Simple Aggregation in MapReduce

MsgSizeAggregateMapReduce.java

```
import java.io.IOException;
import java.util.Iterator;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
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;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class MsgSizeAggregateMapReduce extends Configured
implements Tool {

     public static void main(String[] args) throws Exception {
          int res = ToolRunner.run(new Configuration(),
                    new MsgSizeAggregateMapReduce(), args);
          System.exit(res);
     }

     @Override
     public int run(String[] args) throws Exception {

          if (args.length != 2) {
               System.err.println("Usage:  <input_path>
<output_path>");
               System.exit(-1);
          }
          /* input parameters */
          String inputPath = args[0];
          String outputPath = args[1];

          Job job = Job.getInstance(getConf(),
"WebLogMessageSizeAggregator");
          job.setJarByClass(MsgSizeAggregateMapReduce.class);
          job.setMapperClass(AMapper.class);
          job.setReducerClass(AReducer.class);
          job.setNumReduceTasks(1);

          job.setOutputKeyClass(Text.class);
          job.setOutputValueClass(IntWritable.class);
          FileInputFormat.setInputPaths(job, new Path(inputPath));
          FileOutputFormat.setOutputPath(job, new
Path(outputPath));

          int exitStatus = job.waitForCompletion(true) ? 0 : 1;
          return exitStatus;
     }

     public static class AMapper extends Mapper<Object, Text, Text,
IntWritable> {
          public static final Pattern httplogPattern = Pattern
                    .compile("([^\\s]+) - - \\[(.+)\\] \"([^\\s]+)
(/[^\\s]*) HTTP/[^\\s]+\" [^\\s]+ ([0-9]+)");

          public void map(Object key, Text value, Context context)
                    throws IOException, InterruptedException {
               Matcher matcher =
httplogPattern.matcher(value.toString());
               if (matcher.matches()) {
                    int size = Integer.parseInt(matcher.group(5));
                    context.write(new Text("msgSize"), new
IntWritable(size));
               }
          }
     }

     public static class AReducer extends
               Reducer<Text, IntWritable, Text, IntWritable> {
          public void reduce(Text key, Iterable<IntWritable>
values,
                    Context context) throws IOException,
InterruptedException {
               double tot = 0;
               int count = 0;
               int min = Integer.MAX_VALUE;
               int max = 0;
               Iterator<IntWritable> iterator = values.iterator();
               while (iterator.hasNext()) {
                    int value = iterator.next().get();
                    tot = tot + value;
                    count++;
                    if (value < min) {
                         min = value;
                    }
                    if (value > max) {
                         max = value;
                    }
               }
               context.write(new Text("Mean"), new
IntWritable((int) tot / count));
               context.write(new Text("Max"), new
IntWritable(max));
               context.write(new Text("Min"), new
IntWritable(min));
          }
     }
}
```

```
[ashesh@MISTRI Aggregate] $ wget
ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz

[ashesh@MISTRI Aggregate] $ gunzip -k NASA_access_log_Jul95.gz

[ashesh@MISTRI Aggregate] $ head -3 NASA_access_log_Jul95
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET
/history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET
/shuttle/countdown/ HTTP/1.0" 200 3985
199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET
/shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085

[ashesh@MISTRI Aggregate] $ hadoop fs -mkdir /input

[ashesh@MISTRI Aggregate] $ hadoop fs -put
NASA_access_log_Jul95 /input

[ashesh@MISTRI Aggregate] $ hadoop com.sun.tools.javac.Main
*.java; jar cf Aggregate.jar *.class

[ashesh@MISTRI Aggregate] $ hadoop jar Aggregate.jar
MsgSizeAggregateMapReduce /input /output

[ashesh@MISTRI Aggregate] $ hadoop fs -cat /output/part*
Mean 1150
Max  6823936
Min  0
```