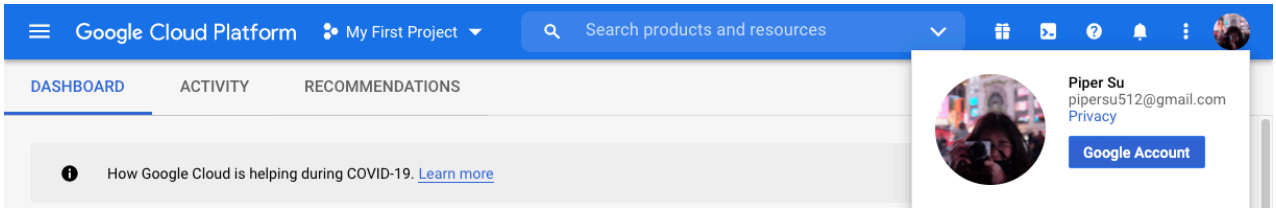
GCP homework

GCP account (25%):



Source code (25%):

```
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

    public static class TokenizerMapper
            extends Mapper<Object, Text, Text, IntWritable>{

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context
        ) throws IOException, InterruptedException {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
            }
        }
```

```
    }

    public static class IntSumReducer
            extends Reducer<Text,IntWritable,Text,IntWritable> {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values,
                           Context context
        ) throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            result.set(sum);
            context.write(key, result);
        }
    }

    public static void main(String[] args) throws Exception {
        // record start time
        long startTime = System.currentTimeMillis();
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        if (job.waitForCompletion(true)) {
            // record end time if job ended successfully
            long endTime = System.currentTimeMillis();
            System.out.println("Execution time: " + (endTime-startTime) + "
ms");
            System.exit(0);
        } else {
            System.exit(1);
        }
    }
}
```

Run WordCount.jar on GCP (50%)

```
yarn jar WordCount1.jar WordCount /wordcountfiles/ /tmp/wordCount_result_new/
```

Execution time: 23312 ms

```
pipersu512@cluster-afb3-m:~$ yarn jar WordCount1.jar WordCount /wordcountfiles/ /tmp/wordCou
nt_result_new/
20/10/23 05:20:21 INFO client.RMProxy: Connecting to ResourceManager at cluster-afb3-m/10.12
8.0.2:8032
20/10/23 05:20:21 INFO client.AHSProxy: Connecting to Application History server at cluster-
afb3-m/10.128.0.2:10200
20/10/23 05:20:21 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
 performed. Implement the Tool interface and execute your application with ToolRunner to rem
edy this.
20/10/23 05:20:21 INFO input.FileInputFormat: Total input files to process : 1
20/10/23 05:20:22 INFO mapreduce.JobSubmitter: number of splits:1
20/10/23 05:20:22 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publis
her.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/10/23 05:20:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1603256724313_
0006
20/10/23 05:20:22 INFO impl.YarnClientImpl: Submitted application application_1603256724313_
0006
20/10/23 05:20:22 INFO mapreduce.Job: The url to track the job: http://cluster-afb3-m:8088/p
roxy/application_1603256724313_0006/
20/10/23 05:20:22 INFO mapreduce.Job: Running job: job_1603256724313_0006
20/10/23 05:20:28 INFO mapreduce.Job: Job job_1603256724313_0006 running in uber mode : fals
e
20/10/23 05:20:28 INFO mapreduce.Job:  map 0% reduce 0%
20/10/23 05:20:32 INFO mapreduce.Job:  map 100% reduce 0%
20/10/23 05:20:39 INFO mapreduce.Job:  map 100% reduce 29%
20/10/23 05:20:40 INFO mapreduce.Job:  map 100% reduce 43%
20/10/23 05:20:41 INFO mapreduce.Job:  map 100% reduce 71%
20/10/23 05:20:42 INFO mapreduce.Job:  map 100% reduce 86%
20/10/23 05:20:43 INFO mapreduce.Job:  map 100% reduce 100%
20/10/23 05:20:43 INFO mapreduce.Job: Job job_1603256724313_0006 completed successfully
20/10/23 05:20:43 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=66
                FILE: Number of bytes written=1660763
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=374
                HDFS: Number of bytes written=18
                HDFS: Number of read operations=38
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=1
                Launched reduce tasks=7
                Data-local map tasks=1
```

```
                HDFS: Number of write operations=21
        Job Counters
                Killed reduce tasks=1
                Launched map tasks=1
                Launched reduce tasks=7
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7404
                Total time spent by all reduces in occupied slots (ms)=100251
                Total time spent by all map tasks (ms)=2468
                Total time spent by all reduce tasks (ms)=33417
                Total vcore-milliseconds taken by all map tasks=2468
                Total vcore-milliseconds taken by all reduce tasks=33417
                Total megabyte-milliseconds taken by all map tasks=7581696
                Total megabyte-milliseconds taken by all reduce tasks=102657024
        Map-Reduce Framework
                Map input records=22
                Map output records=44
                Map output bytes=440
                Map output materialized bytes=66
                Input split bytes=111
                Combine input records=44
                Combine output records=2
                Reduce input groups=2
                Reduce shuffle bytes=66
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =7
                Failed Shuffles=0
                Merged Map outputs=7
                GC time elapsed (ms)=912
                CPU time spent (ms)=5740
                Physical memory (bytes) snapshot=2418044928
                Virtual memory (bytes) snapshot=34936770560
                Total committed heap usage (bytes)=2003304448
        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=263
        File Output Format Counters
                Bytes Written=18
Execution time: 23312 ms
```

The input file "input.txt" contains only "Hello World", results are as expected.

```
pipersu512@cluster-afb3-m:~/tmp$ cd wordCount_result_new/
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ ls
_SUCCESS  part-r-00000  part-r-00001  part-r-00002  part-r-00003  part-r-00004  part-r-00005  part-r-00006
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00000
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00001
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00002
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00003
Hello   22
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00004
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00005
World   22
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$ more part-r-00006
pipersu512@cluster-afb3-m:~/tmp/wordCount_result_new$
```