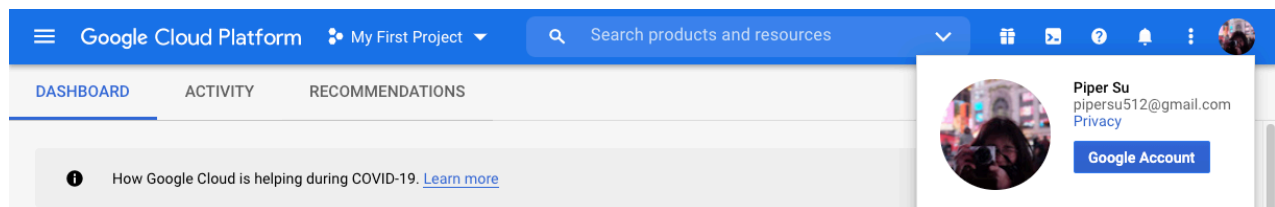


GCP homework

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]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
        // record end time
        long endTime = System.currentTimeMillis();
        // print out execution time duration
        System.out.println("Execution time: " + (endTime-startTime) + " ms");
    }
}

```

Run WordCount.jar on GCP (50%)

```
yarn jar WordCount.jar WordCount /wordcountfiles/ /tmp/wordCount_result/
```

Execution time: 927 ms

```
pipersu512@cluster-afb3-m:~$ pwd
/home/pipersu512
pipersu512@cluster-afb3-m:~$ yarn jar WordCount.jar WordCount /wordcountfiles/ /tmp/wordCount_result/
Execution time: 927 ms
```

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

```
pipersu512@cluster-afb3-m:~$ hdfs dfs -get /tmp/wordCount_result* ./tmp/.
pipersu512@cluster-afb3-m:~$ cd tmp
pipersu512@cluster-afb3-m:~/tmp$ ls
wordCount_result
pipersu512@cluster-afb3-m:~/tmp$ cd wordCount_result/
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ ls
_SUCCESS part-r-00000 part-r-00001 part-r-00002 part-r-00003 part-r-00004 part-r-00005 part-r-0
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00000
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00001
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00002
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00003
Hello 22
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00004
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00005
World 22
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$ more part-r-00006
pipersu512@cluster-afb3-m:~/tmp/wordCount_result$
```