

UNIVERSITY OF SCIENCE  
VIETNAM NATIONAL UNIVERSITY  
FACULTY OF INFORMATION TECHNOLOGY  
COMPUTER SCIENCE  
REPORT LAB02  
GROUP: KAFKA



# REPORT LAB02

# HADOOP MAP-REDUCE

BIG DATA | 2020

**Instructors:** Lê Ngọc Thành  
Nguyễn Ngọc Thảo

<b><u>Students:</u></b>	Nguyễn Huỳnh Xuân Mai	1712091
	Nguyễn Phúc Khôi Nguyên	1712106
	Đào Đức Anh	1712270
	Nguyễn Thành Nhân	1712631
	Phan Hữu Tú	1712861

## MỤC LỤC

THÔNG TIN NHÓM .....	2
PHÂN CÔNG CÔNG VIỆC.....	2
ASSIGNMENTS.....	3
ASSIGNMENT 1 – WORD COUNT PROGRAM.....	3
ASSIGNMENT 2 – WORD SIZE WORD COUNT PROGRAM.....	6
ASSIGNMENT 3 – WEATHER DATA PROGRAM .....	10
ASSIGNMENT 4 – PATENT PROGRAM .....	13
ASSIGNMENT 5 – MAX TEMP PROGRAM .....	16
ASSIGNMENT 6 – AVERAGE SALARY PROGRAM.....	19
ASSIGNMENT 8 – MUSIC TRACK PROGRAM .....	23
ASSIGNMENT 5 – TELECOM CALL DATA RECORD PROGRAM.....	30
TÀI LIỆU THAM KHẢO .....	32

## I. THÔNG TIN NHÓM

<i>STT</i>	<i>MSSV</i>	<i>Họ và tên</i>	<i>Email</i>
1	1712091	Nguyễn Huỳnh Xuân Mai	<a href="mailto:1712091@student.hcmus.edu.vn">1712091@student.hcmus.edu.vn</a>
2	1712106	Nguyễn Phúc Khôi Nguyên	<a href="mailto:1712106@student.hcmus.edu.vn">1712106@student.hcmus.edu.vn</a>
3	1712270	Đào Đức Anh	<a href="mailto:1712270@student.hcmus.edu.vn">1712270@student.hcmus.edu.vn</a>
4	1712631	Nguyễn Thành Nhân	<a href="mailto:1712631@student.hcmus.edu.vn">1712631@student.hcmus.edu.vn</a>
5	1712861	Phan Hữu Tú	<a href="mailto:1712861@student.hcmus.edu.vn">1712861@student.hcmus.edu.vn</a>

## II. PHÂN CÔNG CÔNG VIỆC

<i>MSSV</i>	<i>Công việc</i>	<i>Hoàn thành</i>
1712091	Tổng hợp báo cáo	100%
1712106	Assignment 4, 6	100%
1712270	Assignment 1, 5	100%
1712631	Assignment 8, 9	100%
1712861	Assignment 2, 3	100%
Assignment 7		0%

### III. ASSIGNMENTS

#### a. Assignment 1 – Wordcount Program

- Tạo file **WordCount.java** chứa source code thực hiện chương trình Hadoop MapReduce

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

    private final static IntWritable one = new IntWritable(1); // giá trị 1 cho mỗi value
    private Text word = new Text();

    public void map(Object key, Text value, Context context
        ) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString()); // tách các token từ value nhận được
        while (itr.hasMoreTokens()) {
            word.set(itr.nextToken()); // word được set giá trị là mỗi token
            context.write(word, one); // Với mỗi token kết quả trả về là token đó và giá trị 1
        }
    }
}

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(); // với input value nhận được ta tính tổng các value
        }
        result.set(sum);
        context.write(key, result); // kết quả trả về là key ban đầu và tổng các values
    }
}

public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "word count"); // Tạo job object
    job.setJarByClass(WordCount.class); // set jar file mà mỗi node sẽ tìm Mapper và Reducer class
    job.setMapperClass(TokenizerMapper.class); // set mapper class
    job.setCombinerClass(IntSumReducer.class); // set combiner class
    job.setReducerClass(IntSumReducer.class); // set reducer class
    job.setOutputKeyClass(Text.class); // set key class cho output
    job.setOutputValueClass(IntWritable.class); // set value class cho output
    FileInputFormat.addInputPath(job, new Path(args[0])); // thêm một đường dẫn vào danh sách input cho MapReduce job
    FileOutputFormat.setOutputPath(job, new Path(args[1])); // Đường dẫn đến thư mục chứa kết quả
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

- File **wordcount.txt**: ta sử dụng file wordcount.txt từ mẫu dataset đã có.

- Put file **wordcount.txt** vào HDFS:

```
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/Some Sample Dataset$ ls
CDRlog.txt LastFMlog.txt pg201.txt temps.csv weather_data.txt wordcount.txt
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/Some Sample Dataset$ cp wordcount.txt ../WordCount/
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/Some Sample Dataset$ cd ../WordCount/
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ ls
WordCount.java wordcount.txt
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ hadoop fs -put wordcount.txt /sriMR
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$
```

- Kiểm tra quá trình put lên HDFS:

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

### Browse Directory

/sriMR

Show  entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	ducanh	supergroup	1.27 KB	Nov 30 12:06	1	128 MB	wordcount.txt

Showing 1 to 1 of 1 entries

Previous **1** Next

Hadoop, 2020.

- Từ file **WordCount.java** và build thành các class và build file jar:

```
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ hadoop com.sun.tools.javac.Main WordCount.java
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ jar cf wordcount.jar WordCount*.class
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ ls
WordCount.class WordCount$IntSumReducer.class wordcount.jar WordCount.java WordCount$TokenizerMapper.class wordcount.txt
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$
```

- Thực thi file **wordcount.jar**:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ hadoop jar wordcount.jar WordCount /sriMR/wordcount.txt /sriMR/WordCount/output
20/11/30 12:17:23 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/11/30 12:17:23 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/11/30 12:17:24 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/11/30 12:17:24 INFO input.FileInputFormat: Total input files to process : 1
20/11/30 12:17:24 INFO mapreduce.JobSubmitter: number of splits:1
20/11/30 12:17:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1029704211_0001
20/11/30 12:17:25 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/30 12:17:25 INFO mapreduce.Job: Running job: job_local1029704211_0001
20/11/30 12:17:25 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/11/30 12:17:25 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 12:17:25 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 12:17:25 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/11/30 12:17:25 INFO mapred.LocalJobRunner: Waiting for map tasks
20/11/30 12:17:25 INFO mapred.LocalJobRunner: Starting task: attempt_local1029704211_0001_m_000000_0
20/11/30 12:17:25 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 12:17:25 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 12:17:25 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
20/11/30 12:17:25 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/sriMR/wordcount.txt:0+1305
20/11/30 12:17:26 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/11/30 12:17:26 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/11/30 12:17:26 INFO mapred.MapTask: soft limit at: 83860800
20/11/30 12:17:26 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/11/30 12:17:26 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/11/30 12:17:26 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
20/11/30 12:17:26 INFO mapred.LocalJobRunner:
20/11/30 12:17:26 INFO mapred.MapTask: Starting flush of map output
20/11/30 12:17:26 INFO mapred.MapTask: Spilling map output
20/11/30 12:17:26 INFO mapred.MapTask: bufstart = 0; bufend = 2269; bufvoid = 104857600
20/11/30 12:17:26 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213416(104853664); length = 981/6553600
20/11/30 12:17:26 INFO mapreduce.Job: Job job_local1029704211_0001 running in uber mode : false
20/11/30 12:17:26 INFO mapred.MapTask: map 0% reduce 0%
20/11/30 12:17:26 INFO mapred.MapTask: Finished spill 0
20/11/30 12:17:26 INFO mapred.Task: Task attempt_local1029704211_0001_m_000000_0 is done. And is in the process of committing.

File System Counters
  FILE: Number of bytes read=10040
  FILE: Number of bytes written=994423
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=2610
  HDFS: Number of bytes written=1184
  HDFS: Number of read operations=13
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4

Map-Reduce Framework
  Map input records=9
  Map output records=246
  Map output bytes=2269
  Map output materialized bytes=1755
  Input split bytes=106
  Combine input records=246
  Combine output records=142
  Reduce input groups=142
  Reduce shuffle bytes=1755
  Reduce input records=142
  Reduce output records=142
  Spilled Records=284
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=493879296

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=1305

File Output Format Counters
  Bytes Written=1184
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ _

```

### - Kết quả thực thi của chương trình:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ hadoop fs -ls /sriMR/WordCount/output
Found 2 items
-rw-r--r-- 1 ducanh supergroup          0 2020-11-30 12:17 /sriMR/WordCount/output/_SUCCESS
-rw-r--r-- 1 ducanh supergroup       1184 2020-11-30 12:17 /sriMR/WordCount/output/part-r-00000
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$

```

```

cat: /srimr/wordcount/output/part-r-000000: No such file or directory
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/WordCount$ hadoop fs -cat /srimr/WordCount/output/part-r-000000
In      1
Infinite,      1
Nobody  1
This      1
We        1
When      1
Whether   1
Worry,    1
Years     1
Youth     2
a         11
adventure      1
aerials 2
and        8
appetite    1
appetite,   1
are        4
as         3
at         2
back       1
beauty,    1
being's    1
body       1
bows       1
but        2
by         2
catch      1
center     1
cheeks,    1
cheer,     1
child-like      1
courage    2
covered    1

```

### b. Assignment 2 – WordSizeWordCount Program

- Tạo file *WordSizeWordCount.java* chứa source code chạy Hadoop MapReduce:

```

45 public void map(Object key,
46                 Text value,
47                 Context context)
48                 throws IOException, InterruptedException {
49     /*
50     * Parameters:
51     *   :Key: (Object)
52     *   :value: (Text) Input need to pass through mapping method
53     *   :context: (Context)
54     */
55
56     //Converting the record (single line) to String and storing it in a String variable
57     String line = value.toString();
58
59     //StringTokenizer is breaking the record (line) into words
60     StringTokenizer tokenizer = new StringTokenizer(line);
61
62     //iterating through all the words available in that line and forming the key value pair
63     while (tokenizer.hasMoreTokens()) {
64         String thisH = tokenizer.nextToken();
65
66         //finding the length of each token(word)
67         count = new IntWritable(thisH.length());
68
69         word.set(thisH);
70         //Sending to output collector which in turn passes the same to reducer
71         //So in this case the output from mapper will be the length of a word and the word
72         context.write(count, word);
73     }
74 }
75
76
77 //Reducer
78 public static class Reduce extends Reducer < IntWritable, Text, IntWritable, IntWritable> {
79     /**
80     * @method reduce
81     * <p>This method takes the input as key and list of values pair from mapper, it does
82     * based on keys and produces the final output.
83     * @method arguments key, values, output, reporter
84     * @return void
85     */
86
87     /*
88     * (non-Javadoc)
89     * @see org.apache.hadoop.mapred.Reducer#reduce(java.lang.Object, java.util.Iterator,
90     * org.apache.hadoop.mapred.OutputCollector, org.apache.hadoop.mapred.Reporter)
91     */
92
93     @Override
94     public void reduce(IntWritable key,
95                       Iterable < Text > values,
96                       Context context)
97                       throws IOException, InterruptedException {
98
99         //Defining a local variable sum of type int
100         int sum = 0;
101
102         /*
103         * Iterates through all the values available with a key and add them together and

```

- Tạo file word\_size\_count.txt chứa đoạn text input:

```

tuhp@tuhp: ~/study/DDL/doan2/src/dataset 87x39
tuhp@tuhp:~/study/DDL/doan2/src/dataset$
tuhp@tuhp:~/study/DDL/doan2/src/dataset$
tuhp@tuhp:~/study/DDL/doan2/src/dataset$ touch word_size_count.txt
tuhp@tuhp:~/study/DDL/doan2/src/dataset$ vi word_size_count.txt

```



[illegible]

- Copy dữ liệu vào file system của Hadoop, sau đó kiểm tra trong hệ thống đã có file hay chưa:

```
tuhp@tuhp:~/study/DDL/doan2/src$ hadoop fs -put /home/tuhp/study/DDL/doan2/src/dataset/word_size_count.txt /sriMR
tuhp@tuhp:~/study/DDL/doan2/src$
```

```

tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -ls /sriMR
Found 3 items
drwxr-xr-x  - tuhp supergroup          0 2020-11-29 09:45 /sriMR/result_wc
-rw-r--r--  1 tuhp supergroup    41881 2020-11-29 10:26 /sriMR/weather_data.txt
-rw-r--r--  1 tuhp supergroup    1305 2020-11-29 09:33 /sriMR/wordcount.txt
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -ls /sriMR
Found 4 items
drwxr-xr-x  - tuhp supergroup          0 2020-11-29 09:45 /sriMR/result_wc
-rw-r--r--  1 tuhp supergroup    41881 2020-11-29 10:26 /sriMR/weather_data.txt
-rw-r--r--  1 tuhp supergroup    9123 2020-11-29 10:36 /sriMR/word_size_count.txt
-rw-r--r--  1 tuhp supergroup    1305 2020-11-29 09:33 /sriMR/wordcount.txt
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$

```

- Tạo file **WordSizeWordCount.java** và build thành các class:

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL 1: bash
tuhp@tuhp:~/study/DDDL/doan2/src/WordSizeWordCount$ hadoop com.sun.tools.javac.Main WordSizeWordCount.java
Note: WordSizeWordCount.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
tuhp@tuhp:~/study/DDDL/doan2/src/WordSizeWordCount$

```

- Build thành file **.jar** từ các class vừa tạo:

```

tuhp@tuhp:~/study/DDDL/doan2/src/WordSizeWordCount$ jar cf word_size_count.jar WordSize*.class
tuhp@tuhp:~/study/DDDL/doan2/src/WordSizeWordCount$ ls
word_size_count.jar      WordSizeWordCount.java      'WordSizeWordCount$Reduce.class'
WordSizeWordCount.class  'WordSizeWordCount$Map.class'

```

- Chạy file jar trong môi trường Hadoop với Input là file **word\_size\_count.txt** vừa được copy từ local:

```

tuhp@tuhp:~/study/DDL/doan2/src/WordSizeWordCount$ hadoop jar word_size_count.jar WordSizeWordCount /sriMR/word_s
ize_count.txt /sriMR/result_WordSizeCount
20/11/29 10:47:53 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/11/29 10:47:53 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/11/29 10:47:53 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
20/11/29 10:47:53 INFO input.FileInputFormat: Total input files to process : 1
20/11/29 10:47:53 INFO mapreduce.JobSubmitter: number of splits:1
20/11/29 10:47:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1701560011_0001
20/11/29 10:47:53 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/29 10:47:53 INFO mapreduce.Job: Running job: job_local1701560011_0001
20/11/29 10:47:53 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/11/29 10:47:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/29 10:47:53 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outp
ut directory:false, ignore cleanup failures: false
20/11/29 10:47:53 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutpu
tCommitter
20/11/29 10:47:53 INFO mapred.LocalJobRunner: Waiting for map tasks
20/11/29 10:47:53 INFO mapred.LocalJobRunner: Starting task: attempt_local1701560011_0001_m_000000_0
20/11/29 10:47:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/29 10:47:53 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outp
ut directory:false, ignore cleanup failures: false
20/11/29 10:47:53 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
20/11/29 10:47:53 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/sriMR/word_size_count.txt:0+9123
20/11/29 10:47:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/11/29 10:47:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/11/29 10:47:53 INFO mapred.MapTask: soft limit at 83886080
20/11/29 10:47:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/11/29 10:47:53 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/11/29 10:47:53 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuf
fer
20/11/29 10:47:53 INFO mapred.LocalJobRunner:
20/11/29 10:47:53 INFO mapred.MapTask: Starting flush of map output
20/11/29 10:47:53 INFO mapred.MapTask: Spilling map output
20/11/29 10:47:53 INFO mapred.MapTask: bufstart = 0; bufend = 14682; bufvoid = 104857600
20/11/29 10:47:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26208800(104835200); length = 5597/
6553600
20/11/29 10:47:53 INFO mapred.MapTask: Finished spill 0
20/11/29 10:47:53 INFO mapred.Task: Task:attempt_local1701560011_0001_m_000000_0 is done. And is in the process o
f committing
20/11/29 10:47:53 INFO mapred.LocalJobRunner: map
20/11/29 10:47:53 INFO mapred.Task: Task 'attempt_local1701560011_0001_m_000000_0' done.
20/11/29 10:47:53 INFO mapred.Task: Final Counters for attempt_local1701560011_0001_m_000000_0: Counters: 22

```

- Kiểm tra kết quả:

```

tuhp@tuhp:~/study/DDL/doan2/src/WordSizeWordCount$ hadoop fs -cat /sriMR/result_WordSizeCount/part-r-00000
1      66
2     162
3     208
4     204
5     126
6     163
7     127
8     108
9      73
10     77
11     32
12     20
13     22
14      8
15      2
16      2

```

→ Kết quả có format: *<kích cỡ chữ> <số lượng chữ có kích cỡ tương ứng>*

### c. Assignment 3 – WeatherData Program

- Tạo file *WordSizeWordCount.java* chứa source code chạy Hadoop MapReduce:

```

WeatherData.java x WeatherData.jar
WeatherData > WeatherData.java
1  import java.util.Iterator;
2  import java.io.IOException;
3  import java.util.StringTokenizer;
4  import org.apache.hadoop.conf.Configuration;
5  import org.apache.hadoop.fs.Path;
6  import org.apache.hadoop.io.IntWritable;
7  import org.apache.hadoop.io.LongWritable;
8  import org.apache.hadoop.io.Text;
9  import org.apache.hadoop.mapreduce.Job;
10 import org.apache.hadoop.mapreduce.Mapper;
11 import org.apache.hadoop.mapreduce.Reducer;
12 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
13 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
14 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
15 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
16
17
18 public class WeatherData {
19     public static class MaxTemperatureMapper extends Mapper < Object, Text, Text, Text > {
20         @Override
21         public void map(Object arg0, Text Value, Context context) throws IOException, InterruptedException {
22
23             String line = Value.toString();
24             // Example of Input
25             // Date Max Min
26             // 25380 20130101 2.514 -135.69 58.43 8.3 1.1 4.7 4.9 5.6 0.01 C 1.0 -0.1 0.4 97.3 36.0 69.4
27             // -99.000 -99.000 -99.000 -99.000 -99.000 -9999.0 -9999.0 -9999.0 -9999.0 -9999.0
28             String date = line.substring(6, 14);
29             float temp_Max = Float.parseFloat(line.substring(39, 45).trim());
30             float temp_Min = Float.parseFloat(line.substring(47, 53).trim());
31
32             if (temp_Max > 40.0) {
33                 // Hot day
34                 context.write(new Text("Hot Day " + date), new Text(String.valueOf(temp_Max)));
35                 // output.collect(new Text("Hot Day " + date), new Text(String.valueOf(temp_Max)));
36             }
37             if (temp_Min < 10) {
38                 // Cold day
39                 context.write(new Text("Cold Day " + date), new Text(String.valueOf(temp_Min)));
40                 // output.collect(new Text("Cold Day " + date), new Text(String.valueOf(temp_Min)));
41             }
42         }
43     }
44 }

```

- Copy dữ liệu vào file system của Hadoop, sau đó kiểm tra trong hệ thống đã có file hay chưa:

```

52      * :value: (Text) Input need to pass through mapping method
53      * :context: (Context)
54      */
55
56      //Converting the record (single line) to String and storing it in a String variable
57      String line = value.toString();

```

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
1: bash
tuhp@tuhp:~/study/DDL/doan2/src$
tuhp@tuhp:~/study/DDL/doan2/src$
tuhp@tuhp:~/study/DDL/doan2/src$ cd
dataset/ WeatherData/ WordSizeWordCount/
tuhp@tuhp:~/study/DDL/doan2/src$ cd
dataset/ WeatherData/ WordSizeWordCount/
tuhp@tuhp:~/study/DDL/doan2/src$ cd
dataset/ WeatherData/ WordSizeWordCount/
tuhp@tuhp:~/study/DDL/doan2/src$ hadoop fs -put /home/tuhp/study/DDL/doan2/src/dataset/weather_data.txt /sriMR
tuhp@tuhp:~/study/DDL/doan2/src$

```

```

tuhp@tuhp: ~/Downloads/hadoop/hadoop-2.10.1
tuhp@tuhp: ~/Downloads/hadoop/hadoop-2.10.1 87x14
drwxr-xr-x - tuhp supergroup 0 2020-11-29 09:31 /sriMR
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -ls /srv/
ls: '/srv/': No such file or directory
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -ls /sriMR
Found 1 items
-rw-r--r-- 1 tuhp supergroup 1305 2020-11-29 09:33 /sriMR/wordcount.txt
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -ls /sriMR
Found 3 items
drwxr-xr-x - tuhp supergroup 0 2020-11-29 09:45 /sriMR/result_wc
-rw-r--r-- 1 tuhp supergroup 41881 2020-11-29 10:26 /sriMR/weather_data.txt
-rw-r--r-- 1 tuhp supergroup 1305 2020-11-29 09:33 /sriMR/wordcount.txt
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$

```

- Tạo file *WordSizeWordCount.java* và build thành các class và file *.jar* từ các class vừa tạo:

```

tuhp@tuhp:~/study/DDI/doan2/src/WeatherData$ hadoop com.sun.tools.javac.Main WeatherData.java
Note: WeatherData.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
tuhp@tuhp:~/study/DDI/doan2/src/WeatherData$ jar cf WeatherData.jar Weather*.class
tuhp@tuhp:~/study/DDI/doan2/src/WeatherData$ ls
WeatherData.class WeatherData.jar WeatherData.java 'WeatherData$MaxTemperatureMapper.class' 'WeatherData$MaxTemperatureReducer.class'
tuhp@tuhp:~/study/DDI/doan2/src/WeatherData$

```

- Chạy file jar trong môi trường Hadoop với Input là file *weather\_data.txt* vừa được copy từ local:

```

tuhp@tuhp:~/study/DDI/doan2/src/WeatherData$ hadoop jar WeatherData.jar WeatherData /sriMR/weather_data.txt /sriMR/result_weather
20/11/30 10:14:58 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/11/30 10:14:58 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/11/30 10:14:58 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/11/30 10:14:58 INFO input.FileInputFormat: Total input files to process : 1
20/11/30 10:14:58 INFO mapreduce.JobSubmitter: number of splits:1
20/11/30 10:14:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local326571137_0001
20/11/30 10:14:59 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/30 10:14:59 INFO mapreduce.Job: Running job: job_local326571137_0001
20/11/30 10:14:59 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/11/30 10:14:59 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 10:14:59 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 10:14:59 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/11/30 10:14:59 INFO mapred.LocalJobRunner: Waiting for map tasks
20/11/30 10:14:59 INFO mapred.LocalJobRunner: Starting task: attempt_local326571137_0001_m_000000_0
20/11/30 10:14:59 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 10:14:59 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders

```

- Kiểm tra kết quả:

```

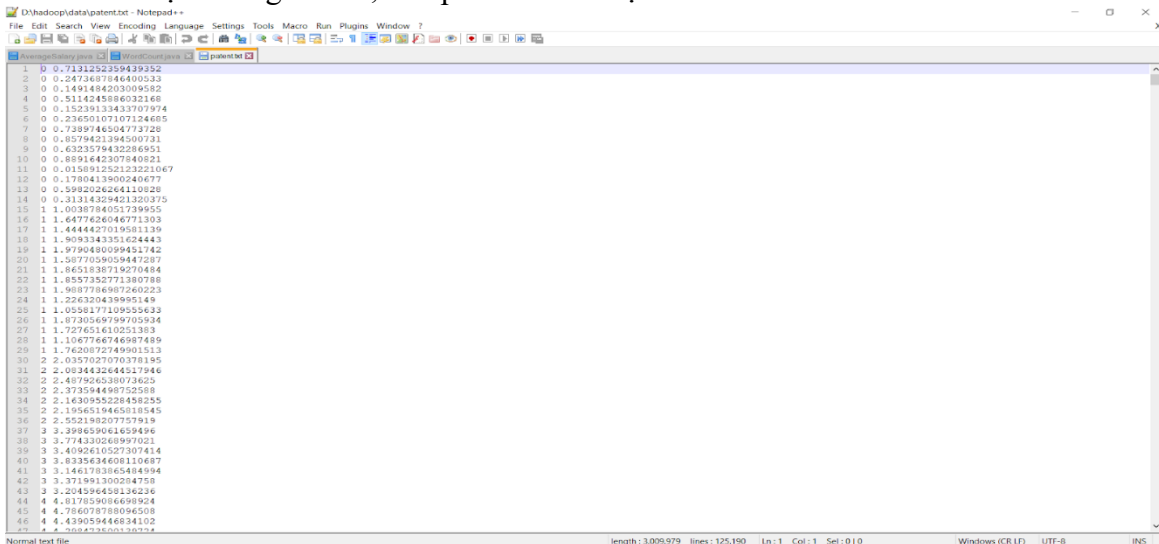
tuhp@tuhp:~/Downloads/hadoop/hadoop-2.10.1$ hadoop fs -cat /sriMR/result_weather/part-r-00000
Cold Day 20150101      1.401298464324817E-45
Cold Day 20150102      1.2999999523162842
Cold Day 20150103      2.299999952316284
Cold Day 20150104      1.401298464324817E-45
Cold Day 20150105      1.401298464324817E-45
Cold Day 20150106      2.90000000953674316
Cold Day 20150107      1.401298464324817E-45
Cold Day 20150108      1.401298464324817E-45
Cold Day 20150109      0.10000000149011612
Cold Day 20150110      1.401298464324817E-45
Cold Day 20150111      1.401298464324817E-45
Cold Day 20150112      1.399999976158142
Cold Day 20150113      1.401298464324817E-45
Cold Day 20150114      0.8999999761581421
Cold Day 20150115      1.2000000476837158
Cold Day 20150116      3.5
Cold Day 20150117      5.0
Cold Day 20150118      7.599999904632568
Cold Day 20150119      6.699999809265137
Cold Day 20150120      9.5
Cold Day 20150121      6.900000095367432

```

→ Kết quả có format: *<(Cold Day/Hot Day) yyyyymmdd> <max temperature>*

#### d. Assignment 4 – Patent Program

- Do dữ liệu không có sẵn, nên phát sinh dữ liệu mới:



- Đưa data vào dfs. Gồm các bước tạo đường dẫn patent/input.
- Đưa file patent.txt từ máy cục bộ lên dfs.
- Kiểm tra file trên dfs đã tồn tại chưa.



```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls
Found 3 items
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:45 input
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:54 output
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 15:11 wordcount
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -mkdir patent
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -mkdir patent/input
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls
Found 4 items
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:45 input
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:54 output
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 11:27 patent
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 15:11 wordcount
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ ls ../data/
CDRlog.txt LastFMlog.txt patent.txt pg201.txt salary.txt temps.csv weather_data.txt wordcount.txt
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -put ../data/patent.txt patent/input
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls patent/input
Found 1 items
-rw-r--r-- 1 nguyennp supergroup 3009979 2020-11-30 11:28 patent/input/patent.txt
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$

```

- Biên dịch `patent.java` và nén các class thành file *patent.jar*:

```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ ls
LICENSE.txt NOTICE.txt Patent.java README.txt bin lib NOTICE.txt LICENSE.txt README.txt bin
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hadoop com.sun.tools.javac.Main Patent.java
Note: Patent.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ ;
-bash: syntax error near unexpected token ';'
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ ls
LICENSE.txt NOTICE.txt "Patent$Map.class" Patent.class Patent.java README.txt bin lib NOTICE.txt LICENSE.txt README.txt bin
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ jar cf pt.jar Patent.class
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ ls
LICENSE.txt NOTICE.txt "Patent$Map.class" "Patent$Reduce.class" Patent.class Patent.java README.txt bin lib NOTICE.txt LICENSE.txt README.txt bin pt.jar
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$

```

- Chạy file jar theo cú pháp *bin/Hadoop jar <tên file jar> <tên class chính> <đường dẫn input> <đường dẫn output>*

```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hadoop jar pt.jar Patent patent/input patent/output
20/11/30 12:27:55 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/11/30 12:27:55 INFO jvm.MetricRegistry: Initializing JVM Metrics with processName:hadoop, sessionId:
20/11/30 12:27:55 INFO mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/11/30 12:27:55 INFO mapreduce.JobSubmitter: total input files to process : 1
20/11/30 12:27:56 INFO mapreduce.JobSubmitter: number of splits:1
20/11/30 12:27:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1364822371_0001
20/11/30 12:27:56 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/30 12:27:56 INFO mapreduce.Job: Running job: job_local1364822371_0001
20/11/30 12:27:56 INFO mapreduce.LocalJobRunner: OutputCommitter set in config null
20/11/30 12:27:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 12:27:56 INFO mapreduce.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/11/30 12:27:56 INFO mapreduce.LocalJobRunner: Waiting for map tasks
20/11/30 12:27:56 INFO mapreduce.LocalJobRunner: Starting task: attempt_local1364822371_0001_m_000000_0
20/11/30 12:27:56 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 12:27:56 INFO output.FileOutputCommitter: skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 12:27:56 INFO mapreduce.Task: Using ResourceCalculatorProcessFree [ ]
20/11/30 12:27:56 INFO mapreduce.MapTask: Processing split: hdfs://localhost:8080/user/nguyennp/patent/input/patent.txt_0-3009979
20/11/30 12:27:56 INFO mapreduce.MapTask: (EQM100) & kv1 20214106(104857504)
20/11/30 12:27:56 INFO mapreduce.MapTask: mapreduce.task.io.sort.mb: 100
20/11/30 12:27:56 INFO mapreduce.MapTask: soft limit at 83886080
20/11/30 12:27:56 INFO mapreduce.MapTask: bufstart = 0; bufvoid = 104857500
20/11/30 12:27:56 INFO mapreduce.MapTask: kvstart = 20214106; length = 4833600
20/11/30 12:27:56 INFO mapreduce.MapTask: Map output collector class = org.apache.hadoop.mapreduce.MapTask$MapOutputBuffer
20/11/30 12:27:57 INFO mapreduce.LocalJobRunner:
20/11/30 12:27:57 INFO mapreduce.MapTask: Starting flush of map output
20/11/30 12:27:57 INFO mapreduce.MapTask: Spilling map output
20/11/30 12:27:57 INFO mapreduce.MapTask: bufstart = 0; bufend = 2884790; bufvoid = 104857500
20/11/30 12:27:57 INFO mapreduce.MapTask: kvstart = 20214106(104857504); kvend = 25713644(302854576); length = 500751/5553600
20/11/30 12:27:57 INFO mapreduce.MapTask: Finished spill 0
20/11/30 12:27:57 INFO mapreduce.Task: Task:attempt_local1364822371_0001_m_000000_0 is done. And is in the process of committing
20/11/30 12:27:57 INFO mapreduce.Task: Task:attempt_local1364822371_0001_m_000000_0 done.
20/11/30 12:27:57 INFO mapreduce.Task: Final Counters for attempt_local1364822371_0001_m_000000_0: Counters: 22
File System Counters
FILE: Number of bytes read=1257
FILE: Number of bytes written=3635271
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=3009979
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Map-Reduce framework
Map input records=125189
Map output records=125189
Map output bytes=2884790
Map output materialized bytes=31135376
Input split bytes=124

```

```

nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1
Failed Shuffle=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=292271296
File Input Format Counters
  Bytes Read=3009979
20/11/30 12:27:57 INFO mapreduce.Job: Job job_local1364822371_0001 running in uber mode : false
20/11/30 12:27:57 INFO mapred.LocalJobRunner: finishing task: attempt_local1364822371_0001_m_000000_0
20/11/30 12:27:57 INFO mapreduce.Job: map 100% reduce 0%
20/11/30 12:27:57 INFO mapred.LocalJobRunner: map task executor complete.
20/11/30 12:27:57 INFO mapred.LocalJobRunner: Waiting for reduce tasks
20/11/30 12:27:57 INFO mapred.LocalJobRunner: Starting task: attempt_local1364822371_0001_r_000000_0
20/11/30 12:27:57 INFO reduce.EventFetcher: attempt_local1364822371_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
20/11/30 12:27:57 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup, temporary folders under output directory=false, ignore cleanup failures: false
20/11/30 12:27:57 INFO mapred.Task: Using ResourceCalculatorProcessImpl: [ ]
20/11/30 12:27:57 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@18670660
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: PurgeManager memoryLeft=131188664, maxMergeSize=1073548128, mergeThreshold=22861392, logSortFactor=10, mergeMapMergeOutputsThreshold=10
20/11/30 12:27:57 INFO reduce.EventFetcher: attempt_local1364822371_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
20/11/30 12:27:57 INFO reduce.LocalFetcher: LocalFetcher#1 about to shuffle output of map attempt_local1364822371_0001_m_000000_0 decom: 1135170 len: 1135170 to REDUCE
20/11/30 12:27:57 INFO reduce.InMemoryMapOutput: Read 1135170 bytes from map-output for attempt_local1364822371_0001_m_000000_0
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: closeInMemoryFile -> map-output of size: 1135170, InMemoryMapOutputs.size() -> 1, commitMemory -> 0, useMemory -> 1135170
20/11/30 12:27:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: finalPurge called with 1 in-memory map-outputs and 0 on-disk map-outputs
20/11/30 12:27:57 INFO mapred.Merger: Merging 1 sorted segments
20/11/30 12:27:57 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1135166 bytes
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: Merged 1 segments, 1135170 bytes to disk to satisfy reduce memory limit
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: Merging 1 files, 1135170 bytes from disk
20/11/30 12:27:57 INFO reduce.PurgeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
20/11/30 12:27:57 INFO mapred.Merger: Merging 1 sorted segments
20/11/30 12:27:57 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1135166 bytes
20/11/30 12:27:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 12:27:57 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
20/11/30 12:27:57 INFO mapred.Task: Task:attempt_local1364822371_0001_r_000000_0 is done. And is in the process of committing
20/11/30 12:27:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 12:27:57 INFO mapred.Task: Task attempt_local1364822371_0001_r_000000_0 is allowed to commit now
20/11/30 12:27:57 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1364822371_0001_r_000000_0' to hdfs://localhost:9000/user/nguyennp/patent/output/_temporary/0/task_local1364822371_0001_r_000000_0
20/11/30 12:27:57 INFO mapred.LocalJobRunner: reduce a reduce
20/11/30 12:27:57 INFO mapred.Task: Task 'attempt_local1364822371_0001_r_000000_0' done.
20/11/30 12:27:57 INFO mapred.Task: Final Counters for attempt_local1364822371_0001_r_000000_0: Counters: 29
File System Counters
  FILE: Number of bytes read=6273637
  FILE: Number of bytes written=6720867
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
HDFS: Number of bytes read=3009979
HDFS: Number of bytes written=75796
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Map-Reduce Framework
  Map input records=125189
  Map output records=125189
  Map output bytes=2884790
  Map output materialized bytes=1135174
  Input split bytes=124
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000
  Reduce shuffle bytes=1135174
  Reduce input records=125189
  Reduce output records=10000
  Spilled Records=250378
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=594542592
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDX=0
File Input Format Counters
  Bytes Read=3009979
File Output Format Counters
  Bytes Written=75796
20/11/30 12:27:58 INFO mapred.LocalJobRunner: Finishing task: attempt_local1364822371_0001_r_000000_0
20/11/30 12:27:58 INFO mapred.LocalJobRunner: reduce task executor complete.
20/11/30 12:27:58 INFO mapreduce.Job: map 100% reduce 100%
20/11/30 12:27:58 INFO mapreduce.Job: Job job_local1364822371_0001 completed successfully
20/11/30 12:27:58 INFO mapreduce.Job: Counters: 35
File System Counters
  FILE: Number of bytes read=6276894
  FILE: Number of bytes written=58606920
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
HDFS: Number of bytes read=6010954
HDFS: Number of bytes written=75796
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Map-Reduce Framework
  Map input records=125189
  Map output records=125189
  Map output bytes=2884790
  Map output materialized bytes=1135174
  Input split bytes=124
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000
  Reduce shuffle bytes=1135174
  Reduce input records=125189
  Reduce output records=10000
  Spilled Records=250378
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=594542592
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDX=0
File Input Format Counters
  Bytes Read=3009979
File Output Format Counters
  Bytes Written=75796

```

- Kiểm tra đường dẫn output:

```

nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls patent/output
Found 2 items
-rw-r--r-- 1 nguyennp supergroup 0 2020-11-30 12:27 patent/output/_SUCCESS
-rw-r--r-- 1 nguyennp supergroup 75796 2020-11-30 12:27 patent/output/part-r-000000
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$

```



```

1 15
2 7
3 6
4 7
5 9
6 8
7 8
8 14
9 19
10 16
11 10
12 5
13 10
14 5
15 19
16 8
17 15
18 5
19 9
20 16
21 18
22 4
23 9
24 18
25 19
26 15
27 12
28 20
29 13
30 15
31 16
32 13
33 18
34 12
35 12
36 10
37 9
38 8
39 16
40 17
41 14
42 10
43 20
44 10
45 19
46 9
47 16
48 7
49 19
50 9
51 13

```

### e. Assignment 5 – MaxTemp Program

- Tạo file *MaxTemp.java* chứa source code thực hiện chương trình Hadoop MapReduce:

```

public static class Map extends Mapper<LongWritable, Text, Text, FloatWritable>
{
    private Text year = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
    {
        StringTokenizer tokens = new StringTokenizer(value.toString(), ","); // Chuyển value thành String, delimiter là ","

        year.set(tokens.nextToken().trim() + ","); // Tách token đầu để set giá trị Year
        float temp = Float.parseFloat(tokens.nextToken().trim()); // Chuyển nhiệt độ từ dạng String sang Float từ token tiếp theo

        context.write(year, new FloatWritable(temp)); // kết quả key là year và value là Temp
    }
}

public static class Reduce extends Reducer<Text, FloatWritable, Text, FloatWritable>
{
    public void reduce(Text key, Iterable<FloatWritable> values, Context context) throws IOException, InterruptedException
    {
        float MaxTemp = Float.MIN_VALUE;

        for(FloatWritable value: values)
        {
            float temp = value.get();
            // Tìm ra giá trị Temp lớn nhất
            if (temp > MaxTemp)
            {
                MaxTemp = temp;
            }
        }

        context.write(key, new FloatWritable(MaxTemp)); // kết quả trả về là Key và giá trị nhiệt độ lớn nhất
    }
}

```

```

public static void main(String[] args) throws Exception
{
    Configuration conf = new Configuration();
    Job job = new Job(conf, "MaxTemp"); // Tạo job object
    job.setJarByClass(MaxTemp.class); // set jar file mà mỗi node sẽ tìm Mapper và Reducer class

    job.setInputFormatClass(TextInputFormat.class); // set input format
    job.setOutputFormatClass(TextOutputFormat.class); // set output format
    FileInputFormat.addInputPath(job, new Path(args[0])); // thêm một đường dẫn vào danh sách input cho MapReduce job
    FileOutputFormat.setOutputPath(job, new Path(args[1])); // Đường dẫn đến thư mục chứa kết quả

    job.setMapperClass(Map.class); // set mapper class
    job.setCombinerClass(Reduce.class); // set combiner class
    job.setReducerClass(Reduce.class); // set reducer class

    job.setOutputKeyClass(Text.class); // set key class cho output
    job.setOutputValueClass(FloatWritable.class); // set value class cho output

    System.exit(job.waitForCompletion(true) ? 0 : 1);
}

```

- File *temps.txt*: ta sẽ sử dụng lại file *temps.csv* trong Dataset mẫu những sẽ drop 2 cột dữ liệu giữ lại 2 cột dữ liệu đầu. Trong đó cột đầu ta sẽ lấy 4 ký tự cuối đại diện cho giá trị Years:

```

hduser@DESKTOP-3HTU97K: ~/Hadoop_MapReduce/bai5$ cat temps.csv
2000,-4
2000,-5
2000,-5
2000,-3
2000,-0.8
2000,-3.1
2000,-2.2
2000,-3.4
2000,-2.5
2000,-1.3
2000,-1
2000,-3
2000,-5.6
2000,-1.3
2000,-2
2000,-4.1
2000,-0.6
2000,3
2000,-1
2000,-3
2000,-2.2
2000,-0.8
2000,0.3
2000,-1.4
2000,-5
2000,-7.2
2000,-6
2000,-2.8
2000,-4
2000,-2.9
2000,-0.1
2000,-0.5
2000,6.6
2000,4.8
2000,3
2000,1
2000,0.6
2000,-1
2000,4.8
2000,5
2000,1
2000,-0.2
2000,1

```

- Put file *temps.txt* vào HDFS:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop fs -put temps.csv /sriMR
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop fs -ls /sriMR
Found 3 items
drwxr-xr-x  - ducanh supergroup          0 2020-11-30 12:17 /sriMR/WordCount
-rw-r--r--  1 ducanh supergroup       47007 2020-11-30 13:54 /sriMR/temps.csv
-rw-r--r--  1 ducanh supergroup        1305 2020-11-30 12:06 /sriMR/wordcount.txt
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$

```

- Từ file *WordCount.java* và build thành các class và build file jar:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop com.sun.tools.javac.Main MaxTemp.java
Note: MaxTemp.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ ls
MaxTemp.class  MaxTemp.java  MaxTemp$Map.class  MaxTemp$Reduce.class  temps.csv  wordcount.jar
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ jar cf MaxTemp.jar MaxTemp*.class
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ ls
MaxTemp.class  MaxTemp.jar  MaxTemp.java  MaxTemp$Map.class  MaxTemp$Reduce.class  temps.csv  wordcount.jar
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$

```

- Thúc thi file *MaxTemp.jar*:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop jar MaxTemp.jar MaxTemp /sriMR/temps.csv /sri/MaxTemp/output
20/11/30 13:58:57 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/11/30 13:58:57 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/11/30 13:58:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/11/30 13:58:57 INFO input.FileInputFormat: Total input files to process : 1
20/11/30 13:58:58 INFO mapreduce.JobSubmitter: number of splits:1
20/11/30 13:58:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1189177109_0001
20/11/30 13:58:59 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/30 13:58:59 INFO mapreduce.Job: Running job: job_local1189177109_0001
20/11/30 13:58:59 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/11/30 13:58:59 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 13:58:59 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 13:58:59 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/11/30 13:58:59 INFO mapred.LocalJobRunner: Waiting for map tasks
20/11/30 13:58:59 INFO mapred.LocalJobRunner: Starting task: attempt_local1189177109_0001_m_000000_0
20/11/30 13:58:59 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 13:58:59 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 13:58:59 INFO mapred.Task: Using ResourceCalculatorProcessTree: [ ]
20/11/30 13:58:59 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/sriMR/temps.csv:0+47007
20/11/30 13:59:00 INFO mapreduce.Job: Job job_local1189177109_0001 running in uber mode : false
20/11/30 13:59:00 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/11/30 13:59:00 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/11/30 13:59:00 INFO mapred.MapTask: soft limit at 83886080
20/11/30 13:59:00 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/11/30 13:59:00 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/11/30 13:59:00 INFO mapreduce.Job: map 0% reduce 0%
20/11/30 13:59:00 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
20/11/30 13:59:00 INFO mapred.LocalJobRunner:
20/11/30 13:59:00 INFO mapred.MapTask: Starting flush of map output
20/11/30 13:59:00 INFO mapred.MapTask: Spilling map output
20/11/30 13:59:00 INFO mapred.MapTask: bufstart = 0; bufend = 50300; bufvoid = 104857600
20/11/30 13:59:00 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26194280(104777120); length = 20117/6553600
20/11/30 13:59:00 INFO mapred.MapTask: Finished spill 0
20/11/30 13:59:00 INFO mapred.Task: Task:attempt_local1189177109_0001_m_000000_0 is done. And is in the process of committing
20/11/30 13:59:00 INFO mapred.LocalJobRunner: map

```

```

File System Counters
  FILE: Number of bytes read=6844
  FILE: Number of bytes written=991454
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=94014
  HDFS: Number of bytes written=154
  HDFS: Number of read operations=13
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
Map-Reduce Framework
  Map input records=5030
  Map output records=5030
  Map output bytes=50300
  Map output materialized bytes=174
  Input split bytes=102
  Combine input records=5030
  Combine output records=14
  Reduce input groups=14
  Reduce shuffle bytes=174
  Reduce input records=14
  Reduce output records=14
  Spilled Records=28
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=437256192
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=47007
File Output Format Counters
  Bytes Written=154
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$

```

- Kết quả chương trình:

```

ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop fs -ls /sri/MaxTemp/output
Found 2 items
-rw-r--r-- 1 ducanh supergroup      0 2020-11-30 13:59 /sri/MaxTemp/output/_SUCCESS
-rw-r--r-- 1 ducanh supergroup    154 2020-11-30 13:59 /sri/MaxTemp/output/part-r-00000
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$ hadoop fs -cat /sri/MaxTemp/output/part-r-00000
2000, 23.0
2001, 23.0
2002, 24.7
2003, 26.0
2004, 23.3
2005, 24.5
2006, 25.0
2007, 21.9
2008, 22.0
2009, 25.0
2010, 24.0
2011, 23.0
2012, 25.0
2013, 24.0
ducanh@DESKTOP-3HTU97K:~/Hadoop_MapReduce/MaxTemp$

```

## f. Assignment 6 – AverageSalary Program

- Đưa data vào dfs:

```

nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls
Found 4 items
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:45 input
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:54 output
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 12:27 patent
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 11:54 wordcount
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -mkdir average_salary
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -mkdir average_salary/input
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls
Found 5 items
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 13:08 average_salary
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:45 input
drwxr-xr-x - nguyennp supergroup          0 2020-11-29 00:54 output
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 12:27 patent
drwxr-xr-x - nguyennp supergroup          0 2020-11-30 11:54 wordcount
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls average_salary
Found 1 items
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -put ../data/salary.txt average_salary/input
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls average_salary/input
Found 1 items
-rw-r--r-- 1 nguyennp supergroup      488647 2020-11-30 13:10 average_salary/input/salary.txt
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$

```

- Biên dịch code *AverageSalary.java* và nén các class thành file *.jar*

```

nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hadoop com.sun.tools.javac.Main AverageSalary.java
Note: AverageSalary.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ jar cf as.jar AverageSalary*.class
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ ls
AverageSalary$AvgMapper.class  AverageSalary.java  README.txt  etc  libexec  sbin
AverageSalary$AvgReducer.class  LICENSE.txt  as.jar  include  logs  share
AverageSalary.class  NOTICE.txt  bin  lib  output
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$

```

- Chạy chương trình trên dfs:

```

nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -rmr average_salary/output
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ bin/hadoop com.sun.tools.javac.Main AverageSalary.java
Note: AverageSalary.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nguyennp@DESKTOP-H7AOLMK: /mnt/d/hadoop/hadoop-2.10.1$ jar cf as.jar AverageSalary*.class
20/11/30 13:30:54 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session.id
20/11/30 13:30:54 INFO jvm.MetricSink: Initializing JVM Metrics with processName=hadoop-tracker, sessionId=
20/11/30 13:30:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/11/30 13:30:54 INFO mapreduce.JobSubmitter: number of splits=1
20/11/30 13:30:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local196355271_0001
20/11/30 13:30:54 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/11/30 13:30:54 INFO mapreduce.Job: Running job: job_local196355271_0001
20/11/30 13:30:54 INFO mapreduce.LocalJobRunner: OutputCommitter set in config null
20/11/30 13:30:54 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 13:30:54 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 13:30:54 INFO mapreduce.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/11/30 13:30:54 INFO mapreduce.LocalJobRunner: Waiting for map tasks
20/11/30 13:30:54 INFO mapreduce.LocalJobRunner: Starting task: attempt_local196355271_0001_m_000000_0
20/11/30 13:30:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
20/11/30 13:30:55 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
20/11/30 13:30:55 INFO mapreduce.Task: Using ResourceCalculatorProcessTree: [ ]
20/11/30 13:30:55 INFO mapreduce.MapTask: Processing split: hdfs://localhost:9000/user/nguyennp/average_salary/input/salary.txt:0-488647
20/11/30 13:30:55 INFO mapreduce.MapTask: (EQS4704) 0 kv: 2021426(10485764)
20/11/30 13:30:55 INFO mapreduce.MapTask: mapreduce.task.io.sort.ab: 100
20/11/30 13:30:55 INFO mapreduce.MapTask: soft limit at 83886080
20/11/30 13:30:55 INFO mapreduce.MapTask: bufstart = 0; bufused = 104857600
20/11/30 13:30:55 INFO mapreduce.MapTask: kvstart = 2021426; length = 6553600
20/11/30 13:30:55 INFO mapreduce.MapTask: Map output collector class = org.apache.hadoop.mapreduce.MapTask$MapOutputBuffer
20/11/30 13:30:55 INFO mapreduce.LocalJobRunner:
20/11/30 13:30:55 INFO mapreduce.MapTask: Starting flush of map output
20/11/30 13:30:55 INFO mapreduce.MapTask: Spilling map output
20/11/30 13:30:55 INFO mapreduce.MapTask: bufstart = 0; bufend = 312342; bufused = 104857600
20/11/30 13:30:55 INFO mapreduce.MapTask: kvstart = 2021426(10485764); kverid = 20073676(104254704); length = 1407216553600
20/11/30 13:30:55 INFO mapreduce.MapTask: Finished split 0
20/11/30 13:30:55 INFO mapreduce.Task: Task(attempt_local196355271_0001_m_000000_0) is done. And is in the process of committing
20/11/30 13:30:55 INFO mapreduce.LocalJobRunner: map
20/11/30 13:30:55 INFO mapreduce.Task: Task(attempt_local196355271_0001_m_000000_0) done.
20/11/30 13:30:55 INFO mapreduce.Task: Final Counters for attempt_local196355271_0001_m_000000_0: Counters: 23
File System Counters
FILE: Number of bytes read=1037
FILE: Number of bytes written=408199
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=488647
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
Map-Reduce Framework

```

```

nguyenvp@DESKTOP-HTAZUMK:/mnt/hadoop/hadoop-2.7.0.1:
HDFS: Number of read operations=1
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
Map-Reduce Framework
  Map input records=35181
  Map output records=35181
  Map output bytes=322342
  Map output materialized bytes=100096
  Input split bytes=212
  Combine input records=35181
  Combine output records=10000
  Spilled Records=10000
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=207271266
  File Input Format Counters
    Bytes Read=400647
20/11/30 13:30:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local196355271_0001_m_000000_0
20/11/30 13:30:55 INFO mapred.LocalJobRunner: map task executor complete.
20/11/30 13:30:55 INFO mapred.LocalJobRunner: Waiting for reduce tasks.
20/11/30 13:30:55 INFO mapred.LocalJobRunner: Starting task: attempt_local196355271_0001_r_000000_0
20/11/30 13:30:55 INFO output.FileOutputCommitter: File OutputCommitter Algorithm version is 1
20/11/30 13:30:55 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup, temporary folders under output directory=false, ignore cleanup failures: false
20/11/30 13:30:55 INFO mapred.Task: Using ResourceCalculatorProcessFree: [ ]
20/11/30 13:30:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle95212a30
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: PurgeManager: memoryLeft=134338664, maxSingleShuffleLimit=43564416, mergeThreshold=238661392, ioSortFactor=10, numPurgeOutputsThreshold=10
20/11/30 13:30:55 INFO reduce.ReduceFetcher: attempt_local196355271_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion events
20/11/30 13:30:55 INFO reduce.ReduceFetcher: localFetcher#1 about to shuffle output of map attempt_local196355271_0001_m_000000_0 decomp: 100092 len: 100096 to PRIORITY
20/11/30 13:30:55 INFO reduce.ReduceTask: Read 100092 bytes from map-output for attempt_local196355271_0001_m_000000_0
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: closeInMemoryFile -> map-output of size: 100092, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, useMemory -> 100092
20/11/30 13:30:55 INFO reduce.ReduceTask: reduce: localFetcher: EventFetcher is interrupted.. Returning
20/11/30 13:30:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: FinalPurge called with 1 in-memory map-outputs and 0 on-disk map-outputs
20/11/30 13:30:55 INFO mapred.ReduceTask: Purging 1 sorted segments
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: Done to the last merge pass, with 1 segments left of total size: 100088 bytes
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: Purged 1 segments, 100092 bytes to disk to satisfy reduce memory limit
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: Purging 1 files, 100096 bytes from disk
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: Purging 0 segments, 0 bytes from memory into reduce
20/11/30 13:30:55 INFO mapred.ReduceTask: Purging 1 sorted segments
20/11/30 13:30:55 INFO reduce.PurgeManagerImpl: Done to the last merge pass, with 1 segments left of total size: 100088 bytes
20/11/30 13:30:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 13:30:55 INFO mapreduce.Job: map 100% reduce 0%
20/11/30 13:30:55 INFO Configuration.deprecation: mapred.skip or is deprecated. Instead, use mapreduce.job.skiprecords
20/11/30 13:30:56 INFO mapred.Task: Task attempt_local196355271_0001_r_000000_0 is done. And is in the process of committing
20/11/30 13:30:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
20/11/30 13:30:56 INFO mapred.Task: Task attempt_local196355271_0001_r_000000_0 is allowed to commit now
20/11/30 13:30:56 INFO output.FileOutputCommitter: Saved output of task 'attempt_local196355271_0001_r_000000_0' to hdfs://localhost:9000/user/nguyenvp/average_salary/output/1_temporary/0/task_local196355271_0001_r_000000
20/11/30 13:30:56 INFO mapred.LocalJobRunner: reduce > reduce
20/11/30 13:30:56 INFO mapred.Task: Task 'attempt_local196355271_0001_r_000000_0' done.
20/11/30 13:30:56 INFO mapred.Task: Final Counters for attempt_local196355271_0001_r_000000_0: Counters: 29
  File System Counters
    FILE: Number of bytes read=223821
    FILE: Number of bytes written=217095
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=400647
    HDFS: Number of bytes written=100096
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=1
  Map-Reduce Framework
    Combine input records=0
    Combine output records=0
    Reduce input groups=10000
    Reduce shuffle bytes=100096
    Reduce input records=10000
    Reduce output records=10000
    Spilled Records=10000
    Shuffled Maps=1
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=13
    Total committed heap usage (bytes)=207271266
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    MRIOO_LENGTH=0
    MRIOO_SIZE=0
    MRIOO_TIMEOUT=0
    WORMHOLE_EXPIRED=0
  File Output Format Counters
    Bytes Written=100096
20/11/30 13:30:56 INFO mapred.LocalJobRunner: Finishing task: attempt_local196355271_0001_r_000000_0
20/11/30 13:30:56 INFO mapred.LocalJobRunner: reduce task executor complete.
20/11/30 13:30:56 INFO mapreduce.Job: map 100% reduce 100%
20/11/30 13:30:56 INFO mapreduce.Job: Job job_local196355271_0001 completed successfully.
20/11/30 13:30:56 INFO mapreduce.Job: Counters: 35
  File System Counters
    FILE: Number of bytes read=224218
    FILE: Number of bytes written=223234
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=977264
    HDFS: Number of bytes written=242096
    HDFS: Number of read operations=11
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=0
  Map-Reduce Framework

```



```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$
2020/11/30 11:30:36 INFO mapred.LocalJobRunner: reduce task executor complete.
2020/11/30 11:30:36 INFO mapreduce.Job: map 100% reduce 100%
2020/11/30 11:30:36 INFO mapreduce.Job: Job job_local196335271_0001 completed successfully
2020/11/30 11:30:36 INFO mapreduce.Job: Counter: 35
File System Counters
  FILE: Number of bytes read=224218
  FILE: Number of bytes written=1325294
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
HDFS: Number of bytes read=977234
HDFS: Number of bytes written=148890
HDFS: Number of read operations=23
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
Map-Reduce Framework
  Map input records=35181
  Map output records=35181
  Map output bytes=327742
  Map output materialized bytes=588896
  Input split bytes=332
  Combine input records=35181
  Combine output records=10000
  Reduce input groups=10000
  Reduce shuffle bytes=108896
  Reduce input records=10000
  Reduce output records=10000
  Spilled Records=20000
  Shuffled Maps=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=11
  Total committed heap usage (bytes)=594542592
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  MISSING_TOKEN=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=488647
File Output Format Counters
  Bytes Written=148890
nguyennp@DESKTOP-H7AOLMK: ~$

```

- Kiểm tra file output

```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -ls average_salary/output
Found 2 items
-rw-r--r-- 1 nguyennp supergroup          0 2020-11-30 13:30 average_salary/output/_SUCCESS
-rw-r--r-- 1 nguyennp supergroup    148890 2020-11-30 13:30 average_salary/output/part-r-00000
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$

```

- In ra màn hình:

```

nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$
Note: AverageSalary.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
nguyennp@DESKTOP-H7AOLMK:/mnt/d/hadoop/hadoop-2.10.1$ bin/hdfs dfs -cat average_salary/output/*
550523.0
1 5691731.5
10 5580989.5
100 5485311.5
1000 5720692.0
10001 5438334.0
10002 5281194.5
10003 5177776.0
10004 5676880.0
10005 5322640.5
10006 5566971.5
10007 5634173.0
10008 5620362.0
10009 5706172.0
1001 5906040.5
10010 5387683.0
10011 5306607.0
10012 5680851.5
10013 5616622.5
10014 5489477.0
10015 5323622.5
10016 5622933.5
10017 5664125.5
10018 5307721.0
10019 5763780.0
1002 5472363.0
10020 5525321.0
10021 5597401.0
10022 5487273.0
10023 5532187.0
10024 5507319.5
10025 5585332.0
10026 5784894.0
10027 5232237.5
10028 5415307.5
10029 5726486.5
1003 5573335.5
10030 5645319.5
10031 5508847.5
10032 5487920.5
10033 5522372.0
10034 5908427.0
10035 5980289.5
10036 5367468.5
10037 5668678.5
10038 5479065.0
10039 5320731.5
1004 5483268.5
10040 5730319.0

```

## g. Assignment 8 – Music Track Program

### i. Number of unique listeners

- File *UniqueListeners.java* chứa code của chương trình:

```
public class UniqueListeners
{
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
    {
        ...
    }

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        ...
    }

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

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
{
    private Text TrackId = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
    {
        // Tách Token theo dấu ,
        StringTokenizer tokens = new StringTokenizer(value.toString(), ",");
        // Lay 2 token dau tien (trackID va UserID)
        int UserId = Integer.parseInt(tokens.nextToken().trim());
        TrackId.set(tokens.nextToken().trim() + ",");
        context.write(TrackId, new IntWritable(UserId));
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
{
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
    {
        Set<Integer> UserIdSet = new HashSet<Integer>();

        for(IntWritable value:values)
        {
            // Add userID vào Set
            UserIdSet.add(value.get());
        }
        // Return size of Set
        context.write(key, new IntWritable(UserIdSet.size()));
    }
}
```



```

public static void main(String[] args) throws Exception
{
    Configuration conf = new Configuration();
    Job job = new Job(conf, "UniqueListeners");
    job.setJarByClass(UniqueListeners.class);

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    System.exit(job.waitForCompletion(true) ? 0 : 1);
}

```

- Biên dịch chương trình:

```

$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ ls
UniqueListeners.java
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ hadoop com.sun.tools.javac.Main UniqueListeners.java
Note: UniqueListeners.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ ls
UniqueListeners$Map.class UniqueListeners$Reduce.class UniqueListeners.class UniqueListeners.java
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners

```

- Compile thành file **.jar**

```

$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ jar cf UniqueListeners.jar *.class
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ ls
UniqueListeners$Map.class UniqueListeners.class UniqueListeners.java
UniqueListeners$Reduce.class UniqueListeners.jar
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners

```

- Chạy trên dfs:

```

$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ hadoop jar UniqueListeners.jar UniqueListeners /sriMR/MusicTrackProgram.csv /sriMR/UniqueListenersOutput

```

- Kết quả trả về:

```

Bytes Written=357
$ cd /mnt/c:/U/N/Hadoop_MapReduce/Mu/UniqueListeners
$ hadoop fs -cat /sriMR/UniqueListenersOutput/part-r-00000
250, 5
251, 2
252, 2
253, 6
254, 3
255, 6
256, 4
257, 7
258, 3
259, 5
260, 6
261, 4

```

ii. Number of times the track was shared with others

- File java chứa chương trình:

```

16 |
17 | public class NumOfShared
18 | {
19 |     public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
20 |     { ...
34 |     }
35 |
36 |     public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
37 |     { ...
50 |     }
51 |
52 |     public static void main(String[] args) throws Exception
53 |     { ...
70 |     }
71 | }

```

---

```

    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
    {
        private Text TrackId = new Text();

        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
        {
            StringTokenizer tokens = new StringTokenizer(value.toString(), ",");

            // Bỏ qua token đầu tiên (UserID)
            tokens.nextToken();
            // Token thứ 2-3 là TrackID và số lượt share (shared)
            TrackId.set(tokens.nextToken().trim() + ",");
            int shared = Integer.parseInt(tokens.nextToken().trim());
            context.write(TrackId, new IntWritable(shared));
        }
    }

```

---

```

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
        {
            int count = 0;

            for(IntWritable value: values)
            {
                // Tính tổng số lượt Share (shared) của mỗi trackID
                count = count + value.get();
            }

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

```

---

```

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
        {
            int count = 0;

            for(IntWritable value: values)
            {
                // Tính tổng số lượt Share (shared) của mỗi trackID
                count = count + value.get();
            }

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

```

- Biên dịch và nén thành file **.jar**:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 hadoop com.sun.tools.javac.Main NumOfShared.java
Note: NumOfShared.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 ls
NOS.txt 'NumOfShared$Map.class' 'NumOfShared$Reduce.class' NumOfShared.class NumOfShared.java
$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 jar cf NumOfShared.jar *.class
$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 ls
NOS.txt 'NumOfShared$Map.class' 'NumOfShared$Reduce.class' NumOfShared.class NumOfShared.jar NumOfShared.java
$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 |

```

- Chạy trên dfs:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 hadoop jar NumOfShared.jar NumOfShared /sriMR/MusicTrack^
kProgram.csv /sriMR/numOfSharedOutputs
2020-11-30 19:23:28,993 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-30 19:23:29,063 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-30 19:23:29,063 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-30 19:23:29,320 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool
al interface and execute your application with ToolRunner to remedy this.

```

- Kết quả:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfShared $ master +5 !8 hadoop fs -cat /sriMR/numOfSharedOutputs/part-r-00000
250, 0
251, 2
252, 2
253, 6
254, 2
255, 7
256, 2
257, 8
258, 1
259, 3
260, 2
261, 4

```

iii. Number of times the track was listened to on the radio

- File java chứa code của chương trình:

```

public class NumOfListenOnRadio
{
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>{...
    }
    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>{ ...
    }
    public static void main(String[] args) throws Exception{...
    }
}

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>{
    private Text TrackId = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException,InterruptedException
    {
        // Tách token theo dấu ,
        StringTokenizer tokens = new StringTokenizer(value.toString(), ",");
        // Bỏ qua 1 token
        tokens.nextToken();
        // Set trackID là Token thứ 2
        TrackId.set(tokens.nextToken().trim() + ",");
        // Bỏ qua token thứ 3
        tokens.nextToken();
        // Set IsListenOnRadio là token thứ 4
        int IsListenOnRadio = Integer.parseInt(tokens.nextToken().trim());
        // Return trackID và IsListenOnRadio
        context.write(TrackId, new IntWritable(IsListenOnRadio));
    }
}

```

```

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>{
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
    {
        int count = 0;

        for(IntWritable value:values)
        {
            // Tính tổng số lượt nghe trên Radio (IsListenOnRadio) của TrackID
            count = count + value.get();
        }

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

```

- Biên dịch chương trình và nén thành file **.jar**:

```

$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !8 hadoop com.sun.tools.javac.Main NumOfListenOnRadio.java
Note: NumOfListenOnRadio.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !11 jar cf NumOfListenOnRadio.jar *.class
$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !12 ls
NOLR.txt                               'NumOfListenOnRadio$Reduce.class'   NumOfListenOnRadio.jar
'NumOfListenOnRadio$Map.class'        NumOfListenOnRadio.class             NumOfListenOnRadio.java
$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !12 |

```

- Chạy trên dfs:

```

$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !12 hadoop jar NumOfListenOnRadio.jar NumOfListenOnRadio /sriMR/MusicTrackProgram.csv /sriMR/numOfListeneOnRadioOutput
2020-11-30 19:27:42,196 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-30 19:27:42,257 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-30 19:27:42,257 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-30 19:27:42,420 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2020-11-30 19:27:42,546 INFO input.FileInputFormat: Total input files to process : 1

```

- Kết quả:

```

$ /mnt/c:/N/Hadoop_MapReduce/NumOfListenOnRadio master +5 !12 hadoop fs -cat /sriMR/numOfListeneOnRadioOutput
/part-r-00000
250, 0
251, 2
252, 2
253, 3
254, 4
255, 7
256, 2
257, 4
258, 3
259, 3
260, 3

```

#### iv. Number of times the track was listened to in total

- File java chứa code của chương trình:

```

public class NumOfListenInTotal
{
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
    {
        ...
    }

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        ...
    }

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

```



```

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
{
    private Text TrackId = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
    {
        // Tách token theo dấu ,
        StringTokenizer tokens = new StringTokenizer(value.toString(), ",");
        // gan UserID va Track ID la 2 token dau tien
        int UserID = Integer.parseInt(tokens.nextToken().trim());
        TrackId.set(tokens.nextToken().trim() + ",");
        // return UserID va TrackID
        context.write(TrackId, new IntWritable(UserID));
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
{
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
    {
        int count = 0;

        for(IntWritable value:values)
        {
            // Tính tổng số Lượt nghe của bài hát (UserID)
            count = count + 1;
        }
        context.write(key, new IntWritable(count));
    }
}

```

- Biên dịch và tạo file **.jar**:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfListenInTotal
$ javac -d . NumOfListenInTotal.java
Note: NumOfListenInTotal.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
$ jar cf NumOfListenInTotal.jar *.class
ls
NumOfListenInTotal$Map.class  NumOfListenInTotal.class  NumOfListenInTotal.java
$

```

- Chạy trên dfs:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfListenInTotal
$ hadoop jar NumOfListenInTotal.jar NumOfListenInTotal /sriMR/MusicTrackProgram.csv /sriMR/numOfListeneInTotalOuput
2020-11-30 19:36:19,180 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-30 19:36:19,245 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-30 19:36:19,245 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-30 19:36:19,415 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2020-11-30 19:36:19,536 INFO input.FileInputFormat: Total input files to process : 1

```

- Kết quả:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/Mu/NumOfListenInTotal
$ hadoop fs -cat /sriMR/numOfListeneInTotalOuput/part-r-00000
250, 5
251, 3
252, 4
253, 9
254, 5
255, 12
256, 4
257, 10

```

v. Number of times the track was skipped on the radio

- File java chứa code của chương trình:

```

public class NumOfSkipOnRadio
{
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
    {
        ...
    }

    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        ...
    }

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

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable>
{
    private Text TrackId = new Text();

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
    {
        StringTokenizer tokens = new StringTokenizer(value.toString(), ",");
        // bỏ qua token đầu tiên
        tokens.nextToken();
        TrackId.set(tokens.nextToken().trim() + ",");
        // bỏ qua token thứ 3 và 4
        tokens.nextToken();
        tokens.nextToken();
        int skipped = Integer.parseInt(tokens.nextToken().trim());

        context.write(TrackId, new IntWritable(skipped));
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
{
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
    {
        int count = 0;

        for(IntWritable value: values)
        {
            count = count + 1;
        }

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

```

#### - Biên dịch chương trình:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/NumOfSkipOnRadio
$ hadoop com.sun.tools.javac.Main NumOfSkipOnRadio.java
Note: NumOfSkipOnRadio.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
$ cd /mnt/c:/N/Hadoop_MapReduce/NumOfSkipOnRadio
$ jar cf NumOfSkipOnRadio.jar *.class
$ cd /mnt/c:/N/Hadoop_MapReduce/NumOfSkipOnRadio
$ ls
NOSR.txt      'NumOfSkipOnRadio$Reduce.class'  NumOfSkipOnRadio.jar
'NumOfSkipOnRadio$Map.class'      NumOfSkipOnRadio.class          NumOfSkipOnRadio.java
$ cd /mnt/c:/N/Hadoop_MapReduce/NumOfSkipOnRadio
$

```

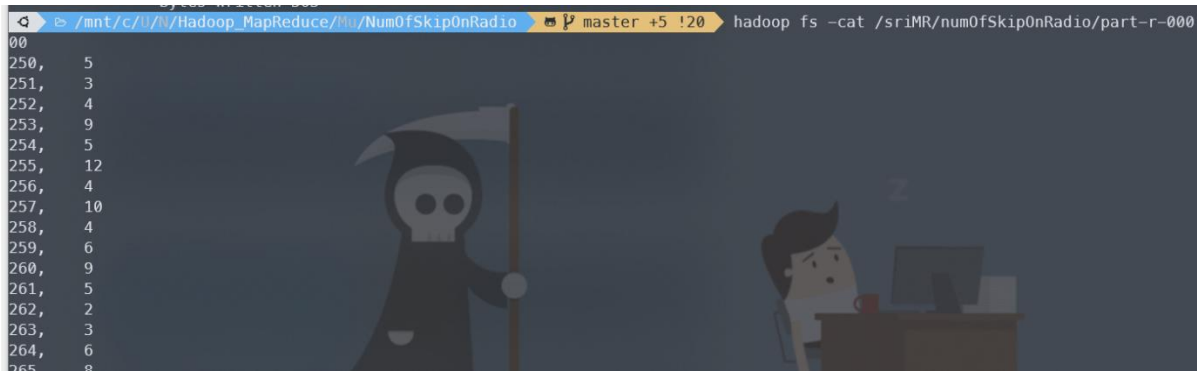
#### - Chạy trên dfs:

```

$ cd /mnt/c:/N/Hadoop_MapReduce/NumOfSkipOnRadio
$ hadoop jar NumOfSkipOnRadio.jar NumOfSkipOnRadio
2020-11-30 19:44:59,466 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-30 19:44:59,556 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-30 19:44:59,556 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-30 19:44:59,728 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2020-11-30 19:44:59,814 INFO input.FileInputFormat: Total input files to process : 1
2020-11-30 19:44:59,864 INFO mapreduce.JobSubmitter: number of splits:1

```

#### - Kết quả:



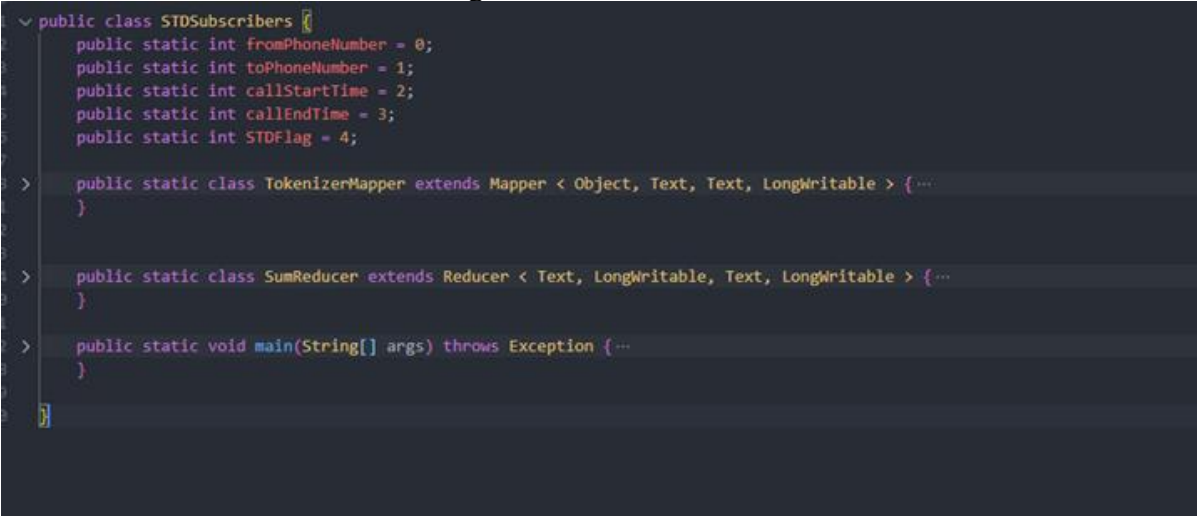
```

hadoop fs -cat /sriMR/numOfSkipOnRadio/part-r-000
00
250, 5
251, 3
252, 4
253, 9
254, 5
255, 12
256, 4
257, 10
258, 4
259, 6
260, 9
261, 5
262, 2
263, 3
264, 6
265, 8

```

## h. Assignment 9 - Telecom Call Data Record Program

- File Java chứa code của chương trình:



```

public class STDSubscribers {
    public static int fromPhoneNumber = 0;
    public static int toPhoneNumber = 1;
    public static int callStartTime = 2;
    public static int callEndTime = 3;
    public static int STDFlag = 4;

    public static class TokenizerMapper extends Mapper < Object, Text, Text, LongWritable > { ...
    }

    public static class SumReducer extends Reducer < Text, LongWritable, Text, LongWritable > { ...
    }

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

```

```

public static class TokenizerMapper extends Mapper < Object, Text, Text, LongWritable > {
    Text phoneNumber = new Text();
    LongWritable durationInMinutes = new LongWritable();
    public void map(Object key, Text value, Mapper < Object, Text, Text, LongWritable > .Context context)
        throws IOException,
        InterruptedException {
        // tách token theo kí tự |
        String[] parts = value.toString().split("[|]");

        // Xet nhung mau du lieu co STDFlag = 1
        if (parts[STDSubscribers.STDFlag].equalsIgnoreCase("1")) {
            // gan cac gia tri cho cac bien phoneNumber, callStartTime, callEndTime
            phoneNumber.set(parts[STDSubscribers.fromPhoneNumber]);
            String callEndTime = parts[STDSubscribers.callEndTime];
            String callStartTime = parts[STDSubscribers.callStartTime];
            // Tinh duration bang ham toMillis
            long duration = toMillis(callEndTime) - toMillis(callStartTime);
            // Chuyen ms thanh minutes
            durationInMinutes.set(duration / (1000 * 60));
            context.write(phoneNumber, durationInMinutes);
        }
    }
    // ham toMillis chuyen date thanh ms
    private long toMillis(String date) {
        SimpleDateFormat format = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
        Date dateFrm = null;
        try {
            dateFrm = format.parse(date);
        } catch (ParseException e) {
            e.printStackTrace();
        }
        return dateFrm.getTime();
    }
}

public static class SumReducer extends Reducer < Text, LongWritable, Text, LongWritable > {
    private LongWritable result = new LongWritable();
    public void reduce(Text key, Iterable < LongWritable > values, Reducer < Text, LongWritable, Text, LongWritable > .Context context)
        throws IOException,
        InterruptedException {
        long sum = 0;
        for (LongWritable val: values) {
            // tinh tong so phut goi di (durationInMinutes) cua moi phoneNumber
            sum += val.get();
        }
        this.result.set(sum);
        if (sum >= 60) {
            // Ghi cac phoneNumber co tong so phut (durationInMinutes) >= 60
            context.write(key, this.result);
        }
    }
}

```

#### - Biên dịch chương trình và tạo file **.jar**:

```

~/hadoop/hadoop-3.3.0/A9 hadoop com.sun.tools.javac.Main STDSubscribers.java
Note: STDSubscribers.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
~/hadoop/hadoop-3.3.0/A9 jar cf CDR.jar *.class
~/hadoop/hadoop-3.3.0/A9 ls
CDR.jar 'STDSubscribers$SumReducer.class' STDSubscribers.class
CDRlog.txt 'STDSubscribers$TokenizerMapper.class' STDSubscribers.java
~/hadoop/hadoop-3.3.0/A9 |

```

#### - Put dữ liệu lên hdfs:



```
~/hadoop/hadoop-3.3.0/A9 hadoop fs -put CDRlog.txt /sriMR/CDR
~/hadoop/hadoop-3.3.0/A9 |
```

- Chạy trên dfs:

```
~/hadoop/hadoop-3.3.0/A9 hadoop jar CDR.jar STDSubscribers /sriMR/CDR /sriMR/STDSubscribersOutput
2020-11-30 19:51:54,585 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-30 19:51:54,687 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-30 19:51:54,687 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-30 19:51:54,821 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
ol interface and execute your application with ToolRunner to remedy this.
2020-11-30 19:51:54,880 INFO input.FileInputFormat: Total input files to process : 1
2020-11-30 19:51:54,928 INFO mapreduce.JobSubmitter: number of splits:1
2020-11-30 19:51:55,003 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local337252371_0001
2020-11-30 19:51:55,004 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-11-30 19:51:55,118 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
```

- Kết quả:

```
~/hadoop/hadoop-3.3.0/A9 hadoop fs -cat /sriMR/STDSubscribersOutput/part-r-00000
9665128505 68
9665128506 64
9665128507 64
~/hadoop/hadoop-3.3.0/A9 |
```

## IV. TÀI LIỆU THAM KHẢO

- Sriram Balasubramanian, *Hadoop MapReduce Lab*, Cloudera, 2016.
- Một số Dataset mẫu.
- Source code