

Project-1

CSE-6332

Swarag Reddy Pingili – 1002158023

REPORT

INTRODUCTION

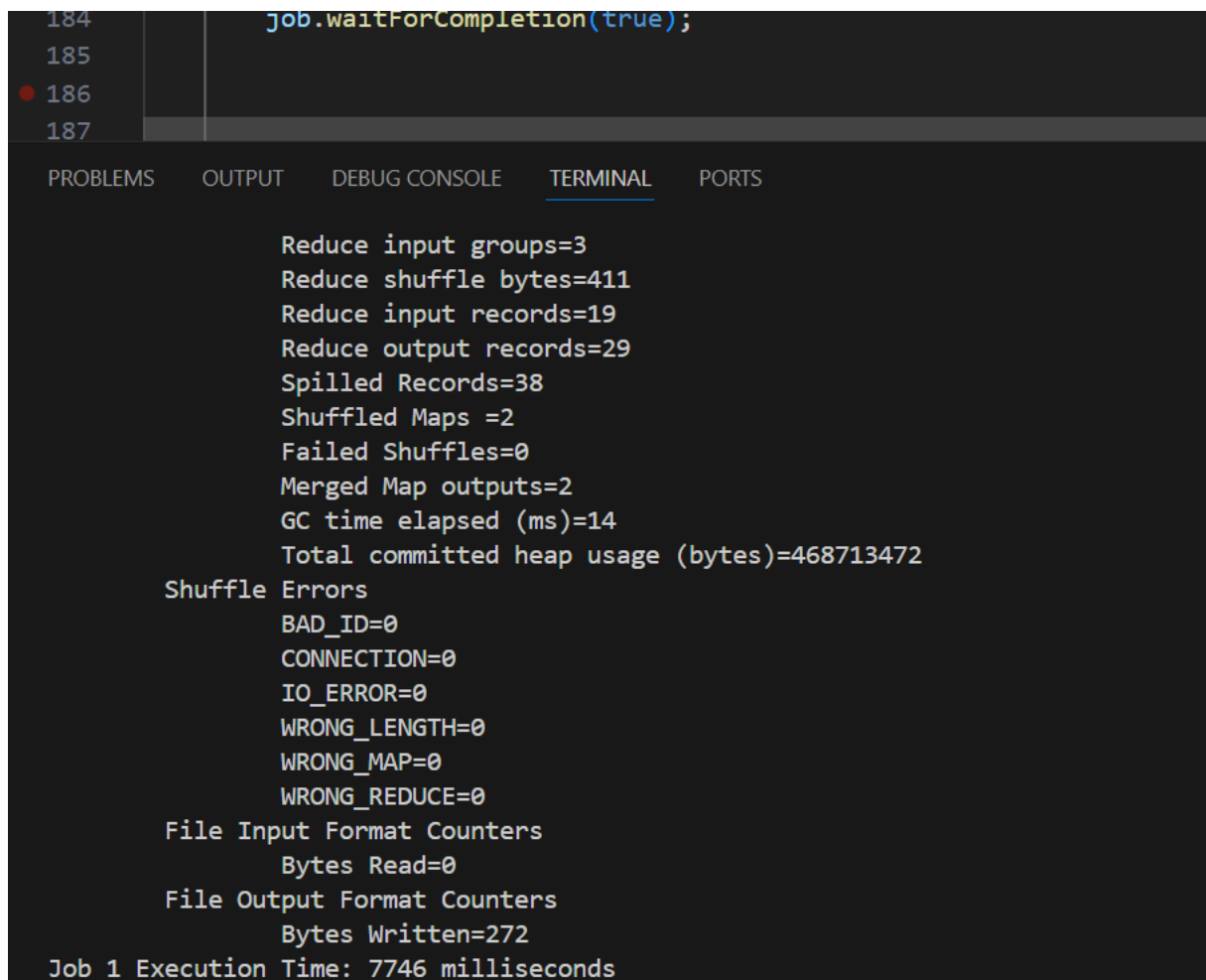
In this project, I have implemented a MapReduce program to multiply two given sparse matrices. The program consists of two jobs. The first job employs two mappers and one reducer and the second job utilizes one mapper and one reducer.

The output below corresponds to small sparse matrices provided in the file.

Job – 1

The job one produces the intermediate output which is passed as an argument to job 2.

Job 1 Execution Time – 7746 milliseconds.



```
184      job.waitForCompletion(true);
185
186
187
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

    Reduce input groups=3
    Reduce shuffle bytes=411
    Reduce input records=19
    Reduce output records=29
    Spilled Records=38
    Shuffled Maps =2
    Failed Shuffles=0
    Merged Map outputs=2
    GC time elapsed (ms)=14
    Total committed heap usage (bytes)=468713472
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=0
File Output Format Counters
    Bytes Written=272
Job 1 Execution Time: 7746 milliseconds
```

Job – 2

Job 2 takes the output of job1 as input and produces the final result.

Job 2 Execution Time – 2924 milliseconds.

```
183 //job.setNumReduceTasks(5);
184 job.waitForCompletion(true);
185
186
187
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Reduce input groups=12
Reduce shuffle bytes=324
Reduce input records=29
Reduce output records=12
Spilled Records=58
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=16
Total committed heap usage (bytes)=524288000
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=276
File Output Format Counters
  Bytes Written=117
Job 2 Execution Time: 2924 milliseconds
```

Output of small matrices:

```
output > ≡ part-r-00000
1 0,0,7.0
2 0,1,-43.0
3 0,2,1.0
4 1,0,45.0
5 1,1,-9.0
6 1,2,1.0
7 2,0,59.0
8 2,1,21.0
9 2,2,-1.0
10 3,0,49.0
11 3,1,56.0
12 3,2,0.0
13
```

Large Sparse Matrices

Datasets:

Dataset1: <https://sparse.tamu.edu/HB/bcsstk08> (rows: 1074, columns: 1074)

Dataset2: <https://sparse.tamu.edu/HB/bcsstk11> (rows: 1473, columns: 1473)

Observations:

size of input splits: 128MB, Reduce Tasks: 10

Total Execution Time: 17120ms

```
src > main > java > J Multiply.java
 49  ~ public class Multiply {
164  ~     public static void main ( String[] args ) throws Exception {
165      ~         Configuration conf = new Configuration();
166
167      ~         conf.set("mapreduce.output.textoutputformat.separator", ",");
168      ~         conf.set("mapreduce.input.fileinputformat.split.maxsize", "134217728");
169
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

      Reduce input records=79546
      Reduce output records=50001
      Spilled Records=159092
      Shuffled Maps =100
      Failed Shuffles=0
      Merged Map outputs=100
      GC time elapsed (ms)=46
      Total committed heap usage (bytes)=5515509760
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=2360378
File Output Format Counters
      Bytes Written=1490770
Job 1 Execution Time: 9547 milliseconds
Job 2 Execution Time: 7573 milliseconds
Total Job Execution Time for 128MB and 10 reduce tasks: 17120 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 64MB, Reduce Tasks: 10
Total Execution Time: 18180ms

```
src > main > java > J Multiply.java
49  public class Multiply {
164      public static void main ( String[] args ) throws Exception {
166
167      conf.set("mapreduce.output.textoutputformat.separator", ",");
168      conf.set("mapreduce.input.fileinputformat.split.maxsize", "67108864");
169  }
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Reduce input records=79546
Reduce output records=50001
Spilled Records=159092
Shuffled Maps =100
Failed Shuffles=0
Merged Map outputs=100
GC time elapsed (ms)=79
Total committed heap usage (bytes)=5599395840
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=2360378
File Output Format Counters
Bytes Written=1490770
Job 1 Execution Time: 9603 milliseconds
Job 2 Execution Time: 8577 milliseconds
Total Job Execution Time for 64MB and 10 reduce tasks: 18180 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 32MB, Reduce Tasks: 10
Total Execution Time: 19379ms

```
src > main > java > J Multiply.java
49  public class Multiply {
164      public static void main ( String[] args ) throws Exception {
165          Configuration conf = new Configuration();
166
167          conf.set("mapreduce.output.textoutputformat.separator", ",");
168          conf.set("mapreduce.input.fileinputformat.split.maxsize", "33554432");
...
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS
Reduce input records=79546
Reduce output records=50001
Spilled Records=159092
Shuffled Maps =100
Failed Shuffles=0
Merged Map outputs=100
GC time elapsed (ms)=135
Total committed heap usage (bytes)=5620367360
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=2360378
File Output Format Counters
Bytes Written=1490770
Job 1 Execution Time: 10777 milliseconds
Job 2 Execution Time: 8602 milliseconds
Total Job Execution Time for 32MB and 10 reduce tasks: 19379 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 16MB, Reduce Tasks: 10
Total Execution Time: 20508ms

```
src > main > java > J Multiply.java
 49  public class Multiply {
164      public static void main ( String[] args ) throws Exception {
166
167          conf.set("mapreduce.output.textoutputformat.separator", ",");
168          conf.set("mapreduce.input.fileinputformat.split.maxsize", "16777216");
169
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

      Reduce input records=79546
      Reduce output records=50001
      Spilled Records=159092
      Shuffled Maps =100
      Failed Shuffles=0
      Merged Map outputs=100
      GC time elapsed (ms)=47
      Total committed heap usage (bytes)=5536481280
  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=2360378
  File Output Format Counters
      Bytes Written=1490770
Job 1 Execution Time: 10821 milliseconds
Job 2 Execution Time: 9687 milliseconds
Total Job Execution Time for 16MB and 10 reduce tasks: 20508 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 128MB, Reduce Tasks: 5
Total Execution Time: 11040ms

```
src > main > java > J Multiply.java
 49  public class Multiply {
164      public static void main ( String[] args ) throws Exception {
166
167          conf.set("mapreduce.output.textoutputformat.separator", ",");
168          conf.set("mapreduce.input.fileinputformat.split.maxsize", "134217728");
169      }
  }

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

      Reduce input records=79546
      Reduce output records=50001
      Spilled Records=159092
      Shuffled Maps =25
      Failed Shuffles=0
      Merged Map outputs=25
      GC time elapsed (ms)=25
      Total committed heap usage (bytes)=3061841920

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=2360310
File Output Format Counters
      Bytes Written=1490694
Job 1 Execution Time: 6650 milliseconds
Job 2 Execution Time: 4390 milliseconds
Total Job Execution Time for 128MB and 5 reduce tasks: 11040 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```


size of input splits: 64MB, Reduce Tasks: 5
Total Execution Time: 11900ms

```
src > main > java > J Multiply.java
  49  public class Multiply {
164      public static void main ( String[] args ) throws Exception {
165          Configuration conf = new Configuration();
166
167          conf.set("mapreduce.output.textoutputformat.separator", ",");
168          conf.set("mapreduce.input.fileinputformat.split.maxsize", "67108864");
169      }
  }

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

      Reduce input records=79546
      Reduce output records=50001
      Spilled Records=159092
      Shuffled Maps =25
      Failed Shuffles=0
      Merged Map outputs=25
      GC time elapsed (ms)=42
      Total committed heap usage (bytes)=3156213760
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=2360310
File Output Format Counters
      Bytes Written=1490694
Job 1 Execution Time: 6493 milliseconds
Job 2 Execution Time: 5407 milliseconds
Total Job Execution Time for 64MB and 5 reduce tasks: 11900 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 32MB, Reduce Tasks: 5
Total Execution Time: 16378ms

```
src > main > java > J Multiply.java
 49  public class Multiply {
164  public static void main ( String[] args ) throws Exception {
165      Configuration conf = new Configuration();
166
167      conf.set("mapreduce.output.textoutputformat.separator", ",");
168      conf.set("mapreduce.input.fileinputformat.split.maxsize", "33554432");
169  }
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Reduce input records=79546
Reduce output records=50001
Spilled Records=159092
Shuffled Maps =25
Failed Shuffles=0
Merged Map outputs=25
GC time elapsed (ms)=36
Total committed heap usage (bytes)=3533701120

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=2360310

File Output Format Counters
Bytes Written=1490694

Job 1 Execution Time: 10563 milliseconds
Job 2 Execution Time: 5815 milliseconds
Total Job Execution Time for 32MB and 5 reduce tasks: 16378 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

size of input splits: 16MB, Reduce Tasks: 5
Total Execution Time: 18585ms

```
49  ∨ public class Multiply {
164  ∨      public static void main ( String[] args ) throws Exception {
165          Configuration conf = new Configuration();
166
167          conf.set("mapreduce.output.textoutputformat.separator", ",");
168          conf.set("mapreduce.input.fileinputformat.split.maxsize", "16777216");
169      }
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

      Reduce input records=79546
      Reduce output records=50001
      Spilled Records=159092
      Shuffled Maps =25
      Failed Shuffles=0
      Merged Map outputs=25
      GC time elapsed (ms)=46
      Total committed heap usage (bytes)=3051356160
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=2360310
File Output Format Counters
      Bytes Written=1490694
Job 1 Execution Time: 10803 milliseconds
Job 2 Execution Time: 7782 milliseconds
Total Job Execution Time for 16MB and 5 reduce tasks: 18585 milliseconds
swarag@Anurag-Windows:/mnt/c/Users/Anurag Reddy/Desktop/MatMult$
```

Conclusion:

The very first observation is that as the size of input splits (split.maxsize) decreases, the total execution time increases. When the size of the split was 128MB and the number of reduce tasks was 10, it gave an execution time of 17120ms. As the size of the split decreased, the execution time increased. When the size was 16MB, the execution time was 20508ms.

The second observation is that as the number of reduce tasks was changed to 5 from 10, the execution time further decreased. When the size of the split was 128MB and the number of reduce tasks was 5, it gave the least execution time of 11040ms. This result was similar for all the split sizes; when the reduce tasks were 5, the execution time decreased.