

# DAS839-NoSQL Assignment-2 (Part B) Report

Siddhesh Deshpande-IMT2022080  
Abhinav Kumar-IMT2022079  
Jinesh Pagaria-IMT2022044  
Krish Patel-IMT2022097

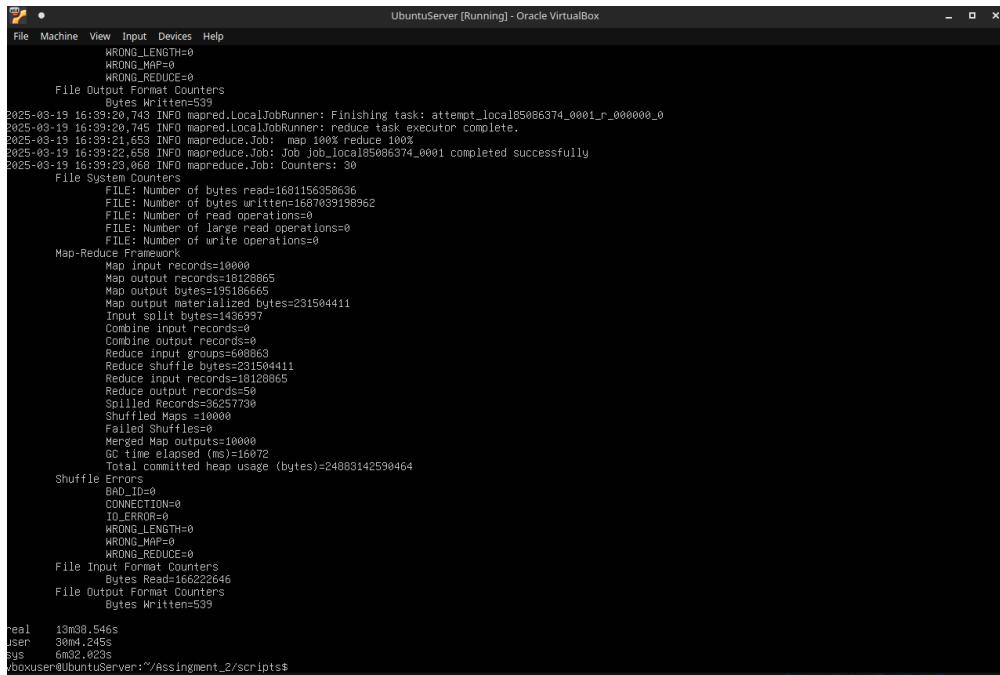
March 23, 2025

## Problem 4: Co-occurring Word Matrix Generation

### a) Top 50 Most Frequently Occurring Words (Pairs Approach)

1. Distributed Cache was used to make stopwords.txt file available to all the mappers so that stopwords can be ignored. Inserted all the stopwords in a hashset so that we can efficiently check if the word is a stopword or not.
2. Min-Heap Data Structure was used in the reducer side to keep track of top 50 most frequently occurring words efficiently.

#### Runtime and Output:



The screenshot shows a terminal window titled "UbuntuServer [Running] - Oracle VirtualBox". The window displays the output of a Hadoop job. The logs include statistics for Map and Reduce tasks, such as record counts, bytes processed, and shuffle metrics. The final output shows the total committed heap usage and the time taken for the job to complete. The terminal prompt at the bottom indicates the user is on an Ubuntu server.

```
File Machine View Input Devices Help
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written:539
2025-03-19 16:39:20 INFO mapred.LocalJobRunner: Finishing task: attempt_local85066374_0001_r_000000_0
2025-03-19 16:39:20,745 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-19 16:39:21,653 INFO mapreduce.Job: map 100% reduce 100%
2025-03-19 16:39:22,658 INFO mapreduce.Job: Job Job_local85066374_0001 completed successfully
2025-03-19 16:39:23,068 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=168156350636
FILE: Number of bytes written=1687039198962
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=10000
Map output records=18128865
Map output bytes=195186665
Map output materialized bytes=231504411
Input split bytes=1436997
Combine input records=0
Combine output records=0
Reduce input records=10000
Reduce shuffle bytes=231504411
Reduce input records=18128865
Reduce output records=50
Spilled Records=96257730
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=15672
Total time committed heap usage (bytes)=24003142590464
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:166222646
File Output Format Counters
Bytes Written:539
real    1m38.546s
user    30m4.245s
sys     6m32.023s
vboxuser@UbuntuServer:~/Assignment_2/scripts$
```

Figure 1: Q4.a

### b) Co-occurring Word Matrix (Pairs Approach)

**Objective:** Create a co-occurring word matrix using the pairs approach for distances  $d = \{1, 2, 3, 4\}$ .

## Runtime Analysis:

- $d = 1$ :
- $d = 2$ :
- $d = 3$ :
- $d = 4$ :

```

PROBLEMS OUTPUT DEBUGCONSOLE TERMINAL PORTS
Reduce shuffle bytes=11220633
Reduce input records=2826165
Reduce output records>2219
Spilled Records=3159315
Shuffled Maps=10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=41
Total committed heap usage (bytes)=2529165312
Shuffle Errors: BAD_ID=0, CONNECTION=0, ID_ERROR=0, WRONG_LENGTH=0, WRONG_MAP=0, WRONG_REDUCE=0
File System Counters: FILE: Number of bytes read=16668793595, FILE: Number of bytes written=16668793595, FILE: Number of read operations=1000000, FILE: Number of large read operations=0, FILE: Number of write operations=0
Map-Reduce Framework: Map Input records=10000, Map output Records=2826165, Map output bytes=10000000, Map output materialized bytes=41328633, Input split bytes=1666997, Combine output records=0, Reduce input groups=219, Reduce input records=2826163, Reduce input records=2826165, Reduce output records=2219, Spilled Records=3159315, Shuffled Maps=10000, Failed Shuffles=0, Merged Map outputs=10000, GC time elapsed (ms)=5757, Total committed heap usage (bytes)=2511803128764
Shuffle Errors: BAD_ID=0, CONNECTION=0, ID_ERROR=0, WRONG_LENGTH=0, WRONG_MAP=0, WRONG_REDUCE=0
File Input Format Counters: Bytes Read=166222646, File Output Format Counters: Bytes Written=25088
real: 1m56.278s
user: 5m13.778s
sys: 0m18.033s
[abhi@abhi ~]$ cd /Downloads/hdp/part2/q4-sid/Assignment_2/scripts
[abhi@abhi ~]$ ./q4b.sh

```

Figure 2: Execution time for  $Q4.b$  with  $d = 1$

```

PROBLEMS OUTPUT DEBUGCONSOLE TERMINAL PORTS
Reduce shuffle bytes=1666997
Reduce input records=3159315
Reduce output records>2412
Spilled Records=43158033
Shuffled Maps=10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=42
Total committed heap usage (bytes)=254699648
Shuffle Errors: BAD_ID=0, CONNECTION=0, ID_ERROR=0, WRONG_LENGTH=0, WRONG_MAP=0, WRONG_REDUCE=0
File System Counters: FILE: Number of bytes read=166696663381, FILE: Number of bytes written=353600000000, FILE: Number of read operations=1000000, FILE: Number of large read operations=0, FILE: Number of write operations=0
Map-Reduce Framework: Map Input records=10000, Map output Records=3159315, Map output bytes=93917505, Map output materialized bytes=45693571, Input split bytes=1666997, Combine input records=0, Combine output records=0, Reduce input groups=2412, Reduce shuffle bytes=1666997, Reduce input records=3159315, Reduce output records=2412, Spilled Records=43166666, Shuffled Maps=10000, Failed Shuffles=0, Merged Map outputs=10000, GC time elapsed (ms)=4964, Total committed heap usage (bytes)=25383689781248
Shuffle Errors: BAD_ID=0, CONNECTION=0, ID_ERROR=0, WRONG_LENGTH=0, WRONG_MAP=0, WRONG_REDUCE=0
File Input Format Counters: Bytes Read=166222646, File Output Format Counters: Bytes Written=29296
real: 1m56.735s
user: 5m39.896s
sys: 0m18.773s
[abhi@abhi ~]$ cd /Downloads/hdp/part2/q4-sid/Assignment_2/scripts
[abhi@abhi ~]$ ./q4b.sh

```

Figure 3: Execution time for  $Q4.b$  with  $d = 2$

```

File Edit Selection View Go Run Terminal Help
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reduce shuffle bytes=50894173
Reduce input records=531845
Reduced Maps +10000
Spilled Records+351845
Shuffled Maps +10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=39
Total committed heap usage (bytes)=2533359616
Shuffle Errors:
BAD ID#0
CONNECTION=0
ID ERROR=0
WRONG LENGTH=0
WRONG MAP=0
WRONG REDUCE=0
File Output Format Counters
Bytes Written=30676
2022-03-20 18:55:19,031 INFO mapred.LocalJobRunner: Finishing task: attempt local089296054_0001_r_000008
2022-03-20 18:55:19,031 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-03-20 18:55:19,031 INFO mapred.LocalJobRunner: map 10000 reduce 1000
2022-03-20 18:55:19,702 INFO mapreduce.Job: Job job_local089296054_0001 completed successfully
2022-03-20 18:55:19,903 INFO mapreduce.Job: Counters: 30
File Input Format Counters
FILE: Number of bytes read=169670703495
FILE: Number of bytes written=30895420000
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map Input records=10000
Map Input bytes=169670703495
Map Output bytes=3797883
Map output materialized bytes=50894173
Input split bytes=1666997
Combine input records=0
Combine output records=0
Reduce input records=50894173
Reduce shuffle bytes=50894173
Reduce input records=50894173
Reduce output records=2485
Spilled Records+7036290
Shuffled Maps +10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=39
Total committed heap usage (bytes)=25046631365632
Shuffle Errors:
BAD ID#0
CONNECTION=0
ID ERROR=0
WRONG LENGTH=0
WRONG MAP=0
WRONG REDUCE=0
File Input Format Counters
Bytes Read=169670703495
File Output Format Counters
Bytes Written=30676
real: 1m57.032s
user: 5m37.001s
sys: 0m18.051s
[tab@abhi ~]$ cd /Downloads/hdp/part2/q4-sid/Assignment_2/scripts
[tab@abhi ~]$ ./q4b.sh

```

Figure 4: Execution time for  $Q4.b$  with  $d = 3$

```

File Edit Selection View Go Run Terminal Help
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reduce shuffle bytes=55069236
Reduce input records=3854319
Reduce output records=2485
SPLITTING=0
Shuffled Maps +10000
Failed Shuffles=0
GC time elapsed (ms)=48
Total committed heap usage (bytes)=2523922432
Shuffle Errors:
BAD ID#0
CONNECTION=0
ID ERROR=0
WRONG LENGTH=0
WRONG MAP=0
WRONG REDUCE=0
File Output Format Counters
Bytes Written=1295
2022-03-20 18:49:38,141 INFO mapred.LocalJobRunner: Finishing task: attempt local1783123510_0001_r_000008
2022-03-20 18:49:38,141 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-03-20 18:49:39,302 INFO mapreduce.Job: map 10000 reduce 1000
2022-03-20 18:49:39,103 INFO mapreduce.Job: job job_local1783123510_0001 completed successfully
2022-03-20 18:49:39,103 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=169670703495
FILE: Number of bytes written=48129751249
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map Input records=10000
Map output records=3854319
Map output bytes=48031598
Map output materialized bytes=55860236
Input split bytes=1666997
Combine input records=0
Combine output records=0
Reduce input groups=2485
Reduce input records=3854319
Reduce output records=2485
SPLITTING=0
Shuffled Maps +10000
Failed Shuffles=0
Merged Map outputs=10000
GC time elapsed (ms)=379
Total committed heap usage (bytes)=25018469076992
Shuffle Errors:
BAD ID#0
CONNECTION=0
ID ERROR=0
WRONG LENGTH=0
WRONG MAP=0
WRONG REDUCE=0
File Input Format Counters
Bytes Read=169670703495
File Output Format Counters
Bytes Written=31205
real: 2m 9.97s
user: 5m55.931s
sys: 0m18.032s
[tab@abhi ~]$ cd /Downloads/hdp/part2/q4-sid/Assignment_2/scripts
[tab@abhi ~]$ ./q4b.sh

```

Figure 5: Execution time for  $Q4.b$  with  $d = 4$

### c) Co-occurring Word Matrix (Stripe Algorithm)

**Objective:** Construct the co-occurring word matrix using the stripe algorithm with distances  $d = \{1, 2, 3, 4\}$ .  
**Runtime Analysis:**

- $d = 1$ :
- $d = 2$ :

- $d = 3$ :
  - $d = 4$ :

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS ①
Reduce shuffle bytes=54961709
Reduce input records=2350594
Reduce output records=219
Spilled Records>350594
Shuffled Maps >10000
Failed Shuffles<1
Merged Map outputs=18000
Local Map outputs=18000
Total committed heap usage (bytes)=2531262464

Shuffle Errors
  IO ERROR=0
  CONNECTION=0
  ID ERROR=0
  HDFS LENGTH=0
  WRONG LENGTH=0
  WRONG MAP=0
  WRONG REDUCE=0
File Input Format Counters
  Bytes Written=25888

2025-03-20 19:51:33.102 INFO mapred.LocalJobRunner: Finishing task attempt_local108455334_0001_r_0000000_0
2025-03-20 19:51:33.102 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-20 19:51:53.430 INFO mapred.JobClient: map 10% reduce 100%
2025-03-20 19:51:53.430 INFO mapred.JobClient: Job completed successfully
2025-03-20 19:51:54.688 INFO mapred.JobClient: Counters: 36
  File System Counters
    FILE Number of bytes read=169071516957
    FILE Number of bytes written=418892177149
    FILE: Number of read operations=8
    FILE: Number of read/write operations=8
    FILE: Number of write operations=0
Map-Reduce Framework
  Map input records=18000
  Map output records=2350594
  Map output bytes=1066997
  Map output materialized bytes=54961709
  Input split bytes=1066997
  Combine output records=0
  Reduce input groups=50
  Reduce input records=2350594
  Reduce output records=219
  Spilled Records>350594
  Shuffled Maps >10000
  Failed Shuffles<1
  Merged Map outputs=18000
  GC Time elapsed (ms)=5602
  Total committed heap usage (bytes)=251696250880000
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  HDFS_LENGTH=0
  WRONG_WRONG=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Written=2529646
File Output Format Counters
  Bytes Written=25888

real    1m59.846s
user    5644.920s
sys     0.010s
cpu    5644.920s
shredded@shredded:/Downloads/hdp_part2/q4-sid/Assignment_2/scripts
```

Figure 6: Execution time for  $Q4.c$  with  $d = 1$

The screenshot shows a terminal window at the top with the title bar "File Edit Selection View Go Run Terminal Help" and the URL "hdp\_port [SSH: 172.16.23.34]". The terminal content displays Hadoop logs for a mapreduce job, including counts for various metrics like Input/Output bytes and records, and errors like "Failed Shuffles". Below the terminal is an "EXPLORER" sidebar from an IDE, showing a tree view of project files including "Assignment\_2", "partA", "partB", "scripts", and "WordCount.java". The bottom of the screen features a navigation bar with icons for file operations, search, and help.

```
PROBLEMS 0 OUTPUT DEBUG CONSOLE TERMINAL PROFILE ①

Reduce input records=235094
Reduce output records=9412
SPLIT input records=235094
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=18000
GC time elapsed (ms)=43
Total committed heap usage (bytes)=2545942528

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=162223640
    Bytes Written=25298
2025-03-20 19:47:44 INFO mapred.LocalJobRunner: Finishing task: attempt_local330325191_0001_r_000000_0
2025-03-20 19:47:44 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-20 19:47:44 INFO mapred.Task: map 0% reduce 0%
2025-03-20 19:47:56 INFO mapreduce.Job: Job job_local330325191_0001 completed successfully
2025-03-20 19:47:57 INFO mapreduce.Job: Counters: 38
  File Input Format Counters
    FILE: Number of bytes read=1069719424535
    FILE: Number of bytes written=43884066003
    FILE: Number of records read=100000000
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=18000
    Map output bytes=235094
    Map output bytes=52327960
    Map output materialized bytes=57809148
    Input split bytes=10000000
    Combine input records=0
    Combiner output records=0
    Reducer input records=18000
    Reduce shuffle bytes=57089148
    Reduce output records=9412
    Reduced output records=9412
    Spilled Records=701188
    Shuffled Maps =10000
    Failed Shuffles=0
    Merged Map outputs=18000
    GC time elapsed (ms)=43
    Total committed heap usage (bytes)=25276662725632

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=162223640
    Bytes Written=25298
real    1m58s 2.655
user    5m37.798s
sys     0m18.026s
_vsc_update_cwd/printfln: write error: interrupt

[snip]
[hdp@hdp1 ~]$ ./download/hdp_port/q4-sid/Assignment_2/scripts
```

Figure 7: Execution time for *Q4.c* with  $d = 2$

```

File Edit Selection View Go Run Terminal Help
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reduce shuffle bytes=9393720
Reduce input records=2356594
Reduce output records=2485
Spilled Records=2356594
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=18000
GC time elapsed (ms)=11
Total committed heap usage (bytes)=2530211888
Shuffle Errors:
  BAD ID#0
  CONNECTION=0
  IO ERROR=0
  WRONG LENGTH=0
  WRONG MAP=0
  WRONG REDUCE=0
  File Output Format Counters
    Bytes Written=30676
2022-03-20 19:44:50 INFO org.apache.hadoop.mapreduce.LocalJobRunner: Finishing task: attempt local00738352_0001_r_000000_0
2022-03-20 19:44:50.632 INFO org.apache.hadoop.mapreduce.LocalJobRunner: reduce task executor complete.
2022-03-20 19:44:51.000 INFO org.apache.hadoop.mapreduce.Job: Job local00738352_0001 completed successfully
2022-03-20 19:44:51.494 INFO org.apache.hadoop.mapreduce.Job: Job local00738352_0001 completed successfully
FILE Number of bytes read=169672392079
FILE Number of bytes written=497508440
FILE Number of large read operations=8
FILE Number of write operations=8
Map-Reduce Framework
  Map Input records=18000
  Map Input bytes=9393720
  Map Output bytes=2356594
  Map Output records=2485
  Map output materialized bytes=59337328
  Input bytes=169672392079
  Combine input records=0
  Combine output records=0
  Reduce Input records=18000
  Reduce Input bytes=9393720
  Reduce shuffle bytes=9393720
  Reduce output records=2485
  Spilled Records=479188
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=18000
  GC time elapsed (ms)=525
  Total committed heap usage (bytes)=2529535354764
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=169672392079
  File Output Format Counters
    Bytes Written=30676
real    1m58.002s
user    5m17.926s
sys     0m18.916s
[tab:bash] ~ /Downloads/hdp_part2/q4-sid/Assignment_2/scripts
yuan@yuan-OptiPlex-5090: ~ /Downloads/hdp_part2/q4-sid/Assignment_2/scripts

```

Figure 8: Execution time for  $Q4.c$  with  $d = 3$

```

File Edit Selection View Go Run Terminal Help
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reduce shuffle bytes=1457369
Reduce input records=2356594
Reduce output records=2485
Spilled Records=2356594
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=18000
GC time elapsed (ms)=65
Total committed heap usage (bytes)=2545942528
Shuffle Errors:
  BAD ID#0
  CONNECTION=0
  IO ERROR=0
  WRONG LENGTH=0
  WRONG MAP=0
  WRONG REDUCE=0
  File Output Format Counters
    Bytes Written=1205
2022-03-20 19:41:18.524 INFO org.apache.hadoop.mapreduce.LocalJobRunner: Finishing task: attempt local200823503_0001_r_000000_0
2022-03-20 19:41:18.524 INFO org.apache.hadoop.mapreduce.LocalJobRunner: reduce task executor complete.
2022-03-20 19:41:19.271 INFO org.apache.hadoop.mapreduce.Job: map 100% reduce 100%
2022-03-20 19:41:19.272 INFO org.apache.hadoop.mapreduce.Job: job_local200823503_0001 completed successfully
2022-03-20 19:41:19.272 INFO org.apache.hadoop.mapreduce.Job: Counters: 36
  File System Counters
    FILE Number of bytes read=16967239097
    FILE Number of bytes written=2164237
    FILE Number of read operations=8
    FILE Number of large read operations=8
    FILE Number of write operations=8
Map-Reduce Framework
  Map Input records=18000
  Map Input bytes=9393720
  Map Output bytes=2356594
  Map Output records=2485
  Input bytes=16967239097
  Combine output records=0
  Reduce Input groups=0
  Reduce Input records=1457369
  Reduce Input bytes=2356594
  Reduce output records=2485
  Spilled Records=2356594
  Shuffled Maps =10000
  Failed Shuffles=0
  Merged Map outputs=18000
  GC time elapsed (ms)=5689
  Total committed heap usage (bytes)=25378692268032
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=16967239246
  File Output Format Counters
    Bytes Written=31205
real    2m11.602s
user    5m46.296s
sys     0m18.284s
[tab:bash] ~ /Downloads/hdp_part2/q4-sid/Assignment_2/scripts
yuan@yuan-OptiPlex-5090: ~ /Downloads/hdp_part2/q4-sid/Assignment_2/scripts

```

Figure 9: Execution time for  $Q4.c$  with  $d = 4$

## d) Local Aggregation Performance Comparison

**Objective:** Compare performance between Map-class and Map-function level aggregation.

# 1 Results for Q4.d

## 1.1 Function Level Pairs

The screenshot shows an IDE interface with several tabs: File, Edit, Selection, View, Go, Run, Terminal, Help, and a central terminal window titled "Assignment\_2 [SSH: 172.16.238.243]". The terminal displays the execution of a MapReduce job. The log output includes:

```
File Output Format Counters
Bytes Written=25888
2025-03-21 12:08:57,414 INFO mapped LocalJobRunner: Finishing task: attempt_local1729613340_0001_r_000000_0
2025-03-21 12:08:57,415 INFO mapped LocalJobRunner: reduce task executor complete.
2025-03-21 12:08:57,420 INFO mapreduce.Job: map 100% reduce 100%
2025-03-21 12:08:57,420 INFO mapreduce.Job: Job Job_172.16.238.243_0001 completed successfully
2025-03-21 12:08:59,759 INFO mapreduce.Job: Counters: 38
Map-Reduce Framework
  Map input records=10000
  Map output records=99101
  Map output bytes=105338
  Map output bytestotal=105338
  File System Counters
    FILE: Number of bytes read=160000180538
    FILE: Number of bytes written=66787244743
    FILE: Number of read operations=8
    FILE: Number of large read operations=8
    FILE: Number of write operations=8
  Map-Reduce Framework
    Map input records=10000
    Map output records=99101
    Map output bytes=105338
    Map output bytestotal=105338
    Input split bytes=7185687
    Input split records=1436997
    Combine input records=8
    Combine output records=8
    Reduce input records=2219
    Reduce shuffle bytes=7105687
    Reduce input records=491081
    Reduce output bytes=623532
    Spilled Records=623532
    Shuffled Maps =10000
    Failed Shuffles=
    Merged Map outputs=10000
    GC time elapsed (ms)=18376
    Total committed heap usage (bytes)=25282639568794
Shuffle Errors
  IOError=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=160222046
  File Output Format Counters
  Bytes Written=25888
System
  total 9001 394s
  user 15m0.335s
  sys 1m4.852s
```

The bottom status bar shows "SSH 172.16.238.243" and "Java Ready".

Figure 10: Function - level pairs  $d = 1$

The screenshot shows an IDE interface with several tabs: File, Edit, Selection, View, Go, Run, Terminal, Help, and a central terminal window titled "Assignment\_2 [SSH: 172.16.238.243]". The terminal displays the execution of a MapReduce job. The log output includes:

```
File Output Format Counters
Bytes Written=25900
2025-03-21 12:08:59,243 INFO mapped LocalJobRunner: Finishing task: attempt_local1958294589_0001_r_000000_0
2025-03-21 12:08:59,244 INFO mapped LocalJobRunner: reduce task executor complete.
2025-03-21 12:08:59,356 INFO mapreduce.Job: map 100% reduce 100%
2025-03-21 12:08:59,370 INFO mapreduce.Job: Job Job_172.16.238.243_0001 completed successfully
2025-03-21 12:08:59,370 INFO mapreduce.Job: Counters: 38
Map-Reduce Framework
  Map input records=10000
  Map output records=99101
  Map output bytes=105338
  Map output bytestotal=105338
  Input split bytes=7185687
  Input split records=1436997
  Combine input records=8
  Combine output records=8
  Reduce input records=2212
  Reduce shuffle bytes=8971982
  Reduce input records=623536
  Reduce output records=9412
  Spilled Records=1247872
  Shuffled Maps =10000
  Failed Shuffles=
  Merged Map outputs=10000
  GC time elapsed (ms)=5644
  Total committed heap usage (bytes)=25405251168256
Shuffle Errors
  IOError=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=160222046
  File Output Format Counters
  Bytes Written=25900
System
  total 9004 1105s
  user 10m9.307s
  sys 1m6.464s
```

The bottom status bar shows "SSH 172.16.238.243" and "Java Ready".

Figure 11: Function - level pairs  $d = 2$

```

File Output Format Counters
Bytes Written=38676
2025-03-21 11:50:45,147 INFO mapred.LocalJobRunner: Finishing task: attempt_local224908054_0001_r_000000_0
2025-03-21 11:50:45,147 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-21 11:50:49,737 INFO mapreduce.Job: Map 100% reduce 100%
2025-03-21 11:50:49,737 INFO mapreduce.Job: Job [Job ID: job_123456789012345678901] completed successfully
2025-03-21 11:50:49,328 INFO mapreduce.Job: Counters: 36
File System Counters
FILE Number of bytes read=18088080000074
FILE Number of bytes written=91854389854
FILE: Number of read operations=91854389854
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=100000
Map output records=761443
Map input bytes=18088080000074
Map output bytes=91854389854
Map output materialized bytes=1808808355
Input split bytes=1436997
Combine input records=0
Combine output records=0
Reduce input bytes=1808808355
Reduce shuffle bytes=1808808355
Reduce input records=761443
Reduce output bytes=245
Spilled Records=1522288
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC Time elapsed (ms)=11010
Total committed heap usage (bytes)=2491535280256
Shuffle Errors
  IO=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_SEQ=0
  WRONG_CHECKSUM=0
  WRONG_REDUCE=0
File Input Format Counters
Bytes Read=16622866
Bytes Written=38676
File Output Format Counters
Bytes Written=38676
real    0m15.539s
user   1m14.028s
sys    1m13.025s
vboxuser@UbuntuServer:~/Assignment_2/scripts$ 

```

Figure 12: Function - level pairs  $d = 3$

```

File Output Format Counters
Bytes Written=1097108
2025-03-21 11:46:57,915 INFO mapred.LocalJobRunner: Finishing task: attempt_local1422436608_0001_r_000000_0
2025-03-21 11:46:57,935 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-21 11:46:58,830 INFO mapreduce.Job: map 100% reduce 100%
2025-03-21 11:46:58,830 INFO mapreduce.Job: Job [Job ID: job_123456789012345678901] completed successfully
2025-03-21 11:47:00,469 INFO mapreduce.Job: Counters: 36
File System Counters
FILE Number of bytes read=1808801967108
FILE Number of bytes written=918543710981
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=100000
Map output records=888449
Map output bytes=18749870
Map output materialized bytes=12569968
Input split bytes=1436997
Combine input records=0
Combine output records=0
Reduce input bytes=1808801967108
Reduce shuffle bytes=12569968
Reduce input records=888449
Reduce output records=245
Spilled Records=1522288
Shuffled Maps =10000
Failed Shuffles=0
Merged Map outputs=10000
GC Time elapsed (ms)=1106
Total committed heap usage (bytes)=24717812503968
Shuffle Errors
  IO=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_SEQ=0
  WRONG_CHECKSUM=0
  WRONG_REDUCE=0
File Input Format Counters
Bytes Read=16622846
Bytes Written=3205
File Output Format Counters
Bytes Written=3205
real    0m15.603s
user   1m15.894s
sys    1m15.775s
vboxuser@UbuntuServer:~/Assignment_2/scripts$ 

```

Figure 13: Function - level pairs  $d = 4$

## 1.2 Class Level Pairs

The screenshot shows a Java development environment with the following details:

- File Explorer:** Shows project structure with files like J\_ClassLevelPairs.java, J\_ClassLevelPairs.java, J\_FunctionsLevelPairs.java, J\_Part.java, J\_Part.java, and stopword.txt.
- Code Editor:** Displays the content of J\_ClassLevelPairs.java, which contains code for a TokenizerMapper class that implements MapObject, Text, myPair, and IntWritable interfaces. It includes logic for reading tokens from input, mapping them to pairs, and writing them to output.
- Terminal:** Shows HDFS metrics for Assignment\_2 (blk\_wrt\_1024.16.238.243). The metrics include:
  - Reduce shuffle bytes: 7185687
  - Reduce input records: 491181
  - Reduce output records: 2219
  - Split Input Records: 922802
  - Shuffled Maps: 1000000000
  - Failed Shuffles: 0
  - Merged Map outputs: 10000
  - GC time elapsed (ms): 10350
  - Total committed heap usage (bytes): >2541817677856
- Bottom Status Bar:** Shows the terminal command \$ hbase shell, the current line number (Ln 127), and other system information.

Figure 14: Class - level pairs  $d = 1$

Figure 15: Class - level pairs  $d = 2$

The screenshot shows the Eclipse IDE interface. In the Explorer view, several Java files are listed under the 'ASSIGNMENT\_2' project, including ClassLevelPairs.java, ClassLevelPairsTest.java, FunctionLevelPairs.java, PartA.java, PartB.java, and PartC.java. The terminal window displays the execution of the program, showing the number of input records (761443), the number of mappers (10000), and the number of reducers (1000). It also lists various error types (IO, CONNECTION, IO\_ERROR, etc.) and provides detailed file input and output statistics.

```

src > J ClassLevelPairs.java > J ClassLevelPairs > TokenizerMapper > map(Object,Text,Context)
  35 public class ClassLevelPairs
  36 {
  37     public static class TokenizerMapper extends Mapper<Object, Text, mypair, IntWritable>
  38     {
  39         private void parseMyFile(String filename)
  40         {
  41             System.out.println("Reading file: " + filename);
  42             try
  43             {
  44                 BufferedReader br = new BufferedReader(new InputStreamReader(new FileInputStream(filename)));
  45                 String line;
  46                 while ((line = br.readLine()) != null)
  47                 {
  48                     StringTokenizer st = new StringTokenizer(line);
  49                     mypair words = new mypair();
  50                     words.setWord(st.nextToken());
  51                     words.setCount(1);
  52                     int i = 0;
  53                     while (st.hasMoreTokens())
  54                     {
  55                         words.incrCount();
  56                         i++;
  57                     }
  58                     map(words, new IntWritable(i));
  59                 }
  60             }
  61             catch (IOException e)
  62             {
  63                 e.printStackTrace();
  64             }
  65         }
  66     }
  67     @Override
  68     public void map(Object key,Text value,Context context) throws IOException, InterruptedException
  69     {
  70         String line = value.toString().toLowerCase();
  71         String tokens[] = line.split("\\W+");
  72         int distance = 3;// Can be changed to 2,3,4
  73         for(int i=0;i<tokens.length;i++)
  74         {
  75             if(words.contains(tokens[i]))
  76             {
  77                 for(int j=i+1;j<tokens.length;j++)
  78                 {
  79                     if(Math.abs(i-j)<distance)
  80                     {
  81                         if(tokens[i].equals(tokens[j]))
  82                         {
  83                             reduce(new mypair(tokens[i], 1), 1, context);
  84                         }
  85                     }
  86                 }
  87             }
  88         }
  89     }
  90     @Override
  91     protected void reduce(mypair key,IntWritable value,Context context) throws IOException, InterruptedException
  92     {
  93         context.write(key, value);
  94     }
  95     public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException
  96     {
  97         Job job = new Job();
  98         job.setJarByClass(ClassLevelPairs.class);
  99         job.setMapperClass(TokenizerMapper.class);
  100        job.setCombinerClass(NullReducer.class);
  101        job.setReducerClass(NullReducer.class);
  102        job.setInputFormatClass(TextInputFormat.class);
  103        job.setOutputFormatClass(TextOutputFormat.class);
  104        job.setMapOutputKeyClass(mypair.class);
  105        job.setMapOutputValueClass(IntWritable.class);
  106        job.setNumReduceTasks(1);
  107        FileInputFormat.addInputPath(job, new Path(args[0]));
  108        FileOutputFormat.setOutputPath(job, new Path(args[1]));
  109        job.waitForCompletion(true);
  110    }
  111}
  112
  113
  114
  115
  116
  117
  118
  119
  120
  121
  122
  123
  124
  125
  126
  127
  128
  129
  130
  131
  132
  133
  134
  135
  136
  137
  138
  139
  140
  141
  142
  143
  144
  145
  146
  147
  148
  149
  150
  151
  152
  153
  154
  155
  156
  157
  158
  159
  160
  161
  162
  163
  164
  165
  166
  167
  168
  169
  170
  171
  172
  173
  174
  175
  176
  177
  178
  179
  180
  181
  182
  183
  184
  185
  186
  187
  188
  189
  190
  191
  192
  193
  194
  195
  196
  197
  198
  199
  200
  201
  202
  203
  204
  205
  206
  207
  208
  209
  210
  211
  212
  213
  214
  215
  216
  217
  218
  219
  220
  221
  222
  223
  224
  225
  226
  227
  228
  229
  230
  231
  232
  233
  234
  235
  236
  237
  238
  239
  240
  241
  242
  243
  244
  245
  246
  247
  248
  249
  250
  251
  252
  253
  254
  255
  256
  257
  258
  259
  260
  261
  262
  263
  264
  265
  266
  267
  268
  269
  270
  271
  272
  273
  274
  275
  276
  277
  278
  279
  280
  281
  282
  283
  284
  285
  286
  287
  288
  289
  290
  291
  292
  293
  294
  295
  296
  297
  298
  299
  300
  301
  302
  303
  304
  305
  306
  307
  308
  309
  310
  311
  312
  313
  314
  315
  316
  317
  318
  319
  320
  321
  322
  323
  324
  325
  326
  327
  328
  329
  330
  331
  332
  333
  334
  335
  336
  337
  338
  339
  340
  341
  342
  343
  344
  345
  346
  347
  348
  349
  350
  351
  352
  353
  354
  355
  356
  357
  358
  359
  360
  361
  362
  363
  364
  365
  366
  367
  368
  369
  370
  371
  372
  373
  374
  375
  376
  377
  378
  379
  380
  381
  382
  383
  384
  385
  386
  387
  388
  389
  390
  391
  392
  393
  394
  395
  396
  397
  398
  399
  400
  401
  402
  403
  404
  405
  406
  407
  408
  409
  410
  411
  412
  413
  414
  415
  416
  417
  418
  419
  420
  421
  422
  423
  424
  425
  426
  427
  428
  429
  430
  431
  432
  433
  434
  435
  436
  437
  438
  439
  440
  441
  442
  443
  444
  445
  446
  447
  448
  449
  450
  451
  452
  453
  454
  455
  456
  457
  458
  459
  460
  461
  462
  463
  464
  465
  466
  467
  468
  469
  470
  471
  472
  473
  474
  475
  476
  477
  478
  479
  480
  481
  482
  483
  484
  485
  486
  487
  488
  489
  490
  491
  492
  493
  494
  495
  496
  497
  498
  499
  500
  501
  502
  503
  504
  505
  506
  507
  508
  509
  510
  511
  512
  513
  514
  515
  516
  517
  518
  519
  520
  521
  522
  523
  524
  525
  526
  527
  528
  529
  530
  531
  532
  533
  534
  535
  536
  537
  538
  539
  540
  541
  542
  543
  544
  545
  546
  547
  548
  549
  550
  551
  552
  553
  554
  555
  556
  557
  558
  559
  560
  561
  562
  563
  564
  565
  566
  567
  568
  569
  570
  571
  572
  573
  574
  575
  576
  577
  578
  579
  580
  581
  582
  583
  584
  585
  586
  587
  588
  589
  590
  591
  592
  593
  594
  595
  596
  597
  598
  599
  600
  601
  602
  603
  604
  605
  606
  607
  608
  609
  610
  611
  612
  613
  614
  615
  616
  617
  618
  619
  620
  621
  622
  623
  624
  625
  626
  627
  628
  629
  630
  631
  632
  633
  634
  635
  636
  637
  638
  639
  640
  641
  642
  643
  644
  645
  646
  647
  648
  649
  650
  651
  652
  653
  654
  655
  656
  657
  658
  659
  660
  661
  662
  663
  664
  665
  666
  667
  668
  669
  670
  671
  672
  673
  674
  675
  676
  677
  678
  679
  680
  681
  682
  683
  684
  685
  686
  687
  688
  689
  690
  691
  692
  693
  694
  695
  696
  697
  698
  699
  700
  701
  702
  703
  704
  705
  706
  707
  708
  709
  710
  711
  712
  713
  714
  715
  716
  717
  718
  719
  720
  721
  722
  723
  724
  725
  726
  727
  728
  729
  730
  731
  732
  733
  734
  735
  736
  737
  738
  739
  740
  741
  742
  743
  744
  745
  746
  747
  748
  749
  750
  751
  752
  753
  754
  755
  756
  757
  758
  759
  760
  761
  762
  763
  764
  765
  766
  767
  768
  769
  770
  771
  772
  773
  774
  775
  776
  777
  778
  779
  780
  781
  782
  783
  784
  785
  786
  787
  788
  789
  790
  791
  792
  793
  794
  795
  796
  797
  798
  799
  800
  801
  802
  803
  804
  805
  806
  807
  808
  809
  810
  811
  812
  813
  814
  815
  816
  817
  818
  819
  820
  821
  822
  823
  824
  825
  826
  827
  828
  829
  830
  831
  832
  833
  834
  835
  836
  837
  838
  839
  840
  841
  842
  843
  844
  845
  846
  847
  848
  849
  850
  851
  852
  853
  854
  855
  856
  857
  858
  859
  860
  861
  862
  863
  864
  865
  866
  867
  868
  869
  870
  871
  872
  873
  874
  875
  876
  877
  878
  879
  880
  881
  882
  883
  884
  885
  886
  887
  888
  889
  890
  891
  892
  893
  894
  895
  896
  897
  898
  899
  900
  901
  902
  903
  904
  905
  906
  907
  908
  909
  910
  911
  912
  913
  914
  915
  916
  917
  918
  919
  920
  921
  922
  923
  924
  925
  926
  927
  928
  929
  930
  931
  932
  933
  934
  935
  936
  937
  938
  939
  940
  941
  942
  943
  944
  945
  946
  947
  948
  949
  950
  951
  952
  953
  954
  955
  956
  957
  958
  959
  960
  961
  962
  963
  964
  965
  966
  967
  968
  969
  970
  971
  972
  973
  974
  975
  976
  977
  978
  979
  980
  981
  982
  983
  984
  985
  986
  987
  988
  989
  990
  991
  992
  993
  994
  995
  996
  997
  998
  999
  1000
  1001
  1002
  1003
  1004
  1005
  1006
  1007
  1008
  1009
  1010
  1011
  1012
  1013
  1014
  1015
  1016
  1017
  1018
  1019
  1020
  1021
  1022
  1023
  1024
  1025
  1026
  1027
  1028
  1029
  1030
  1031
  1032
  1033
  1034
  1035
  1036
  1037
  1038
  1039
  1040
  1041
  1042
  1043
  1044
  1045
  1046
  1047
  1048
  1049
  1050
  1051
  1052
  1053
  1054
  1055
  1056
  1057
  1058
  1059
  1060
  1061
  1062
  1063
  1064
  1065
  1066
  1067
  1068
  1069
  1070
  1071
  1072
  1073
  1074
  1075
  1076
  1077
  1078
  1079
  1080
  1081
  1082
  1083
  1084
  1085
  1086
  1087
  1088
  1089
  1090
  1091
  1092
  1093
  1094
  1095
  1096
  1097
  1098
  1099
  1100
  1101
  1102
  1103
  1104
  1105
  1106
  1107
  1108
  1109
  1110
  1111
  1112
  1113
  1114
  1115
  1116
  1117
  1118
  1119
  1120
  1121
  1122
  1123
  1124
  1125
  1126
  1127
  1128
  1129
  1130
  1131
  1132
  1133
  1134
  1135
  1136
  1137
  1138
  1139
  1140
  1141
  1142
  1143
  1144
  1145
  1146
  1147
  1148
  1149
  1150
  1151
  1152
  1153
  1154
  1155
  1156
  1157
  1158
  1159
  1160
  1161
  1162
  1163
  1164
  1165
  1166
  1167
  1168
  1169
  1170
  1171
  1172
  1173
  1174
  1175
  1176
  1177
  1178
  1179
  1180
  1181
  1182
  1183
  1184
  1185
  1186
  1187
  1188
  1189
  1190
  1191
  1192
  1193
  1194
  1195
  1196
  1197
  1198
  1199
  1200
  1201
  1202
  1203
  1204
  1205
  1206
  1207
  1208
  1209
  1210
  1211
  1212
  1213
  1214
  1215
  1216
  1217
  1218
  1219
  1220
  1221
  1222
  1223
  1224
  1225
  1226
  1227
  1228
  1229
  1230
  1231
  1232
  1233
  1234
  1235
  1236
  1237
  1238
  1239
  1240
  1241
  1242
  1243
  1244
  1245
  1246
  1247
  1248
  1249
  1250
  1251
  1252
  1253
  1254
  1255
  1256
  1257
  1258
  1259
  1260
  1261
  1262
  1263
  1264
  1265
  1266
  1267
  1268
  1269
  1270
  1271
  1272
  1273
  1274
  1275
  1276
  1277
  1278
  1279
  1280
  1281
  1282
  1283
  1284
  1285
  1286
  1287
  1288
  1289
  1290
  1291
  1292
  1293
  1294
  1295
  1296
  1297
  1298
  1299
  1300
  1301
  1302
  1303
  1304
  1305
  1306
  1307
  1308
  1309
  1310
  1311
  1312
  1313
  1314
  1315
  1316
  1317
  1318
  1319
  1320
  1321
  1322
  1323
  1324
  1325
  1326
  1327
  1328
  1329
  1330
  1331
  1332
  1333
  1334
  1335
  1336
  1337
  1338
  1339
  1340
  1341
  1342
  1343
  1344
  1345
  1346
  1347
  1348
  1349
  1350
  1351
  1352
  1353
  1354
  1355
  1356
  1357
  1358
  1359
  1360
  1361
  1362
  1363
  1364
  1365
  1366
  1367
  1368
  1369
  1370
  1371
  1372
  1373
  1374
  1375
  1376
  1377
  1378
  1379
  1380
  1381
  1382
  1383
  1384
  1385
  1386
  1387
  1388
  1389
  1390
  1391
  1392
  1393
  1394
  1395
  1396
  1397
  1398
  1399
  1400
  1401
  1402
  1403
  1404
  1405
  1406
  1407
  1408
  1409
  1410
  1411
  1412
  1413
  1414
  1415
  1416
  1417
  1418
  1419
  1420
  1421
  1422
  1423
  1424
  1425
  1426
  1427
  1428
  1429
  1430
  1431
  1432
  1433
  1434
  1435
  1436
  1437
  1438
  1439
  1440
  1441
  1442
  1443
  1444
  1445
  1446
  1447
  1448
  1449
  1450
  1451
  1452
  1453
  1454
  1455
  1456
  1457
  1458
  1459
  1460
  1461
  1462
  1463
  1464
  1465
  1466
  1467
  1468
  1469
  1470
  1471
  1472
  1473
  1474
  1475
  1476
  1477
  1478
  1479
  1480
  1481
  1482
  1483
  1484
  1485
  1486
  1487
  1488
  1489
  1490
  1491
  1492
  1493
  1494
  1495
  1496
  1497
  1498
  1499
  1500
  1501
  1502
  1503
  1504
  1505
  1506
  1507
  1508
  1509
  1510
  1511
  1512
  1513
  1514
  1515
  1516
  1517
  1518
  1519
  1520
  1521
  1522
  1523
  1524
  1525
  1526
  1527
  1528
  1529
  1530
  1531
  1532
  1533
  1534
  1535
  1536
  1537
  1538
  1539
  1540
  1541
  1542
  1543
  1544
  1545
  1546
  1547
  1548
  1549
  1550
  1551
  1552
  1553
  1554
  1555
  1556
  1557
  1558
  1559
  1560
  1561
  1562
  1563
  1564
  1565
  1566
  1567
  1568
  1569
  1570
  1571
  1572
  1573
  1574
  1575
  1576
  1577
  1578
  1579
  1580
  1581
  1582
  1583
  1584
  1585
  1586
  1587
  1588
  1589
  1590
  1591
  1592
  1593
  1594
  1595
  1596
  1597
  1598
  1599
  1600
  1601
  1602
  1603
  1604
  1605
  1606
  1607
  1608
  1609
  1610
  1611
  1612
  1613
  1614
  1615
  1616
  1617
  1618
  1619
  1620
  1621
  1622
  1623
  1624
  1625
  1626
  1627
  1628
  1629
  1630
  1631
  1632
  1633
  1634
  1635
  1636
  1637
  1638
  1639
  1640
  1641
  1642
  1643
  1644
  1645
  1646
  1647
  1648
  1649
  1650
  1651
  1652
  1653
  1654
  1655
  1656
  1657
  1658
  1659
  1660
  1661
  1662
  1663
  1664
  1665
  1666
  1667
  1668
  1669
  1670
  1671
  1672
  1673
  1674
  1675
  1676
  1677
  1678
  1679
  1680
  1681
  1682
  1683
  1684
  1685
  1686
  1687
  1688
  1689
  1690
  1691
  1692
  1693
  1694
  1695
  1696
  1697
  1698
  1699
  1700
  1701
  1702
  1703
  1704
  1705
  1706
  1707
  1708
  1709
  1710
  1711
  1712
  1713
 
```

### 1.3 Function Level Stripes

The screenshot shows a Java IDE interface with the following details:

- File Path:** file:///Assignment\_2/Assignment\_2/src/j/FunctionLevelStrips.java
- Code Editor Content:** The code defines a class `TokenizerMapper` that extends `Mapper<Object, Text, Text, MapWritable>`. It includes methods for parsing a cached file and mapping objects to text values.
- Output Terminal:** Shows the execution of the code, including statistics like records processed, spilled records, and memory usage.
- Bottom Status Bar:** Displays the current time (Ln 116, Col 29), space usage (Space: 4), and terminal status (Ready).

Figure 18: Function - level stripes  $d = 1$

The screenshot shows an IDE interface with several tabs and panes. The main pane displays the Java code for a MapReduce job named `J_FunctionLevelStripes.java`. The code includes imports for `Mapper`, `Text`, `Context`, and `MapWritable`. It defines a class `FunctionLevelStripes` that extends `Mapper<Object, Text, Text, MapWritable>`. The `map` method reads tokens from a file, splits them into words, and maps them to a key-value pair where the value is the word converted to uppercase. The `reduce` method then concatenates all values for each key. Below the code, the terminal output shows the execution of the job, including shuffle statistics, record counts, and memory usage.

```
File Edit Selection View Go Run Terminal Help
... J_FunctionLevelStripes.java 9+ Assignment_2 [SSH: 172.16.232.209]
EXPLORER
ASSIGNMENT_2 [SSH: 172.16.232.209]
data
res
outputs
scripts
src
stopwords.txt

src J_FunctionLevelStripes.java > J_FunctionLevelStripes > TokenizerMapper > Mapper<Object, Text, Context>
27 public class FunctionLevelStripes
28 {
29     public static class TokenizerMapper extends Mapper<Object, Text, Text, MapWritable>
30     {
31         private void parseSkipFile(String filename)
32         {
33             try
34             {
35                 BufferedReader reader = new BufferedReader(new FileReader(filename));
36                 String line;
37                 while ((line = reader.readLine()) != null)
38                 {
39                     StringTokenizer tokens = new StringTokenizer(line);
40                     while (tokens.hasMoreTokens())
41                     {
42                         String token = tokens.nextToken();
43                         if (!token.equals(" "))
44                         {
45                             System.out.println(token);
46                         }
47                     }
48                 }
49             }
50             catch (IOException ioe)
51             {
52                 System.err.println("Caught exception while parsing the cached file: " + StringUtils.stringifyException(ioe));
53             }
54         }
55     }
56     @Override
57     public void map(Object key, Text value, Context context) throws IOException, InterruptedException
58     {
59         String line = value.toString().toLowerCase();
60         String tokens[] = line.split("(\\W)+");
61         int distance = 2;
62         Map<String, Map<String, Text>> associative_array = new TreeMap<>();
63         for (int i = 0; i < tokens.length; i++)
64         {
65             if (tokens.contains(tokens[i]))
66             {
67                 Shuffle shuffle = new Shuffle();
68                 shuffle.setInputRecordId(0);
69                 shuffle.setOutputRecordId(0);
70                 shuffle.setSpilledRecords(0);
71                 shuffle.setShuffledMaps(0);
72                 shuffle.setTotalInputRecords(0);
73                 shuffle.setMergedMapOutputs(0);
74                 GC time elapsed (ms)=12851
75                 Total committed heap usage (bytes)=24758987411456
76                 BAD_ID=0
77                 CONNECTED=0
78                 DL_BLOCKED=0
79                 WRONG_LENGTH=0
80                 WRONG_MAP=0
81                 WRONG_SENSE=0
82                 File Input Format Counters
83                 Bytes Read=166222646
84                 File Output Format Counters
85                 Bytes Written=29296
86             real 5m33.036s
87             user 1m55s.683s
88             sys 1m6.479s
89             o 0m0user@kurtisdesver:~/Assignment_2/scripts [ ]
```

Figure 19: Function - level stripes  $d = 2$

```

    ...
    public class FunctionLevelStripes {
        ...
        private void parseStopFile(String filename) {
            ...
        }
        ...
        @Override
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            ...
            String line = value.toString().toLowerCase();
            String tokens[] = line.split("\\s+");
            int distance = 3;
            ...
            Map<Text, Text> associative_array = new TreeMap<>();
            for(int i = 0; i < tokens.length; i++) {
                ...
                if(words.contains(tokens[i])) {
                    ...
                }
            }
        }
        ...
        Reduce shuffle bytes=11384596
        Reduce input records=318522
        Reduce output bytes=159475
        Spilled Records=337944
        Shuffled Maps =10000
        Failed Shuffles=
        Merged Shuffles=10000
        GC Time elapsed (ms)=19842
        Total committed heap usage (bytes)=23313175142400
        Shuffle Errors:
        IO=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_TYPE=0
        WRONG_VALUE=0
        WRONG_REDUCE=0
        File Input Format Counters
        Bytes Read=106222646
        File Output Format Counters
        Bytes Written=386576
        ...
        real 5m2.649s
        user 15m9.596s
        sys 1m7.932s
        vboxuser@UbuntuServer:~/Assignment_2/scripts$ 
  
```

Figure 20: Function - level stripes  $d = 3$

```

    ...
    public class FunctionLevelStripes {
        ...
        private void parseStopFile(String filename) {
            ...
        }
        ...
        @Override
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            ...
            String line = value.toString().toLowerCase();
            String tokens[] = line.split("\\s+");
            int distance = 3;
            ...
            Map<Text, Text> associative_array = new TreeMap<>();
            for(int i = 0; i < tokens.length; i++) {
                ...
                if(words.contains(tokens[i])) {
                    ...
                }
            }
        }
        ...
        Reduce shuffle bytes=11384596
        Reduce input records=318522
        Reduce output bytes=159475
        Spilled Records=337944
        Shuffled Maps =10000
        Failed Shuffles=
        Merged Shuffles=10000
        GC Time elapsed (ms)=19842
        Total committed heap usage (bytes)=23313175142400
        Shuffle Errors:
        IO=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_TYPE=0
        WRONG_VALUE=0
        WRONG_REDUCE=0
        File Input Format Counters
        Bytes Read=106222646
        File Output Format Counters
        Bytes Written=386576
        ...
        real 5m2.649s
        user 15m9.596s
        sys 1m7.932s
        vboxuser@UbuntuServer:~/Assignment_2/scripts$ 
  
```

Figure 21: Function - level stripes  $d = 4$

## 1.4 Class Level Stripes

The screenshot shows an IDE interface with the title bar "Assignment\_2 [SSH: 172.16.232.209]". The left sidebar shows a project structure for "ASSIGNMENT\_2 [SSH: 172.16.232.209]" containing "data", "files", "outputs", "scripts", "src", and "stopwords.txt". The main editor window displays Java code for "ClassLevelStripes.java". The code defines a public class "ClassLevelStripes" with a static inner class "TokenizerMapper" that implements the "Mapper<Object, Text, Context>" interface. The "map" method reads tokens from a line of text and puts them into an associative array. The "reduce" method then processes these tokens. The code also includes imports for "java.util", "org.apache.hadoop.mapreduce", and "org.apache.hadoop.io". Below the code, the terminal output shows Hadoop logs for a map-reduce job, indicating successful execution with 100% map and reduce completion. The bottom status bar shows the terminal session is ready.

Figure 22: Class - level stripes  $d = 1$

The screenshot shows an IDE interface with the title bar "Assignment\_2 [SSH: 172.16.232.209]". The left sidebar shows a project structure for "ASSIGNMENT\_2 [SSH: 172.16.232.209]" containing "data", "files", "outputs", "scripts", "src", and "stopwords.txt". The main editor window displays Java code for "ClassLevelStripes.java". The code defines a public class "ClassLevelStripes" with a static inner class "TokenizerMapper" that implements the "Mapper<Object, Text, Context>" interface. The "map" method reads tokens from a line of text and puts them into an associative array. The "reduce" method then processes these tokens. The code also includes imports for "java.util", "org.apache.hadoop.mapreduce", and "org.apache.hadoop.io". Below the code, the terminal output shows Hadoop logs for a map-reduce job, indicating successful execution with 100% map and reduce completion. The bottom status bar shows the terminal session is ready.

Figure 23: Class - level stripes  $d = 2$

```

File Output Format Counters
Bytes Written=38676
2025-03-21 15:28:42,170 INFO mapred.LocalJobRunner: Finishing task: attempt_local100154919_0001_r_000000
2025-03-21 15:28:42,171 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-21 15:28:42,172 INFO mapred.MrAppMaster: mapreduce.job.maps=1
2025-03-21 15:28:42,173 INFO mapred.MrAppMaster: mapreduce.job.reduces=1
2025-03-21 15:28:44,573 INFO mapreduce.Job: Job job_100154919_0001 completed successfully
2025-03-21 15:28:44,957 INFO mapreduce.Counters: Counters: 30
Map-Reduce Framework
Map input records=10000
Map output records=118522
Map input bytes=11384506
Map output bytes=11384506
Input split bytes=1436997
Combine input records=0
Combine output records=0
Reduce input groups=50
Reduce shuffle bytes=11384506
Reduce input records=318522
Reduce output records=54745
Spilled Records=337944
Shuffled Maps =10000
Failed Shuffles=
Merged Map output=10000
GC time elapsed (ms)=10405
Total committed heap usage (bytes)=25544239349768
Shuffle Errors
  IO=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_TYPE=0
  WRONG_REDUCE=0
File Input Format Counters
Bytes Read=10622266
File Output Format Counters
Bytes Written=38676
real 5m3.43s
user 15m9.96s
sys 1m8.99s
vboxuser@UbuntuServer:~/Assignment_2/scripts$ 

```

Figure 24: Class - level stripes  $d = 3$

```

File Output Format Counters
Bytes Written=31205
2025-03-21 15:28:42,170 INFO mapred.LocalJobRunner: Finishing task: attempt_local100154919_0001_r_000000
2025-03-21 15:28:42,171 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-21 15:28:42,172 INFO mapred.MrAppMaster: mapreduce.job.maps=1
2025-03-21 15:28:42,173 INFO mapred.MrAppMaster: mapreduce.job.reduces=1
2025-03-21 15:28:44,573 INFO mapreduce.Job: Job job_100154919_0001 completed successfully
2025-03-21 15:28:44,957 INFO mapreduce.Counters: Counters: 30
Map-Reduce Framework
Map input records=10000
Map output records=1258763
Map input bytes=11384506
Map output bytes=11384506
Input split bytes=1436997
Combine input records=0
Combine output records=0
Reduce input groups=50
Reduce shuffle bytes=11384506
Reduce input records=318522
Reduce output records=54745
Spilled Records=337944
Shuffled Maps =10000
Failed Shuffles=
Merged Map output=10000
GC time elapsed (ms)=10405
Total committed heap usage (bytes)=2492688869120
Shuffle Errors
  IO=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_TYPE=0
  WRONG_REDUCE=0
File Input Format Counters
Bytes Read=10622266
File Output Format Counters
Bytes Written=31205
real 5m3.42s
user 16m11.13s
sys 1m4.01s
vboxuser@UbuntuServer:~/Assignment_2/scripts$ 

```

Figure 25: Class - level stripes  $d = 4$

## Problem 5: Indexing Documents via Hadoop

### a) Document Frequency (DF) Calculation

**Objective:** Implement MapReduce to calculate the DF of each distinct term, filtering out stop words.  
**Execution Steps:**

1. Preprocess the Wikipedia dump.
2. Use MapReduce to count DF.
3. Store the DF values in a TSV file.

## Output Format:

TERM<TAB>DF
data 3456
hadoop 2890
map 2678

**Map :**

- We read *stopwords.txt* and store all the words in the file in a Set for filtering.
  - We read the input, split the line into words and stem using *PorterStemmer* from *openlp-tools* and convert to lowercase, We allow those words which are not in *stopwords.txt* and only distinct words using Set datastructure.
  - We filter words using sets by sending (word,1) when word occurs the first time (not in set) and add it to the set i.e. we send a word only if it is not in the set and also stopwords are filtered using sets.

**Reduce :**

- We declare a priority queue inorder to obtain top 100 words with highest document frequency.
  - We receive values for a key which we sum the values (we get 1 for each document) and store the key and sum in the priority queue and if the number of elements in priority queue is more than 100 we poll it i.e. the key value pair with least document frequency is popped.
  - We output all the key sum pairs stored in the priority queue as it contains only top 100 words which have highest document frequency.

Figure 26: Q5.a

### b) Term Frequency-Inverse Document Frequency (TF-IDF)

**Objective:** Calculate the TF-IDF score using MapReduce with stripes algorithm.

### **Execution Steps:**

- ## 1. Preprocess the Wikipedia dump.

2. Use MapReduce with stripes to compute TF.
3. Cache the DF file using `Job.addCacheFile`.
4. Calculate TF-IDF:

$$\text{SCORE} = \text{TF} \times \log \left( \frac{10000}{\text{DF} + 1} \right)$$

5. Store results in a TSV file.

#### **Output Format:**

```
ID<TAB>TERM<TAB>SCORE
123      data      0.678
124      hadoop    0.592
125      map       0.502
```

#### **Map :**

- We extract the id of file from its name. We then read the output file from previous map reduce ans store the words in a set (top 100 words containing highest document frequency).
- we read the lines from input files ans split them into words and stem them using *PorterStemmer* from *opennlp-tools* and convert them to lower case and if the word is in the set having words with highest document frequency the we store it in a Map (if map already has the word we add 1 to its Term Frequency or else we put (word,1) in the map).
- we release the map (having word, tf) as value and document id as key.

#### **Reduce :**

- We then read the output file from previous map reduce ans store the words and their document frequency in a Map (top 100 words containing highest document frequency).
- for a key(id) we receive a Map (stripes) we iterate over the map (value) and calculates its score using formula :

$$\text{SCORE} = \text{TF} \times \log \left( \frac{10000}{\text{DF} + 1} \right)$$

and here DF is document frequency is obtained from map containing words with highest document frequency.

- for each word we release a key value pair as (id, (word,score)) (word is obtained from map (value) format : ID<TAB>TERM<TAB>SCORE).

#### **Execution Time:**

```

File Edit Selection View Go Run Terminal Help
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reduce shuffle bytes=6622577
Reduce Input Records=10000
Reduced Output Records=10000
Spilled Records=10000
Shuffled Maps=10000
Failed Shuffles=0
Merged Map outputs=10000
Or 1.00% of total (0.1)=10000
Total committed heap usage (bytes)=2539651072
Shuffle Errors:
BAD IDB
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_PARTITION=0
File Output Format Counters
Bytes Written=18586938
2025-03-20 22:51:43.797 INFO mapred.LocalJobRunner: Finishing task: attempt_local001_023391_0001_r_000000_0
2025-03-20 22:51:43.797 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-03-20 22:51:43.802 INFO mapred.Task: Task finished in 10ms
2025-03-20 22:51:43.802 INFO mapreduce.Job: Job [job_local001_023391_0001] completed successfully
2025-03-20 22:51:45.446 INFO mapreduce.Job: Counters: 30
File Input Format Counters
FILE: Number of bytes read=167910589482
FILE: Number of bytes written=20304289375
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map Input records=10000
Map Input bytes=10000000
Map Output bytes=6522685
Map output materialized bytes=6622577
Input File(s) read=1
Combine Input records=0
Combine output records=0
Partial key output records=0
Reduce shuffle bytes=6622577
Reduced Input Records=10000
Reduced Output Records=10000
Spilled Records=20000
Shuffled Maps=10000
Failed Shuffles=0
Merged Map outputs=10000
Or 1.00% of total (0.1)=10000
Total committed heap usage (bytes)=25151466897408
Shuffle Errors:
BAD_IDB
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_PARTITION=0
File Input Format Counters
Bytes Read=10000000
File Output Format Counters
Bytes Written=18586938
real: 2m5.738s
user: 7m25.826s
sys: 0m25.296s
pi@elab01: ~/Downloads/hdp_part2

```

Figure 27: Q5.b

## System Configuration

The Following is the configuration of the system that was used for running the mapreduce task for problem 4 and 5.

1. RAM:10000 MB
2. No of Cores : 9 Cores
3. CPU : Intel i7 12th gen

**Note-**These are the configurations of the virtual machine that we used for running mapreduce tasks.