

CS585: **BIG DATA MANAGEMENT**

PROJECT 1 **(Report for Query 2)**

Group Members:
Gayathri Harilal
Xing Liu

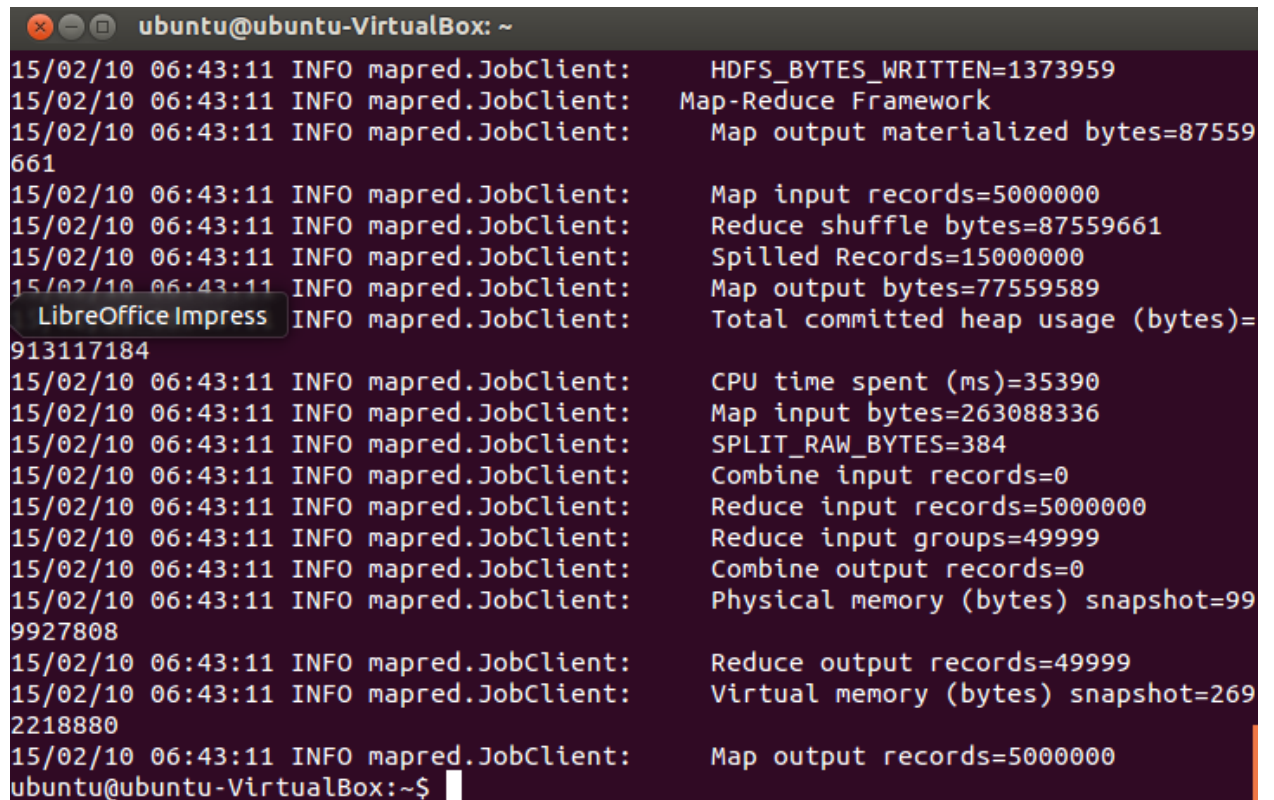
The two pictures below compares the performance between the two cases: Map-Reduce job with a combiner and Map-Reduce job without a combiner

(A) **WITHOUT COMBINER:**

```
ubuntu@ubuntu-VirtualBox: ~  
15/02/10 04:41:53 INFO mapred.JobClient: Map-Reduce Framework  
15/02/10 04:41:53 INFO mapred.JobClient: Map output materialized bytes=87559661  
15/02/10 04:41:53 INFO mapred.JobClient: Map input records=5000000  
15/02/10 04:41:53 INFO mapred.JobClient: Reduce shuffle bytes=87559661  
15/02/10 04:41:53 INFO mapred.JobClient: Spilled Records=15000000  
15/02/10 04:41:53 INFO mapred.JobClient: Map output bytes=77559589  
15/02/10 04:41:53 INFO mapred.JobClient: Total committed heap usage (bytes)=913195008  
15/02/10 04:41:53 INFO mapred.JobClient: CPU time spent (ms)=77670  
15/02/10 04:41:53 INFO mapred.JobClient: Map input bytes=263088336  
15/02/10 04:41:53 INFO mapred.JobClient: SPLIT_RAW_BYTES=384  
15/02/10 04:41:53 INFO mapred.JobClient: Combine input records=0  
15/02/10 04:41:53 INFO mapred.JobClient: Reduce input records=5000000  
15/02/10 04:41:53 INFO mapred.JobClient: Reduce input groups=49999  
15/02/10 04:41:53 INFO mapred.JobClient: Combine output records=0  
15/02/10 04:41:53 INFO mapred.JobClient: Physical memory (bytes) snapshot=988053504  
15/02/10 04:41:53 INFO mapred.JobClient: Reduce output records=49999  
15/02/10 04:41:53 INFO mapred.JobClient: Virtual memory (bytes) snapshot=2690228224  
15/02/10 04:41:53 INFO mapred.JobClient: Map output records=5000000  
ubuntu@ubuntu-VirtualBox:~$
```

Fig: Query 2 without using combiner

(B) **WITH COMBINER:**



```
ubuntu@ubuntu-VirtualBox: ~
15/02/10 06:43:11 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=1373959
15/02/10 06:43:11 INFO mapred.JobClient: Map-Reduce Framework
15/02/10 06:43:11 INFO mapred.JobClient: Map output materialized bytes=87559661
15/02/10 06:43:11 INFO mapred.JobClient: Map input records=5000000
15/02/10 06:43:11 INFO mapred.JobClient: Reduce shuffle bytes=87559661
15/02/10 06:43:11 INFO mapred.JobClient: Spilled Records=15000000
15/02/10 06:43:11 INFO mapred.JobClient: Map output bytes=77559589
15/02/10 06:43:11 INFO mapred.JobClient: Total committed heap usage (bytes)=913117184
15/02/10 06:43:11 INFO mapred.JobClient: CPU time spent (ms)=35390
15/02/10 06:43:11 INFO mapred.JobClient: Map input bytes=263088336
15/02/10 06:43:11 INFO mapred.JobClient: SPLIT_RAW_BYTES=384
15/02/10 06:43:11 INFO mapred.JobClient: Combine input records=0
15/02/10 06:43:11 INFO mapred.JobClient: Reduce input records=5000000
15/02/10 06:43:11 INFO mapred.JobClient: Reduce input groups=49999
15/02/10 06:43:11 INFO mapred.JobClient: Combine output records=0
15/02/10 06:43:11 INFO mapred.JobClient: Physical memory (bytes) snapshot=9927808
15/02/10 06:43:11 INFO mapred.JobClient: Reduce output records=49999
15/02/10 06:43:11 INFO mapred.JobClient: Virtual memory (bytes) snapshot=2692218880
15/02/10 06:43:11 INFO mapred.JobClient: Map output records=5000000
ubuntu@ubuntu-VirtualBox:~$
```

Fig: Query 2 using combiner

In these two cases, the CPU time spent went from 77670 ms to 35390 ms. The file system stats and shuffle stats are also reduced significantly. From these two pictures, we can see that combiner can optimize/minimize the number of key value pairs that will be shuffled across the network between mappers and reducers and thus to save as most bandwidth as possible.