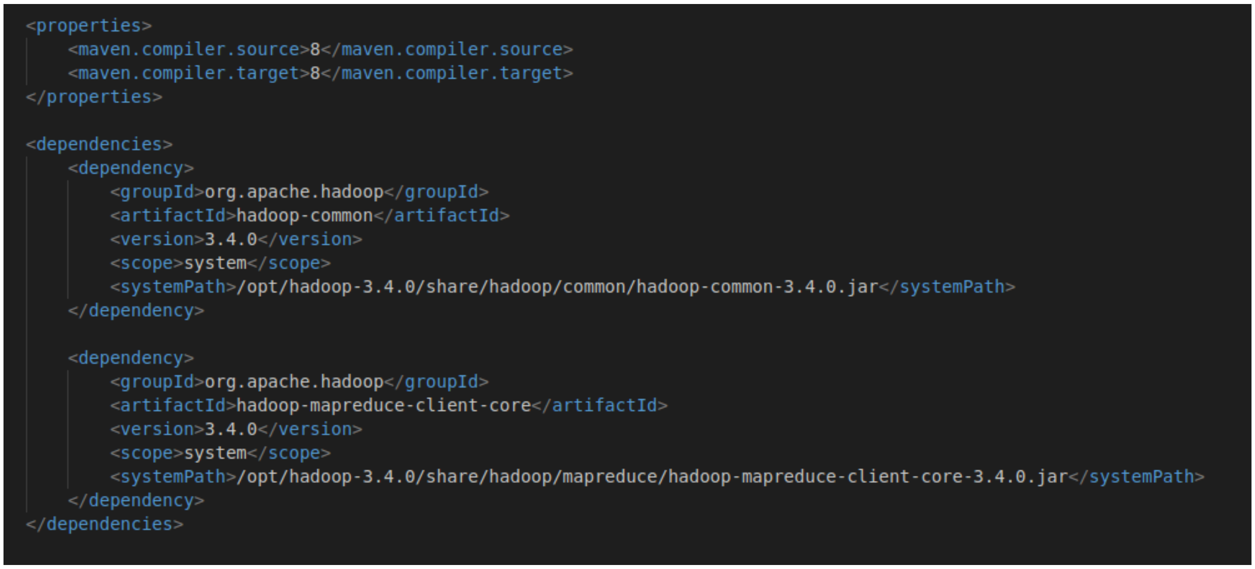
NOSQL SYSTEMS

Assignment-2
REPORT

Introduction

Here, I tried to run the Wikipedia-EN-20120601_ARTICLES.tar.gz input-file. But, unfortunately, my laptop has only 8GB RAM, so it wouldn't work out. So, I used Wikipedia-50-ARTICLES.tar for running the code.

I created the project using Maven, and in the pom.xml file, I added these lines of codes:

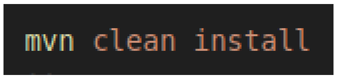


```xml
<properties>
    <maven.compiler.source>8</maven.compiler.source>
    <maven.compiler.target>8</maven.compiler.target>
</properties>

<dependencies>
    <dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-common</artifactId>
        <version>3.4.0</version>
        <scope>system</scope>
        <systemPath>/opt/hadoop-3.4.0/share/hadoop/common/hadoop-common-3.4.0.jar</systemPath>
    </dependency>

    <dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-mapreduce-client-core</artifactId>
        <version>3.4.0</version>
        <scope>system</scope>
        <systemPath>/opt/hadoop-3.4.0/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.4.0.jar</systemPath>
    </dependency>
</dependencies>
```

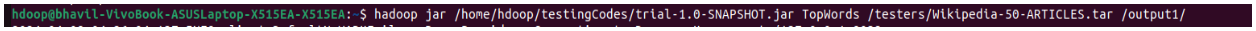
Then, copy the relevant java file to the source folder, and run:



```
mvn clean install
```

Now, we have the jar file for the java file. Now, in my system, I had a separate user called “hdoop”, which had hadoop in it. Now, in that user, I have to copy the jar file generated. Also, in the HDFS, I have to put the required input file.

Now, run the hadoop command:



```
hdoop@bhavil-VivoBook-ASUSLaptop-X515EA-X515EA:~$ hadoop jar /home/hdoop/testingCodes/trial-1.0-SNAPSHOT.jar TopWords /testers/Wikipedia-50-ARTICLES.tar /output1/
```

To run hadoop on the jar file, and get the outputs. We can also get some screenshots from localhost:8088.

Co-occurring word matrix generation: Pairs & Stripes and local aggregation

PART-A

Program Logic:

The program aims to identify the top 50 most frequently occurring words from a given input dataset while excluding stop words. It utilizes Hadoop's MapReduce framework to distribute the computation across multiple nodes in a cluster.

Pseudocode Explanation:

Mapper Phase:

Load stop words from the distributed cache.

For each input record:

Tokenize the text into words.

Convert each word to lowercase and check if it is not a stop word.

Emit key-value pairs where the word is the key and the count is set to 1.

Reducer Phase:

Initialize a priority queue to store the top 50 words based on their counts.

For each word received:

Sum up the counts for the same word.

Add the word and its count to the priority queue.

If the queue size exceeds 50, remove the least frequent word.

After processing all words, emit the top 50 words from the priority queue.

Runtime Analysis:

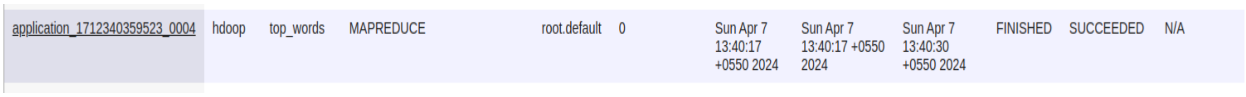
The program utilizes Hadoop's MapReduce framework, which is highly scalable and can handle large datasets efficiently by distributing the workload across multiple nodes.

The Mapper and Reducer tasks are executed in parallel across the cluster, leveraging the distributed computing capabilities of Hadoop.

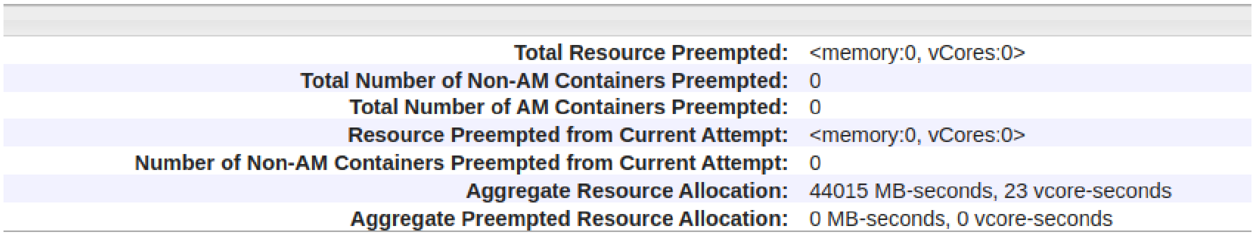
The runtime performance depends on various factors such as the size of the input dataset, the number of nodes in the cluster, the hardware configuration, and the complexity of the processing logic.

Generally, Hadoop MapReduce jobs are suitable for processing large-scale datasets, and the runtime can scale linearly with the size of the dataset and the cluster's capacity.

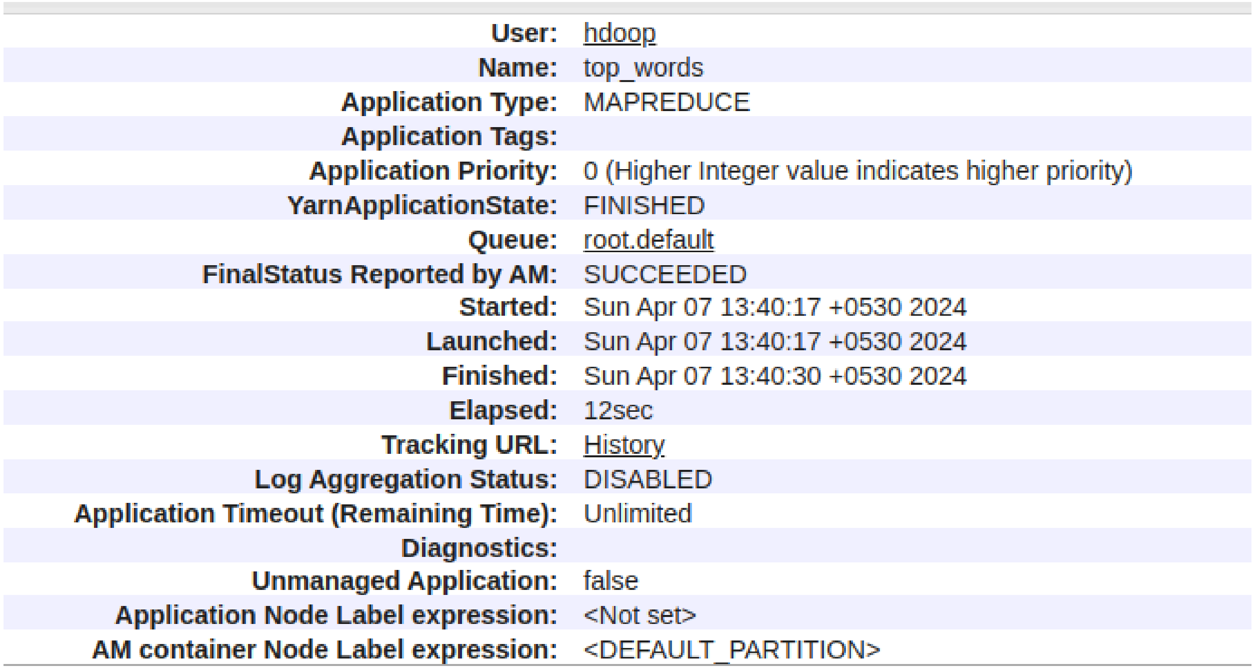
Relevant Screenshots:



application_1712340359523_0004	hdoop	top_words	MAPREDUCE	root.default	0	Sun Apr 7 13:40:17 +0550 2024	Sun Apr 7 13:40:17 +0550 2024	Sun Apr 7 13:40:30 +0550 2024	FINISHED	SUCCEEDED	N/A



Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	44015 MB-seconds, 23 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds



User:	hdoop
Name:	top_words
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Apr 07 13:40:17 +0530 2024
Launched:	Sun Apr 07 13:40:17 +0530 2024
Finished:	Sun Apr 07 13:40:30 +0530 2024
Elapsed:	12sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

PART-B

Program Logic:

The program aims to construct a co-occurring word matrix based on the input dataset, considering different word distances (d). It utilizes Hadoop's MapReduce framework to distribute the computation across multiple nodes in a cluster.

Pseudocode Explanation:

Mapper Phase:

Tokenize each input record into words.

For each word in the record:

Determine the word distance (d) from the configuration.

Generate pairs of words within the specified distance (excluding the word itself).

Emit key-value pairs where the key is a combination of the word and its co-occurring word, and the value is 1.

Reducer Phase:

Sum up the counts for each word pair.

Emit the word pair and its total count.

Main Function:

Iterate over different word distances specified in the 'distances' array.

Set the word distance in the configuration.

Configure and run the MapReduce job for each word distance.

Measure the runtime for each word distance and print the results.

Runtime Analysis:

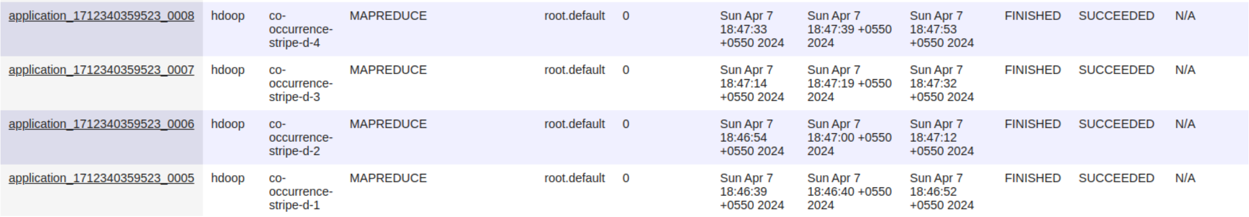
The program runs multiple MapReduce jobs, each with a different word distance (d), specified in the 'distances' array.

The runtime performance depends on factors such as the size of the input dataset, the number of nodes in the Hadoop cluster, the complexity of the processing logic, and the specified word distances.

As the word distance increases, the number of word pairs considered for co-occurrence also increases, potentially leading to longer execution times.

The program utilizes Hadoop's distributed processing capabilities to handle large-scale datasets efficiently, with runtime scaling linearly with the dataset size and cluster capacity.

Relevant Screenshots:



application_1712340359523_0008	hdoop	co-occurrence-stripe-d-4	MAPREDUCE	root.default	0	Sun Apr 7 18:47:33 +0550 2024	Sun Apr 7 18:47:39 +0550 2024	Sun Apr 7 18:47:53 +0550 2024	FINISHED	SUCCEEDED	N/A
application_1712340359523_0007	hdoop	co-occurrence-stripe-d-3	MAPREDUCE	root.default	0	Sun Apr 7 18:47:14 +0550 2024	Sun Apr 7 18:47:19 +0550 2024	Sun Apr 7 18:47:32 +0550 2024	FINISHED	SUCCEEDED	N/A
application_1712340359523_0006	hdoop	co-occurrence-stripe-d-2	MAPREDUCE	root.default	0	Sun Apr 7 18:46:54 +0550 2024	Sun Apr 7 18:47:00 +0550 2024	Sun Apr 7 18:47:12 +0550 2024	FINISHED	SUCCEEDED	N/A
application_1712340359523_0005	hdoop	co-occurrence-stripe-d-1	MAPREDUCE	root.default	0	Sun Apr 7 18:46:39 +0550 2024	Sun Apr 7 18:46:40 +0550 2024	Sun Apr 7 18:46:52 +0550 2024	FINISHED	SUCCEEDED	N/A

PART-C

Program Logic:

The program aims to construct a co-occurrence stripe matrix based on the input dataset, considering different word distances (d). It utilizes Hadoop's MapReduce framework to distribute the computation across multiple nodes in a cluster.

Pseudocode Explanation:

Mapper Phase:

Tokenize each input record into words.

For each word in the record:

Determine the word distance (d) from the configuration.

Generate a stripe (map) of co-occurring words within the specified distance (excluding the word itself).

Emit key-value pairs where the key is the word, and the value is the stripe represented as a string.

Reducer Phase:

Merge the stripes for each word, aggregating the counts for co-occurring words.

Emit the word and its consolidated stripe.

Main Function:

Iterate over different word distances specified in the 'distances' array.

Set the word distance in the configuration.

Configure and run the MapReduce job for each word distance.

Measure the runtime for each word distance and print the results.

Runtime Analysis:

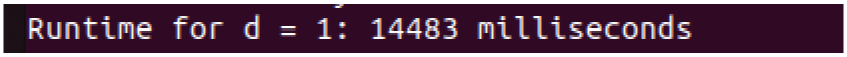
The program runs multiple MapReduce jobs, each with a different word distance (d), specified in the 'distances' array.

Similar to the previous example, the runtime performance depends on factors such as the size of the input dataset, cluster resources, and specified word distances.

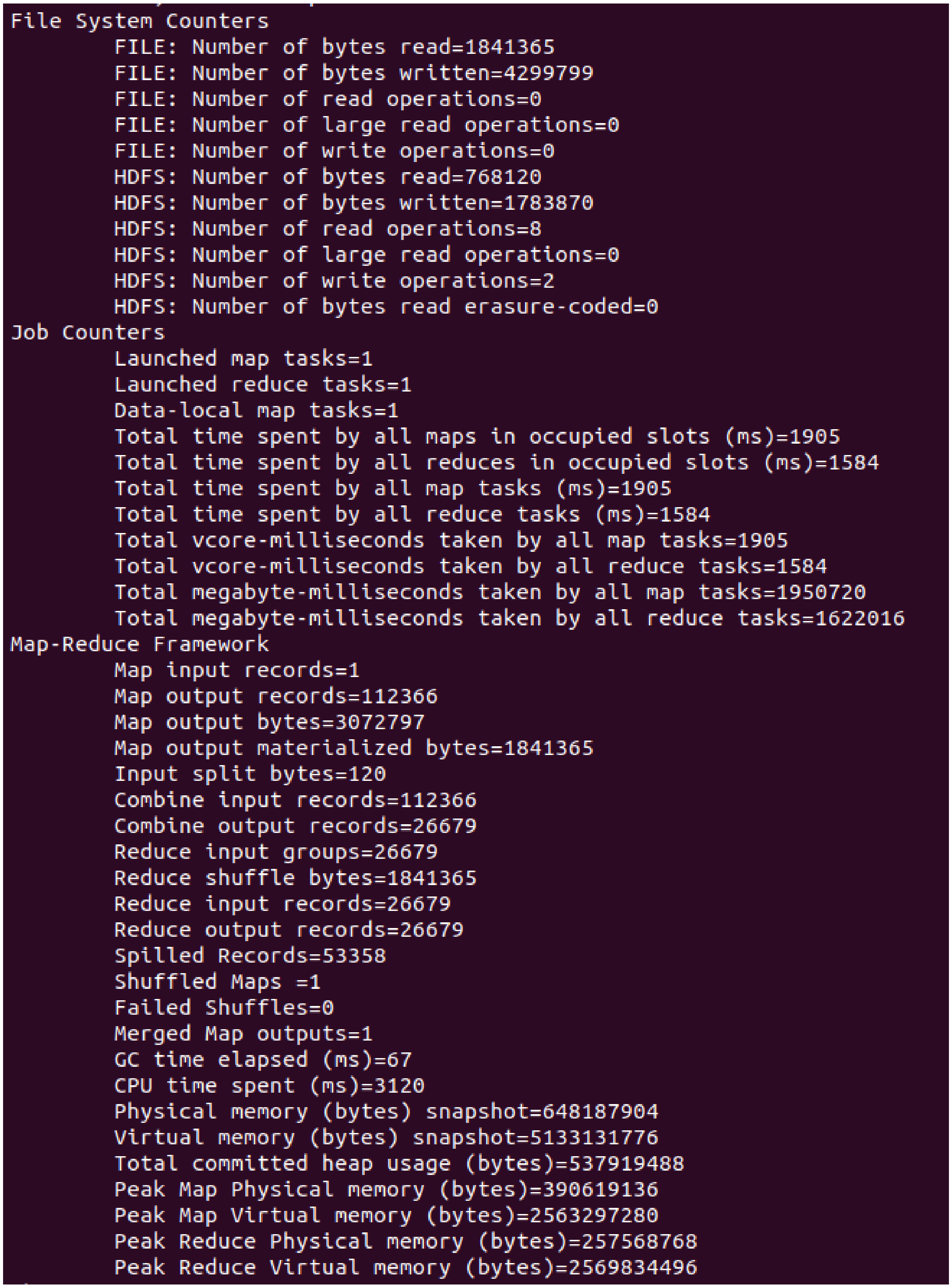
As the word distance increases, the number of co-occurring words considered in the stripe also increases, potentially leading to longer execution times.

The program utilizes Hadoop's distributed processing capabilities to efficiently handle large-scale datasets, with runtime scaling linearly with the dataset size and cluster capacity.

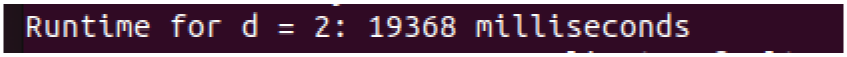
Relevant Screenshots:



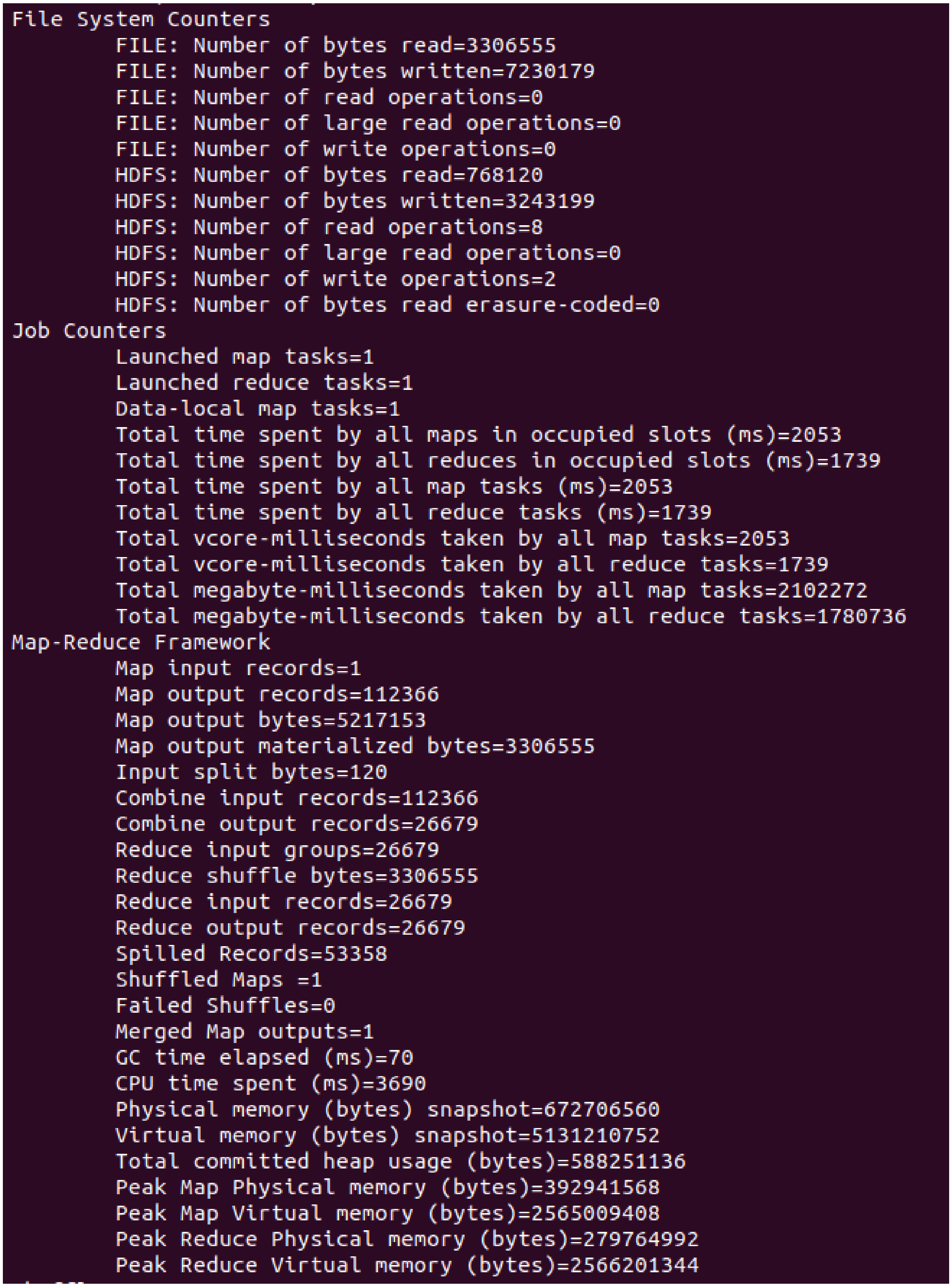
```
Runtime for d = 1: 14483 milliseconds
```



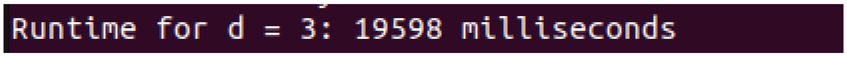
```
File System Counters
        FILE: Number of bytes read=1841365
        FILE: Number of bytes written=4299799
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=1783870
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=1905
        Total time spent by all reduces in occupied slots (ms)=1584
        Total time spent by all map tasks (ms)=1905
        Total time spent by all reduce tasks (ms)=1584
        Total vcore-milliseconds taken by all map tasks=1905
        Total vcore-milliseconds taken by all reduce tasks=1584
        Total megabyte-milliseconds taken by all map tasks=1950720
        Total megabyte-milliseconds taken by all reduce tasks=1622016
Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=3072797
        Map output materialized bytes=1841365
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=1841365
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=67
        CPU time spent (ms)=3120
        Physical memory (bytes) snapshot=648187904
        Virtual memory (bytes) snapshot=5133131776
        Total committed heap usage (bytes)=537919488
        Peak Map Physical memory (bytes)=390619136
        Peak Map Virtual memory (bytes)=2563297280
        Peak Reduce Physical memory (bytes)=257568768
        Peak Reduce Virtual memory (bytes)=2569834496
```



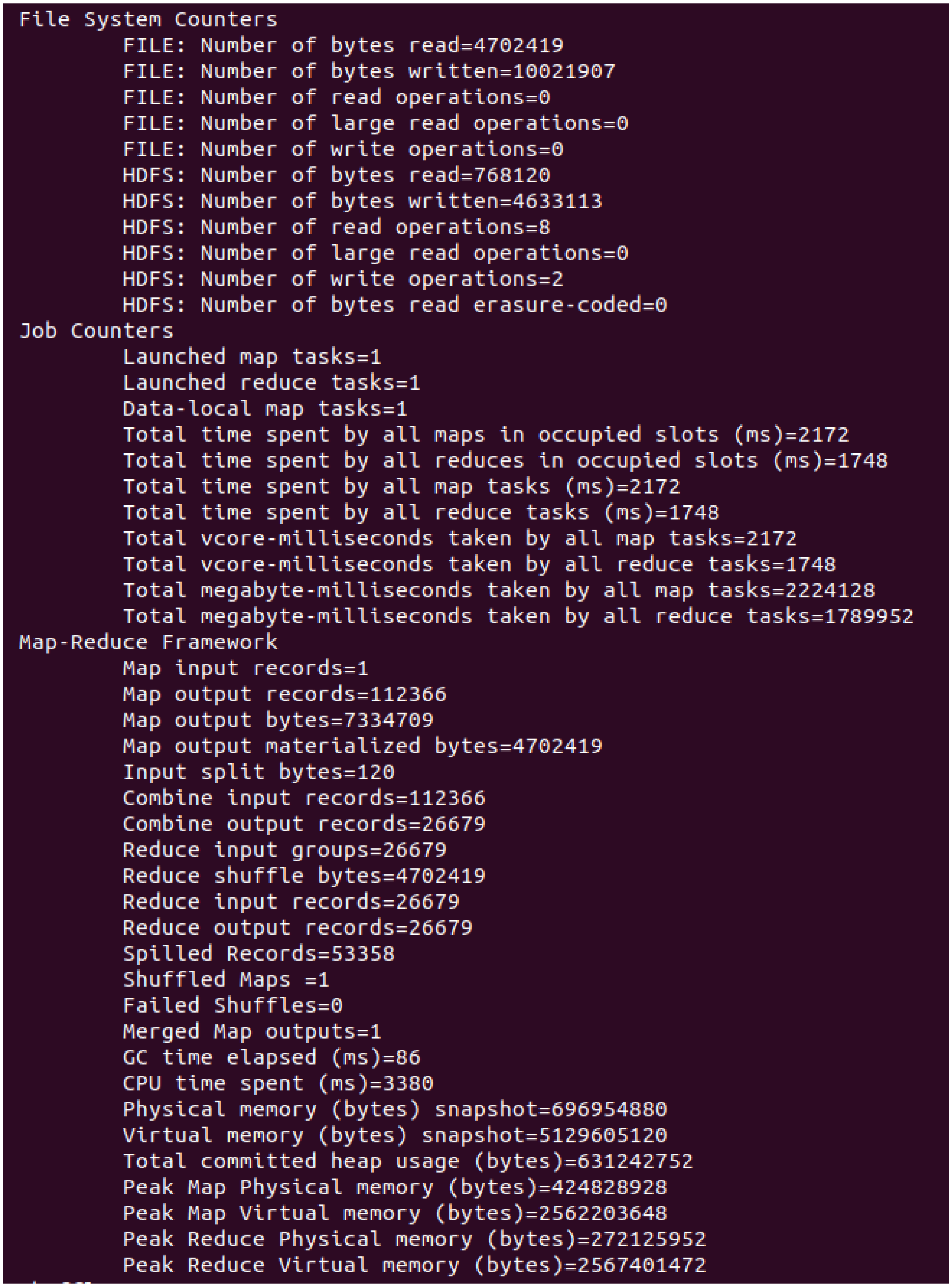
```
Runtime for d = 2: 19368 milliseconds
```



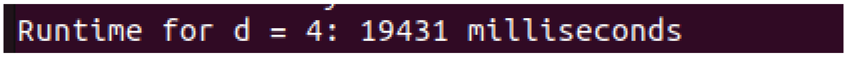
```
File System Counters
        FILE: Number of bytes read=3306555
        FILE: Number of bytes written=7230179
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=3243199
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=2053
        Total time spent by all reduces in occupied slots (ms)=1739
        Total time spent by all map tasks (ms)=2053
        Total time spent by all reduce tasks (ms)=1739
        Total vcore-milliseconds taken by all map tasks=2053
        Total vcore-milliseconds taken by all reduce tasks=1739
        Total megabyte-milliseconds taken by all map tasks=2102272
        Total megabyte-milliseconds taken by all reduce tasks=1780736
Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=5217153
        Map output materialized bytes=3306555
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=3306555
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=70
        CPU time spent (ms)=3690
        Physical memory (bytes) snapshot=672706560
        Virtual memory (bytes) snapshot=5131210752
        Total committed heap usage (bytes)=588251136
        Peak Map Physical memory (bytes)=392941568
        Peak Map Virtual memory (bytes)=2565009408
        Peak Reduce Physical memory (bytes)=279764992
        Peak Reduce Virtual memory (bytes)=2566201344
```



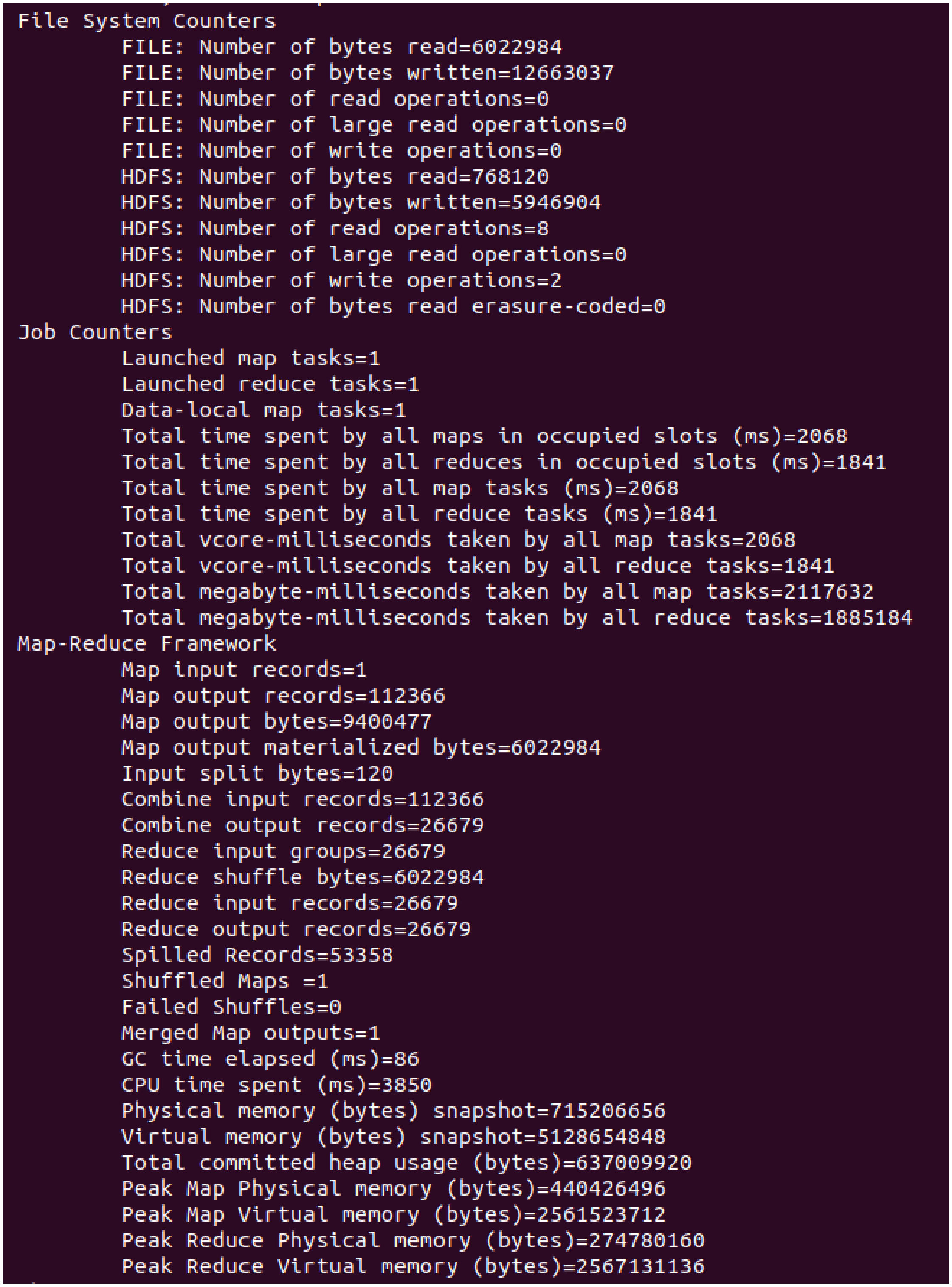
```
Runtime for d = 3: 19598 milliseconds
```



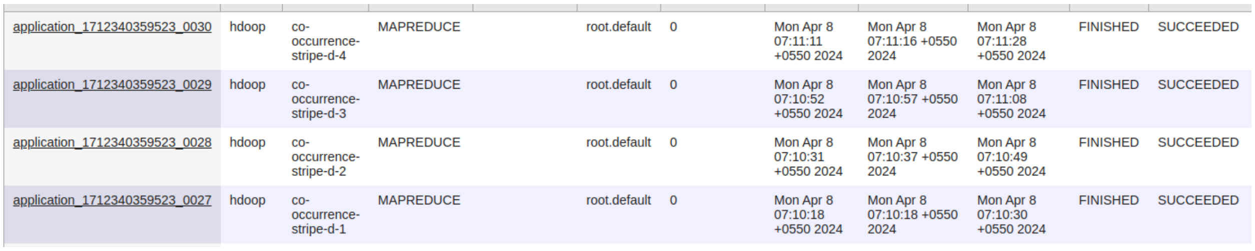
```
 File System Counters
        FILE: Number of bytes read=4702419
        FILE: Number of bytes written=10021907
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=4633113
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
 Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=2172
        Total time spent by all reduces in occupied slots (ms)=1748
        Total time spent by all map tasks (ms)=2172
        Total time spent by all reduce tasks (ms)=1748
        Total vcore-milliseconds taken by all map tasks=2172
        Total vcore-milliseconds taken by all reduce tasks=1748
        Total megabyte-milliseconds taken by all map tasks=2224128
        Total megabyte-milliseconds taken by all reduce tasks=1789952
 Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=7334709
        Map output materialized bytes=4702419
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=4702419
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=86
        CPU time spent (ms)=3380
        Physical memory (bytes) snapshot=696954880
        Virtual memory (bytes) snapshot=5129605120
        Total committed heap usage (bytes)=631242752
        Peak Map Physical memory (bytes)=424828928
        Peak Map Virtual memory (bytes)=2562203648
        Peak Reduce Physical memory (bytes)=272125952
        Peak Reduce Virtual memory (bytes)=2567401472
```



```
Runtime for d = 4: 19431 milliseconds
```



```
File System Counters
        FILE: Number of bytes read=6022984
        FILE: Number of bytes written=12663037
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=5946904
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=2068
        Total time spent by all reduces in occupied slots (ms)=1841
        Total time spent by all map tasks (ms)=2068
        Total time spent by all reduce tasks (ms)=1841
        Total vcore-milliseconds taken by all map tasks=2068
        Total vcore-milliseconds taken by all reduce tasks=1841
        Total megabyte-milliseconds taken by all map tasks=2117632
        Total megabyte-milliseconds taken by all reduce tasks=1885184
Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=9400477
        Map output materialized bytes=6022984
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=6022984
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=86
        CPU time spent (ms)=3850
        Physical memory (bytes) snapshot=715206656
        Virtual memory (bytes) snapshot=5128654848
        Total committed heap usage (bytes)=637009920
        Peak Map Physical memory (bytes)=440426496
        Peak Map Virtual memory (bytes)=2561523712
        Peak Reduce Physical memory (bytes)=274780160
        Peak Reduce Virtual memory (bytes)=2567131136
```



application_1712340359523_0030	hdoop	co-occurrence-stripe-d-4	MAPREDUCE		root.default	0	Mon Apr 8 07:11:11 +0550 2024	Mon Apr 8 07:11:16 +0550 2024	Mon Apr 8 07:11:28 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0029	hdoop	co-occurrence-stripe-d-3	MAPREDUCE		root.default	0	Mon Apr 8 07:10:52 +0550 2024	Mon Apr 8 07:10:57 +0550 2024	Mon Apr 8 07:11:08 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0028	hdoop	co-occurrence-stripe-d-2	MAPREDUCE		root.default	0	Mon Apr 8 07:10:31 +0550 2024	Mon Apr 8 07:10:37 +0550 2024	Mon Apr 8 07:10:49 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0027	hdoop	co-occurrence-stripe-d-1	MAPREDUCE		root.default	0	Mon Apr 8 07:10:18 +0550 2024	Mon Apr 8 07:10:18 +0550 2024	Mon Apr 8 07:10:30 +0550 2024	FINISHED	SUCCEEDED

PART-D

Program Logic:

CoOccurrenceStripe:

Mapper Logic:

Tokenizes input text into individual words.

Constructs a stripe (a map) for each word, where keys are co-occurring words and values are their counts.

Emits each word along with its corresponding stripe.

Reducer Logic:

Receives word-to-stripe mappings from mappers.

Merges the stripes for each word to generate the final co-occurrence counts.

Emits each word along with its merged stripe.

CoOccurrenceMatrix:

Mapper Logic:

Tokenizes input text into individual words.

Constructs pairs of words within the given distance (d) and emits them with a count of 1.

Reducer Logic:

Receives pairs of words and their counts from mappers.

Aggregates the counts for each pair to generate the final co-occurrence counts.

Emits each pair along with its aggregated count.

<u>Pseudocode Explanations:</u>

CoOccurrenceStripe:

```
Mapper:
for each word in the input text:
    construct a stripe for the word
    emit (word, stripe)
Reducer:
for each (word, stripe) received:
    merge the stripes for the word
emit (word, merged stripe)
```

CoOccurrenceMatrix:

```
Mapper:
for each word in the input text:
    for each co-occurring word within distance d:
        emit (word, co-occurring word) with count 1
Reducer:
for each (word, co-occurring word) received:
    sum up the counts for the pair
```

emit (word, co-occurring word) with aggregated count

Runtime Analysis:

CoOccurrenceStripe:

Mapper Runtime: O(n), where n is the number of words in the input text.

Reducer Runtime: O(n), where n is the number of distinct words emitted by mappers.

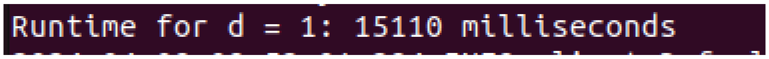
CoOccurrenceMatrix:

Mapper Runtime: O(n), where n is the number of words in the input text.

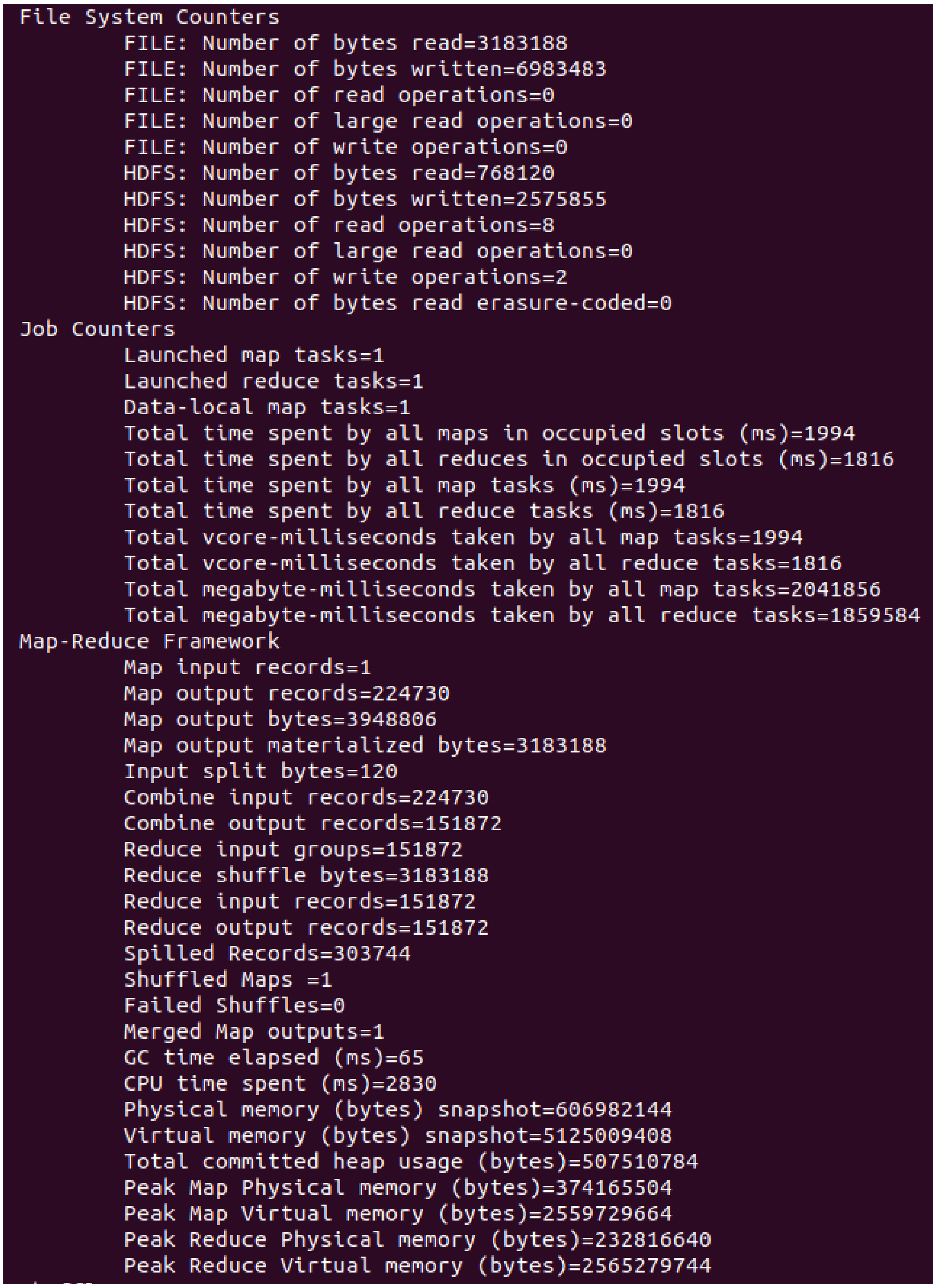
Reducer Runtime: O(n^2), where n is the number of distinct word pairs emitted by mappers.

Relevant Screenshots:

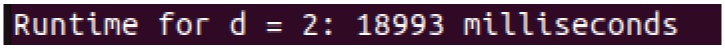
CoOccurrenceMatrix:



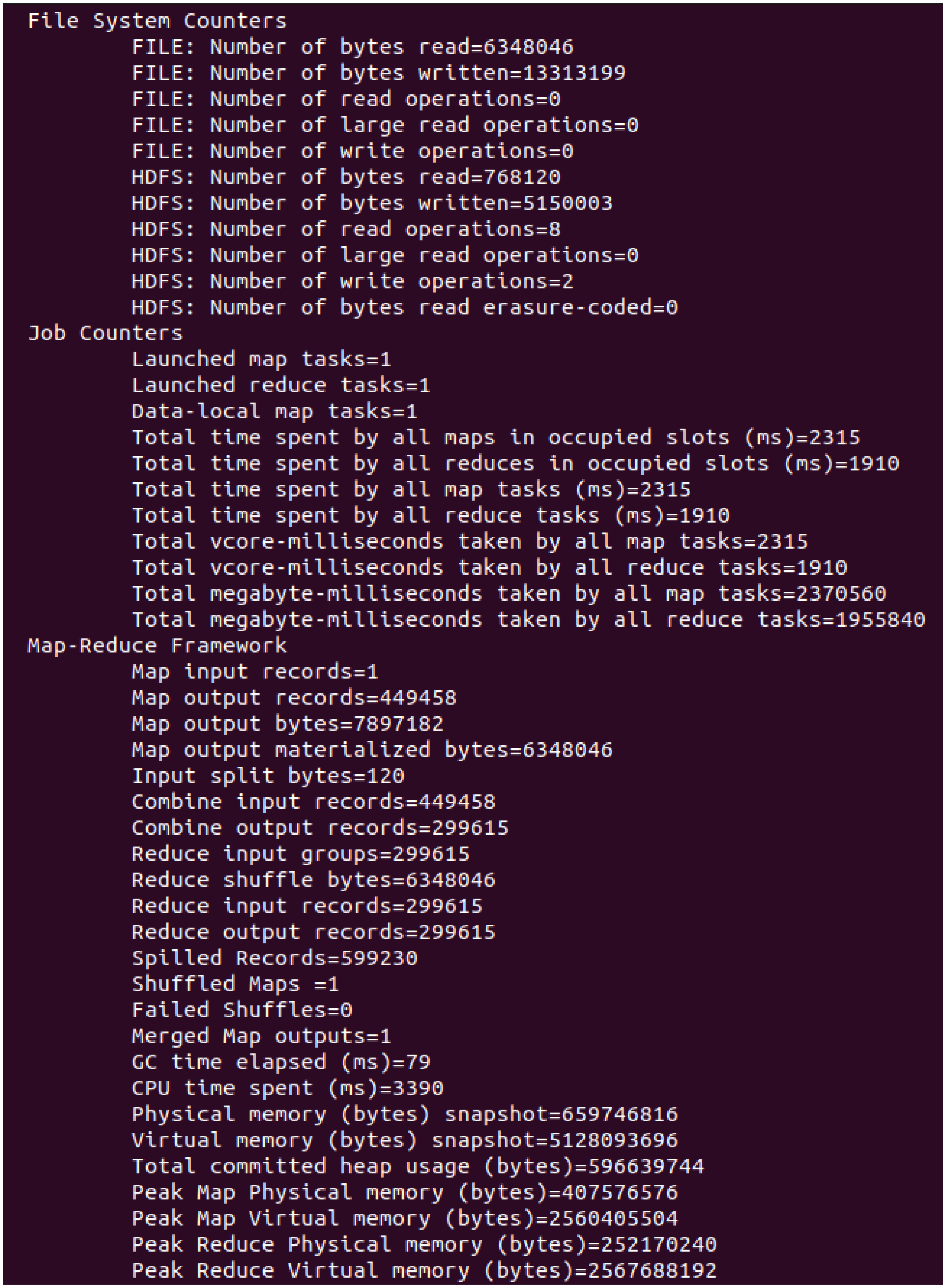
```
Runtime for d = 1: 15110 milliseconds
```



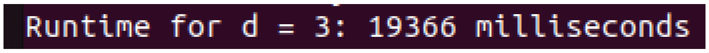
```
File System Counters
        FILE: Number of bytes read=3183188
        FILE: Number of bytes written=6983483
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=2575855
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=1994
        Total time spent by all reduces in occupied slots (ms)=1816
        Total time spent by all map tasks (ms)=1994
        Total time spent by all reduce tasks (ms)=1816
        Total vcore-milliseconds taken by all map tasks=1994
        Total vcore-milliseconds taken by all reduce tasks=1816
        Total megabyte-milliseconds taken by all map tasks=2041856
        Total megabyte-milliseconds taken by all reduce tasks=1859584
Map-Reduce Framework
        Map input records=1
        Map output records=224730
        Map output bytes=3948806
        Map output materialized bytes=3183188
        Input split bytes=120
        Combine input records=224730
        Combine output records=151872
        Reduce input groups=151872
        Reduce shuffle bytes=3183188
        Reduce input records=151872
        Reduce output records=151872
        Spilled Records=303744
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=65
        CPU time spent (ms)=2830
        Physical memory (bytes) snapshot=606982144
        Virtual memory (bytes) snapshot=5125009408
        Total committed heap usage (bytes)=507510784
        Peak Map Physical memory (bytes)=374165504
        Peak Map Virtual memory (bytes)=2559729664
        Peak Reduce Physical memory (bytes)=232816640
        Peak Reduce Virtual memory (bytes)=2565279744
```



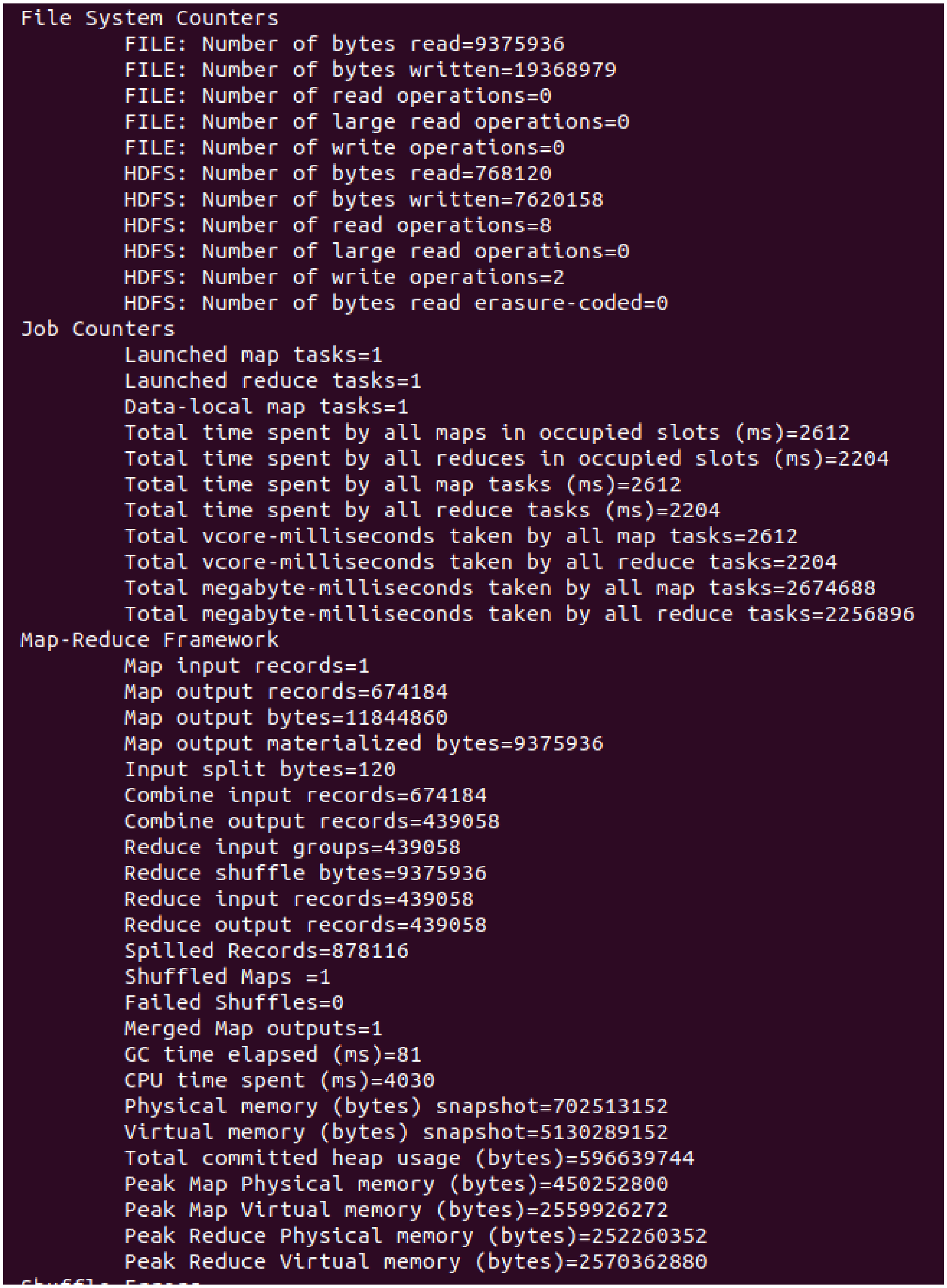
```
Runtime for d = 2: 18993 milliseconds
```



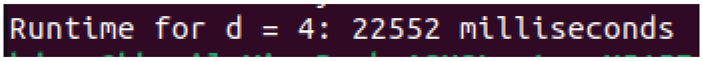
```
  File System Counters
          FILE: Number of bytes read=6348046
          FILE: Number of bytes written=13313199
          FILE: Number of read operations=0
          FILE: Number of large read operations=0
          FILE: Number of write operations=0
          HDFS: Number of bytes read=768120
          HDFS: Number of bytes written=5150003
          HDFS: Number of read operations=8
          HDFS: Number of large read operations=0
          HDFS: Number of write operations=2
          HDFS: Number of bytes read erasure-coded=0
  Job Counters
          Launched map tasks=1
          Launched reduce tasks=1
          Data-local map tasks=1
          Total time spent by all maps in occupied slots (ms)=2315
          Total time spent by all reduces in occupied slots (ms)=1910
          Total time spent by all map tasks (ms)=2315
          Total time spent by all reduce tasks (ms)=1910
          Total vcore-milliseconds taken by all map tasks=2315
          Total vcore-milliseconds taken by all reduce tasks=1910
          Total megabyte-milliseconds taken by all map tasks=2370560
          Total megabyte-milliseconds taken by all reduce tasks=1955840
  Map-Reduce Framework
          Map input records=1
          Map output records=449458
          Map output bytes=7897182
          Map output materialized bytes=6348046
          Input split bytes=120
          Combine input records=449458
          Combine output records=299615
          Reduce input groups=299615
          Reduce shuffle bytes=6348046
          Reduce input records=299615
          Reduce output records=299615
          Spilled Records=599230
          Shuffled Maps =1
          Failed Shuffles=0
          Merged Map outputs=1
          GC time elapsed (ms)=79
          CPU time spent (ms)=3390
          Physical memory (bytes) snapshot=659746816
          Virtual memory (bytes) snapshot=5128093696
          Total committed heap usage (bytes)=596639744
          Peak Map Physical memory (bytes)=407576576
          Peak Map Virtual memory (bytes)=2560405504
          Peak Reduce Physical memory (bytes)=252170240
          Peak Reduce Virtual memory (bytes)=2567688192
```



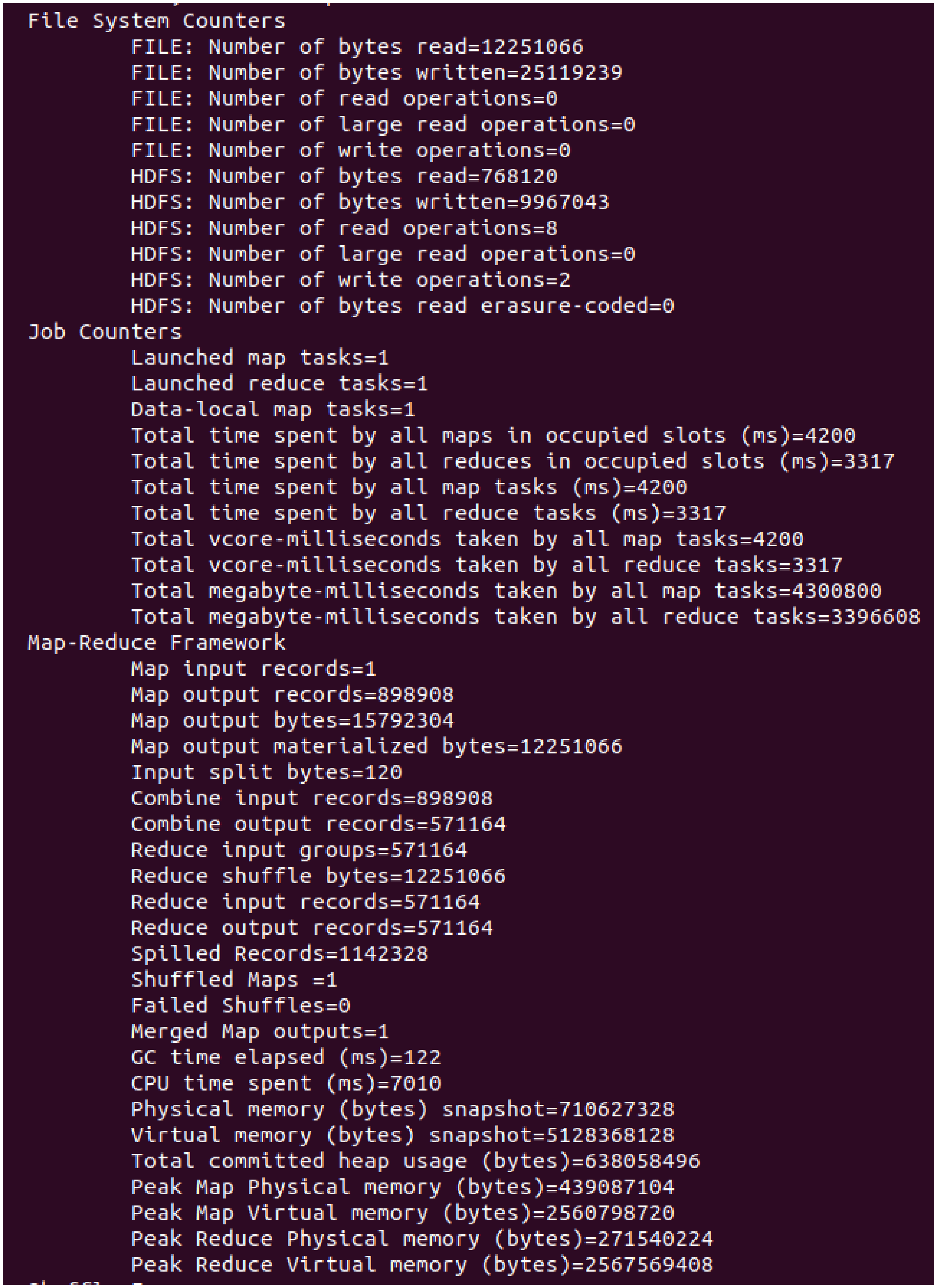
```
Runtime for d = 3: 19366 milliseconds
```



```
 File System Counters
         FILE: Number of bytes read=9375936
         FILE: Number of bytes written=19368979
         FILE: Number of read operations=0
         FILE: Number of large read operations=0
         FILE: Number of write operations=0
         HDFS: Number of bytes read=768120
         HDFS: Number of bytes written=7620158
         HDFS: Number of read operations=8
         HDFS: Number of large read operations=0
         HDFS: Number of write operations=2
         HDFS: Number of bytes read erasure-coded=0
 Job Counters
         Launched map tasks=1
         Launched reduce tasks=1
         Data-local map tasks=1
         Total time spent by all maps in occupied slots (ms)=2612
         Total time spent by all reduces in occupied slots (ms)=2204
         Total time spent by all map tasks (ms)=2612
         Total time spent by all reduce tasks (ms)=2204
         Total vcore-milliseconds taken by all map tasks=2612
         Total vcore-milliseconds taken by all reduce tasks=2204
         Total megabyte-milliseconds taken by all map tasks=2674688
         Total megabyte-milliseconds taken by all reduce tasks=2256896
 Map-Reduce Framework
         Map input records=1
         Map output records=674184
         Map output bytes=11844860
         Map output materialized bytes=9375936
         Input split bytes=120
         Combine input records=674184
         Combine output records=439058
         Reduce input groups=439058
         Reduce shuffle bytes=9375936
         Reduce input records=439058
         Reduce output records=439058
         Spilled Records=878116
         Shuffled Maps =1
         Failed Shuffles=0
         Merged Map outputs=1
         GC time elapsed (ms)=81
         CPU time spent (ms)=4030
         Physical memory (bytes) snapshot=702513152
         Virtual memory (bytes) snapshot=5130289152
         Total committed heap usage (bytes)=596639744
         Peak Map Physical memory (bytes)=450252800
         Peak Map Virtual memory (bytes)=2559926272
         Peak Reduce Physical memory (bytes)=252260352
         Peak Reduce Virtual memory (bytes)=2570362880
```

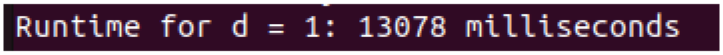


```
Runtime for d = 4: 22552 milliseconds
```

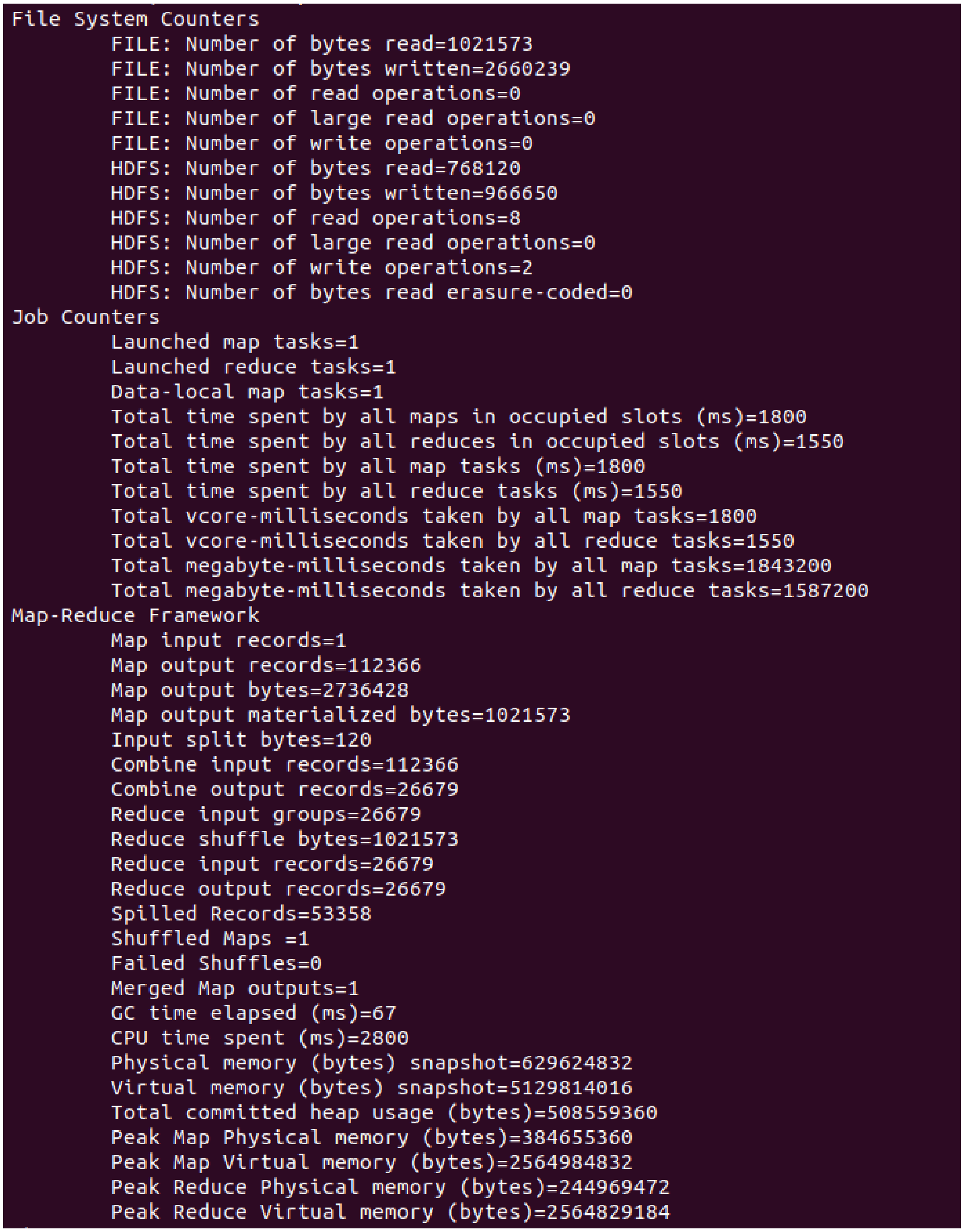


```
  File System Counters
          FILE: Number of bytes read=12251066
          FILE: Number of bytes written=25119239
          FILE: Number of read operations=0
          FILE: Number of large read operations=0
          FILE: Number of write operations=0
          HDFS: Number of bytes read=768120
          HDFS: Number of bytes written=9967043
          HDFS: Number of read operations=8
          HDFS: Number of large read operations=0
          HDFS: Number of write operations=2
          HDFS: Number of bytes read erasure-coded=0
  Job Counters
          Launched map tasks=1
          Launched reduce tasks=1
          Data-local map tasks=1
          Total time spent by all maps in occupied slots (ms)=4200
          Total time spent by all reduces in occupied slots (ms)=3317
          Total time spent by all map tasks (ms)=4200
          Total time spent by all reduce tasks (ms)=3317
          Total vcore-milliseconds taken by all map tasks=4200
          Total vcore-milliseconds taken by all reduce tasks=3317
          Total megabyte-milliseconds taken by all map tasks=4300800
          Total megabyte-milliseconds taken by all reduce tasks=3396608
  Map-Reduce Framework
          Map input records=1
          Map output records=898908
          Map output bytes=15792304
          Map output materialized bytes=12251066
          Input split bytes=120
          Combine input records=898908
          Combine output records=571164
          Reduce input groups=571164
          Reduce shuffle bytes=12251066
          Reduce input records=571164
          Reduce output records=571164
          Spilled Records=1142328
          Shuffled Maps =1
          Failed Shuffles=0
          Merged Map outputs=1
          GC time elapsed (ms)=122
          CPU time spent (ms)=7010
          Physical memory (bytes) snapshot=710627328
          Virtual memory (bytes) snapshot=5128368128
          Total committed heap usage (bytes)=638058496
          Peak Map Physical memory (bytes)=439087104
          Peak Map Virtual memory (bytes)=2560798720
          Peak Reduce Physical memory (bytes)=271540224
          Peak Reduce Virtual memory (bytes)=2567569408
```

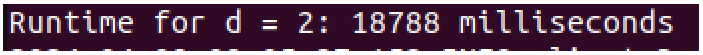
CoOccurrenceStripe:



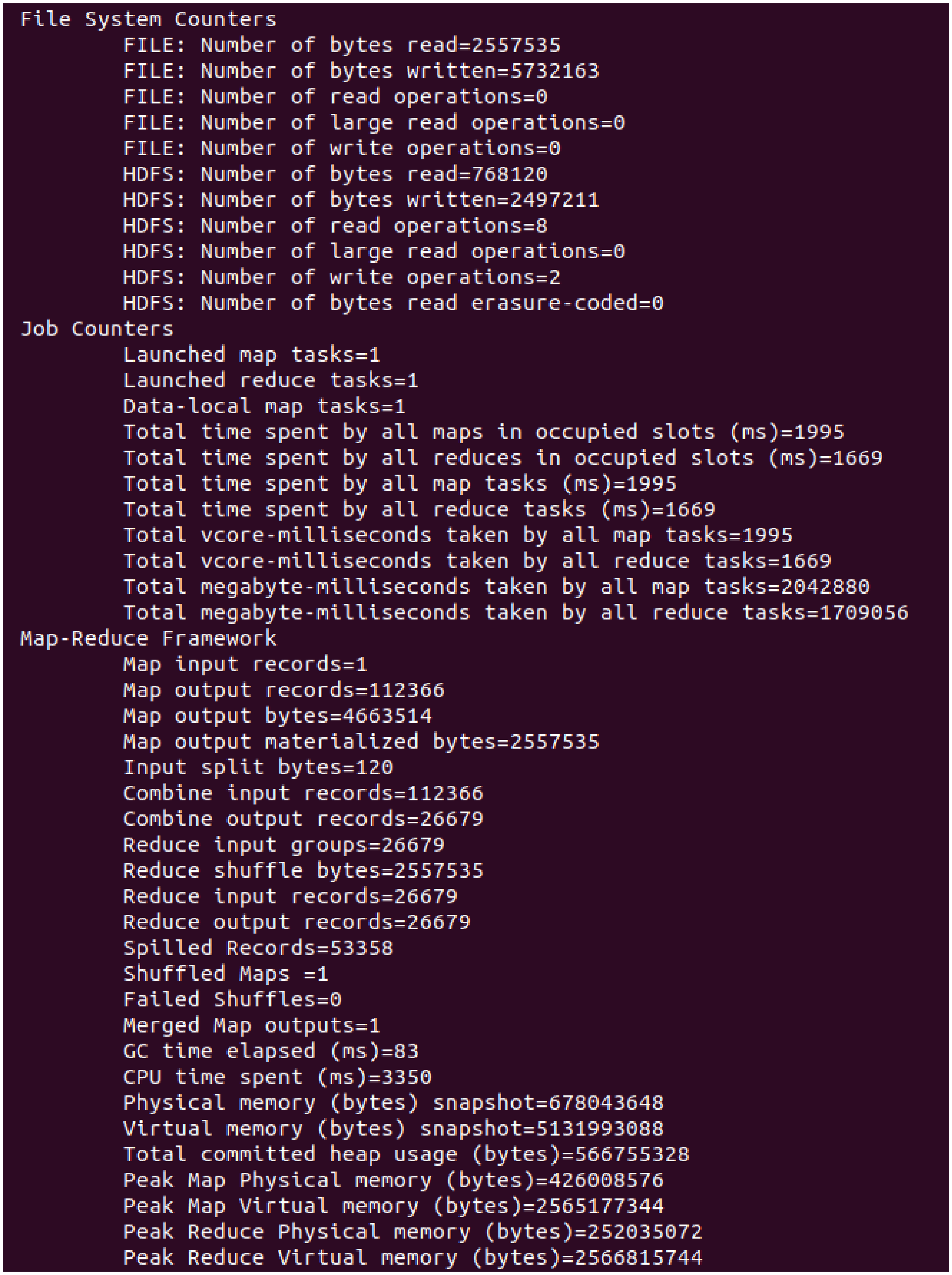
```
Runtime for d = 1: 13078 milliseconds
```



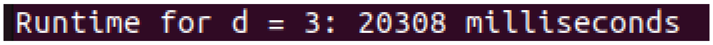
```
File System Counters
        FILE: Number of bytes read=1021573
        FILE: Number of bytes written=2660239
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=966650
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=1800
        Total time spent by all reduces in occupied slots (ms)=1550
        Total time spent by all map tasks (ms)=1800
        Total time spent by all reduce tasks (ms)=1550
        Total vcore-milliseconds taken by all map tasks=1800
        Total vcore-milliseconds taken by all reduce tasks=1550
        Total megabyte-milliseconds taken by all map tasks=1843200
        Total megabyte-milliseconds taken by all reduce tasks=1587200
Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=2736428
        Map output materialized bytes=1021573
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=1021573
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=67
        CPU time spent (ms)=2800
        Physical memory (bytes) snapshot=629624832
        Virtual memory (bytes) snapshot=5129814016
        Total committed heap usage (bytes)=508559360
        Peak Map Physical memory (bytes)=384655360
        Peak Map Virtual memory (bytes)=2564984832
        Peak Reduce Physical memory (bytes)=244969472
        Peak Reduce Virtual memory (bytes)=2564829184
```



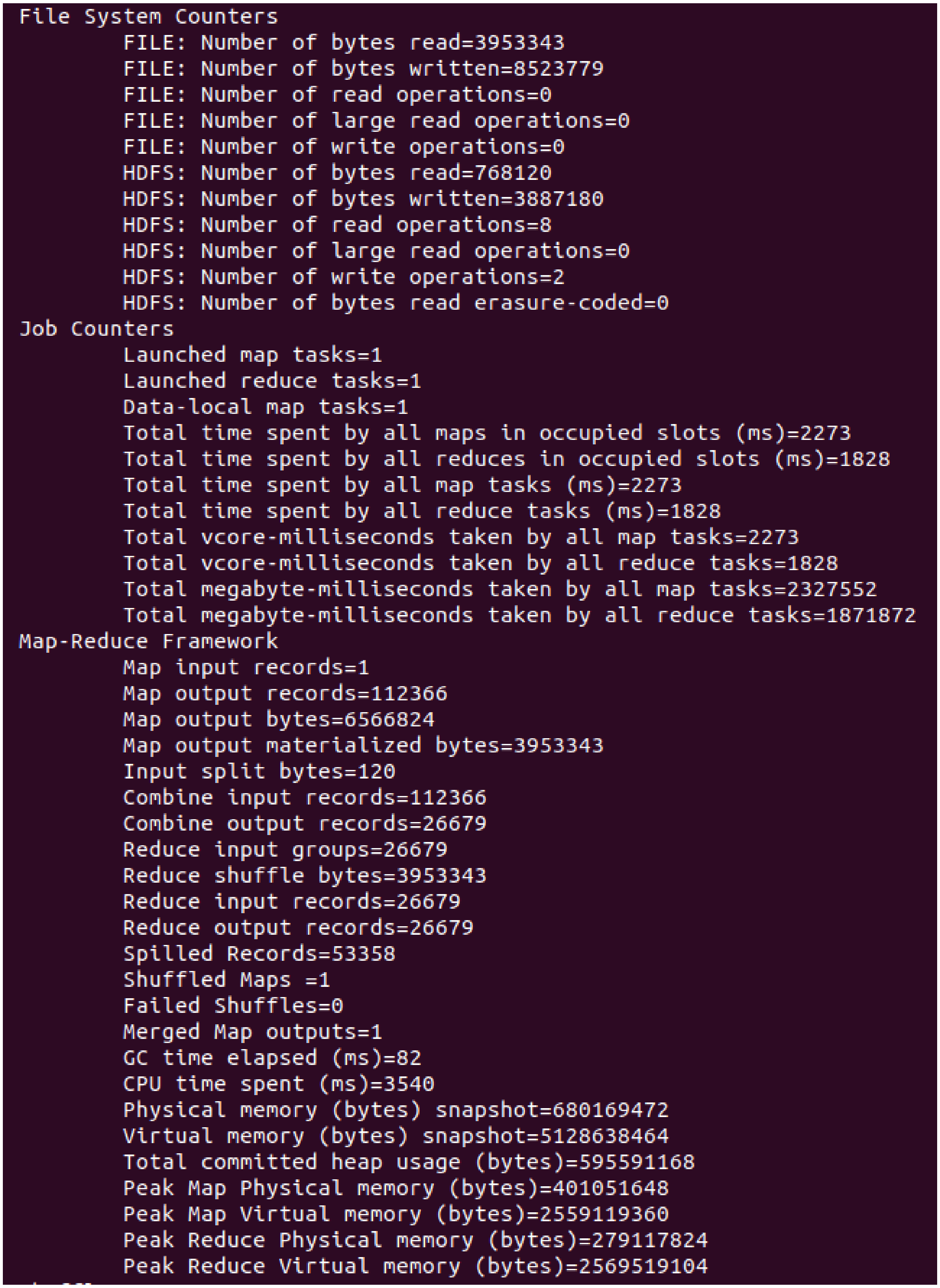
```
Runtime for d = 2: 18788 milliseconds
```



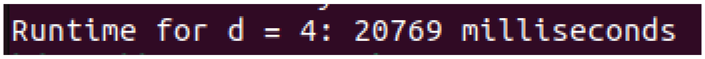
```
 File System Counters
         FILE: Number of bytes read=2557535
         FILE: Number of bytes written=5732163
         FILE: Number of read operations=0
         FILE: Number of large read operations=0
         FILE: Number of write operations=0
         HDFS: Number of bytes read=768120
         HDFS: Number of bytes written=2497211
         HDFS: Number of read operations=8
         HDFS: Number of large read operations=0
         HDFS: Number of write operations=2
         HDFS: Number of bytes read erasure-coded=0
 Job Counters
         Launched map tasks=1
         Launched reduce tasks=1
         Data-local map tasks=1
         Total time spent by all maps in occupied slots (ms)=1995
         Total time spent by all reduces in occupied slots (ms)=1669
         Total time spent by all map tasks (ms)=1995
         Total time spent by all reduce tasks (ms)=1669
         Total vcore-milliseconds taken by all map tasks=1995
         Total vcore-milliseconds taken by all reduce tasks=1669
         Total megabyte-milliseconds taken by all map tasks=2042880
         Total megabyte-milliseconds taken by all reduce tasks=1709056
 Map-Reduce Framework
         Map input records=1
         Map output records=112366
         Map output bytes=4663514
         Map output materialized bytes=2557535
         Input split bytes=120
         Combine input records=112366
         Combine output records=26679
         Reduce input groups=26679
         Reduce shuffle bytes=2557535
         Reduce input records=26679
         Reduce output records=26679
         Spilled Records=53358
         Shuffled Maps =1
         Failed Shuffles=0
         Merged Map outputs=1
         GC time elapsed (ms)=83
         CPU time spent (ms)=3350
         Physical memory (bytes) snapshot=678043648
         Virtual memory (bytes) snapshot=5131993088
         Total committed heap usage (bytes)=566755328
         Peak Map Physical memory (bytes)=426008576
         Peak Map Virtual memory (bytes)=2565177344
         Peak Reduce Physical memory (bytes)=252035072
         Peak Reduce Virtual memory (bytes)=2566815744
```



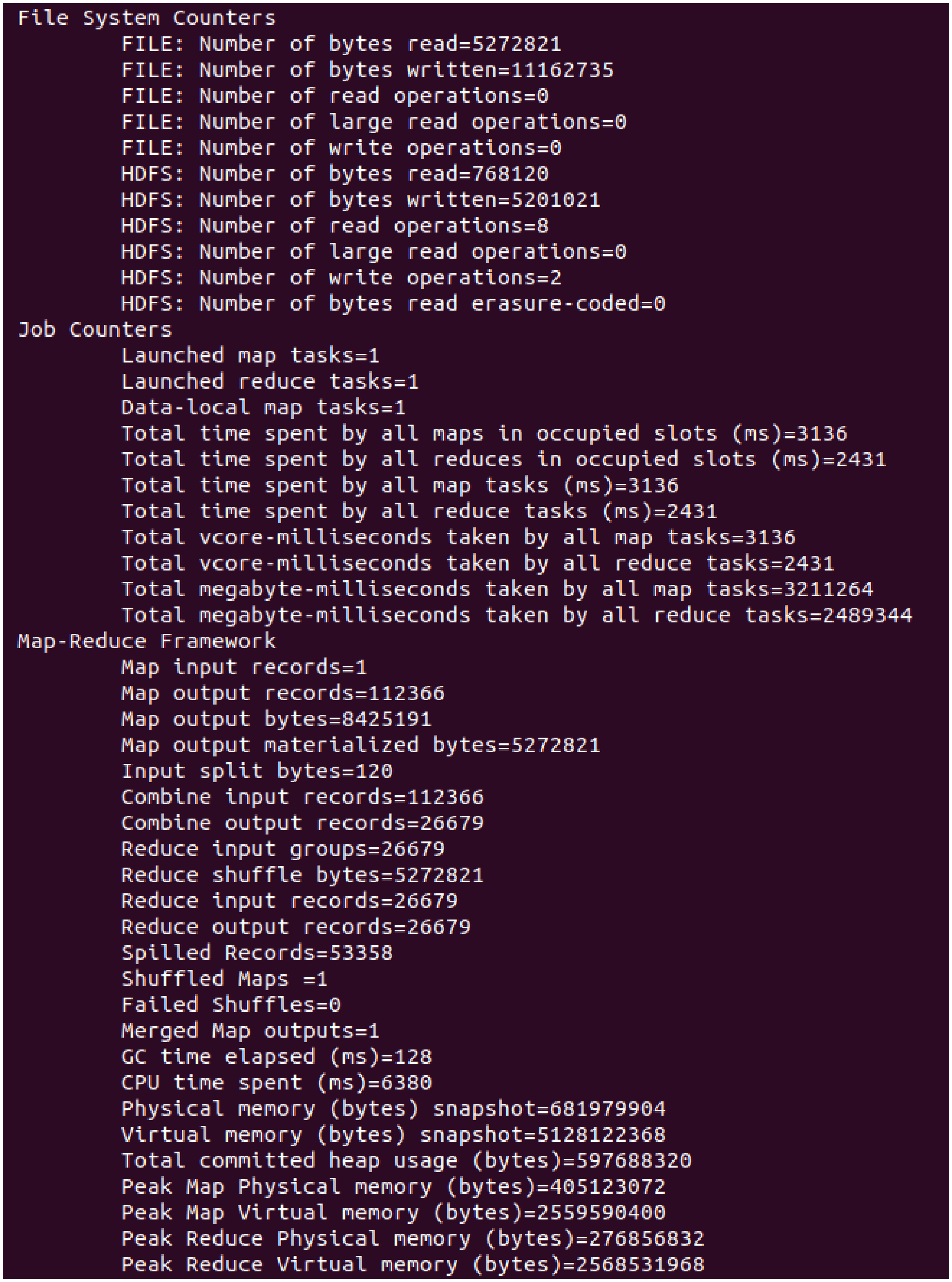
```
Runtime for d = 3: 20308 milliseconds
```



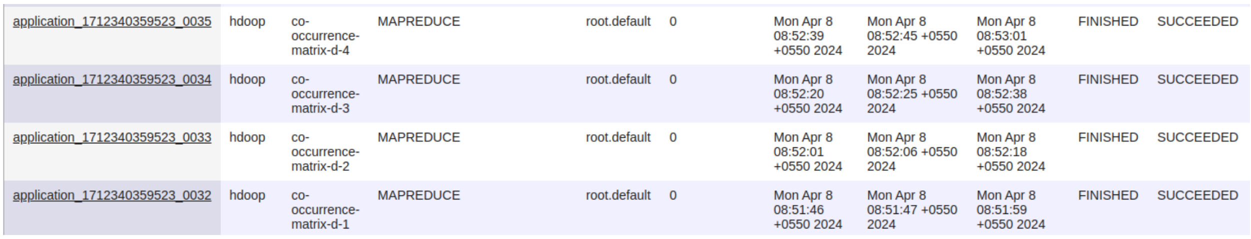
```
 File System Counters
        FILE: Number of bytes read=3953343
        FILE: Number of bytes written=8523779
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=3887180
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
 Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=2273
        Total time spent by all reduces in occupied slots (ms)=1828
        Total time spent by all map tasks (ms)=2273
        Total time spent by all reduce tasks (ms)=1828
        Total vcore-milliseconds taken by all map tasks=2273
        Total vcore-milliseconds taken by all reduce tasks=1828
        Total megabyte-milliseconds taken by all map tasks=2327552
        Total megabyte-milliseconds taken by all reduce tasks=1871872
 Map-Reduce Framework
        Map input records=1
        Map output records=112366
        Map output bytes=6566824
        Map output materialized bytes=3953343
        Input split bytes=120
        Combine input records=112366
        Combine output records=26679
        Reduce input groups=26679
        Reduce shuffle bytes=3953343
        Reduce input records=26679
        Reduce output records=26679
        Spilled Records=53358
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=82
        CPU time spent (ms)=3540
        Physical memory (bytes) snapshot=680169472
        Virtual memory (bytes) snapshot=5128638464
        Total committed heap usage (bytes)=595591168
        Peak Map Physical memory (bytes)=401051648
        Peak Map Virtual memory (bytes)=2559119360
        Peak Reduce Physical memory (bytes)=279117824
        Peak Reduce Virtual memory (bytes)=2569519104
```



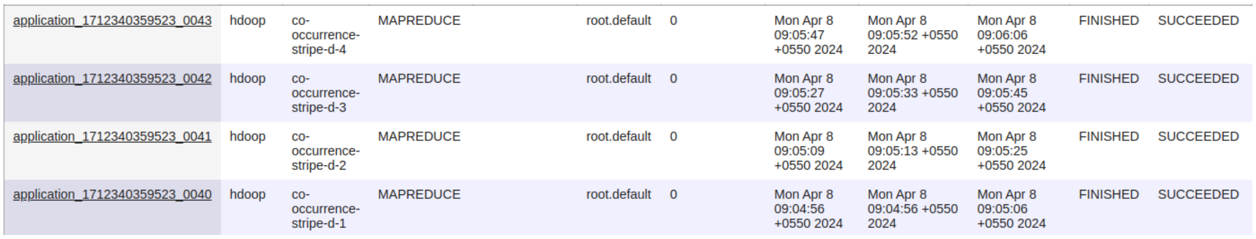
```
Runtime for d = 4: 20769 milliseconds
```



```
 File System Counters
         FILE: Number of bytes read=5272821
         FILE: Number of bytes written=11162735
         FILE: Number of read operations=0
         FILE: Number of large read operations=0
         FILE: Number of write operations=0
         HDFS: Number of bytes read=768120
         HDFS: Number of bytes written=5201021
         HDFS: Number of read operations=8
         HDFS: Number of large read operations=0
         HDFS: Number of write operations=2
         HDFS: Number of bytes read erasure-coded=0
 Job Counters
         Launched map tasks=1
         Launched reduce tasks=1
         Data-local map tasks=1
         Total time spent by all maps in occupied slots (ms)=3136
         Total time spent by all reduces in occupied slots (ms)=2431
         Total time spent by all map tasks (ms)=3136
         Total time spent by all reduce tasks (ms)=2431
         Total vcore-milliseconds taken by all map tasks=3136
         Total vcore-milliseconds taken by all reduce tasks=2431
         Total megabyte-milliseconds taken by all map tasks=3211264
         Total megabyte-milliseconds taken by all reduce tasks=2489344
 Map-Reduce Framework
         Map input records=1
         Map output records=112366
         Map output bytes=8425191
         Map output materialized bytes=5272821
         Input split bytes=120
         Combine input records=112366
         Combine output records=26679
         Reduce input groups=26679
         Reduce shuffle bytes=5272821
         Reduce input records=26679
         Reduce output records=26679
         Spilled Records=53358
         Shuffled Maps =1
         Failed Shuffles=0
         Merged Map outputs=1
         GC time elapsed (ms)=128
         CPU time spent (ms)=6380
         Physical memory (bytes) snapshot=681979904
         Virtual memory (bytes) snapshot=5128122368
         Total committed heap usage (bytes)=597688320
         Peak Map Physical memory (bytes)=405123072
         Peak Map Virtual memory (bytes)=2559590400
         Peak Reduce Physical memory (bytes)=276856832
         Peak Reduce Virtual memory (bytes)=2568531968
```



application_1712340359523_0035	hdoop	co-occurrence-matrix-d-4	MAPREDUCE		root.default	0	Mon Apr 8 08:52:39 +0550 2024	Mon Apr 8 08:52:45 +0550 2024	Mon Apr 8 08:53:01 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0034	hdoop	co-occurrence-matrix-d-3	MAPREDUCE		root.default	0	Mon Apr 8 08:52:20 +0550 2024	Mon Apr 8 08:52:25 +0550 2024	Mon Apr 8 08:52:38 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0033	hdoop	co-occurrence-matrix-d-2	MAPREDUCE		root.default	0	Mon Apr 8 08:52:01 +0550 2024	Mon Apr 8 08:52:06 +0550 2024	Mon Apr 8 08:52:18 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0032	hdoop	co-occurrence-matrix-d-1	MAPREDUCE		root.default	0	Mon Apr 8 08:51:46 +0550 2024	Mon Apr 8 08:51:47 +0550 2024	Mon Apr 8 08:51:59 +0550 2024	FINISHED	SUCCEEDED



application_1712340359523_0043	hdoop	co-occurrence-stripe-d-4	MAPREDUCE		root.default	0	Mon Apr 8 09:05:47 +0550 2024	Mon Apr 8 09:05:52 +0550 2024	Mon Apr 8 09:06:06 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0042	hdoop	co-occurrence-stripe-d-3	MAPREDUCE		root.default	0	Mon Apr 8 09:05:27 +0550 2024	Mon Apr 8 09:05:33 +0550 2024	Mon Apr 8 09:05:45 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0041	hdoop	co-occurrence-stripe-d-2	MAPREDUCE		root.default	0	Mon Apr 8 09:05:09 +0550 2024	Mon Apr 8 09:05:13 +0550 2024	Mon Apr 8 09:05:25 +0550 2024	FINISHED	SUCCEEDED
application_1712340359523_0040	hdoop	co-occurrence-stripe-d-1	MAPREDUCE		root.default	0	Mon Apr 8 09:04:56 +0550 2024	Mon Apr 8 09:04:56 +0550 2024	Mon Apr 8 09:05:06 +0550 2024	FINISHED	SUCCEEDED

Indexing Documents via Hadoop

PART-A

Program Logic:

The program aims to calculate the document frequency of terms in a corpus using Hadoop's MapReduce framework. It preprocesses the input text by removing stop words and performing stemming before emitting each term with its document ID and frequency of occurrence.

Pseudocode Explanation:

Mapper Phase:

Load stop words from the "stopwords.txt" file into a set.

For each input record:

Tokenize the text into words.

Lowercase each word and remove non-alphanumeric characters.

Skip stop words and perform stemming using the Porter Stemmer.

Emit each term with its document ID (key) and value 1.

Reducer Phase:

Aggregate the frequency counts for each term across different documents.

Emit each term along with its total frequency in the corpus.

Main Function:

Configure the MapReduce job with the necessary classes, jar file, and input/output formats.

Set the mapper, combiner (optional), and reducer classes.

Specify the output key and value classes.

Set the input and output paths.

Run the job and exit with status 0 if successful, else exit with status 1.

<u>Runtime Analysis:</u>

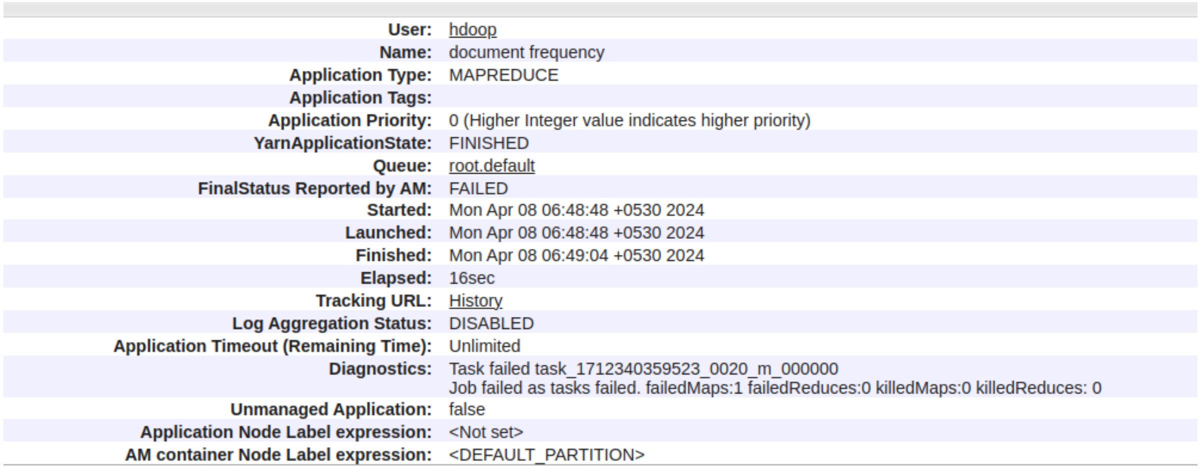
The program's runtime performance depends on factors such as the size of the input corpus, the number of terms, and the efficiency of the stemming algorithm.

Preprocessing steps like stop word removal and stemming may introduce overhead, but they contribute to more accurate document frequency calculations by eliminating noise from the text.

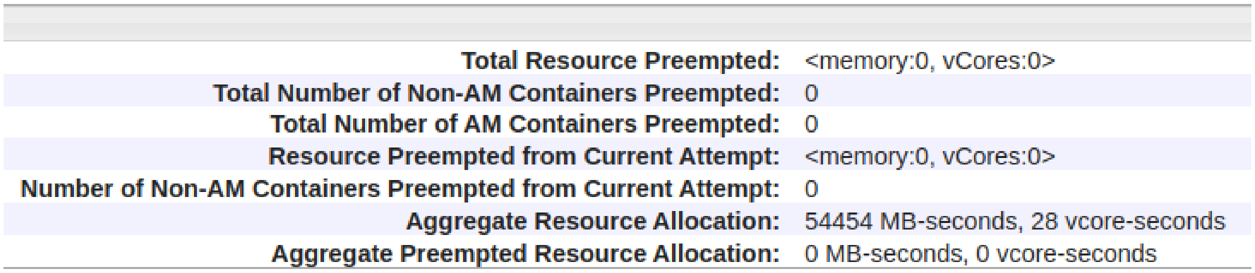
The MapReduce framework enables parallel processing of input data across multiple nodes in a cluster, which helps handle large-scale corpora efficiently.

<u>Relevant Screenshots:</u>

Got a stemmer error, so couldn't get the final output...



User:	hdoop
Name:	document frequency
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	root.default
FinalStatus Reported by AM:	FAILED
Started:	Mon Apr 08 06:48:48 +0530 2024
Launched:	Mon Apr 08 06:48:48 +0530 2024
Finished:	Mon Apr 08 06:49:04 +0530 2024
Elapsed:	16sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	Task failed task_1712340359523_0020_m_000000 Job failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0 killedReduces: 0
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>



Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	54454 MB-seconds, 28 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

PART-B

Program Logic:

The program calculates the TF-IDF (Term Frequency-Inverse Document Frequency) scores for terms in a corpus using the "stripes" approach. It first computes the term frequency (TF) for each term within each document and then calculates the IDF component using the total number of documents in the corpus. Finally, it combines these values to compute the TF-IDF score for each term.

Pseudocode Explanation:

Mapper Phase:

Parse each input record, which consists of a document ID followed by term-count pairs.

Emit each document ID along with its corresponding term-count pair.

Reducer Phase:

For each document ID, aggregate the term-count pairs and calculate the total number of terms in the document.

Compute the TF-IDF score for each term in the document using the formula:

TF-IDF = term frequency / total terms

Emit each document ID along with its corresponding term and TF-IDF score.

Main Function:

Configure the MapReduce job with the necessary classes, jar file, and input/output formats.

Set the mapper and reducer classes.

Specify the output key and value classes.

Set the input and output paths.

Run the job and exit with status 0 if successful, else exit with status 1.

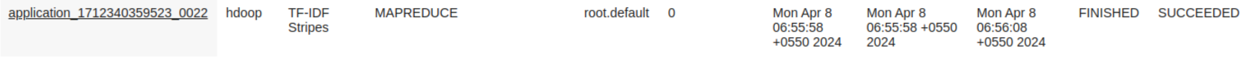
Runtime Analysis:

The program's runtime performance depends on factors such as the size of the corpus, the number of unique terms, and the efficiency of the TF-IDF computation.

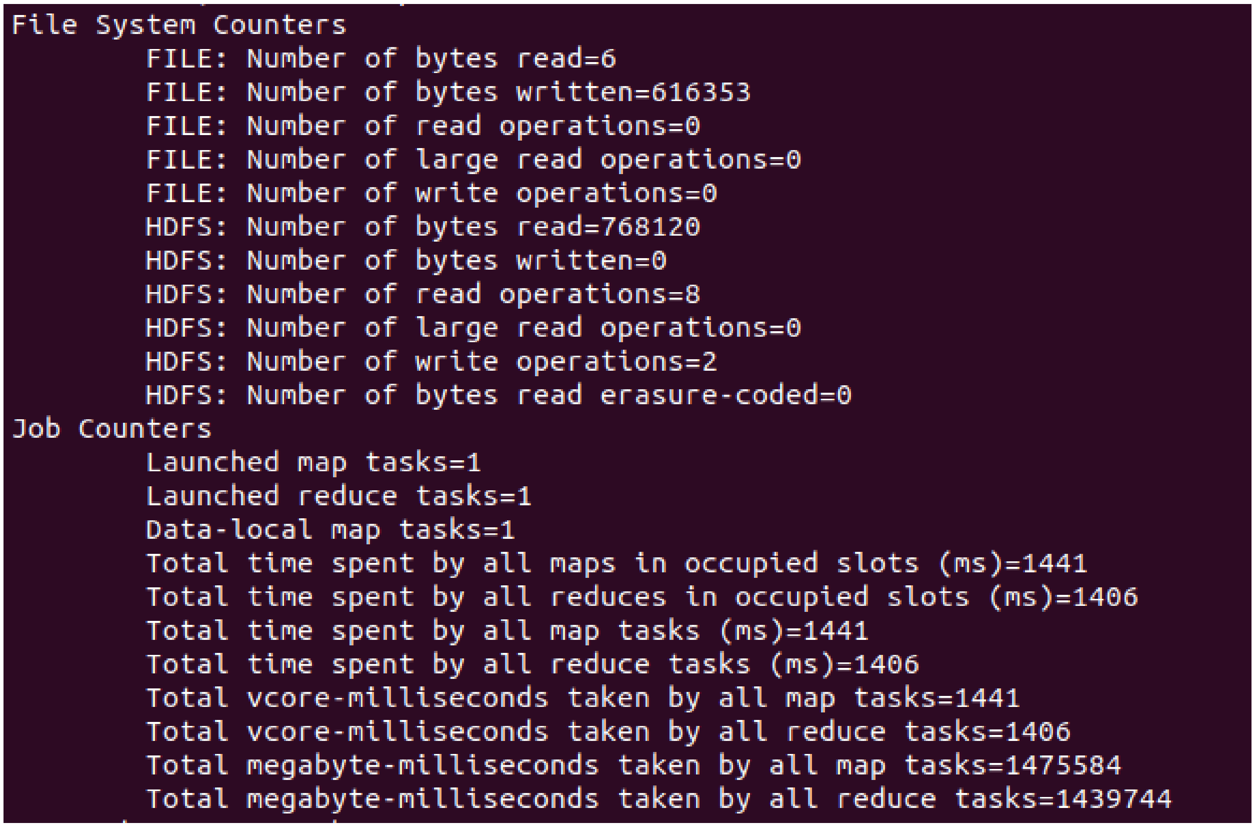
The MapReduce framework enables parallel processing of input data across multiple nodes in a cluster, which helps handle large-scale corpora efficiently.

The complexity of the TF-IDF calculation is linear with respect to the number of unique terms in each document and the total number of documents in the corpus.

Relevant Screenshots:



application_1712340359523_0022	hdoop	TF-IDF Stripes	MAPREDUCE	root.default	0	Mon Apr 8 06:55:58 +0550 2024	Mon Apr 8 06:55:58 +0550 2024	Mon Apr 8 06:56:08 +0550 2024	FINISHED	SUCCEEDED



```
File System Counters
        FILE: Number of bytes read=6
        FILE: Number of bytes written=616353
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=768120
        HDFS: Number of bytes written=0
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=1441
        Total time spent by all reduces in occupied slots (ms)=1406
        Total time spent by all map tasks (ms)=1441
        Total time spent by all reduce tasks (ms)=1406
        Total vcore-milliseconds taken by all map tasks=1441
        Total vcore-milliseconds taken by all reduce tasks=1406
        Total megabyte-milliseconds taken by all map tasks=1475584
        Total megabyte-milliseconds taken by all reduce tasks=1439744
```