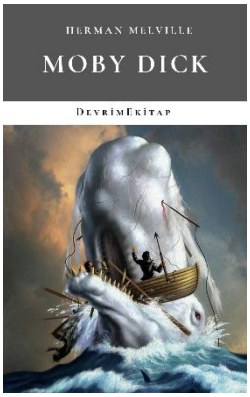
Large Scale Data Management
Athens University of Economics and Business
M.Sc. in Data Science
Programming Project I

Student: Vasileios Dimopoulos (f3352311) Date: 16/02/2024

PART I:

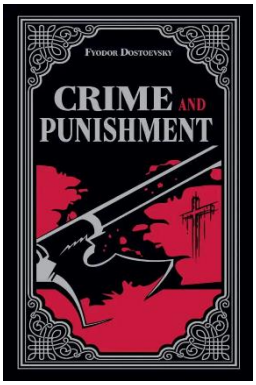
Input file:

Title: **Moby Dick; Or, The Whale**
Author: Melville, Herman
Language: English
Source: https://www.gutenberg.org/ebooks/2701
Lines: 22314
Size: 1.247 KB



Input file:

Title: **Crime and Punishment**
Author: Dostoyevsky, Fyodor
Language: English
Source: https://www.gutenberg.org/ebooks/2554
Lines: 22444
Size: 1.174 KB



Output Logs:

2024-02-01 14:52:56,535 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.3:8032

2024-02-01 14:52:56,978 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.4:10200

2024-02-01 14:52:57,565 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

2024-02-01 14:52:57,619 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1706796164921_0003

2024-02-01 14:52:57,904 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-01 14:52:58,591 INFO input.FileInputFormat: Total input files to process : 1

2024-02-01 14:52:58,689 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-01 14:52:58,729 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-01 14:52:58,766 INFO mapreduce.JobSubmitter: number of splits:1

2024-02-01 14:52:59,134 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-01 14:52:59,168 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1706796164921_0003

2024-02-01 14:52:59,169 INFO mapreduce.JobSubmitter: Executing with tokens: []

2024-02-01 14:52:59,636 INFO conf.Configuration: resource-types.xml not found

2024-02-01 14:52:59,640 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2024-02-01 14:53:00,279 INFO impl.YarnClientImpl: Submitted application application_1706796164921_0003

2024-02-01 14:53:00,518 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1706796164921_0003/

2024-02-01 14:53:00,519 INFO mapreduce.Job: Running job: job_1706796164921_0003

2024-02-01 14:53:16,269 INFO mapreduce.Job: Job job_1706796164921_0003 running in uber mode : false

2024-02-01 14:53:16,274 INFO mapreduce.Job: map 0% reduce 0%

2024-02-01 14:53:28,680 INFO mapreduce.Job: map 100% reduce 0%

2024-02-01 14:53:38,845 INFO mapreduce.Job: map 100% reduce 100%

2024-02-01 14:53:39,887 INFO mapreduce.Job: Job job_1706796164921_0003 completed successfully

2024-02-01 14:53:40,085 INFO mapreduce.Job: Counters: 54

File System Counters

FILE: Number of bytes read=104555

FILE: Number of bytes written=667425

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=1201855

HDFS: Number of bytes written=249695

HDFS: Number of read operations=8

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

HDFS: Number of bytes read erasure-coded=0

Job Counters

Launched map tasks=1

Launched reduce tasks=1

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=34220

Total time spent by all reduces in occupied slots (ms)=54344

Total time spent by all map tasks (ms)=8555

Total time spent by all reduce tasks (ms)=6793

Total vcore-milliseconds taken by all map tasks=8555

Total vcore-milliseconds taken by all reduce tasks=6793

Total megabyte-milliseconds taken by all map tasks=35041280

Total megabyte-milliseconds taken by all reduce tasks=55648256

Map-Reduce Framework

Map input records=22446

Map output records=210692

Map output bytes=2021641

Map output materialized bytes=104547

Input split bytes=126

Combine input records=210692

Combine output records=22357

Reduce input groups=22357

Reduce shuffle bytes=104547

Reduce input records=22357

Reduce output records=22357

Spilled Records=44714

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=213

CPU time spent (ms)=4800

Physical memory (bytes) snapshot=358342656

Virtual memory (bytes) snapshot=13150445568

Total committed heap usage (bytes)=230821888

Peak Map Physical memory (bytes)=237207552

Peak Map Virtual memory (bytes)=4955017216

Peak Reduce Physical memory (bytes)=121135104

Peak Reduce Virtual memory (bytes)=8195428352

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=1201729

File Output Format Counters

Bytes Written=249695

PART II:

Driver.java:

```
package gr.aueb.panagiotisl.mapreduce.wordcount;

import org.apache.hadoop.conf.Configuration;
```

```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

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

        // Set the Hadoop home directory to resolve any compatibility
issues
        System.setProperty("hadoop.home.dir", "/");

        // instantiate a configuration
        Configuration configuration = new Configuration();

        // Instantiate a MapReduce job
        Job job = Job.getInstance(configuration, "Song Danceability");

        // Set the main class containing the MapReduce job
configuration
        job.setJarByClass(SongDanceability.class);

        // Set the Mapper and Reducer classes
        job.setMapperClass(SongDanceability.DanceabilityMapper.class);
        job.setReducerClass(SongDanceability.DanceabilityReducer.class)
;

        // Set the output key and value classes
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);


        // Set input and output paths
        FileInputFormat.addInputPath(job, new
Path("/user/hdfs/input/universal_top_spotify_songs.csv"));
        FileOutputFormat.setOutputPath(job, new
Path("/user/hdfs/output/"));

        // Wait for the job to complete and exit the program
accordingly
        System.exit(job.waitForCompletion(true)? 0 : 1);
    }
}
```

SongDanceability.java:

```
package gr.aueb.panagiotisl.mapreduce.wordcount;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import java.util.logging.Logger;
import java.io.IOException;


public class SongDanceability {

    // Mapper class for Danceability
    public static class DanceabilityMapper extends Mapper<LongWritable,
Text, Text, Text> {

        private static final Logger log =
Logger.getLogger(DanceabilityMapper.class.getName());
        private Text outputKey = new Text();
        private Text outputValue = new Text();

        @Override
        public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {

            // Skip header line
            if (key.get() == 0) {
                return;
            }

            // Split CSV line into columns
            String[] columns =
value.toString().split(",(?=(?:[^\"]*\"[^\"]*\")*[^\"]*$)", -1);

            // Extract relevant columns
            String country = columns[6].trim().replaceAll("\"",""); //
Country column
            if (country.equals(null) || country.isEmpty()){
                return;
            }

            String date = columns[7].trim().replaceAll("\"",""); //
Date column
            if (date.equals(null) || date.isEmpty()){
                return;
            }
```

```
            // Format date to year-month
            String[] date_parts = date.split("-");
            String formatted_date = date_parts[0] + "-" +
date_parts[1];

            String song = columns[1].trim().replaceAll("\"",""); //
Song column
            if (song.equals(null) || song.isEmpty()){
                return;
            }

            String danceability =
columns[13].trim().replaceAll("\"",""); // Danceability column
            if (danceability.equals(null) || danceability.isEmpty()){
                return;
            }


            // Set output key-value pairs
            outputKey.set(country + ": " + formatted_date);
            outputValue.set(song + ";" + danceability);

            // Log output value for debugging
            log.info(outputValue.toString());

            // Emit key-value pair
            context.write(outputKey, outputValue);
        }
    }

    // Reducer class for Danceability
    public static class DanceabilityReducer extends Reducer<Text, Text,
Text, Text> {

        private static final Logger log =
Logger.getLogger(DanceabilityReducer.class.getName());

        @Override
        public void reduce(Text key, Iterable<Text> values, Context
context) throws IOException, InterruptedException {
            // Initialize variables for calculating average
danceability and finding the top song
            float sum = 0;
            float max = 0;
            int count = 0;
            String top_song = null;

            // Iterate through values for the same key
```

```
            for (Text value : values) {
                String value_s = value.toString();
                log.info(value_s.toString());

                // Split value into song and danceability parts
                String[] parts = value_s.split(";");
                String song = parts[0];
                String danceability = parts[1];
                float danceability_i =
Float.parseFloat(danceability.toString());

                // Update sum, count, and find the top song with
maximum danceability
                sum += danceability_i;
                count++;
                if (max < danceability_i) {
                    max = danceability_i;
                    top_song = song;
                }

            }

            // Calculate average danceability
            float avg = sum/count;

            // Output (key, value) pair with top song and its
danceability, along with average danceability
            context.write(key, new Text(top_song + ": " + max + ", avg:
" + avg));
        }
    }
}
```

Logs:

2024-02-11 13:19:19,326 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.3:8032

2024-02-11 13:19:19,820 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.4:10200

2024-02-11 13:19:20,409 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

2024-02-11 13:19:20,449 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1707657500865_0001

2024-02-11 13:19:20,877 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-11 13:19:21,983 INFO input.FileInputFormat: Total input files to process : 1

2024-02-11 13:19:22,104 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-11 13:19:22,207 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-11 13:19:22,232 INFO mapreduce.JobSubmitter: number of splits:1

2024-02-11 13:19:22,680 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

2024-02-11 13:19:22,744 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707657500865_0001

2024-02-11 13:19:22,747 INFO mapreduce.JobSubmitter: Executing with tokens: []

2024-02-11 13:19:23,289 INFO conf.Configuration: resource-types.xml not found

2024-02-11 13:19:23,291 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2024-02-11 13:19:24,163 INFO impl.YarnClientImpl: Submitted application application_1707657500865_0001

2024-02-11 13:19:24,418 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1707657500865_0001/

2024-02-11 13:19:24,426 INFO mapreduce.Job: Running job: job_1707657500865_0001

2024-02-11 13:19:49,462 INFO mapreduce.Job: Job job_1707657500865_0001 running in uber mode : false

2024-02-11 13:19:49,469 INFO mapreduce.Job: map 0% reduce 0%

2024-02-11 13:20:12,014 INFO mapreduce.Job: map 6% reduce 0%

2024-02-11 13:20:18,139 INFO mapreduce.Job: map 15% reduce 0%

2024-02-11 13:20:24,253 INFO mapreduce.Job: map 24% reduce 0%

2024-02-11 13:20:30,374 INFO mapreduce.Job: map 34% reduce 0%

2024-02-11 13:20:36,482 INFO mapreduce.Job: map 41% reduce 0%

2024-02-11 13:20:42,588 INFO mapreduce.Job: map 51% reduce 0%

2024-02-11 13:20:48,650 INFO mapreduce.Job: map 60% reduce 0%

2024-02-11 13:20:54,757 INFO mapreduce.Job: map 100% reduce 0%

2024-02-11 13:21:15,033 INFO mapreduce.Job: map 100% reduce 81%

2024-02-11 13:21:21,168 INFO mapreduce.Job: map 100% reduce 94%

2024-02-11 13:21:24,218 INFO mapreduce.Job: map 100% reduce 100%

2024-02-11 13:21:25,242 INFO mapreduce.Job: Job job_1707657500865_0001 completed successfully

2024-02-11 13:21:25,441 INFO mapreduce.Job: Counters: 54

File System Counters

FILE: Number of bytes read=935616

FILE: Number of bytes written=2329177

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=85676844

HDFS: Number of bytes written=13879

HDFS: Number of read operations=8

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

HDFS: Number of bytes read erasure-coded=0

Job Counters

Launched map tasks=1

Launched reduce tasks=1

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=245792

Total time spent by all reduces in occupied slots (ms)=212672

Total time spent by all map tasks (ms)=61448

Total time spent by all reduce tasks (ms)=26584

Total vcore-milliseconds taken by all map tasks=61448

Total vcore-milliseconds taken by all reduce tasks=26584

Total megabyte-milliseconds taken by all map tasks=251691008

Total megabyte-milliseconds taken by all reduce tasks=217776128

Map-Reduce Framework

Map input records=360798

Map output records=355868

Map output bytes=12472998

Map output materialized bytes=935608

Input split bytes=133

Combine input records=0

Combine output records=0

Reduce input groups=288

Reduce shuffle bytes=935608

Reduce input records=355868

Reduce output records=288

Spilled Records=711736

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=1601

CPU time spent (ms)=67200

Physical memory (bytes) snapshot=376320000

Virtual memory (bytes) snapshot=13150580736

Total committed heap usage (bytes)=230821888

Peak Map Physical memory (bytes)=236457984

Peak Map Virtual memory (bytes)=4955021312

Peak Reduce Physical memory (bytes)=141762560

Peak Reduce Virtual memory (bytes)=8195559424

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=85676711

File Output Format Counters

Bytes Written=13879