Map Reduce

Group 1
Mayank Singh Chauhan 2016CS50394
Atishya Jain 2016CS50393
Mankaran Singh 2016CS50391
Avaljot Singh 2016CS50389

October 20, 2019

1 Introduction

MapReduce is a programming model suitable for processing of huge data. Hadoop is capable of running MapReduce programs written in various languages. These programs are parallel in nature, thus are very useful for performing large scale data analysis using multiple machines in the cluster.
The whole process goes through the following phases:

- Input Splits
- Mapping
- Shuffling
- Reducing

2 Installation

1. Map Reduce framework comes installed with Hadoop File System.

2. Now we need to write the java code for WordCounting and AverageGradeComputing.

3. Compile the java code using:
   ```
   javac AverageGradeCompute.java -cp $(/opt/hadoop/hadoop/bin/hadoop classpath)
   ```

4. Create the jar file using:
   ```
   jar cf ag.jar AverageGrade*.class
   ```

5. Move the input files to the hadoop file system:
   ```
   bin/hadoop fs -copyFromLocal  /records.txt /Assignment_mapreduce/records.txt
   ```

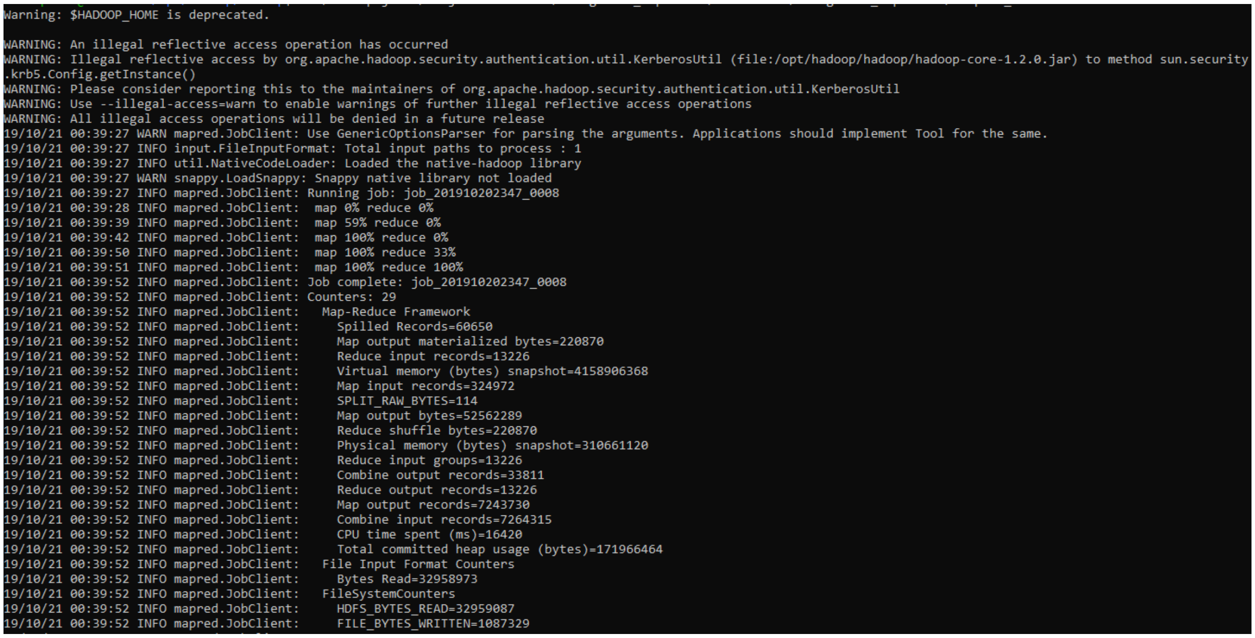
6. Run the program as follows:
   bin/hadoop jar /ag.jar AverageGradeCompute /Assignment_mapreduce/records.txt /Assignment_mapreduce/output2

7. You can view the job at the jobtracker at host:50030

3 Text file word count

In many cases, we need to perform a word count in the text files. However, the large size of the files makes it very difficult to count the words by a single worker. This is where MapReduce comes to rescue. This task can be performed by splitting the large file into smaller chunks. We can then perform the word count operation on each file independently by different workers. This is the map task. As the map workers start writing their individual outputs in their own respective files, the reduce workers need to merge the output in the global output files. This is the reduce step.
The word count application is quite straight forward. The Mapper implementation, via the map method, processes one line at a time, as provided by the specified TextInputFormat. It then splits the line into tokens separated by white spaces, via the StringTokenizer, and emits a key-value pair of (word, 1). This happens at each of the map worker nodes. The output of each map is then passed through the local combiner (which is same as the Reducer as per the job configuration) for local aggregation, after being sorted on the keys. The Reducer implementation, via the reduce method just sums up the values, which are the occurrence counts for each key.

3.1 No shutdown



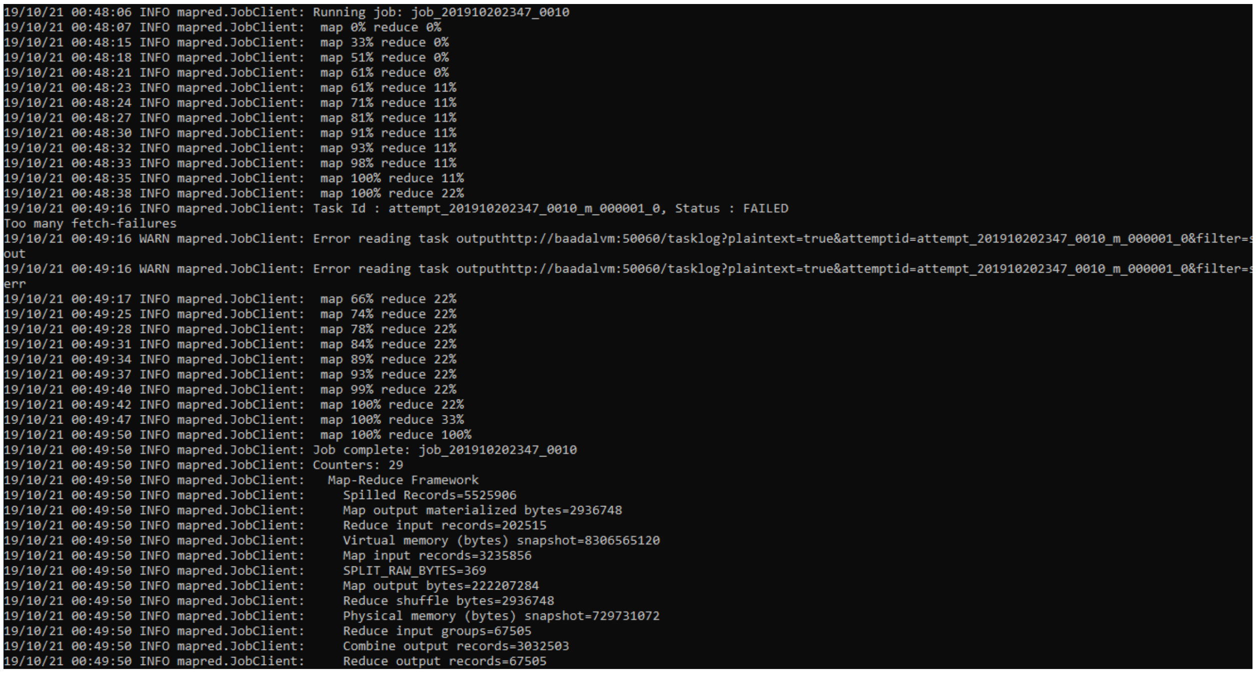
```
Warning: $HADOOP_HOME is deprecated.

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/opt/hadoop/hadoop/hadoop-core-1.2.0.jar) to method sun.security
.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
19/10/21 00:39:27 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
19/10/21 00:39:27 INFO input.FileInputFormat: Total input paths to process : 1
19/10/21 00:39:27 INFO util.NativeCodeLoader: Loaded the native-hadoop library
19/10/21 00:39:27 WARN snappy.LoadSnappy: Snappy native library not loaded
19/10/21 00:39:27 INFO mapred.JobClient: Running job: job_201910202347_0008
19/10/21 00:39:28 INFO mapred.JobClient:  map 0% reduce 0%
19/10/21 00:39:39 INFO mapred.JobClient:  map 59% reduce 0%
19/10/21 00:39:42 INFO mapred.JobClient:  map 100% reduce 0%
19/10/21 00:39:50 INFO mapred.JobClient:  map 100% reduce 33%
19/10/21 00:39:51 INFO mapred.JobClient:  map 100% reduce 100%
19/10/21 00:39:52 INFO mapred.JobClient: Job complete: job_201910202347_0008
19/10/21 00:39:52 INFO mapred.JobClient: Counters: 29
19/10/21 00:39:52 INFO mapred.JobClient:   Map-Reduce Framework
19/10/21 00:39:52 INFO mapred.JobClient:     Spilled Records=60650
19/10/21 00:39:52 INFO mapred.JobClient:     Map output materialized bytes=220870
19/10/21 00:39:52 INFO mapred.JobClient:     Reduce input records=13226
19/10/21 00:39:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=4158906368
19/10/21 00:39:52 INFO mapred.JobClient:     Map input records=324972
19/10/21 00:39:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=114
19/10/21 00:39:52 INFO mapred.JobClient:     Map output bytes=52562289
19/10/21 00:39:52 INFO mapred.JobClient:     Reduce shuffle bytes=220870
19/10/21 00:39:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot=310661120
19/10/21 00:39:52 INFO mapred.JobClient:     Reduce input groups=13226
19/10/21 00:39:52 INFO mapred.JobClient:     Combine output records=33811
19/10/21 00:39:52 INFO mapred.JobClient:     Reduce output records=13226
19/10/21 00:39:52 INFO mapred.JobClient:     Map output records=7243730
19/10/21 00:39:52 INFO mapred.JobClient:     Combine input records=7264315
19/10/21 00:39:52 INFO mapred.JobClient:     CPU time spent (ms)=16420
19/10/21 00:39:52 INFO mapred.JobClient:     Total committed heap usage (bytes)=171966464
19/10/21 00:39:52 INFO mapred.JobClient:   File Input Format Counters
19/10/21 00:39:52 INFO mapred.JobClient:     Bytes Read=32958973
19/10/21 00:39:52 INFO mapred.JobClient:   FileSystemCounters
19/10/21 00:39:52 INFO mapred.JobClient:     HDFS_BYTES_READ=32959087
19/10/21 00:39:52 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=1087329
```

In this experiment, we do not shutdown any map worker. So, the map task gradually rises to 100%. Meanwhile, as the map workers start creating their individual output files, the reduce workers get deployed, and gradually the reduce task also rises to 100%. **Logs**:

- Size of the input file = 32.959087 MB
- Size of the output file = 1.087329 MB
- Total CPU time spent = 16420 ms
- Total time taken for Map task to reach 100% = 4 s

3.2 One shutdown

In this experiment, we shutdown a worker in between. After the worker shuts down, the map task on that worker needs to be recomputed. So, the map task falls from 100% to 66% suddenly. This task now gets redistributed to other map workers and thus, the map task again rises to 100%. The reduce task stops for this transition time and regains pace once the map task again reaches 100%.



```
19/10/21 00:48:06 INFO mapred.JobClient: Running job: job_201910202347_0010
19/10/21 00:48:07 INFO mapred.JobClient:  map 0% reduce 0%
19/10/21 00:48:15 INFO mapred.JobClient:  map 33% reduce 0%
19/10/21 00:48:18 INFO mapred.JobClient:  map 51% reduce 0%
19/10/21 00:48:21 INFO mapred.JobClient:  map 61% reduce 0%
19/10/21 00:48:23 INFO mapred.JobClient:  map 61% reduce 11%
19/10/21 00:48:24 INFO mapred.JobClient:  map 71% reduce 11%
19/10/21 00:48:27 INFO mapred.JobClient:  map 81% reduce 11%
19/10/21 00:48:30 INFO mapred.JobClient:  map 91% reduce 11%
19/10/21 00:48:32 INFO mapred.JobClient:  map 93% reduce 11%
19/10/21 00:48:33 INFO mapred.JobClient:  map 98% reduce 11%
19/10/21 00:48:35 INFO mapred.JobClient:  map 100% reduce 11%
19/10/21 00:48:38 INFO mapred.JobClient:  map 100% reduce 22%
19/10/21 00:49:16 INFO mapred.JobClient: Task Id : attempt_201910202347_0010_m_000001_0, Status : FAILED
Too many fetch-failures
19/10/21 00:49:16 WARN mapred.JobClient: Error reading task outputhttp://baadalvm:50060/tasklog?plaintext=true&attemptid=attempt_201910202347_0010_m_000001_0&filter=s
out
19/10/21 00:49:16 WARN mapred.JobClient: Error reading task outputhttp://baadalvm:50060/tasklog?plaintext=true&attemptid=attempt_201910202347_0010_m_000001_0&filter=s
err
19/10/21 00:49:17 INFO mapred.JobClient:  map 66% reduce 22%
19/10/21 00:49:25 INFO mapred.JobClient:  map 74% reduce 22%
19/10/21 00:49:28 INFO mapred.JobClient:  map 78% reduce 22%
19/10/21 00:49:31 INFO mapred.JobClient:  map 84% reduce 22%
19/10/21 00:49:34 INFO mapred.JobClient:  map 89% reduce 22%
19/10/21 00:49:37 INFO mapred.JobClient:  map 93% reduce 22%
19/10/21 00:49:40 INFO mapred.JobClient:  map 99% reduce 22%
19/10/21 00:49:42 INFO mapred.JobClient:  map 100% reduce 22%
19/10/21 00:49:47 INFO mapred.JobClient:  map 100% reduce 33%
19/10/21 00:49:50 INFO mapred.JobClient:  map 100% reduce 100%
19/10/21 00:49:50 INFO mapred.JobClient: Job complete: job_201910202347_0010
19/10/21 00:49:50 INFO mapred.JobClient: Counters: 29
19/10/21 00:49:50 INFO mapred.JobClient:   Map-Reduce Framework
19/10/21 00:49:50 INFO mapred.JobClient:     Spilled Records=5525906
19/10/21 00:49:50 INFO mapred.JobClient:     Map output materialized bytes=2936748
19/10/21 00:49:50 INFO mapred.JobClient:     Reduce input records=202515
19/10/21 00:49:50 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=8306565120
19/10/21 00:49:50 INFO mapred.JobClient:     Map input records=3235856
19/10/21 00:49:50 INFO mapred.JobClient:     SPLIT_RAW_BYTES=369
19/10/21 00:49:50 INFO mapred.JobClient:     Map output bytes=222207284
19/10/21 00:49:50 INFO mapred.JobClient:     Reduce shuffle bytes=2936748
19/10/21 00:49:50 INFO mapred.JobClient:     Physical memory (bytes) snapshot=729731072
19/10/21 00:49:50 INFO mapred.JobClient:     Reduce input groups=67505
19/10/21 00:49:50 INFO mapred.JobClient:     Combine output records=3032503
19/10/21 00:49:50 INFO mapred.JobClient:     Reduce output records=67505
```

Logs:

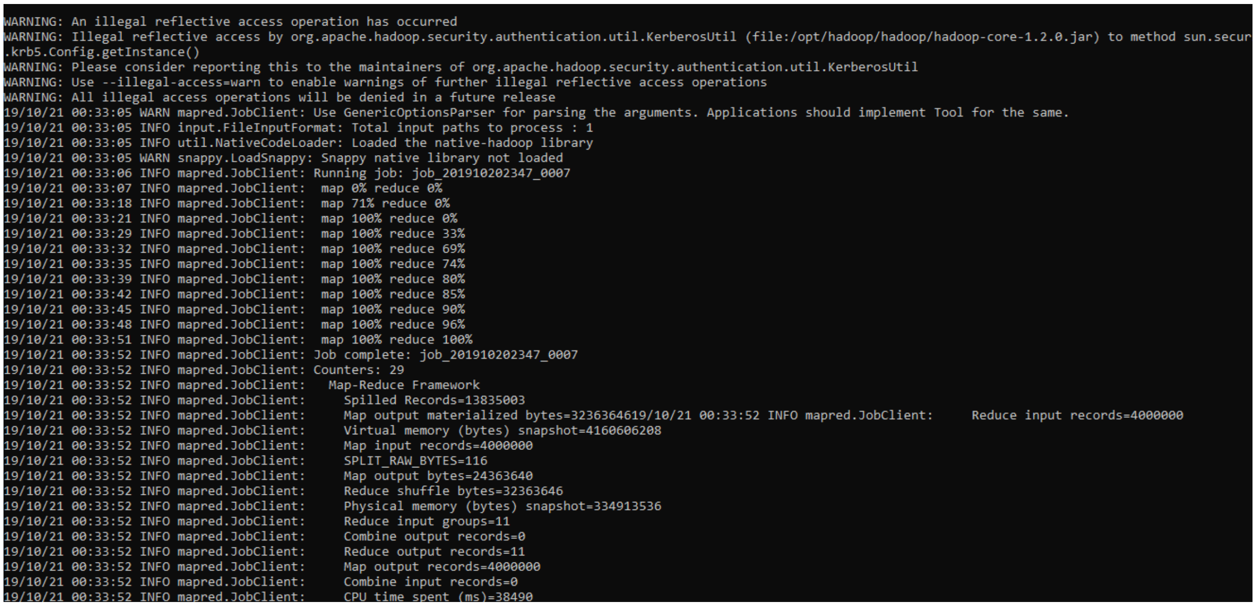
- Size of the input file = 135.34 MB
- Size of the output file = 777.35 KB
- Total CPU time spent = 74080 ms
- Total time taken for Map task to reach 100% = 35 s

4 Average grade of each course

In this section, we create a large file with records in the following format:

$$\text{Roll no.} < \textit{space} > \text{course code} < \textit{space} > \text{grade}$$

This is a large file with 40,00,000 records. This makes it difficult for a single worker to compute the average grade of each roll no.So, we divide the work among different workers using MapReduce.



```
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/opt/hadoop/hadoop/hadoop-core-1.2.0.jar) to method sun.secur
.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
19/10/21 00:33:05 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
19/10/21 00:33:05 INFO input.FileInputFormat: Total input paths to process : 1
19/10/21 00:33:05 INFO util.NativeCodeLoader: Loaded the native-hadoop library
19/10/21 00:33:05 WARN snappy.LoadSnappy: Snappy native library not loaded
19/10/21 00:33:06 INFO mapred.JobClient: Running job: job_201910202347_0007
19/10/21 00:33:07 INFO mapred.JobClient:  map 0% reduce 0%
19/10/21 00:33:18 INFO mapred.JobClient:  map 71% reduce 0%
19/10/21 00:33:21 INFO mapred.JobClient:  map 100% reduce 0%
19/10/21 00:33:29 INFO mapred.JobClient:  map 100% reduce 33%
19/10/21 00:33:32 INFO mapred.JobClient:  map 100% reduce 69%
19/10/21 00:33:35 INFO mapred.JobClient:  map 100% reduce 74%
19/10/21 00:33:39 INFO mapred.JobClient:  map 100% reduce 80%
19/10/21 00:33:42 INFO mapred.JobClient:  map 100% reduce 85%
19/10/21 00:33:45 INFO mapred.JobClient:  map 100% reduce 90%
19/10/21 00:33:48 INFO mapred.JobClient:  map 100% reduce 96%
19/10/21 00:33:51 INFO mapred.JobClient:  map 100% reduce 100%
19/10/21 00:33:52 INFO mapred.JobClient: Job complete: job_201910202347_0007
19/10/21 00:33:52 INFO mapred.JobClient: Counters: 29
19/10/21 00:33:52 INFO mapred.JobClient:   Map-Reduce Framework
19/10/21 00:33:52 INFO mapred.JobClient:     Spilled Records=13835003
19/10/21 00:33:52 INFO mapred.JobClient:     Map output materialized bytes=3236364619/10/21 00:33:52 INFO mapred.JobClient:     Reduce input records=4000000
19/10/21 00:33:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=4160606208
19/10/21 00:33:52 INFO mapred.JobClient:     Map input records=4000000
19/10/21 00:33:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=116
19/10/21 00:33:52 INFO mapred.JobClient:     Map output bytes=24363640
19/10/21 00:33:52 INFO mapred.JobClient:     Reduce shuffle bytes=32363646
19/10/21 00:33:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot=334913536
19/10/21 00:33:52 INFO mapred.JobClient:     Reduce input groups=11
19/10/21 00:33:52 INFO mapred.JobClient:     Combine output records=0
19/10/21 00:33:52 INFO mapred.JobClient:     Reduce output records=11
19/10/21 00:33:52 INFO mapred.JobClient:     Map output records=4000000
19/10/21 00:33:52 INFO mapred.JobClient:     Combine input records=0
19/10/21 00:33:52 INFO mapred.JobClient:     CPU time spent (ms)=38490
```

Logs:

- Size of the input file = 48.5 MB
- Size of the output file = 0.46 KB
- Total CPU time spent = 38490 ms
- Total time taken for Map task to reach 100% = 14 s