HU Extension **Assignment 06** E-63 Big Data Analytics

**Working Java classes and outputs produced are attached in the zip folder. Important code snippets (mapper/reducers) will be explained here along with code outputs and some console outputs.**

**Problem 1**

//map

**public** **void** map(Object key, Text value, Context context

) **throws** IOException, InterruptedException {

//split into comma separated variables

String set[] = value.toString().split(",");

String country=set[4].replace("\"", "");

//take US states as separate countries

**if**(country.equals("US")){country=set[4].replace("\"", "")+"-"+set[5].replace("\"", "");}

String citation = set[0];

context.write(**new** Text(country), **new** Text(citation));

}

//reduce

**public** **void** reduce(Text key, Iterable<Text> values,

Context context

) **throws** IOException, InterruptedException {

String csv = "";

**while** (values.iterator().hasNext()) {

**if** (csv.length() > 0) csv += ",";

csv += values.iterator().next().toString();

}

context.write(key, **new** Text(csv));

}

**Console output**

hadoop jar /mnt/hgfs/DilmiShared/CountryToCitationMRv2.jar edu.hu.bigdata.CountryToCitationMRv2 citations\_input/apat63\_99.txt citations\_output\_problem1.111

15/03/13 18:29:56 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

15/03/13 18:29:57 INFO input.FileInputFormat: Total input paths to process : 1

15/03/13 18:29:57 INFO mapreduce.JobSubmitter: number of splits:2

15/03/13 18:29:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1426159947607\_0024

15/03/13 18:29:57 INFO impl.YarnClientImpl: Submitted application application\_1426159947607\_0024

15/03/13 18:29:57 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1426159947607\_0024/

15/03/13 18:29:57 INFO mapreduce.Job: Running job: job\_1426159947607\_0024

15/03/13 18:30:07 INFO mapreduce.Job: Job job\_1426159947607\_0024 running in uber mode : false

15/03/13 18:30:07 INFO mapreduce.Job: map 0% reduce 0%

15/03/13 18:30:24 INFO mapreduce.Job: map 5% reduce 0%

15/03/13 18:30:26 INFO mapreduce.Job: map 17% reduce 0%

15/03/13 18:30:27 INFO mapreduce.Job: map 23% reduce 0%

15/03/13 18:30:29 INFO mapreduce.Job: map 33% reduce 0%

15/03/13 18:30:31 INFO mapreduce.Job: map 42% reduce 0%

15/03/13 18:30:32 INFO mapreduce.Job: map 53% reduce 0%

15/03/13 18:30:34 INFO mapreduce.Job: map 62% reduce 0%

15/03/13 18:30:36 INFO mapreduce.Job: map 79% reduce 0%

15/03/13 18:30:37 INFO mapreduce.Job: map 83% reduce 0%

15/03/13 18:30:38 INFO mapreduce.Job: map 100% reduce 0%

15/03/13 18:30:50 INFO mapreduce.Job: map 100% reduce 67%

15/03/13 18:31:11 INFO mapreduce.Job: map 100% reduce 68%

15/03/13 18:31:48 INFO mapreduce.Job: map 100% reduce 69%

15/03/13 18:34:25 INFO mapreduce.Job: map 100% reduce 70%

15/03/13 18:39:27 INFO mapreduce.Job: map 100% reduce 71%

15/03/13 18:40:28 INFO mapreduce.Job: map 100% reduce 72%

15/03/13 18:41:44 INFO mapreduce.Job: map 100% reduce 73%

15/03/13 18:42:42 INFO mapreduce.Job: map 100% reduce 74%

15/03/13 18:48:11 INFO mapreduce.Job: map 100% reduce 75%

**Problem 2**

//map

**public** **void** map(Text key, Text value, Context context)

**throws** IOException, InterruptedException {

String citationList[]=value.toString().split(",");

//count values and invert

context.write(**new** IntWritable(citationList.length), key);

}

//reducer

**public** **void** reduce(IntWritable key, Iterable<Text> values, Context context)

**throws** IOException, InterruptedException {

**int** count = 0;

//calculate histogram

**for** (Text val : values) { // Iterable allows for looping

count++;

}

context.write(key, **new** IntWritable(count));

}

hadoop jar /mnt/hgfs/DilmiShared/CountryToCitationHistogramMRv2.jar edu.hu.bigdata.CountryToCitationHistogramMRv2 input/Problem1-output.txt output2.13

15/03/13 19:49:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

15/03/13 19:50:00 INFO input.FileInputFormat: Total input paths to process : 1

15/03/13 19:50:01 INFO mapreduce.JobSubmitter: number of splits:1

15/03/13 19:50:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1426159947607\_0026

15/03/13 19:50:02 INFO impl.YarnClientImpl: Submitted application application\_1426159947607\_0026

15/03/13 19:50:02 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1426159947607\_0026/

15/03/13 19:50:02 INFO mapreduce.Job: Running job: job\_1426159947607\_0026

15/03/13 19:50:18 INFO mapreduce.Job: Job job\_1426159947607\_0026 running in uber mode : false

15/03/13 19:50:18 INFO mapreduce.Job: map 0% reduce 0%

15/03/13 19:50:33 INFO mapreduce.Job: map 100% reduce 0%

15/03/13 19:50:47 INFO mapreduce.Job: map 100% reduce 100%

15/03/13 19:50:47 INFO mapreduce.Job: Job job\_1426159947607\_0026 completed successfully

15/03/13 19:50:48 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=2153

FILE: Number of bytes written=217471

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=23392339

HDFS: Number of bytes written=1656

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=1

Launched reduce tasks=1

Data-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12928

Total time spent by all reduces in occupied slots (ms)=11467

Total time spent by all map tasks (ms)=12928

Total time spent by all reduce tasks (ms)=11467

Total vcore-seconds taken by all map tasks=12928

Total vcore-seconds taken by all reduce tasks=11467

Total megabyte-seconds taken by all map tasks=13238272

Total megabyte-seconds taken by all reduce tasks=11742208

Map-Reduce Framework

Map input records=219

Map output records=219

Map output bytes=1709

Map output materialized bytes=2153

Input split bytes=121

Combine input records=0

Combine output records=0

Reduce input groups=145

Reduce shuffle bytes=2153

Reduce input records=219

Reduce output records=219

Spilled Records=438

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=1439

CPU time spent (ms)=2790

Physical memory (bytes) snapshot=432488448

Virtual memory (bytes) snapshot=1701339136

Total committed heap usage (bytes)=263323648

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=23392218

File Output Format Counters

Bytes Written=1656

**Problem 3**

**Use the same mappers and reducers from above. The master program will change as follows to run chained mapreduce.**

Configuration conf = getConf();

Path in = **new** Path(args[0]);

Path out = **new** Path(args[1]);

Path temp = **new** Path("chain-temp");

//Job 1

Job job = **new** ~~Job~~(conf, "CountryToCitationMRv2");

job.setJarByClass(ChainedHistogramMRv2.**class**);

FileInputFormat.*setInputPaths*(job, in);

FileOutputFormat.*setOutputPath*(job, temp);

job.setMapperClass(CountryToCitationMapper.**class**);

job.setReducerClass(CountryToCitationReducer.**class**);

job.setMapOutputKeyClass(Text.**class**);

job.setMapOutputValueClass(Text.**class**);

//job.setInputFormatClass(KeyValueTextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.**class**);

job.setOutputKeyClass(Text.**class**);

job.setOutputValueClass(Text.**class**);

**boolean** success = job.waitForCompletion(**true**);

**if** (success){

//Job 2

Job job2 = **new** ~~Job~~(conf, "CountryToCitationHistogramMRv2");

job2.setJarByClass(ChainedHistogramMRv2.**class**);

FileInputFormat.*setInputPaths*(job2, temp);

FileOutputFormat.*setOutputPath*(job2, out);

job2.setMapperClass(InverterCounterMapper.**class**);

job2.setReducerClass(HistogramReducer.**class**);

job2.setMapOutputKeyClass(IntWritable.**class**);

job2.setMapOutputValueClass(Text.**class**);

job2.setInputFormatClass(KeyValueTextInputFormat.**class**);

job2.setOutputFormatClass(TextOutputFormat.**class**);

job2.setOutputKeyClass(IntWritable.**class**);

job2.setOutputValueClass(IntWritable.**class**);

success = job2.waitForCompletion(**true**);

**if** (success)

**return** 0;

**else**

**return** 1;

}

**else**

**return** 1;

**Console output**

hadoop jar /mnt/hgfs/DilmiShared/ChainedHistogramMRv2.jar edu.hu.bigdata.ChainedHistogramMRv2 citations\_input/apat63\_99.txt citations\_output\_problem3.2

15/03/13 07:53:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

15/03/13 07:53:42 INFO input.FileInputFormat: Total input paths to process : 1

15/03/13 07:53:42 INFO mapreduce.JobSubmitter: number of splits:2

15/03/13 07:53:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1426159947607\_0020

15/03/13 07:53:43 INFO impl.YarnClientImpl: Submitted application application\_1426159947607\_0020

15/03/13 07:53:43 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1426159947607\_0020/

15/03/13 07:53:43 INFO mapreduce.Job: Running job: job\_1426159947607\_0020

15/03/13 07:53:54 INFO mapreduce.Job: Job job\_1426159947607\_0020 running in uber mode : false

15/03/13 07:53:54 INFO mapreduce.Job: map 0% reduce 0%

15/03/13 07:54:10 INFO mapreduce.Job: map 4% reduce 0%

15/03/13 07:54:13 INFO mapreduce.Job: map 11% reduce 0%

15/03/13 07:54:14 INFO mapreduce.Job: map 20% reduce 0%

15/03/13 07:54:16 INFO mapreduce.Job: map 28% reduce 0%

15/03/13 07:54:17 INFO mapreduce.Job: map 39% reduce 0%

15/03/13 07:54:19 INFO mapreduce.Job: map 49% reduce 0%

15/03/13 07:54:20 INFO mapreduce.Job: map 61% reduce 0%

15/03/13 07:54:22 INFO mapreduce.Job: map 66% reduce 0%

15/03/13 07:54:23 INFO mapreduce.Job: map 67% reduce 0%

15/03/13 07:54:24 INFO mapreduce.Job: map 83% reduce 0%

15/03/13 07:54:25 INFO mapreduce.Job: map 100% reduce 0%

15/03/13 07:54:37 INFO mapreduce.Job: map 100% reduce 67%

15/03/13 07:54:55 INFO mapreduce.Job: map 100% reduce 68%

15/03/13 07:55:32 INFO mapreduce.Job: map 100% reduce 69%

15/03/13 07:58:04 INFO mapreduce.Job: map 100% reduce 70%

15/03/13 08:02:46 INFO mapreduce.Job: map 100% reduce 71%

15/03/13 08:03:46 INFO mapreduce.Job: map 100% reduce 72%

15/03/13 08:05:00 INFO mapreduce.Job: map 100% reduce 73%

15/03/13 08:05:55 INFO mapreduce.Job: map 100% reduce 74%

15/03/13 08:11:15 INFO mapreduce.Job: map 100% reduce 75%

15/03/13 08:26:27 INFO mapreduce.Job: map 100% reduce 76%

15/03/13 08:57:27 INFO mapreduce.Job: map 100% reduce 77%

^[[A^[[B15/03/13 09:16:27 INFO mapreduce.Job: map 100% reduce 78%

15/03/13 09:16:43 INFO mapreduce.Job: map 100% reduce 79%

15/03/13 09:17:40 INFO mapreduce.Job: map 100% reduce 80%

15/03/13 09:21:56 INFO mapreduce.Job: map 100% reduce 81%

15/03/13 09:32:45 INFO mapreduce.Job: map 100% reduce 82%

15/03/13 09:34:54 INFO mapreduce.Job: map 100% reduce 83%

15/03/13 09:35:19 INFO mapreduce.Job: map 100% reduce 84%

15/03/13 09:35:43 INFO mapreduce.Job: map 100% reduce 85%

15/03/13 09:37:17 INFO mapreduce.Job: map 100% reduce 86%

15/03/13 09:37:33 INFO mapreduce.Job: map 100% reduce 87%

15/03/13 09:38:18 INFO mapreduce.Job: map 100% reduce 88%

15/03/13 09:39:05 INFO mapreduce.Job: map 100% reduce 89%

15/03/13 09:39:54 INFO mapreduce.Job: map 100% reduce 90%

15/03/13 09:40:09 INFO mapreduce.Job: map 100% reduce 91%

15/03/13 09:41:37 INFO mapreduce.Job: map 100% reduce 92%

15/03/13 09:42:38 INFO mapreduce.Job: map 100% reduce 93%

15/03/13 09:44:56 INFO mapreduce.Job: map 100% reduce 94%

15/03/13 09:47:48 INFO mapreduce.Job: map 100% reduce 95%

15/03/13 09:48:58 INFO mapreduce.Job: map 100% reduce 96%

15/03/13 09:49:38 INFO mapreduce.Job: map 100% reduce 97%

15/03/13 09:50:42 INFO mapreduce.Job: map 100% reduce 98%

15/03/13 09:51:40 INFO mapreduce.Job: map 100% reduce 99%

15/03/13 09:52:04 INFO mapreduce.Job: map 100% reduce 100%

15/03/13 09:52:13 INFO mapreduce.Job: Job job\_1426159947607\_0020 completed successfully

15/03/13 09:52:13 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=43365978

FILE: Number of bytes written=87051047

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=236907525

HDFS: Number of bytes written=23392218

HDFS: Number of read operations=9

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=2

Launched reduce tasks=1

Data-local map tasks=2

Total time spent by all maps in occupied slots (ms)=57738

Total time spent by all reduces in occupied slots (ms)=7066401

Total time spent by all map tasks (ms)=57738

Total time spent by all reduce tasks (ms)=7066401

Total vcore-seconds taken by all map tasks=57738

Total vcore-seconds taken by all reduce tasks=7066401

Total megabyte-seconds taken by all map tasks=59123712

Total megabyte-seconds taken by all reduce tasks=7235994624

Map-Reduce Framework

Map input records=2923923

Map output records=2923923

Map output bytes=37518126

Map output materialized bytes=43365984

Input split bytes=250

Combine input records=0

Combine output records=0

Reduce input groups=219

Reduce shuffle bytes=43365984

Reduce input records=2923923

Reduce output records=219

Spilled Records=5847846

Shuffled Maps =2

Failed Shuffles=0

Merged Map outputs=2

GC time elapsed (ms)=3916254

CPU time spent (ms)=6878470

Physical memory (bytes) snapshot=695148544

Virtual memory (bytes) snapshot=2548867072

Total committed heap usage (bytes)=466157568

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=236907275

File Output Format Counters

Bytes Written=23392218

15/03/13 09:52:13 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

15/03/13 09:52:14 INFO input.FileInputFormat: Total input paths to process : 1

15/03/13 09:52:14 INFO mapreduce.JobSubmitter: number of splits:1

15/03/13 09:52:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1426159947607\_0021

15/03/13 09:52:14 INFO impl.YarnClientImpl: Submitted application application\_1426159947607\_0021

15/03/13 09:52:14 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application\_1426159947607\_0021/

15/03/13 09:52:14 INFO mapreduce.Job: Running job: job\_1426159947607\_0021

15/03/13 09:52:24 INFO mapreduce.Job: Job job\_1426159947607\_0021 running in uber mode : false

15/03/13 09:52:24 INFO mapreduce.Job: map 0% reduce 0%

15/03/13 09:52:32 INFO mapreduce.Job: map 100% reduce 0%

15/03/13 09:52:40 INFO mapreduce.Job: map 100% reduce 100%

15/03/13 09:52:41 INFO mapreduce.Job: Job job\_1426159947607\_0021 completed successfully

15/03/13 09:52:42 INFO mapreduce.Job: Counters: 49

File System Counters

FILE: Number of bytes read=2153

FILE: Number of bytes written=217443

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=23392337

HDFS: Number of bytes written=952

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=1

Launched reduce tasks=1

Data-local map tasks=1

Total time spent by all maps in occupied slots (ms)=6547

Total time spent by all reduces in occupied slots (ms)=6103

Total time spent by all map tasks (ms)=6547

Total time spent by all reduce tasks (ms)=6103

Total vcore-seconds taken by all map tasks=6547

Total vcore-seconds taken by all reduce tasks=6103

Total megabyte-seconds taken by all map tasks=6704128

Total megabyte-seconds taken by all reduce tasks=6249472

Map-Reduce Framework

Map input records=219

Map output records=219

Map output bytes=1709

Map output materialized bytes=2153

Input split bytes=119

Combine input records=0

Combine output records=0

Reduce input groups=145

Reduce shuffle bytes=2153

Reduce input records=219

Reduce output records=145

Spilled Records=438

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=488

CPU time spent (ms)=2310

Physical memory (bytes) snapshot=434028544

Virtual memory (bytes) snapshot=1701339136

Total committed heap usage (bytes)=263323648

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=23392218

File Output Format Counters

Bytes Written=952

hadoop fs -cat citations\_output\_problem3.2/part-r-00000 | head

1 23

2 13

3 9

4 8

5 1

6 5

7 2

8 5

9 3

10 2

[joe@localhost ~]$

**Problem 4**

Configuration conf = getConf();

Path in = **new** Path(args[0]);

Path out = **new** Path(args[1]);

Path temp = **new** Path("chain-temp");

//Job 1

JobConf job = **new** JobConf(conf, CountryToCitationMRv1.**class**);

FileInputFormat.*setInputPaths*(job, in);

FileOutputFormat.*setOutputPath*(job, temp);

job.setJobName("CountryToCitationMRv1");

job.setMapperClass(CountryToCitationMapper.**class**);

job.setReducerClass(CountryToCitationReducer.**class**);

job.setInputFormat(KeyValueTextInputFormat.**class**);

job.setMapOutputKeyClass(Text.**class**);

job.setMapOutputValueClass(Text.**class**);

job.setOutputFormat(TextOutputFormat.**class**);

job.setOutputKeyClass(Text.**class**);

//job.setOutputValueClass(Text.class);

job.setOutputValueClass(Text.**class**);

job.set("key.value.separator.in.input.line", ",");

RunningJob runningJob = JobClient.*runJob*(job);

**boolean** success = runningJob.isSuccessful();

**if** (success){

//Job 2

JobConf job2 = **new** JobConf(conf, CountryToCitationMRv1.**class**);

FileInputFormat.*setInputPaths*(job2, temp);

FileOutputFormat.*setOutputPath*(job2, out);

job2.setJobName("InverterCounterMRv1");

job2.setMapperClass(InverterCounterMapper.**class**);

job2.setReducerClass(InverterCounterReducer.**class**);

job2.setInputFormat(KeyValueTextInputFormat.**class**);

job2.setMapOutputKeyClass(IntWritable.**class**);

job2.setMapOutputValueClass(Text.**class**);

job2.setOutputFormat(TextOutputFormat.**class**);

job2.setOutputKeyClass(IntWritable.**class**);

//job.setOutputValueClass(Text.class);

job2.setOutputValueClass(IntWritable.**class**);

job2.set("key.value.separator.in.input.line", ",");

runningJob = JobClient.*runJob*(job2);

success = runningJob.isSuccessful();

**if**(success)

**return** 0;

**else**

**return** 1;

}

**else**

**return** 1;

}

**public** **static** **void** main(String[] args) **throws** Exception {

**int** res = ToolRunner.*run*(**new** Configuration(), **new** CountryToCitationMRv1(), args);

System.*exit*(res);

}

**public** **static** **class** CountryToCitationMapper **extends** Mapper<Object, Text, Text, Text> {

**public** **void** map(Object key, Text value, Context context

) **throws** IOException, InterruptedException {

String set[] = value.toString().split(",");

String country=set[4].replace("\"", "");

**if**(country.equals("US")){country=set[4].replace("\"", "")+"-"+set[5].replace("\"", "");}

String citation = set[0];

context.write(**new** Text(country), **new** Text(citation));

}

}

**public** **static** **class** CountryToCitationReducer

**extends** Reducer<Text,Text,Text,Text> {

**public** **void** reduce(Text key, Iterable<Text> values,

Context context

) **throws** IOException, InterruptedException {

String csv = "";

**while** (values.iterator().hasNext()) {

**if** (csv.length() > 0) csv += ",";

csv += values.iterator().next().toString();

}

context.write(key, **new** Text(csv));

}

}

**public** **static** **class** InverterCounterMapper **extends** Mapper<Text, Text, IntWritable, Text> {

**public** **void** map(Text key, Text value, Context context)

**throws** IOException, InterruptedException {

String citationList[]=value.toString().split(",");

context.write(**new** IntWritable(citationList.length), key);

}

}

**public** **static** **class** HistogramReducer **extends** Reducer<IntWritable, Text, IntWritable, IntWritable> {

**public** **void** reduce(IntWritable key, Iterable<Text> values, Context context)

**throws** IOException, InterruptedException {

**int** count = 0;

**for** (Text val : values) { // Iterable allows for looping

count++;

}

context.write(key, **new** IntWritable(count));

}

}

**Console output**

hadoop jar /mnt/hgfs/DilmiShared/CountryToCitationMRv1.jar edu.hu.bigdata.CountryToCitationMRv1 input/apat63\_99.txt output\_problem4

15/03/13 10:32:26 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

15/03/13 10:32:27 INFO mapred.FileInputFormat: Total input paths to process : 1

15/03/13 10:32:27 INFO mapred.JobClient: Running job: job\_201503130856\_0024

15/03/13 10:32:28 INFO mapred.JobClient: map 0% reduce 0%

15/03/13 10:32:44 INFO mapred.JobClient: map 34% reduce 0%

15/03/13 10:32:47 INFO mapred.JobClient: map 66% reduce 0%

15/03/13 10:32:51 INFO mapred.JobClient: map 95% reduce 0%

15/03/13 10:32:53 INFO mapred.JobClient: map 100% reduce 0%

15/03/13 10:33:01 INFO mapred.JobClient: map 100% reduce 67%

15/03/13 10:33:19 INFO mapred.JobClient: map 100% reduce 68%

15/03/13 10:39:29 INFO mapred.JobClient: map 100% reduce 70%

15/03/13 10:40:11 INFO mapred.JobClient: map 100% reduce 71%

15/03/13 10:41:12 INFO mapred.JobClient: map 100% reduce 72%

15/03/13 10:41:18 INFO mapred.JobClient: map 100% reduce 73%

15/03/13 11:36:47 INFO mapred.JobClient: map 100% reduce 77%

15/03/13 11:36:56 INFO mapred.JobClient: map 100% reduce 78%

15/03/13 11:50:08 INFO mapred.JobClient: map 100% reduce 82%

15/03/13 11:50:29 INFO mapred.JobClient: map 100% reduce 83%

15/03/13 11:50:44 INFO mapred.JobClient: map 100% reduce 84%

15/03/13 11:52:17 INFO mapred.JobClient: map 100% reduce 85%

15/03/13 11:52:26 INFO mapred.JobClient: map 100% reduce 86%

15/03/13 11:53:03 INFO mapred.JobClient: map 100% reduce 87%

15/03/13 11:53:09 INFO mapred.JobClient: map 100% reduce 88%

15/03/13 11:54:09 INFO mapred.JobClient: map 100% reduce 89%

15/03/13 11:54:24 INFO mapred.JobClient: map 100% reduce 90%

15/03/13 11:56:07 INFO mapred.JobClient: map 100% reduce 92%

15/03/13 12:00:01 INFO mapred.JobClient: map 100% reduce 94%

15/03/13 12:01:05 INFO mapred.JobClient: map 100% reduce 95%

15/03/13 12:01:11 INFO mapred.JobClient: map 100% reduce 96%

15/03/13 12:02:29 INFO mapred.JobClient: map 100% reduce 97%

15/03/13 12:03:26 INFO mapred.JobClient: map 100% reduce 99%

15/03/13 12:03:41 INFO mapred.JobClient: map 100% reduce 100%

15/03/13 12:03:45 INFO mapred.JobClient: Job complete: job\_201503130856\_0024

15/03/13 12:03:45 INFO mapred.JobClient: Counters: 33

15/03/13 12:03:45 INFO mapred.JobClient: File System Counters

15/03/13 12:03:45 INFO mapred.JobClient: FILE: Number of bytes read=86732022

15/03/13 12:03:45 INFO mapred.JobClient: FILE: Number of bytes written=130829261

15/03/13 12:03:45 INFO mapred.JobClient: FILE: Number of read operations=0

15/03/13 12:03:45 INFO mapred.JobClient: FILE: Number of large read operations=0

15/03/13 12:03:45 INFO mapred.JobClient: FILE: Number of write operations=0

15/03/13 12:03:45 INFO mapred.JobClient: HDFS: Number of bytes read=236904020

15/03/13 12:03:45 INFO mapred.JobClient: HDFS: Number of bytes written=23392218

15/03/13 12:03:45 INFO mapred.JobClient: HDFS: Number of read operations=5

15/03/13 12:03:45 INFO mapred.JobClient: HDFS: Number of large read operations=0

15/03/13 12:03:45 INFO mapred.JobClient: HDFS: Number of write operations=2

15/03/13 12:03:45 INFO mapred.JobClient: Job Counters

15/03/13 12:03:45 INFO mapred.JobClient: Launched map tasks=2

15/03/13 12:03:45 INFO mapred.JobClient: Launched reduce tasks=1

15/03/13 12:03:45 INFO mapred.JobClient: Data-local map tasks=2

15/03/13 12:03:45 INFO mapred.JobClient: Total time spent by all maps in occupied slots (ms)=46275

15/03/13 12:03:45 INFO mapred.JobClient: Total time spent by all reduces in occupied slots (ms)=5450733

15/03/13 12:03:45 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0

15/03/13 12:03:45 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0

15/03/13 12:03:45 INFO mapred.JobClient: Map-Reduce Framework

15/03/13 12:03:45 INFO mapred.JobClient: Map input records=2923923

15/03/13 12:03:45 INFO mapred.JobClient: Map output records=2923923

15/03/13 12:03:45 INFO mapred.JobClient: Map output bytes=37518126

15/03/13 12:03:45 INFO mapred.JobClient: Input split bytes=206

15/03/13 12:03:45 INFO mapred.JobClient: Combine input records=0

15/03/13 12:03:45 INFO mapred.JobClient: Combine output records=0

15/03/13 12:03:45 INFO mapred.JobClient: Reduce input groups=219

15/03/13 12:03:45 INFO mapred.JobClient: Reduce shuffle bytes=43365984

15/03/13 12:03:45 INFO mapred.JobClient: Reduce input records=2923923

15/03/13 12:03:45 INFO mapred.JobClient: Reduce output records=219

15/03/13 12:03:45 INFO mapred.JobClient: Spilled Records=8771769

15/03/13 12:03:45 INFO mapred.JobClient: CPU time spent (ms)=5374660

15/03/13 12:03:45 INFO mapred.JobClient: Physical memory (bytes) snapshot=588517376

15/03/13 12:03:45 INFO mapred.JobClient: Virtual memory (bytes) snapshot=2224742400

15/03/13 12:03:45 INFO mapred.JobClient: Total committed heap usage (bytes)=378077184

15/03/13 12:03:45 INFO mapred.JobClient: org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter

15/03/13 12:03:45 INFO mapred.JobClient: BYTES\_READ=236903179

15/03/13 12:03:45 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

15/03/13 12:03:45 INFO mapred.FileInputFormat: Total input paths to process : 1

15/03/13 12:03:45 INFO mapred.JobClient: Running job: job\_201503130856\_0025

15/03/13 12:03:46 INFO mapred.JobClient: map 0% reduce 0%

15/03/13 12:03:59 INFO mapred.JobClient: map 100% reduce 0%

15/03/13 12:04:04 INFO mapred.JobClient: map 100% reduce 100%

15/03/13 12:04:06 INFO mapred.JobClient: Job complete: job\_201503130856\_0025

15/03/13 12:04:06 INFO mapred.JobClient: Counters: 33

15/03/13 12:04:06 INFO mapred.JobClient: File System Counters

15/03/13 12:04:06 INFO mapred.JobClient: FILE: Number of bytes read=3906

15/03/13 12:04:06 INFO mapred.JobClient: FILE: Number of bytes written=737402

15/03/13 12:04:06 INFO mapred.JobClient: FILE: Number of read operations=0

15/03/13 12:04:06 INFO mapred.JobClient: FILE: Number of large read operations=0

15/03/13 12:04:06 INFO mapred.JobClient: FILE: Number of write operations=0

15/03/13 12:04:06 INFO mapred.JobClient: HDFS: Number of bytes read=23754943

15/03/13 12:04:06 INFO mapred.JobClient: HDFS: Number of bytes written=946

15/03/13 12:04:06 INFO mapred.JobClient: HDFS: Number of read operations=5

15/03/13 12:04:06 INFO mapred.JobClient: HDFS: Number of large read operations=0

15/03/13 12:04:06 INFO mapred.JobClient: HDFS: Number of write operations=2

15/03/13 12:04:06 INFO mapred.JobClient: Job Counters

15/03/13 12:04:06 INFO mapred.JobClient: Launched map tasks=2

15/03/13 12:04:06 INFO mapred.JobClient: Launched reduce tasks=1

15/03/13 12:04:06 INFO mapred.JobClient: Data-local map tasks=2

15/03/13 12:04:06 INFO mapred.JobClient: Total time spent by all maps in occupied slots (ms)=23055

15/03/13 12:04:06 INFO mapred.JobClient: Total time spent by all reduces in occupied slots (ms)=4638

15/03/13 12:04:06 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0

15/03/13 12:04:06 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0

15/03/13 12:04:06 INFO mapred.JobClient: Map-Reduce Framework

15/03/13 12:04:06 INFO mapred.JobClient: Map input records=219

15/03/13 12:04:06 INFO mapred.JobClient: Map output records=219

15/03/13 12:04:06 INFO mapred.JobClient: Map output bytes=3462

15/03/13 12:04:06 INFO mapred.JobClient: Input split bytes=210

15/03/13 12:04:06 INFO mapred.JobClient: Combine input records=0

15/03/13 12:04:06 INFO mapred.JobClient: Combine output records=0

15/03/13 12:04:06 INFO mapred.JobClient: Reduce input groups=144

15/03/13 12:04:06 INFO mapred.JobClient: Reduce shuffle bytes=3912

15/03/13 12:04:06 INFO mapred.JobClient: Reduce input records=219

15/03/13 12:04:06 INFO mapred.JobClient: Reduce output records=144

15/03/13 12:04:06 INFO mapred.JobClient: Spilled Records=438

15/03/13 12:04:06 INFO mapred.JobClient: CPU time spent (ms)=2330

15/03/13 12:04:06 INFO mapred.JobClient: Physical memory (bytes) snapshot=686694400

15/03/13 12:04:06 INFO mapred.JobClient: Virtual memory (bytes) snapshot=2220208128

15/03/13 12:04:06 INFO mapred.JobClient: Total committed heap usage (bytes)=466092032

15/03/13 12:04:06 INFO mapred.JobClient: org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter

15/03/13 12:04:06 INFO mapred.JobClient: BYTES\_READ=23392218

**Problem 5**

//map

**public** **void** map(Text key, Text value,

OutputCollector<Text, IntWritable> output,

Reporter reporter) **throws** IOException {

**try**{

**int** count = Integer.*parseInt*(key.toString());

**int** bucket = (**int**)(count-(count%20))/20;

output.collect(**new** Text(String.*valueOf*((bucket\*20)+1)+"-"+String.*valueOf*(bucket\*20+20)), **new** IntWritable(Integer.*parseInt*(value.toString())));

}

**catch**(NumberFormatException e){

//skip

}

//reduce

**public** **void** reduce(Text key, Iterator<IntWritable> values,

OutputCollector<Text, Text>output,

Reporter reporter) **throws** IOException {

**int** no\_patents\_cited = 0;

**while** (values.hasNext()) {

no\_patents\_cited += values.next().get();

}

**int** numOfStars=(**int**) (Math.*log10*(no\_patents\_cited+1)\*4);

String stars="";

**for**(**int** i=0;i<numOfStars;i++){

stars+="\*";

}

output.collect(key, **new** Text(stars));

}

**Console output**

[joe@localhost DilmiShared]$ hadoop jar /mnt/hgfs/DilmiShared/GraphicalHistogram.jar edu.hu.bigdata.GraphicalHistogram output5.1/part-00000 output\_problem5.13

15/03/13 16:42:38 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

15/03/13 16:42:38 INFO mapred.FileInputFormat: Total input paths to process : 1

15/03/13 16:42:39 INFO mapred.JobClient: Running job: job\_201503052235\_0021

15/03/13 16:42:40 INFO mapred.JobClient: map 0% reduce 0%

15/03/13 16:42:55 INFO mapred.JobClient: map 100% reduce 0%

15/03/13 16:43:01 INFO mapred.JobClient: map 100% reduce 100%

15/03/13 16:43:05 INFO mapred.JobClient: Job complete: job\_201503052235\_0021

15/03/13 16:43:05 INFO mapred.JobClient: Counters: 33

15/03/13 16:43:05 INFO mapred.JobClient: File System Counters

15/03/13 16:43:05 INFO mapred.JobClient: FILE: Number of bytes read=3421

15/03/13 16:43:05 INFO mapred.JobClient: FILE: Number of bytes written=736543

15/03/13 16:43:05 INFO mapred.JobClient: FILE: Number of read operations=0

15/03/13 16:43:05 INFO mapred.JobClient: FILE: Number of large read operations=0

15/03/13 16:43:05 INFO mapred.JobClient: FILE: Number of write operations=0

15/03/13 16:43:05 INFO mapred.JobClient: HDFS: Number of bytes read=2790

15/03/13 16:43:05 INFO mapred.JobClient: HDFS: Number of bytes written=380

15/03/13 16:43:05 INFO mapred.JobClient: HDFS: Number of read operations=5

15/03/13 16:43:05 INFO mapred.JobClient: HDFS: Number of large read operations=0

15/03/13 16:43:05 INFO mapred.JobClient: HDFS: Number of write operations=2

15/03/13 16:43:05 INFO mapred.JobClient: Job Counters

15/03/13 16:43:05 INFO mapred.JobClient: Launched map tasks=2

15/03/13 16:43:05 INFO mapred.JobClient: Launched reduce tasks=1

15/03/13 16:43:05 INFO mapred.JobClient: Data-local map tasks=2

15/03/13 16:43:05 INFO mapred.JobClient: Total time spent by all maps in occupied slots (ms)=27706

15/03/13 16:43:05 INFO mapred.JobClient: Total time spent by all reduces in occupied slots (ms)=6298

15/03/13 16:43:05 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0

15/03/13 16:43:05 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0

15/03/13 16:43:05 INFO mapred.JobClient: Map-Reduce Framework

15/03/13 16:43:05 INFO mapred.JobClient: Map input records=258

15/03/13 16:43:05 INFO mapred.JobClient: Map output records=258

15/03/13 16:43:05 INFO mapred.JobClient: Map output bytes=2899

15/03/13 16:43:05 INFO mapred.JobClient: Input split bytes=208

15/03/13 16:43:05 INFO mapred.JobClient: Combine input records=0

15/03/13 16:43:05 INFO mapred.JobClient: Combine output records=0

15/03/13 16:43:05 INFO mapred.JobClient: Reduce input groups=26

15/03/13 16:43:05 INFO mapred.JobClient: Reduce shuffle bytes=3427

15/03/13 16:43:05 INFO mapred.JobClient: Reduce input records=258

15/03/13 16:43:05 INFO mapred.JobClient: Reduce output records=26

15/03/13 16:43:05 INFO mapred.JobClient: Spilled Records=516

15/03/13 16:43:05 INFO mapred.JobClient: CPU time spent (ms)=1660

15/03/13 16:43:05 INFO mapred.JobClient: Physical memory (bytes) snapshot=531570688

15/03/13 16:43:05 INFO mapred.JobClient: Virtual memory (bytes) snapshot=1923522560

15/03/13 16:43:05 INFO mapred.JobClient: Total committed heap usage (bytes)=378011648

15/03/13 16:43:05 INFO mapred.JobClient: org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter

15/03/13 16:43:05 INFO mapred.JobClient: BYTES\_READ=1721

[joe@localhost DilmiShared]$ hadoop fs -cat output\_problem5.13/part-00000 | head1-20 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

101-120 \*\*\*\*\*\*\*\*\*\*

121-140 \*\*\*\*\*\*\*\*\*

141-160 \*\*\*\*\*\*\*\*

161-180 \*\*\*\*\*\*\*

181-200 \*\*\*\*\*\*

201-220 \*\*\*\*\*

21-40 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

221-240 \*\*\*\*\*

241-260 \*\*\*\*

Formatted output

**1-20 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**21-40 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**41-60 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**61-80 \*\*\*\*\*\*\*\*\*\*\*\*\***

**81-100 \*\*\*\*\*\*\*\*\*\*\*\***

**101-120 \*\*\*\*\*\*\*\*\*\***

**121-140 \*\*\*\*\*\*\*\*\***

**141-160 \*\*\*\*\*\*\*\***

**161-180 \*\*\*\*\*\*\***

**181-200 \*\*\*\*\*\***

**201-220 \*\*\*\*\***

**221-240 \*\*\*\*\***

**241-260 \*\*\*\***

**261-280 \*\*\***

**281-300 \*\*\***

**301-320 \***

**321-340 \***

**361-380 \***

**381-400 \***

**401-420 \***

**601-620 \***

**621-640 \***

**641-660 \***

**661-680 \***

**701-720 \***

**761-780 \***

**Problem 6**

**Input file – list of values N=10000**

|  |  |
| --- | --- |
| 1 | 1 |
| 2 | 1.0009 |
| 3 | 1.0018 |
| 4 | 1.0027 |
| 5 | 1.0036 |
| 6 | 1.0045 |
| 7 | 1.005401 |
| 8 | 1.006301 |
| 9 | 1.007201 |
| 10 | 1.008101 |
| //map |  |

**public** **void** map(Text key, Text value,

OutputCollector<Text, Text> output,

Reporter reporter) **throws** IOException {

**double** deltaX = (10.0-1.0)/(10000-1);

**try**{

**double** x=Double.*parseDouble*(value.toString());

**double** fx = (1.0/x)\*deltaX;

output.collect(**new** Text("Reinmann\_sum"), **new** Text(String.*valueOf*(fx)));

}

**catch**(NumberFormatException e){

//output.collect(new Text("Reinmann\_sum"), new Text("0"));

}

}

//reduce

**public** **void** reduce(Text key, Iterator<Text> values,

OutputCollector<Text, Text>output,

Reporter reporter) **throws** IOException {

**double** reinmann\_sum=0.0;

**while** (values.hasNext()) {

reinmann\_sum+= Double.*parseDouble*(values.next().toString());

}

output.collect(key, **new** Text(String.*valueOf*(reinmann\_sum)));

}

**Console output**

[joe@localhost DilmiShared]$ hadoop jar /mnt/hgfs/DilmiShared/ReinmannSumCalculator.jar edu.hu.bigdata.ReinmannSumCalculator input/ReinmannSumXvalues.txt output\_problem6.9

15/03/13 18:15:51 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

15/03/13 18:15:52 INFO mapred.FileInputFormat: Total input paths to process : 1

15/03/13 18:15:52 INFO mapred.JobClient: Running job: job\_201503052235\_0031

15/03/13 18:15:53 INFO mapred.JobClient: map 0% reduce 0%

15/03/13 18:16:12 INFO mapred.JobClient: map 100% reduce 0%

15/03/13 18:16:19 INFO mapred.JobClient: map 100% reduce 100%

15/03/13 18:16:23 INFO mapred.JobClient: Job complete: job\_201503052235\_0031

15/03/13 18:16:23 INFO mapred.JobClient: Counters: 33

15/03/13 18:16:23 INFO mapred.JobClient: File System Counters

15/03/13 18:16:23 INFO mapred.JobClient: FILE: Number of bytes read=364622

15/03/13 18:16:23 INFO mapred.JobClient: FILE: Number of bytes written=1458972

15/03/13 18:16:23 INFO mapred.JobClient: FILE: Number of read operations=0

15/03/13 18:16:23 INFO mapred.JobClient: FILE: Number of large read operations=0

15/03/13 18:16:23 INFO mapred.JobClient: FILE: Number of write operations=0

15/03/13 18:16:23 INFO mapred.JobClient: HDFS: Number of bytes read=243221

15/03/13 18:16:23 INFO mapred.JobClient: HDFS: Number of bytes written=31

15/03/13 18:16:23 INFO mapred.JobClient: HDFS: Number of read operations=5

15/03/13 18:16:23 INFO mapred.JobClient: HDFS: Number of large read operations=0

15/03/13 18:16:23 INFO mapred.JobClient: HDFS: Number of write operations=2

15/03/13 18:16:23 INFO mapred.JobClient: Job Counters

15/03/13 18:16:23 INFO mapred.JobClient: Launched map tasks=2

15/03/13 18:16:23 INFO mapred.JobClient: Launched reduce tasks=1

15/03/13 18:16:23 INFO mapred.JobClient: Data-local map tasks=2

15/03/13 18:16:23 INFO mapred.JobClient: Total time spent by all maps in occupied slots (ms)=35124

15/03/13 18:16:23 INFO mapred.JobClient: Total time spent by all reduces in occupied slots (ms)=7428

15/03/13 18:16:23 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0

15/03/13 18:16:23 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0

15/03/13 18:16:23 INFO mapred.JobClient: Map-Reduce Framework

15/03/13 18:16:23 INFO mapred.JobClient: Map input records=10000

15/03/13 18:16:23 INFO mapred.JobClient: Map output records=10000

15/03/13 18:16:23 INFO mapred.JobClient: Map output bytes=344616

15/03/13 18:16:23 INFO mapred.JobClient: Input split bytes=224

15/03/13 18:16:23 INFO mapred.JobClient: Combine input records=0

15/03/13 18:16:23 INFO mapred.JobClient: Combine output records=0

15/03/13 18:16:23 INFO mapred.JobClient: Reduce input groups=1

15/03/13 18:16:23 INFO mapred.JobClient: Reduce shuffle bytes=364628

15/03/13 18:16:23 INFO mapred.JobClient: Reduce input records=10000

15/03/13 18:16:23 INFO mapred.JobClient: Reduce output records=1

15/03/13 18:16:23 INFO mapred.JobClient: Spilled Records=20000

15/03/13 18:16:23 INFO mapred.JobClient: CPU time spent (ms)=4190

15/03/13 18:16:23 INFO mapred.JobClient: Physical memory (bytes) snapshot=534528000

15/03/13 18:16:23 INFO mapred.JobClient: Virtual memory (bytes) snapshot=1923813376

15/03/13 18:16:23 INFO mapred.JobClient: Total committed heap usage (bytes)=378011648

15/03/13 18:16:23 INFO mapred.JobClient: org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter

15/03/13 18:16:23 INFO mapred.JobClient: BYTES\_READ=240234

[joe@localhost DilmiShared]$ hadoop fs -cat output\_problem6.9/part-00000 | head

Reinmann\_sum 2.303080209337345