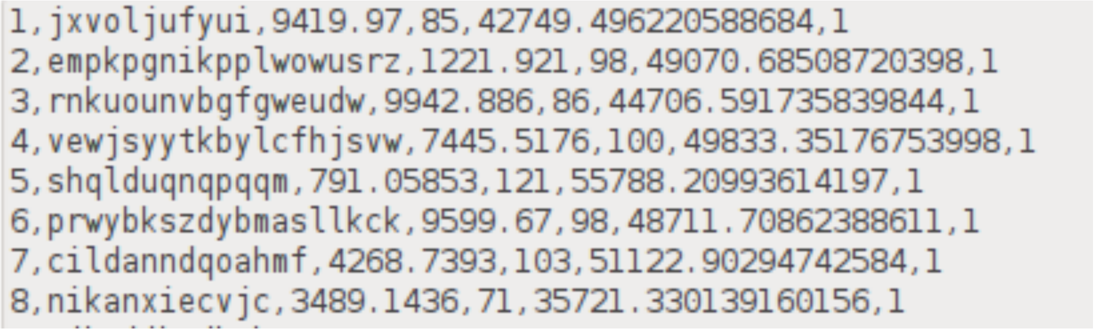
4.Writing Apache Pig Jobs

1)Sample Result:



```
orouujnswevjbena,61
fvywgsfmxkbt,61
```

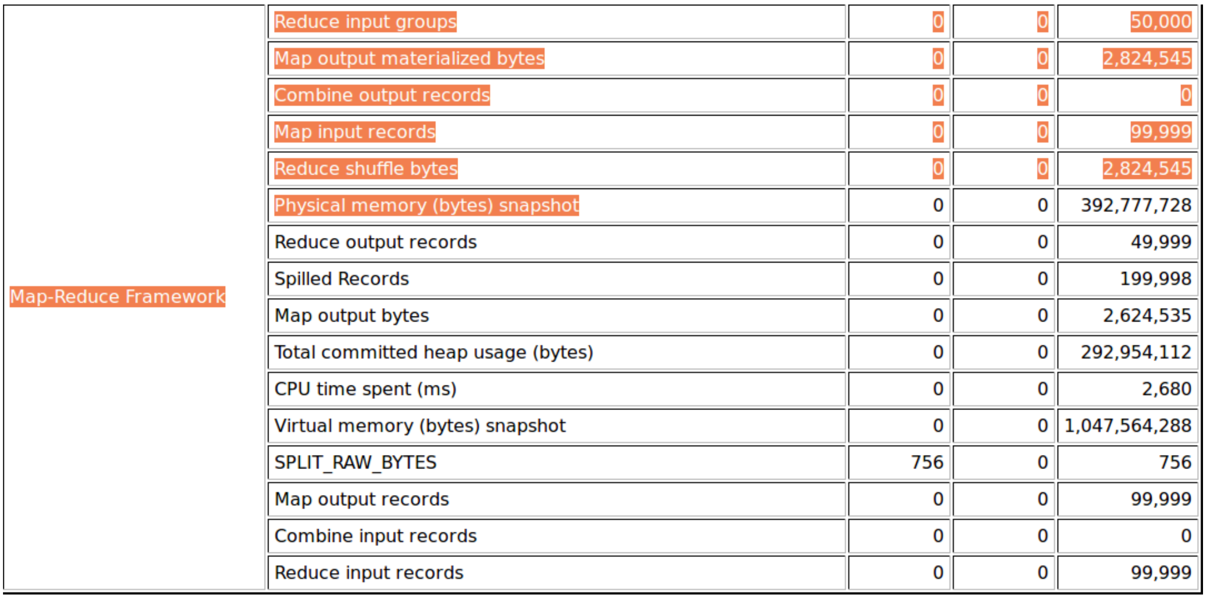
2)Sample Result:



```
1,jxvoljufyui,9419.97,85,42749.496220588684,1
2,empkpgnikpplwowusrz,1221.921,98,49070.68508720398,1
3,rnkuounvbgfgweudw,9942.886,86,44706.591735839844,1
4,vewjsyytkbylcfhjsvw,7445.5176,100,49833.35176753998,1
5,shqlduqnqpqqm,791.05853,121,55788.20993614197,1
6,prwybkszdybmasllkck,9599.67,98,48711.70862388611,1
7,cildanndqoahmf,4268.7393,103,51122.90294742584,1
8,nikanxiecvjc,3489.1436,71,35721.330139160156,1
```

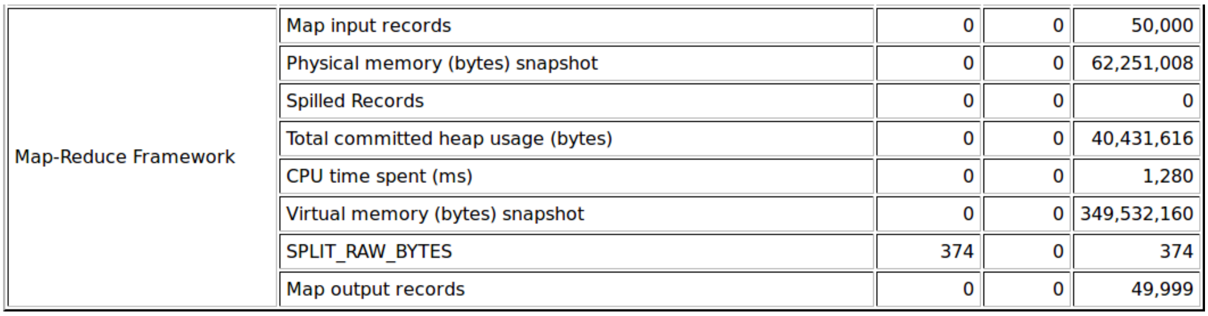
Run Time Analysis between with and without Broadcast Join:

Without Broadcast Join:



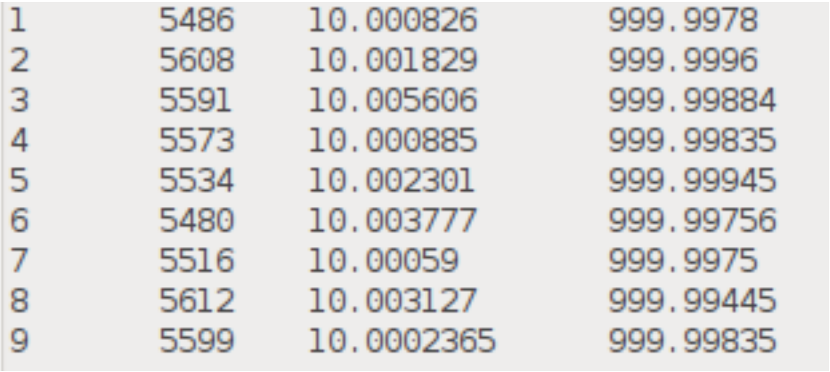
Map-Reduce Framework	Reduce input groups	0	0	50,000
	Map output materialized bytes	0	0	2,824,545
	Combine output records	0	0	0
	Map input records	0	0	99,999
	Reduce shuffle bytes	0	0	2,824,545
	Physical memory (bytes) snapshot	0	0	392,777,728
	Reduce output records	0	0	49,999
	Spilled Records	0	0	199,998
	Map output bytes	0	0	2,624,535
	Total committed heap usage (bytes)	0	0	292,954,112
	CPU time spent (ms)	0	0	2,680
	Virtual memory (bytes) snapshot	0	0	1,047,564,288
	SPLIT_RAW_BYTES	756	0	756
	Map output records	0	0	99,999
	Combine input records	0	0	0
	Reduce input records	0	0	99,999

With Broadcast Join:



Map-Reduce Framework	Map input records	0	0	50,000
	Physical memory (bytes) snapshot	0	0	62,251,008
	Spilled Records	0	0	0
	Total committed heap usage (bytes)	0	0	40,431,616
	CPU time spent (ms)	0	0	1,280
	Virtual memory (bytes) snapshot	0	0	349,532,160
	SPLIT_RAW_BYTES	374	0	374
	Map output records	0	0	49,999

3)Sample Result:



```
1       5486    10.000826     999.9978
2       5608    10.001829     999.9996
3       5591    10.005606     999.99884
4       5573    10.000885     999.99835
5       5534    10.002301     999.99945
6       5480    10.003777     999.99756
7       5516    10.00059      999.9975
8       5612    10.003127     999.99445
9       5599    10.0002365    999.99835
```