Trường Đại học Khoa học Tự nhiên

Khoa Công nghệ thông tin

Môn Dữ liệu lớn

BÁO CÁO LAB 02

CÀI ĐẶT CÁC CHƯƠNG TRÌNH MAPREDUCE TRÊN HADOOP

GVLT: Dr. Nguyễn Ngọc Thảo

GVHD: Msc. Lê Ngọc Thành

TP. Hồ Chí Minh, ngày 10 tháng 10 năm 2019

Mục lục

I. THÔNG TIN NHÓM

Nhóm 1:

- Võ Nhật Vinh – 1612815
- Hồng Thanh Hoài – 1612855
- Huỳnh Minh Huấn – 1612858

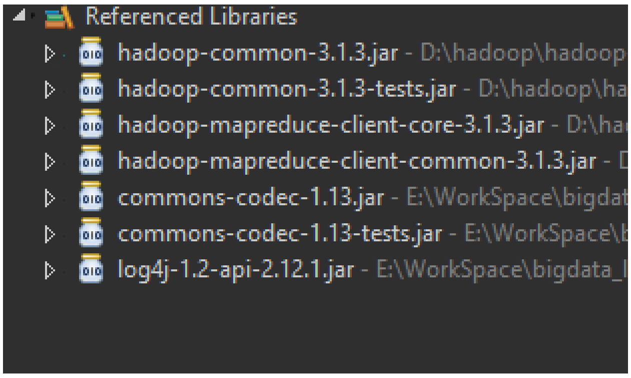
II. BÁO CÁO TỔNG QUÁT

Sinh viên hoàn thành đầy đủ các câu từ 1-9

III. BÁO CÁO CHI TIẾT

Ghi chú: nhóm em không sử dụng Hadoop trên cloudera (Hadoop 2.6) mà sử dụng Hadoop trên máy (Hadoop 3.1)

Các external jars nhóm sử dụng



```
Referenced Libraries
  hadoop-common-3.1.3.jar - D:\hadoop\hadoop
  hadoop-common-3.1.3-tests.jar - D:\hadoop\ha
  hadoop-mapreduce-client-core-3.1.3.jar - D:\ha
  hadoop-mapreduce-client-common-3.1.3.jar - D
  commons-codec-1.13.jar - E:\WorkSpace\bigdat
  commons-codec-1.13-tests.jar - E:\WorkSpace\b
  log4j-1.2-api-2.12.1.jar - E:\WorkSpace\bigdata_l
```

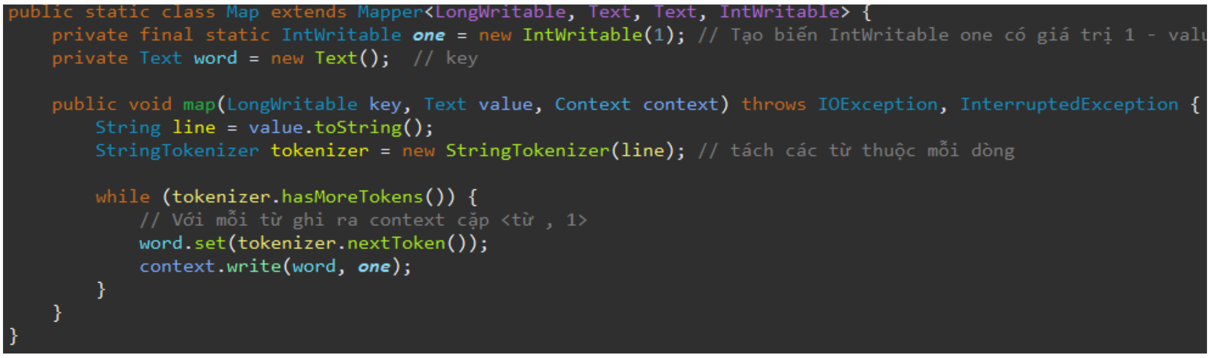
Các files source code - dataset: 16 files source code – 9 files dataset

- Assignment 01: WordCount.java – wordcount.txt
- Assignment 02: WordSizeWordCount.java – alphabets.txt
- Assignment 03: WeatherData.java – weather_data.txt
- Assignment 04: Patent.java – patent.txt
- Assignment 05: MaxTemp.java – MaxTemp.csv
- Assignment 06: AverageSalary.java – salary.csv

- Assignment 07: DeIdentifyHealthcare.java - DeIdentifyHealthcare.txt
  Assignment 08: Assignment08.java - MusicTrackRecord.csv
  LastFMConstant.java
  UniqueListener.java (task 01)
  SharedTrack.java (task 02)
  ListenedTrack.java (task 03)
  ListenedTotalTrack.java (task 04)
  SkippedOnRadioTrack.java (task 05)
- Assignment 09: CallDataRecord.java - CDRlog.txt
  CDRConstants.java

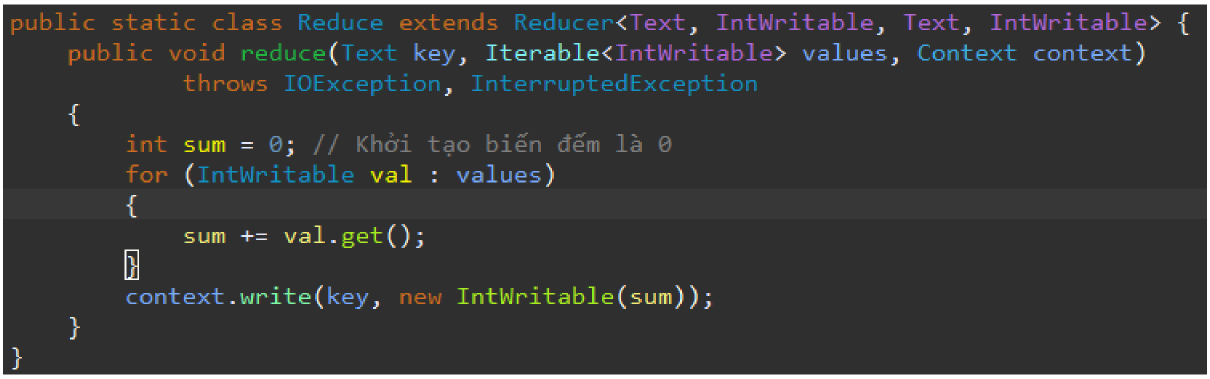
1. Assignment 01 – Word count

- Tạo lớp Map kế thừa từ lớp Mapper của Hadoop.mapreduce. Trong lớp Map, xây dựng phương thức map.



```java
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1); // Tạo biến IntWritable one có giá trị 1 - valu
    private Text word = new Text();  // key

    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line); // tách các từ thuộc mỗi dòng

        while (tokenizer.hasMoreTokens()) {
            // Với mỗi từ ghi ra context cặp <từ , 1>
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
}
```

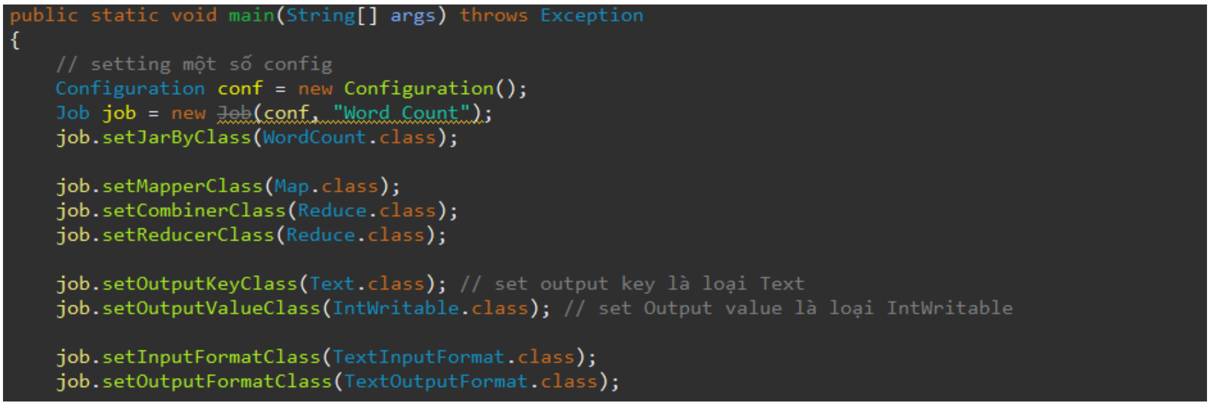
  - Phương thức map có đầu vào value sau đó được chuyển về kiểu chuỗi. mỗi chuỗi ứng với một dòng văn bản.
  - Tiến hành tách các từ trong chuỗi đó bằng Lớp StringTokenizer.
  - Lặp với các token tìm được ở trên, với mỗi token, context trả ra cặp <key, value> là <word, one>. One là biến IntWritable có **giá trị 1** được định nghĩa ở đầu lớp Map.
- Tạo lớp Reduce kế thừa lớp Reducer của Hadoop.mapreduce. Trong lớp Reduce, xây dựng phương thức reduce.



```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
            throws IOException, InterruptedException
    {
        int sum = 0; // Khởi tạo biến đếm là 0
        for (IntWritable val : values)
        {
            sum += val.get();
        }
        context.write(key, new IntWritable(sum));
    }
}
```

- o Phương thức reduce có input là key, values (kiểu Iterable<IntWritable> - Iterable là một interface giúp object thực hiện lệnh forEach loop) nhận từ pha map và context.
- o forEach loop trên values để lấy tổng các giá trị của key tương ứng.
- o context ghi ra cặp <key, value>. Key là từ, value là số lượng ứng với từ đó trong dataset.

- Ở hàm main, ta sẽ khởi tạo các đối tượng cần thiết như
  - o Thiết lập OutputKeyClass là Text.
  - o OutputValueClass là IntWritable.



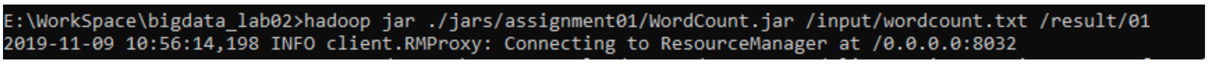
```
public static void main(String[] args) throws Exception
{
    // setting một số config
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Word Count");
    job.setJarByClass(WordCount.class);

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

    job.setOutputKeyClass(Text.class); // set output key là loại Text
    job.setOutputValueClass(IntWritable.class); // set Output value là loại IntWritable

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
```

- Thực thi:
  - o Copy dataset vào hdfs:

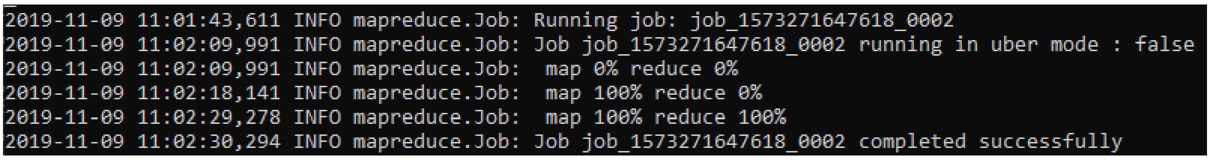
    hadoop fs -copyFromLocal ./Dataset/WordCount.txt /input/wordcount.txt
  - o Gõ lệnh thực thi file jar mapreduce:

  ➔ hadoop jar <đường dẫn file jar WordCount> <input file wordcount.txt> <output>

  ➔ hadoop jar ./jars/assignment01/WordCount.jar /input/wordcount.txt /result/01

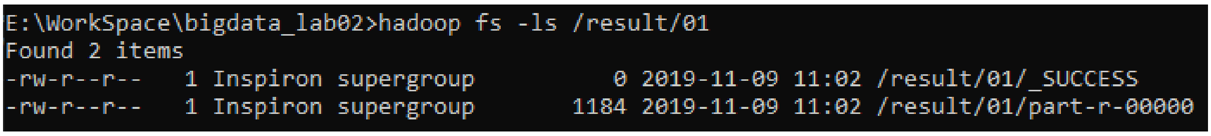


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment01/WordCount.jar /input/wordcount.txt /result/01
2019-11-09 10:56:14,198 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
```

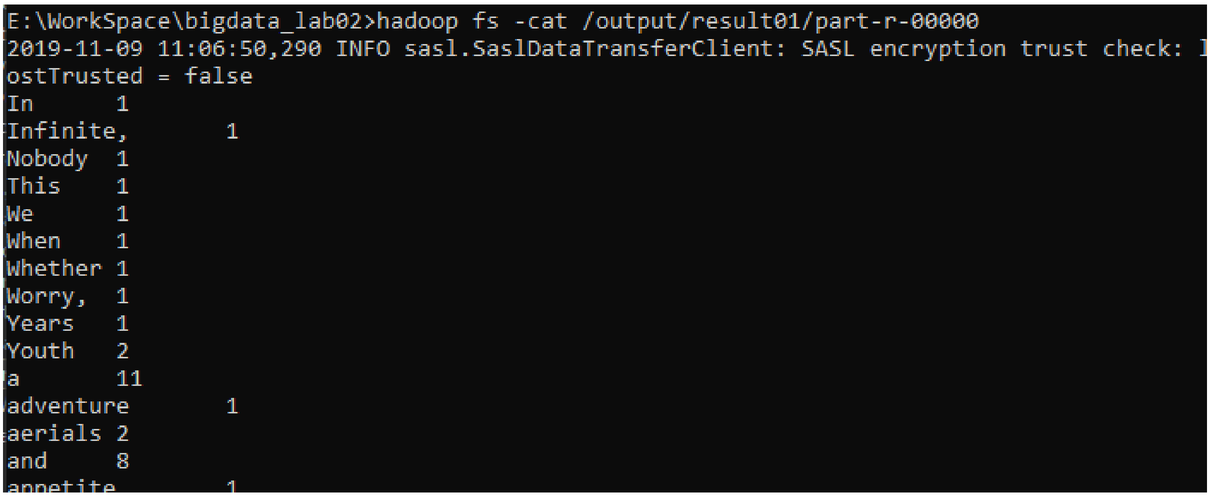


```
2019-11-09 11:01:43,611 INFO mapreduce.Job: Running job: job_1573271647618_0002
2019-11-09 11:02:09,991 INFO mapreduce.Job: Job job_1573271647618_0002 running in uber mode : false
2019-11-09 11:02:09,991 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 11:02:18,141 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 11:02:29,278 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 11:02:30,294 INFO mapreduce.Job: Job job_1573271647618_0002 completed successfully
```

- Kết quả:



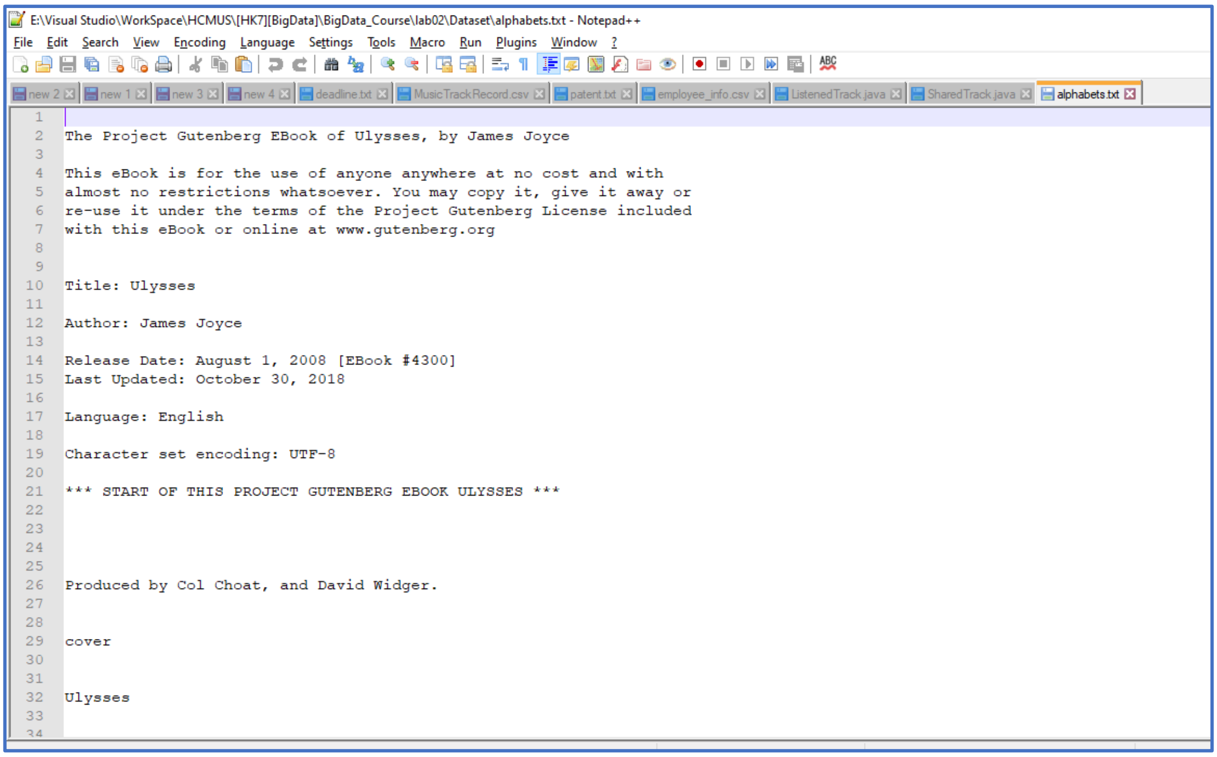
```
E:\WorkSpace\bigdata_lab02>hadoop fs -ls /result/01
Found 2 items
-rw-r--r--   1 Inspiron supergroup          0 2019-11-09 11:02 /result/01/_SUCCESS
-rw-r--r--   1 Inspiron supergroup       1184 2019-11-09 11:02 /result/01/part-r-00000
```



```
E:\WorkSpace\bigdata_lab02>hadoop fs -cat /output/result01/part-r-00000
2019-11-09 11:06:50,290 INFO sasl.SaslDataTransferClient: SASL encryption trust check: l
ostTrusted = false
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
```

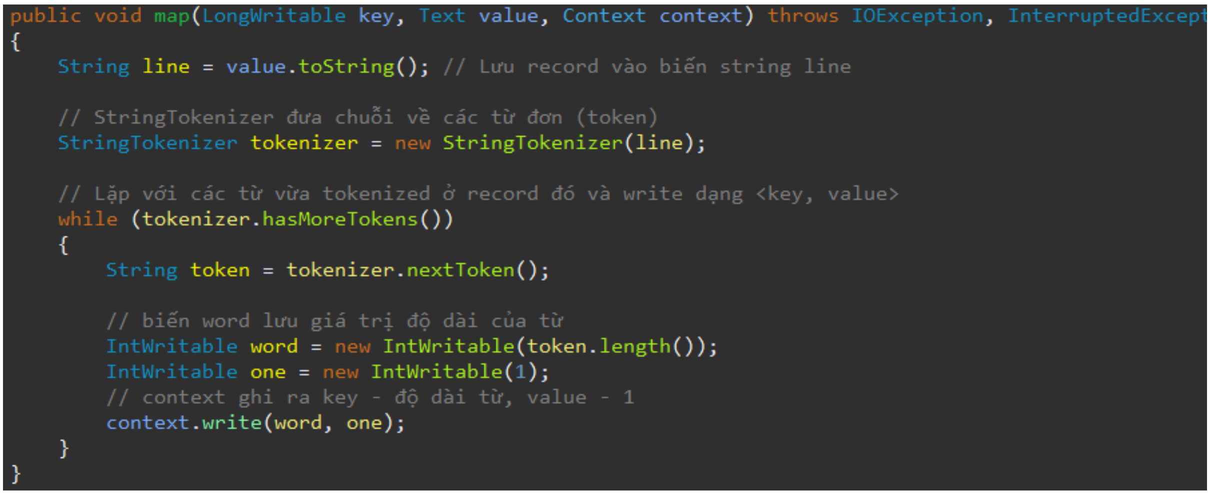
2. Assignment 02 – Word size

- Dataset file "alphabets.txt" (http://www.gutenberg.org/files/4300/4300-0.txt - đổi tên theo đúng đề bài.)


E:\Visual Studio\WorkSpace\HCMUS\[HK7][BigData]\BigData_Course\lab02\Dataset\alphabets.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 2
new 1
new 3
new 4
deadline.txt
MusicTrackRecord.csv
patent.txt
employee_info.csv
ListenedTrack.java
SharedTrack.java
alphabets.txt

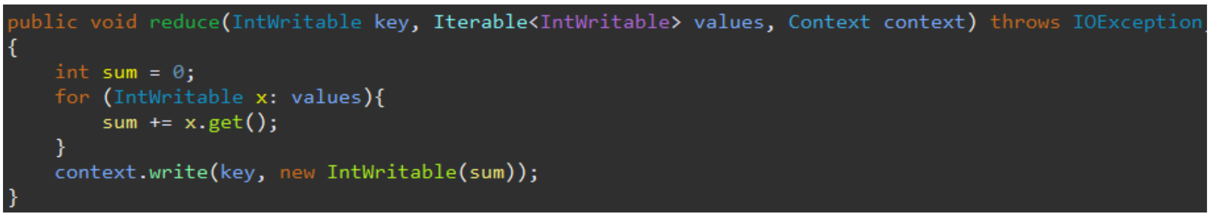
```

The Project Gutenberg EBook of Ulysses, by James Joyce

This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever. You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.org


Title: Ulysses

Author: James Joyce

Release Date: August 1, 2008 [EBook #4300]
Last Updated: October 30, 2018

Language: English

Character set encoding: UTF-8

*** START OF THIS PROJECT GUTENBERG EBOOK ULYSSES ***




Produced by Col Choat, and David Widger.


cover


Ulysses

```

- Phương thức map
  - Tách token với mỗi record
  - Với mỗi token ở trên, context ghi ra key – độ dài từ, value – 1 (1 lần từ có độ dài đó).



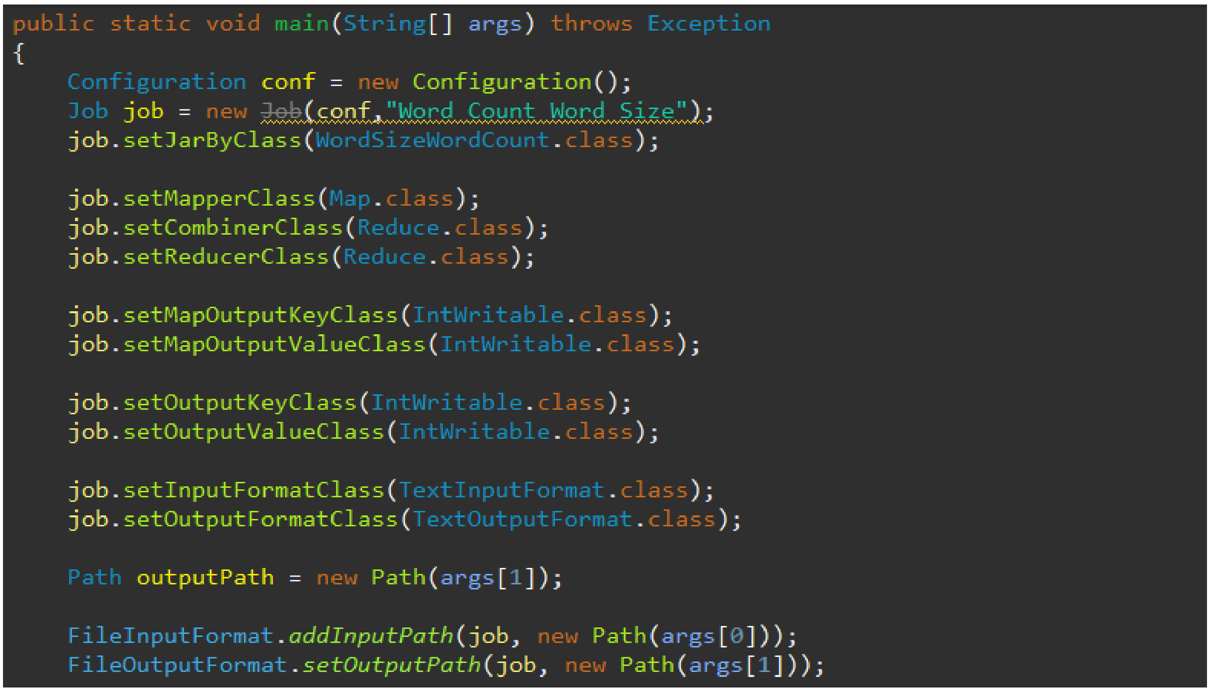
```java
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedExcept
{
    String line = value.toString(); // Lưu record vào biến string line

    // StringTokenizer đưa chuỗi về các từ đơn (token)
    StringTokenizer tokenizer = new StringTokenizer(line);

    // Lặp với các từ vừa tokenized ở record đó và write dạng <key, value>
    while (tokenizer.hasMoreTokens())
    {
        String token = tokenizer.nextToken();

        // biến word lưu giá trị độ dài của từ
        IntWritable word = new IntWritable(token.length());
        IntWritable one = new IntWritable(1);
        // context ghi ra key - độ dài từ, value - 1
        context.write(word, one);
    }
}
```

- Phương thức reduce
  - Với mỗi key, tính tổng số lần xuất hiện độ dài đó.
  - Context ghi ra key, số lần xuất hiện.



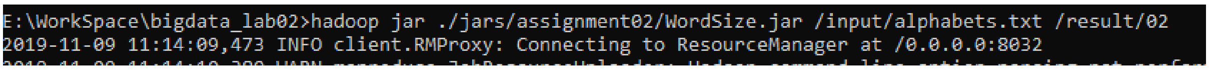
```
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context) throws IOException
{
    int sum = 0;
    for (IntWritable x: values){
        sum += x.get();
    }
    context.write(key, new IntWritable(sum));
}
```

- Hàm main:
    - Thiết lập MapOutputKeyClass là IntWritable, MapOutputValueClass là IntWritable.
    - Thiết lập OutputKeyClass là IntWritable, OutputValueClass là IntWritable.

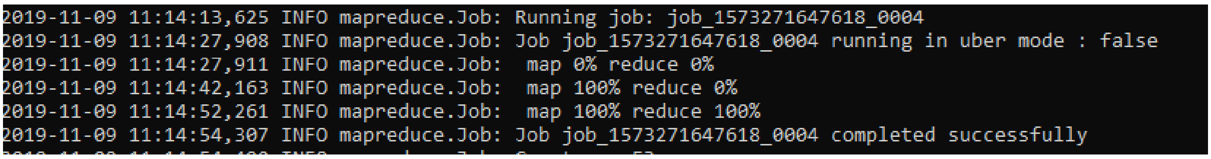


```
public static void main(String[] args) throws Exception
{
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf,"Word Count Word Size");
    job.setJarByClass(WordSizeWordCount.class);

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

    job.setMapOutputKeyClass(IntWritable.class);
    job.setMapOutputValueClass(IntWritable.class);

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

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);

    Path outputPath = new Path(args[1]);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
```

- Thực thi:
    - Copy dataset vào hdfs:
    - ➔ hadoop fs -copyFromLocal ./Dataset/alphabets.txt /input/alphabets.txt
    - Gõ lệnh thực thi file jar thực hiện mapreduce
    - ➔ hadoop jar ./jars/assignment02/WordSize.jar /input/alphabets /result/02

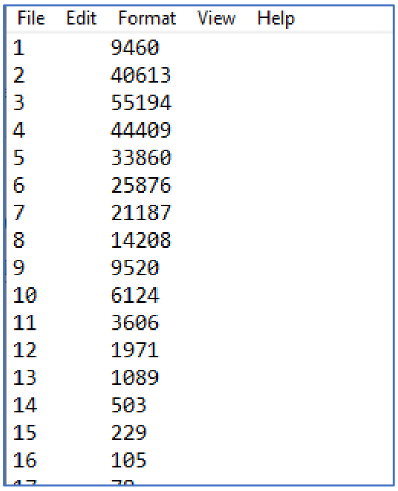


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment02/WordSize.jar /input/alphabets.txt /result/02
2019-11-09 11:14:09,473 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
```



```
2019-11-09 11:14:13,625 INFO mapreduce.Job: Running job: job_1573271647618_0004
2019-11-09 11:14:27,908 INFO mapreduce.Job: Job job_1573271647618_0004 running in uber mode : false
2019-11-09 11:14:27,911 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 11:14:42,163 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 11:14:52,261 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 11:14:54,307 INFO mapreduce.Job: Job job_1573271647618_0004 completed successfully
```

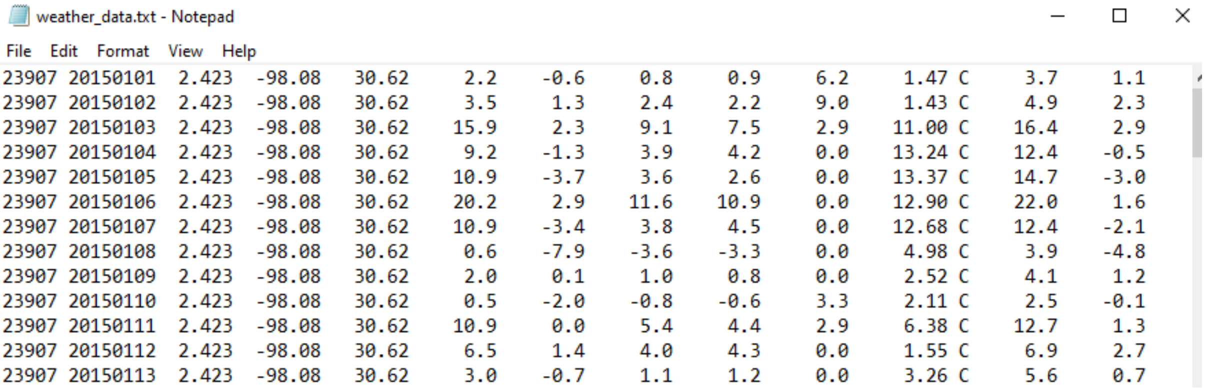
- Kết quả



```
1       9460
2       40613
3       55194
4       44409
5       33860
6       25876
7       21187
8       14208
9       9520
10      6124
11      3606
12      1971
13      1089
14      503
15      229
16      105
```

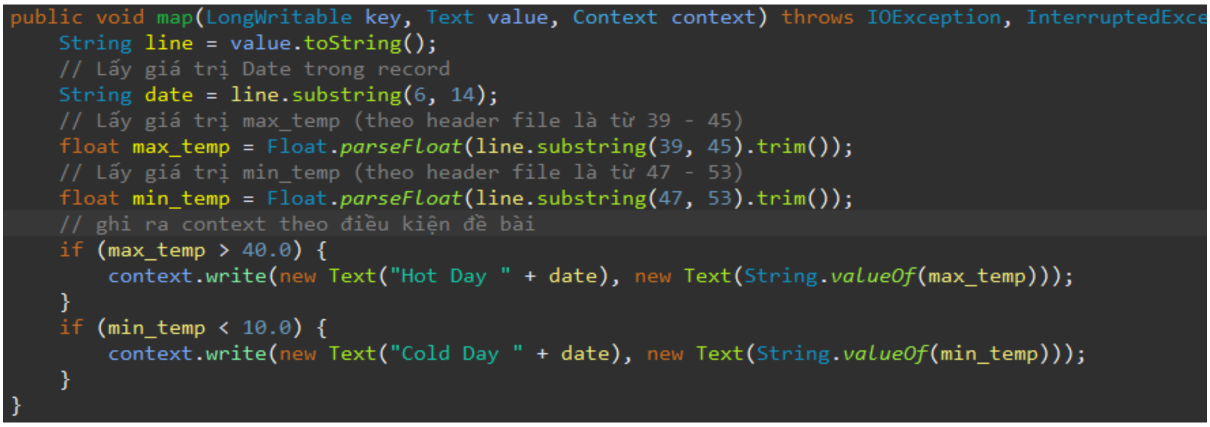
3. Assignment 03 – Weather data

- Dataset: sử dụng data do giáo viên cung cấp.



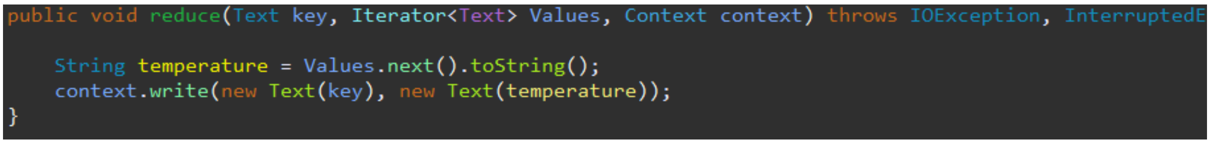
```
23907 20150101  2.423  -98.08   30.62    2.2   -0.6    0.8    0.9    6.2    1.47 C    3.7    1.1
23907 20150102  2.423  -98.08   30.62    3.5    1.3    2.4    2.2    9.0    1.43 C    4.9    2.3
23907 20150103  2.423  -98.08   30.62   15.9    2.3    9.1    7.5    2.9   11.00 C   16.4    2.9
23907 20150104  2.423  -98.08   30.62    9.2   -1.3    3.9    4.2    0.0   13.24 C   12.4   -0.5
23907 20150105  2.423  -98.08   30.62   10.9   -3.7    3.6    2.6    0.0   13.37 C   14.7   -3.0
23907 20150106  2.423  -98.08   30.62   20.2    2.9   11.6   10.9    0.0   12.90 C   22.0    1.6
23907 20150107  2.423  -98.08   30.62   10.9   -3.4    3.8    4.5    0.0   12.68 C   12.4   -2.1
23907 20150108  2.423  -98.08   30.62    0.6   -7.9   -3.6   -3.3    0.0    4.98 C    3.9   -4.8
23907 20150109  2.423  -98.08   30.62    2.0    0.1    1.0    0.8    0.0    2.52 C    4.1    1.2
23907 20150110  2.423  -98.08   30.62    0.5   -2.0   -0.8   -0.6    3.3    2.11 C    2.5   -0.1
23907 20150111  2.423  -98.08   30.62   10.9    0.0    5.4    4.4    2.9    6.38 C   12.7    1.3
23907 20150112  2.423  -98.08   30.62    6.5    1.4    4.0    4.3    0.0    1.55 C    6.9    2.7
23907 20150113  2.423  -98.08   30.62    3.0   -0.7    1.1    1.2    0.0    3.26 C    5.6    0.7
```

- Phương thức map
  - Lấy các giá trị trong record ngày, nhiệt độ cao nhất, nhiệt độ thấp nhất.
  - Context ghi ra theo yêu cầu đề bài.



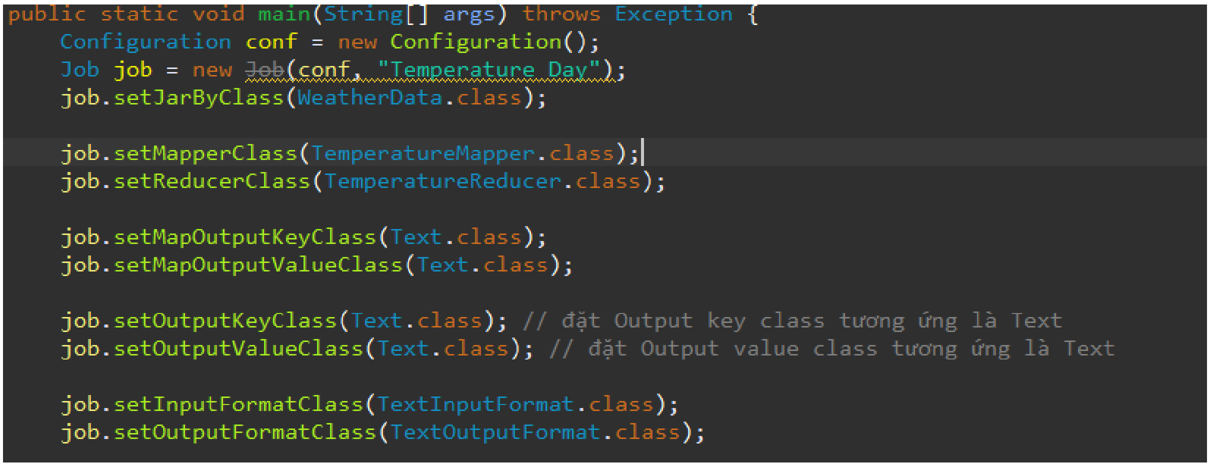
```
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedExce
    String line = value.toString();
    // Lấy giá trị Date trong record
    String date = line.substring(6, 14);
    // Lấy giá trị max_temp (theo header file là từ 39 - 45)
    float max_temp = Float.parseFloat(line.substring(39, 45).trim());
    // Lấy giá trị min_temp (theo header file là từ 47 - 53)
    float min_temp = Float.parseFloat(line.substring(47, 53).trim());
    // ghi ra context theo điều kiện đề bài
    if (max_temp > 40.0) {
        context.write(new Text("Hot Day " + date), new Text(String.valueOf(max_temp)));
    }
    if (min_temp < 10.0) {
        context.write(new Text("Cold Day " + date), new Text(String.valueOf(min_temp)));
    }
}
```

- Phương thức reduce



```
public void reduce(Text key, Iterator<Text> Values, Context context) throws IOException, InterruptedE

    String temperature = Values.next().toString();
    context.write(new Text(key), new Text(temperature));
}
```

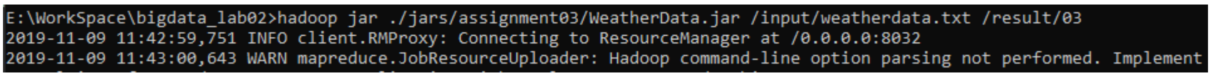
- Hàm main
    - Thiết lập conf và job
    - Thiết lập MapOutputKeyClass là Text, MapOutputValueClass là Text.
    - Thiết lập OutputKeyClass là Text, OutputValueClass là Text.



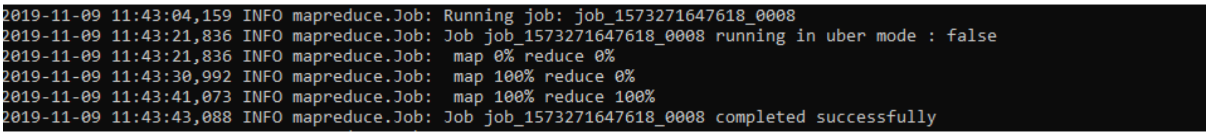
```
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Temperature Day");
    job.setJarByClass(WeatherData.class);

    job.setMapperClass(TemperatureMapper.class);
    job.setReducerClass(TemperatureReducer.class);

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

    job.setOutputKeyClass(Text.class); // đặt Output key class tương ứng là Text
    job.setOutputValueClass(Text.class); // đặt Output value class tương ứng là Text

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
```

- Thực thi:
    - Copy dataset vào hdfs:

    ➔ hadoop fs -copyFromLocal ./Dataset/weather_data.txt /input/weather.txt
    - Lệnh chạy file jar thực thi mapreduce:

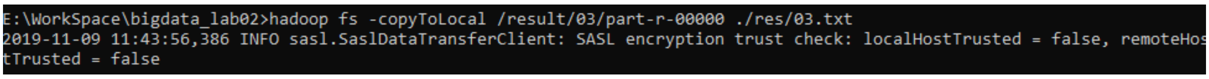


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment03/WeatherData.jar /input/weatherdata.txt /result/03
2019-11-09 11:42:59,751 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 11:43:00,643 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
```

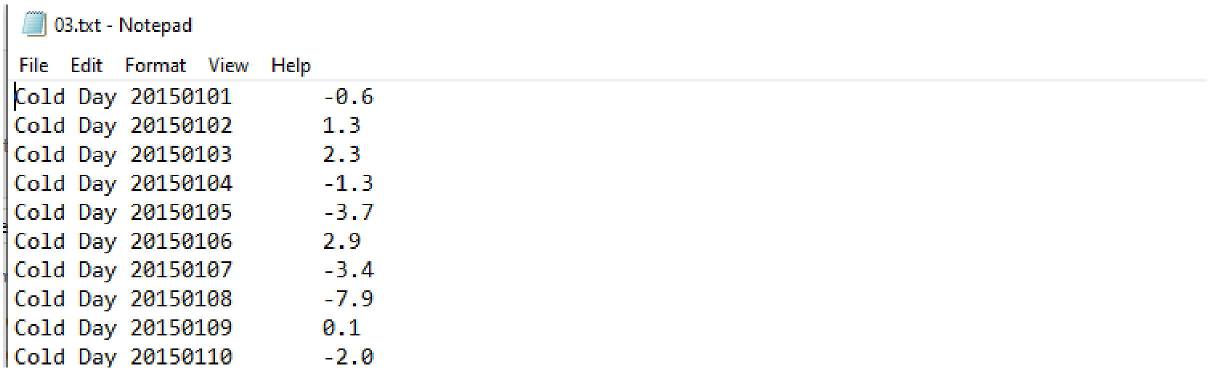


```
2019-11-09 11:43:04,159 INFO mapreduce.Job: Running job: job_1573271647618_0008
2019-11-09 11:43:21,836 INFO mapreduce.Job: Job job_1573271647618_0008 running in uber mode : false
2019-11-09 11:43:21,836 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 11:43:30,992 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 11:43:41,073 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 11:43:43,088 INFO mapreduce.Job: Job job_1573271647618_0008 completed successfully
```

- Kết quả



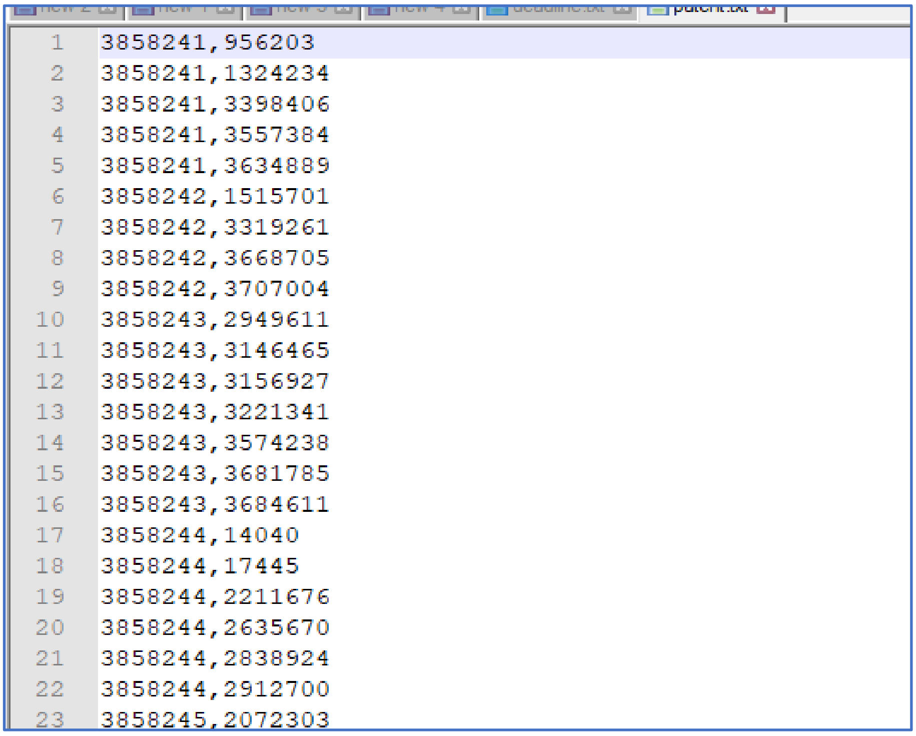
```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/03/part-r-00000 ./res/03.txt
2019-11-09 11:43:56,386 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```



```
Cold Day 20150101	-0.6
Cold Day 20150102	1.3
Cold Day 20150103	2.3
Cold Day 20150104	-1.3
Cold Day 20150105	-3.7
Cold Day 20150106	2.9
Cold Day 20150107	-3.4
Cold Day 20150108	-7.9
Cold Day 20150109	0.1
Cold Day 20150110	-2.0
```

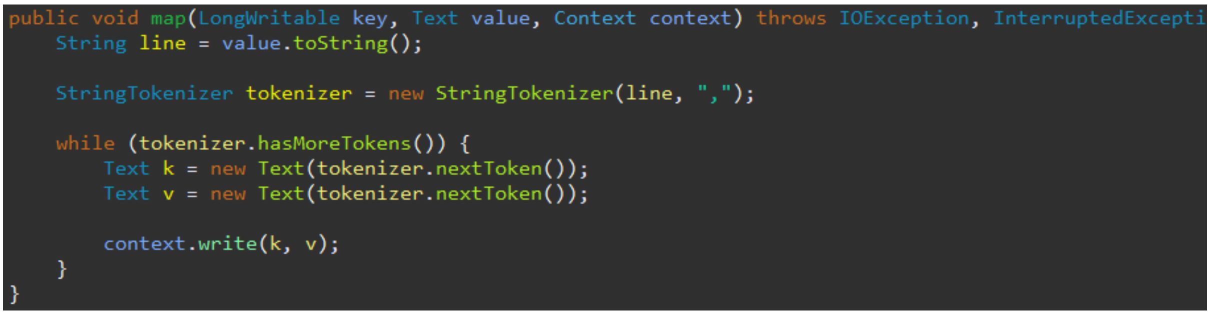
4. Assignment 04 – Patent

- Dataset tìm được (http://data.nber.org/patents/acite75_99.zip - dataset này là về citing patent và cited patent, nhưng nhóm lấy để làm ví dụ tương tự patent dataset trong đề bài)



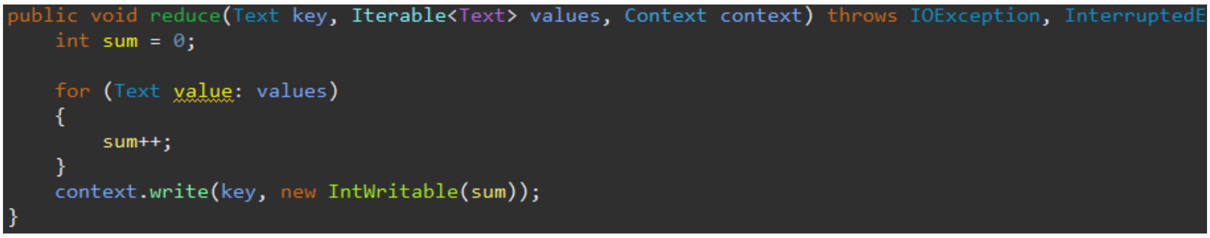
```
3858241,956203
3858241,1324234
3858241,3398406
3858241,3557384
3858241,3634889
3858242,1515701
3858242,3319261
3858242,3668705
3858242,3707004
3858243,2949611
3858243,3146465
3858243,3156927
3858243,3221341
3858243,3574238
3858243,3681785
3858243,3684611
3858244,14040
3858244,17445
3858244,2211676
3858244,2635670
3858244,2838924
3858244,2912700
3858245,2072303
```

- Phương thức map
    - o Tách record thành token
    - o Context ghi ra cặp key value <patent, sub-patent>



```java
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedExcepti
    String line = value.toString();

    StringTokenizer tokenizer = new StringTokenizer(line, ",");

    while (tokenizer.hasMoreTokens()) {
        Text k = new Text(tokenizer.nextToken());
        Text v = new Text(tokenizer.nextToken());

        context.write(k, v);
    }
}
```

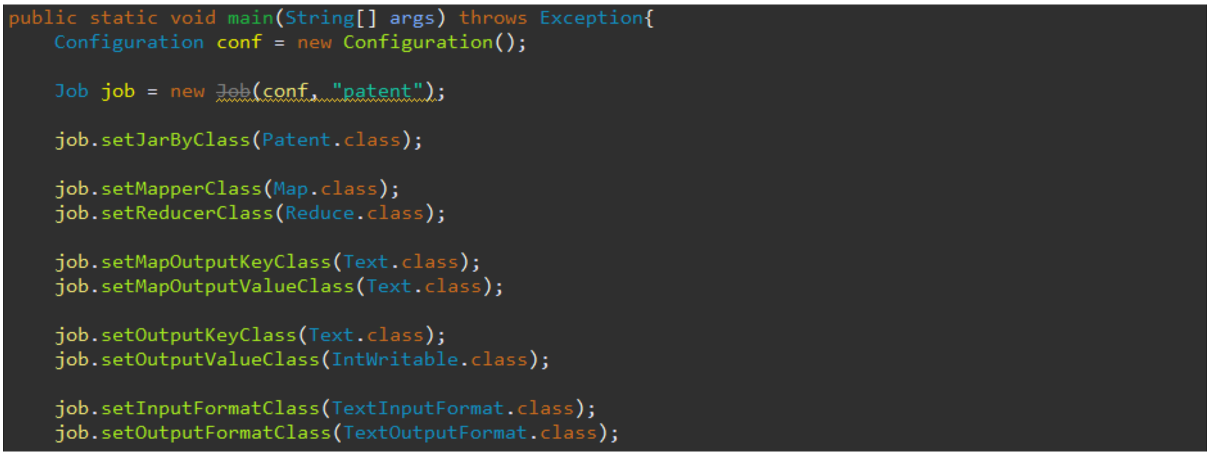
- Phương thức reduce
    - o Với mỗi key <patent>, đếm số lượng sub-patent
    - o Context ghi ra <patent, số lượng sub-patent>



```java
public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedE
    int sum = 0;

    for (Text value: values)
    {
        sum++;
    }
    context.write(key, new IntWritable(sum));
}
```

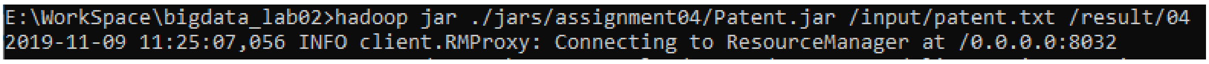
- Hàm main
    - o Thiết lập các conf (Configuration) và job (Job).
    - o Thiết lập MapperClass và ReducerClass.

- Thiết lập MapOutputKeyClass là Text và MapOutputValueClass là Text (<Text, Text>).
- Thiết lập OutputKeyClass là Text, OutputValueClass IntWritable.

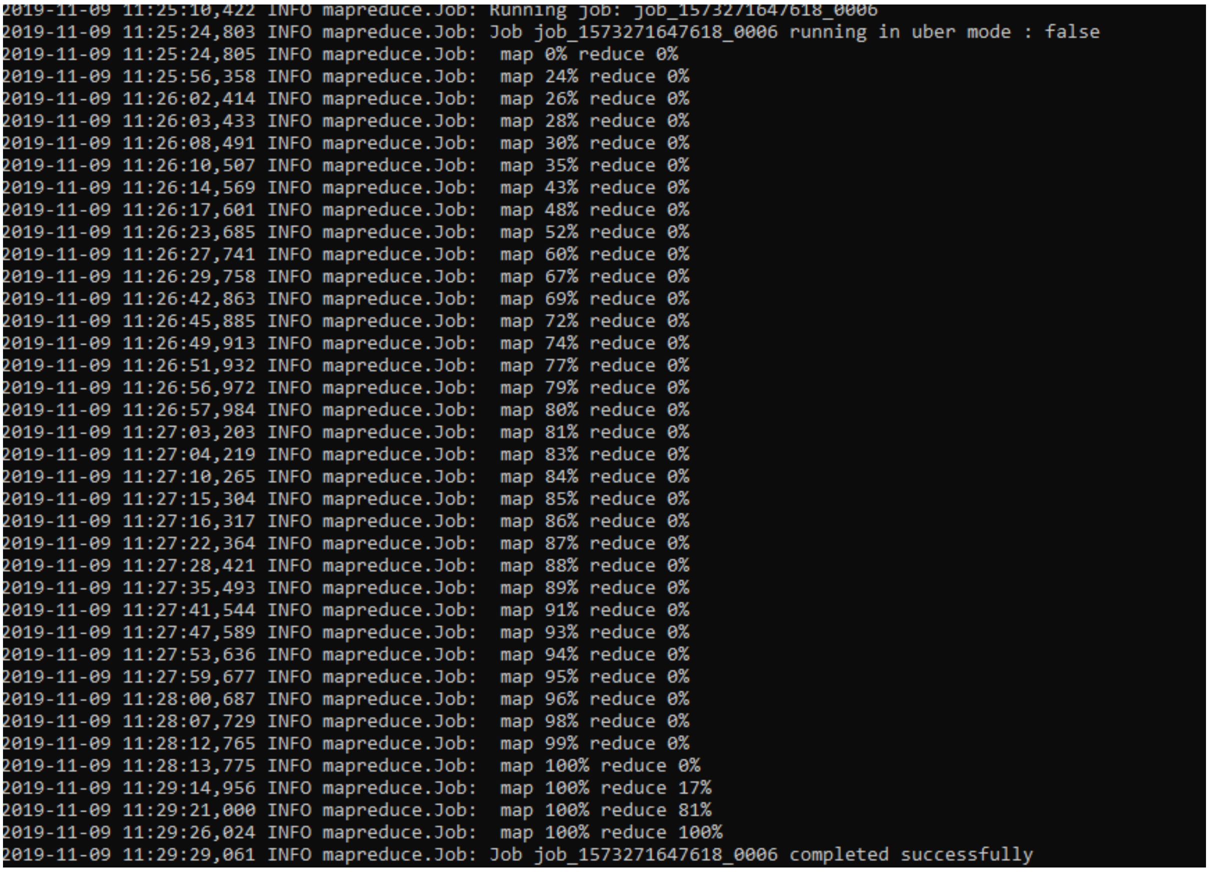


```
public static void main(String[] args) throws Exception{
    Configuration conf = new Configuration();

    Job job = new ~~Job~~(conf, "patent");

    job.setJarByClass(Patent.class);

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

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

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

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
```

- Thực thi:
  - Copy datatset sang hdfs:
  - ➔ hadoop fs -copyFromLocal ./Dataset/patent.txt /input/patent.txt
  - Lệnh thực thi file jar:

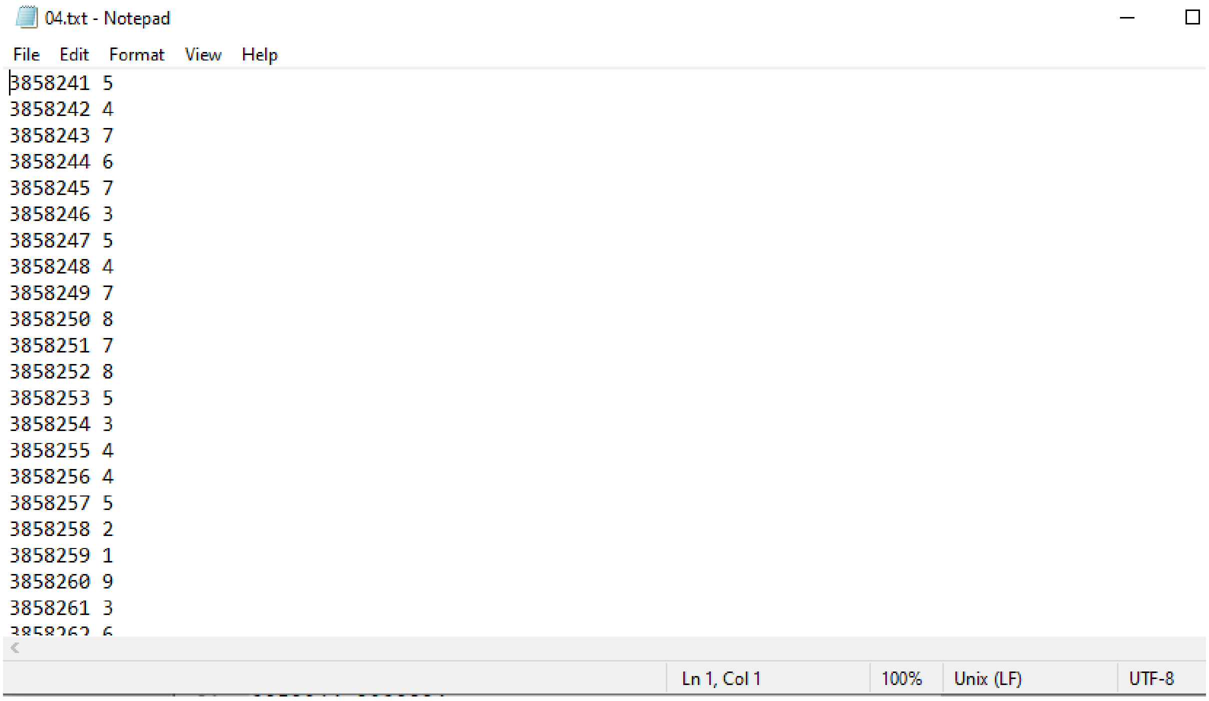


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment04/Patent.jar /input/patent.txt /result/04
2019-11-09 11:25:07,056 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
```



```
2019-11-09 11:25:10,422 INFO mapreduce.Job: Running job: job_1573271647618_0006
2019-11-09 11:25:24,803 INFO mapreduce.Job: Job job_1573271647618_0006 running in uber mode : false
2019-11-09 11:25:24,805 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 11:25:56,358 INFO mapreduce.Job:  map 24% reduce 0%
2019-11-09 11:26:02,414 INFO mapreduce.Job:  map 26% reduce 0%
2019-11-09 11:26:03,433 INFO mapreduce.Job:  map 28% reduce 0%
2019-11-09 11:26:08,491 INFO mapreduce.Job:  map 30% reduce 0%
2019-11-09 11:26:10,507 INFO mapreduce.Job:  map 35% reduce 0%
2019-11-09 11:26:14,569 INFO mapreduce.Job:  map 43% reduce 0%
2019-11-09 11:26:17,601 INFO mapreduce.Job:  map 48% reduce 0%
2019-11-09 11:26:23,685 INFO mapreduce.Job:  map 52% reduce 0%
2019-11-09 11:26:27,741 INFO mapreduce.Job:  map 60% reduce 0%
2019-11-09 11:26:29,758 INFO mapreduce.Job:  map 67% reduce 0%
2019-11-09 11:26:42,863 INFO mapreduce.Job:  map 69% reduce 0%
2019-11-09 11:26:45,885 INFO mapreduce.Job:  map 72% reduce 0%
2019-11-09 11:26:49,913 INFO mapreduce.Job:  map 74% reduce 0%
2019-11-09 11:26:51,932 INFO mapreduce.Job:  map 77% reduce 0%
2019-11-09 11:26:56,972 INFO mapreduce.Job:  map 79% reduce 0%
2019-11-09 11:26:57,984 INFO mapreduce.Job:  map 80% reduce 0%
2019-11-09 11:27:03,203 INFO mapreduce.Job:  map 81% reduce 0%
2019-11-09 11:27:04,219 INFO mapreduce.Job:  map 83% reduce 0%
2019-11-09 11:27:10,265 INFO mapreduce.Job:  map 84% reduce 0%
2019-11-09 11:27:15,304 INFO mapreduce.Job:  map 85% reduce 0%
2019-11-09 11:27:16,317 INFO mapreduce.Job:  map 86% reduce 0%
2019-11-09 11:27:22,364 INFO mapreduce.Job:  map 87% reduce 0%
2019-11-09 11:27:28,421 INFO mapreduce.Job:  map 88% reduce 0%
2019-11-09 11:27:35,493 INFO mapreduce.Job:  map 89% reduce 0%
2019-11-09 11:27:41,544 INFO mapreduce.Job:  map 91% reduce 0%
2019-11-09 11:27:47,589 INFO mapreduce.Job:  map 93% reduce 0%
2019-11-09 11:27:53,636 INFO mapreduce.Job:  map 94% reduce 0%
2019-11-09 11:27:59,677 INFO mapreduce.Job:  map 95% reduce 0%
2019-11-09 11:28:00,687 INFO mapreduce.Job:  map 96% reduce 0%
2019-11-09 11:28:07,729 INFO mapreduce.Job:  map 98% reduce 0%
2019-11-09 11:28:12,765 INFO mapreduce.Job:  map 99% reduce 0%
2019-11-09 11:28:13,775 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 11:29:14,956 INFO mapreduce.Job:  map 100% reduce 17%
2019-11-09 11:29:21,000 INFO mapreduce.Job:  map 100% reduce 81%
2019-11-09 11:29:26,024 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 11:29:29,061 INFO mapreduce.Job: Job job_1573271647618_0006 completed successfully
```

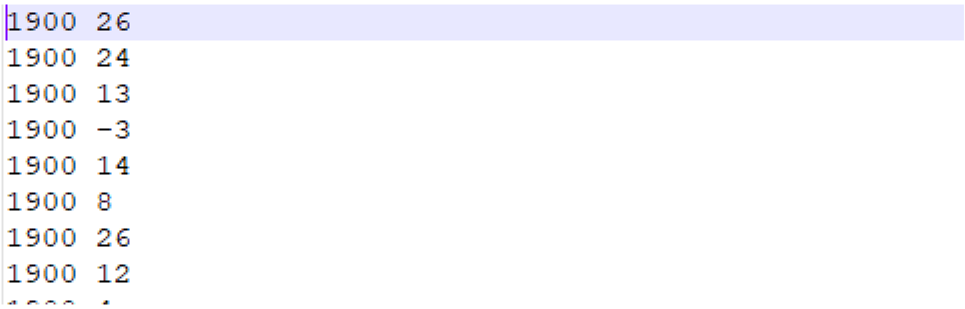
- Kết quả:



```
3858241 5
3858242 4
3858243 7
3858244 6
3858245 7
3858246 3
3858247 5
3858248 4
3858249 7
3858250 8
3858251 7
3858252 8
3858253 5
3858254 3
3858255 4
3858256 4
3858257 5
3858258 2
3858259 1
3858260 9
3858261 3
3858262 6
```

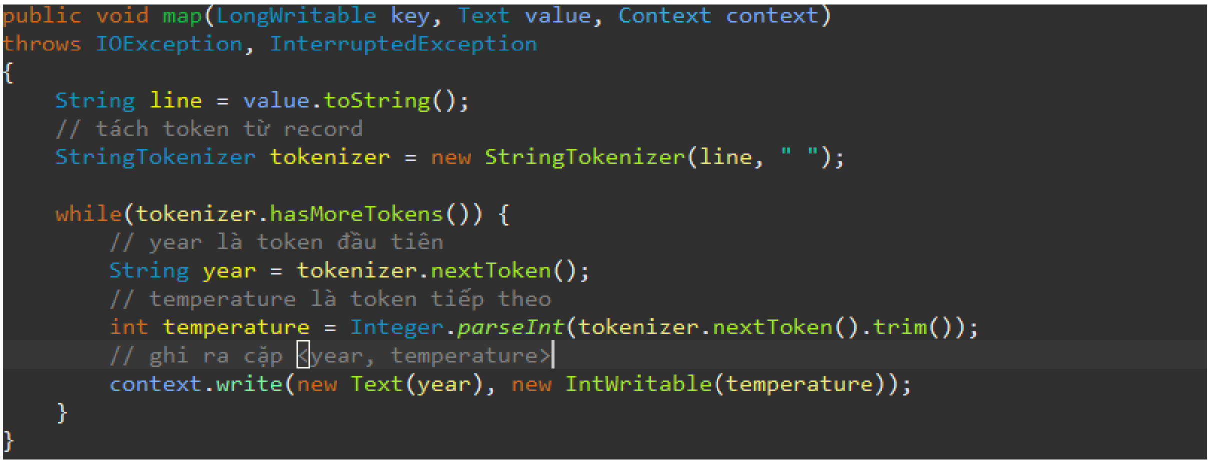
5. Assignment 05 – Max temperature

- Dataset: tự phát sinh (100000 records).



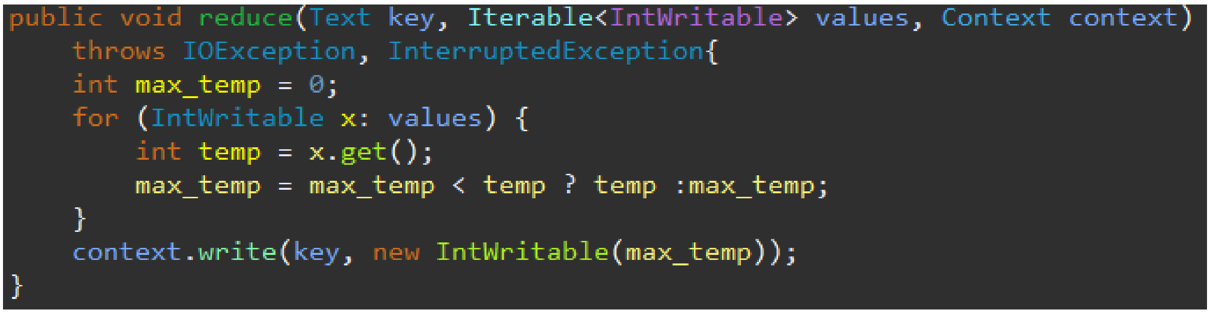
```
1900 26
1900 24
1900 13
1900 -3
1900 14
1900 8
1900 26
1900 12
```

- Phương thức map
    - Tách token từ record.
    - Token thứ nhất là năm, token thứ 2 là nhiệt độ.
    - Context ghi ra key – year (Text), value – temperature (IntWritable).



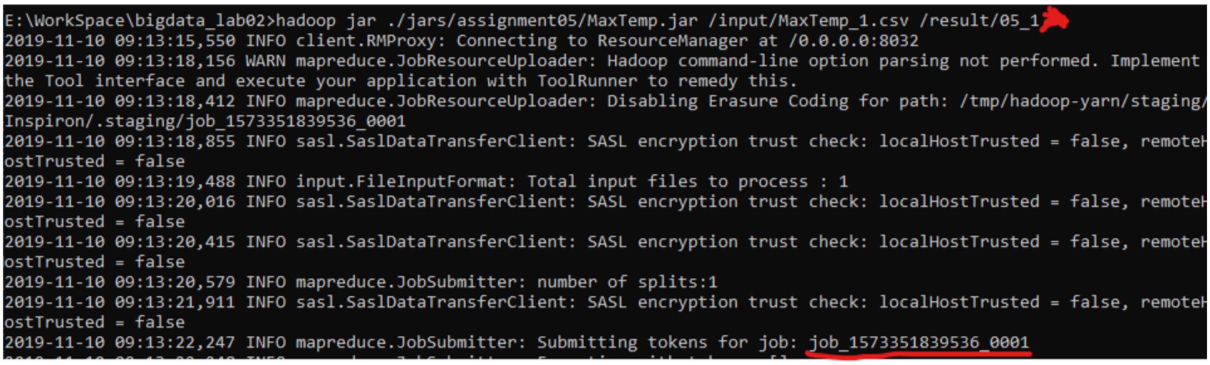
```java
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException
{
    String line = value.toString();
    // tách token từ record
    StringTokenizer tokenizer = new StringTokenizer(line, " ");

    while(tokenizer.hasMoreTokens()) {
        // year là token đầu tiên
        String year = tokenizer.nextToken();
        // temperature là token tiếp theo
        int temperature = Integer.parseInt(tokenizer.nextToken().trim());
        // ghi ra cặp <year, temperature>
        context.write(new Text(year), new IntWritable(temperature));
    }
}
```

- Phương thức reduce
    - Với mỗi key year, tìm nhiệt độ lớn nhất thuộc năm đó.
    - Context ghi ra key – year và value max_temp

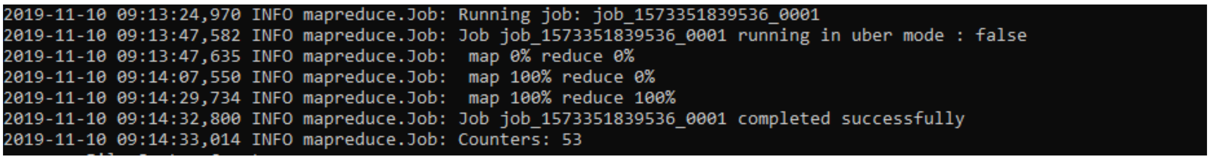


```java
public void reduce(Text key, Iterable<IntWritable> values, Context context)
    throws IOException, InterruptedException{
    int max_temp = 0;
    for (IntWritable x: values) {
        int temp = x.get();
        max_temp = max_temp < temp ? temp :max_temp;
    }
    context.write(key, new IntWritable(max_temp));
}
```

- Thực thi:

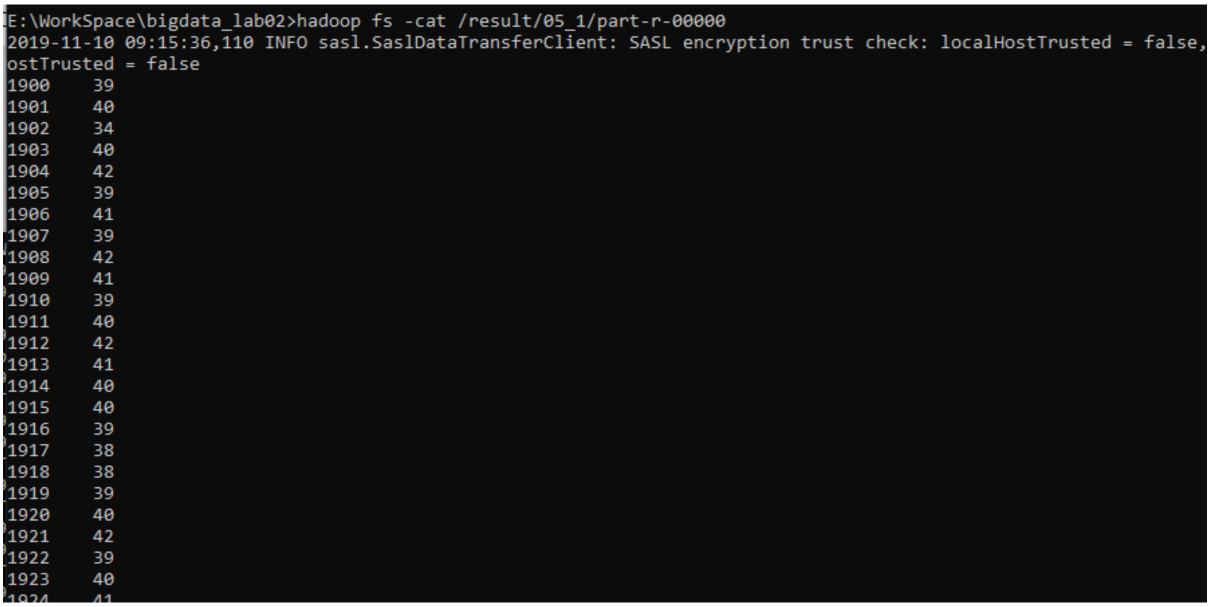


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment05/MaxTemp.jar /input/MaxTemp_1.csv /result/05_1
2019-11-10 09:13:15,550 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-10 09:13:18,156 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2019-11-10 09:13:18,412 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Inspiron/.staging/job_1573351839536_0001
2019-11-10 09:13:18,855 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 09:13:19,488 INFO input.FileInputFormat: Total input files to process : 1
2019-11-10 09:13:20,016 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 09:13:20,415 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 09:13:20,579 INFO mapreduce.JobSubmitter: number of splits:1
2019-11-10 09:13:21,911 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 09:13:22,247 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573351839536_0001
```



```
2019-11-10 09:13:24,970 INFO mapreduce.Job: Running job: job_1573351839536_0001
2019-11-10 09:13:47,582 INFO mapreduce.Job: Job job_1573351839536_0001 running in uber mode : false
2019-11-10 09:13:47,635 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-10 09:14:07,550 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-10 09:14:29,734 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-10 09:14:32,800 INFO mapreduce.Job: Job job_1573351839536_0001 completed successfully
2019-11-10 09:14:33,014 INFO mapreduce.Job: Counters: 53
```

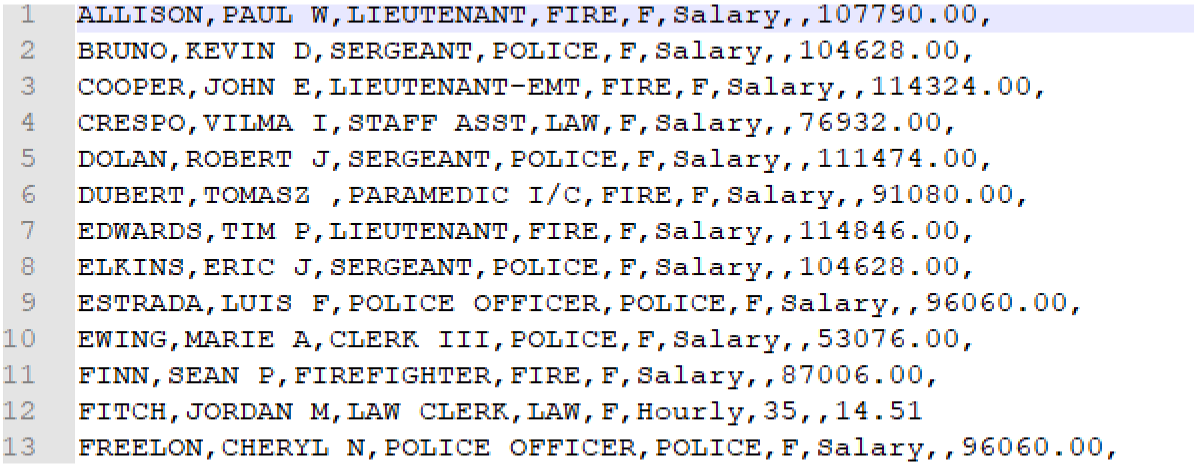
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -cat /result/05_1/part-r-00000
2019-11-10 09:15:36,110 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false,
ostTrusted = false
1900    39
1901    40
1902    34
1903    40
1904    42
1905    39
1906    41
1907    39
1908    42
1909    41
1910    39
1911    40
1912    42
1913    41
1914    40
1915    40
1916    39
1917    38
1918    38
1919    39
1920    40
1921    42
1922    39
1923    40
1924    41
```

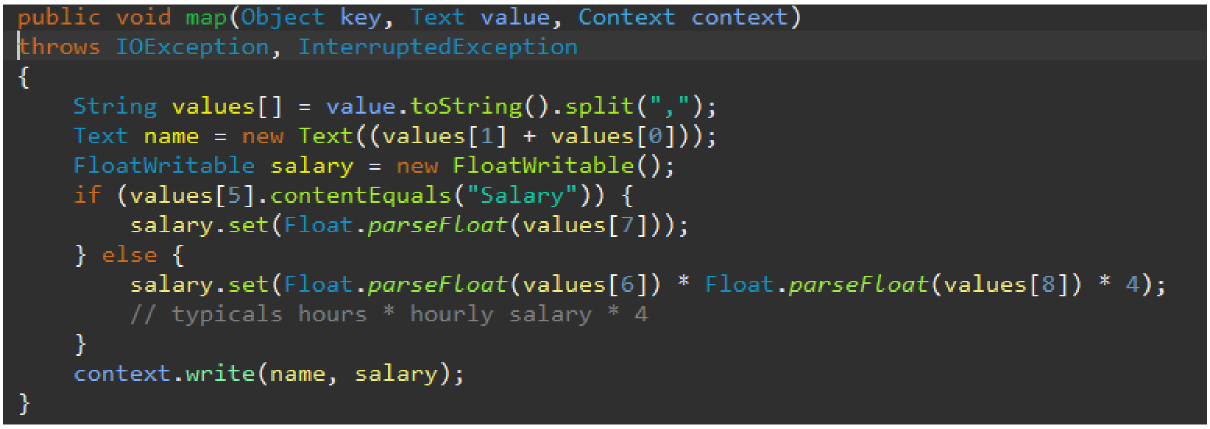
6. Assignment 06 – Average Salary

- Dataset: file mô tả (https://data.cityofchicago.org/api/views/xzkq-xp2w/rows.pdf).



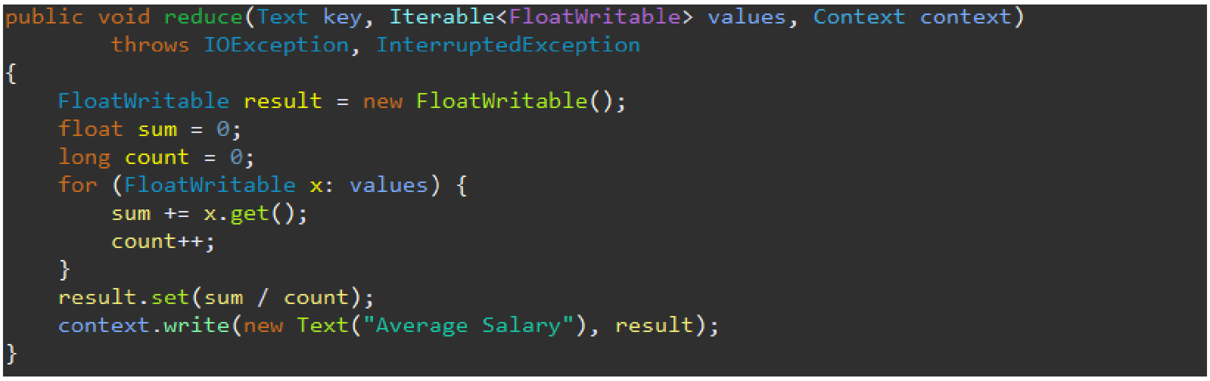
```
ALLISON,PAUL W,LIEUTENANT,FIRE,F,Salary,,107790.00,
BRUNO,KEVIN D,SERGEANT,POLICE,F,Salary,,104628.00,
COOPER,JOHN E,LIEUTENANT-EMT,FIRE,F,Salary,,114324.00,
CRESPO,VILMA I,STAFF ASST,LAW,F,Salary,,76932.00,
DOLAN,ROBERT J,SERGEANT,POLICE,F,Salary,,111474.00,
DUBERT,TOMASZ ,PARAMEDIC I/C,FIRE,F,Salary,,91080.00,
EDWARDS,TIM P,LIEUTENANT,FIRE,F,Salary,,114846.00,
ELKINS,ERIC J,SERGEANT,POLICE,F,Salary,,104628.00,
ESTRADA,LUIS F,POLICE OFFICER,POLICE,F,Salary,,96060.00,
EWING,MARIE A,CLERK III,POLICE,F,Salary,,53076.00,
FINN,SEAN P,FIREFIGHTER,FIRE,F,Salary,,87006.00,
FITCH,JORDAN M,LAW CLERK,LAW,F,Hourly,35,,14.51
FREELON,CHERYL N,POLICE OFFICER,POLICE,F,Salary,,96060.00,
```

- Phương thức map
  - Chuyển record về các chuỗi tương ứng 0 – Last Name, 1 – FirstName, 5 – "Salary"/"Hourly", 7 – salary, 6 – typical hours (per week), 8 – hourly salary.
  - Name = firstName + lastName. Tính toán lương của mỗi người.
  - Context ghi <name, salary>.



```java
public void map(Object key, Text value, Context context)
throws IOException, InterruptedException
{
    String values[] = value.toString().split(",");
    Text name = new Text((values[1] + values[0]));
    FloatWritable salary = new FloatWritable();
    if (values[5].contentEquals("Salary")) {
        salary.set(Float.parseFloat(values[7]));
    } else {
        salary.set(Float.parseFloat(values[6]) * Float.parseFloat(values[8]) * 4);
        // typicals hours * hourly salary * 4
    }
    context.write(name, salary);
}
```

- Phương thức reduce

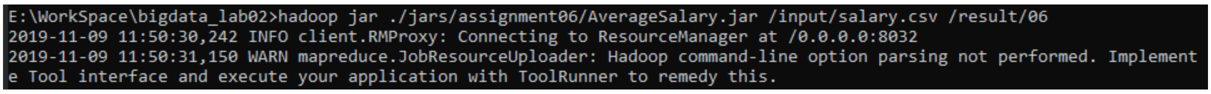


```java
public void reduce(Text key, Iterable<FloatWritable> values, Context context)
        throws IOException, InterruptedException
{
    FloatWritable result = new FloatWritable();
    float sum = 0;
    long count = 0;
    for (FloatWritable x: values) {
        sum += x.get();
        count++;
    }
    result.set(sum / count);
    context.write(new Text("Average Salary"), result);
}
```

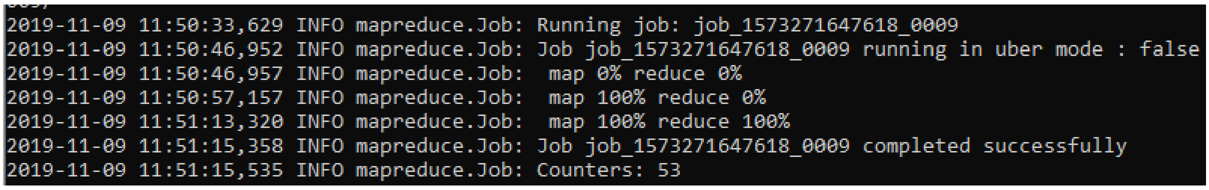
- Thực thi:

- Copy dataset từ local sang hdfs.
- Gọi lệnh jar.

➔ hadoop jar ./jars/assignment06/AverageSalary.jar /input/salary.csv /result/06

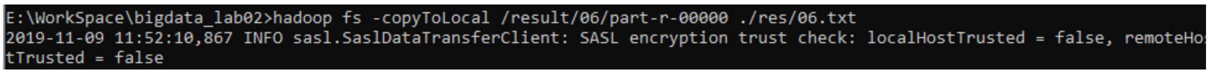


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment06/AverageSalary.jar /input/salary.csv /result/06
2019-11-09 11:50:30,242 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 11:50:31,150 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
e Tool interface and execute your application with ToolRunner to remedy this.
```

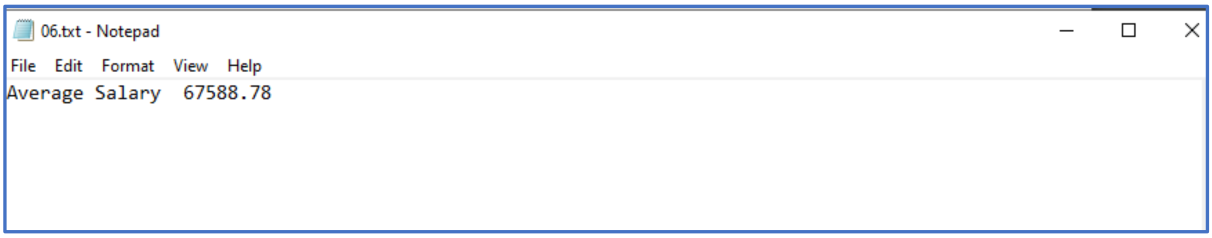


```
2019-11-09 11:50:33,629 INFO mapreduce.Job: Running job: job_1573271647618_0009
2019-11-09 11:50:46,952 INFO mapreduce.Job: Job job_1573271647618_0009 running in uber mode : false
2019-11-09 11:50:46,957 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 11:50:57,157 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 11:51:13,320 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 11:51:15,358 INFO mapreduce.Job: Job job_1573271647618_0009 completed successfully
2019-11-09 11:51:15,535 INFO mapreduce.Job: Counters: 53
```

- Kết quả:



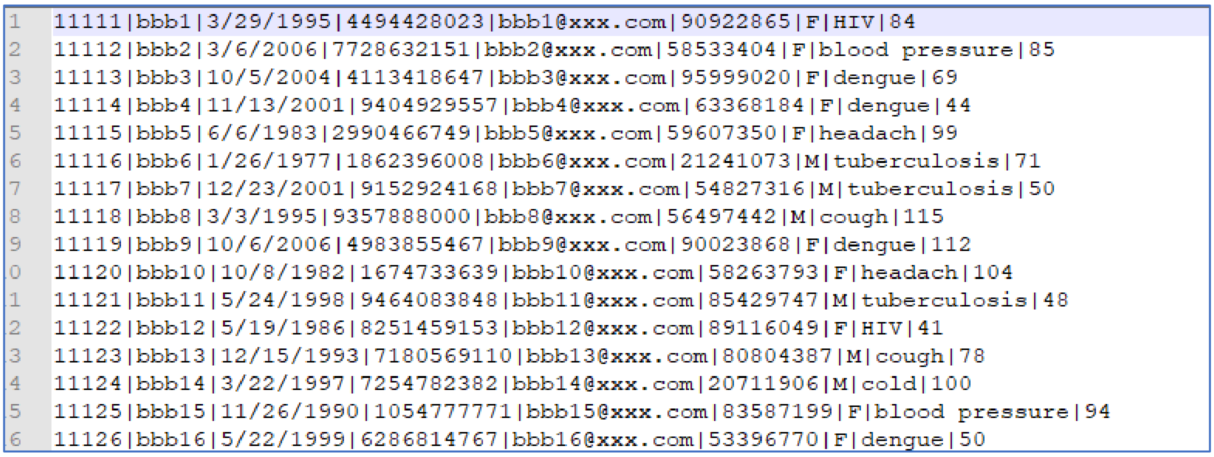
```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/06/part-r-00000 ./res/06.txt
2019-11-09 11:52:10,867 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHo
tTrusted = false
```



```
Average Salary	67588.78
```

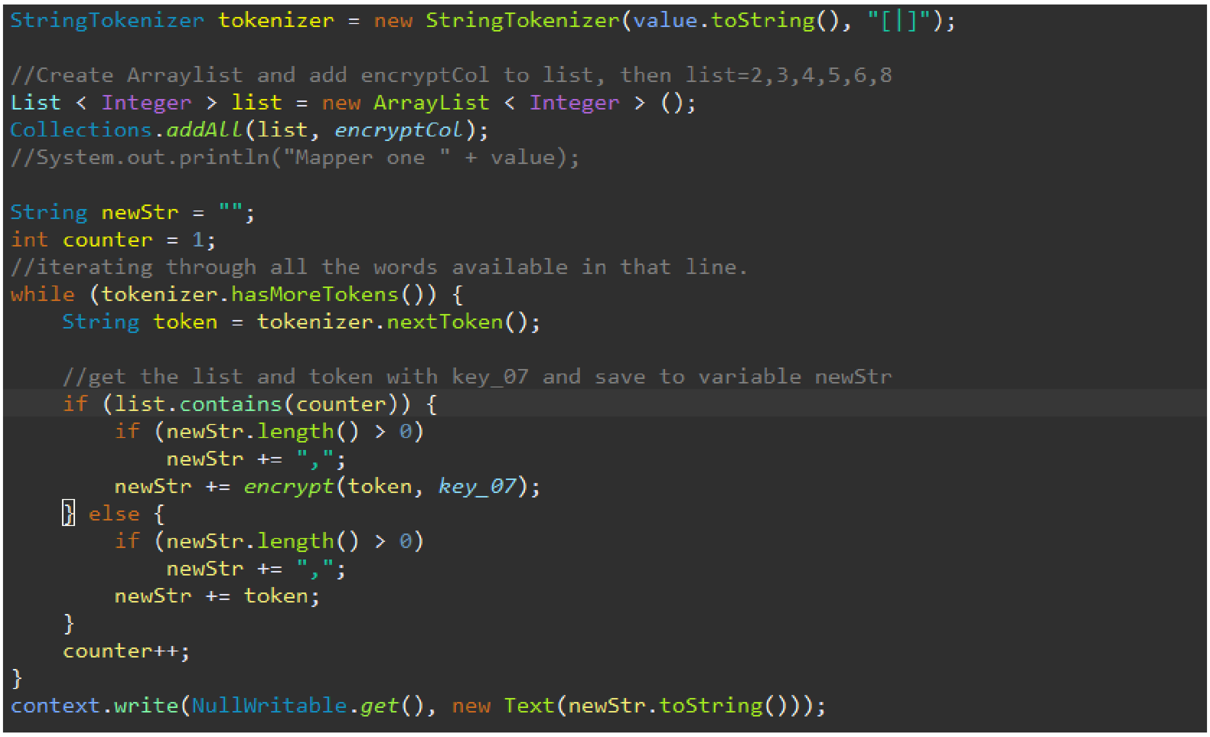
7. Assignment 07 – De Identify Healthcare

- Note: cần tải về và add external jars
    - http://mirrors.viethosting.com/apache/logging/log4j/2.12.1/apache-log4j-2.12.1-bin.zip .
    - http://us.mirrors.quenda.co/apache//commons/codec/binaries/commons-codec-1.13-bin.zip .
- Dataset: tự phát sinh



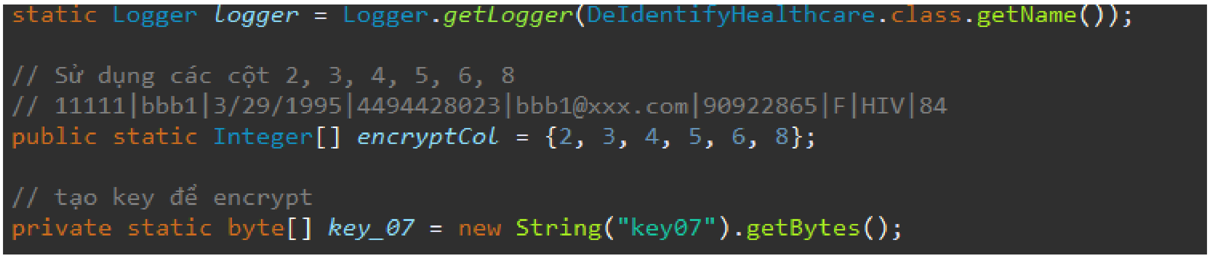
```
11111|bbb1|3/29/1995|4494428023|bbb1@xxx.com|90922865|F|HIV|84
11112|bbb2|3/6/2006|7728632151|bbb2@xxx.com|58533404|F|blood pressure|85
11113|bbb3|10/5/2004|4113418647|bbb3@xxx.com|95999020|F|dengue|69
11114|bbb4|11/13/2001|9404929557|bbb4@xxx.com|63368184|F|dengue|44
11115|bbb5|6/6/1983|2990466749|bbb5@xxx.com|59607350|F|headach|99
11116|bbb6|1/26/1977|1862396008|bbb6@xxx.com|21241073|M|tuberculosis|71
11117|bbb7|12/23/2001|9152924168|bbb7@xxx.com|54827316|M|tuberculosis|50
11118|bbb8|3/3/1995|9357888000|bbb8@xxx.com|56497442|M|cough|115
11119|bbb9|10/6/2006|4983855467|bbb9@xxx.com|90023868|F|dengue|112
11120|bbb10|10/8/1982|1674733639|bbb10@xxx.com|58263793|F|headach|104
11121|bbb11|5/24/1998|9464083848|bbb11@xxx.com|85429747|M|tuberculosis|48
11122|bbb12|5/19/1986|8251459153|bbb12@xxx.com|89116049|F|HIV|41
11123|bbb13|12/15/1993|7180569110|bbb13@xxx.com|80804387|M|cough|78
11124|bbb14|3/22/1997|7254782382|bbb14@xxx.com|20711906|M|cold|100
11125|bbb15|11/26/1990|1054777771|bbb15@xxx.com|83587199|F|blood pressure|94
11126|bbb16|5/22/1999|6286814767|bbb16@xxx.com|53396770|F|dengue|50
```

- Phương thức map
    - Tách các token từ record (ngăn cách bởi dấu |).
    - Với các token ở trên, các token ở cột 2, 3, 4, 5,6 và 8 sẽ được encrypt, các token còn lại giữ nguyên.
    - Cộng chung lại thành chuỗi.
    - Context ghi ra key – null, value – Text (chuỗi được encrypt).



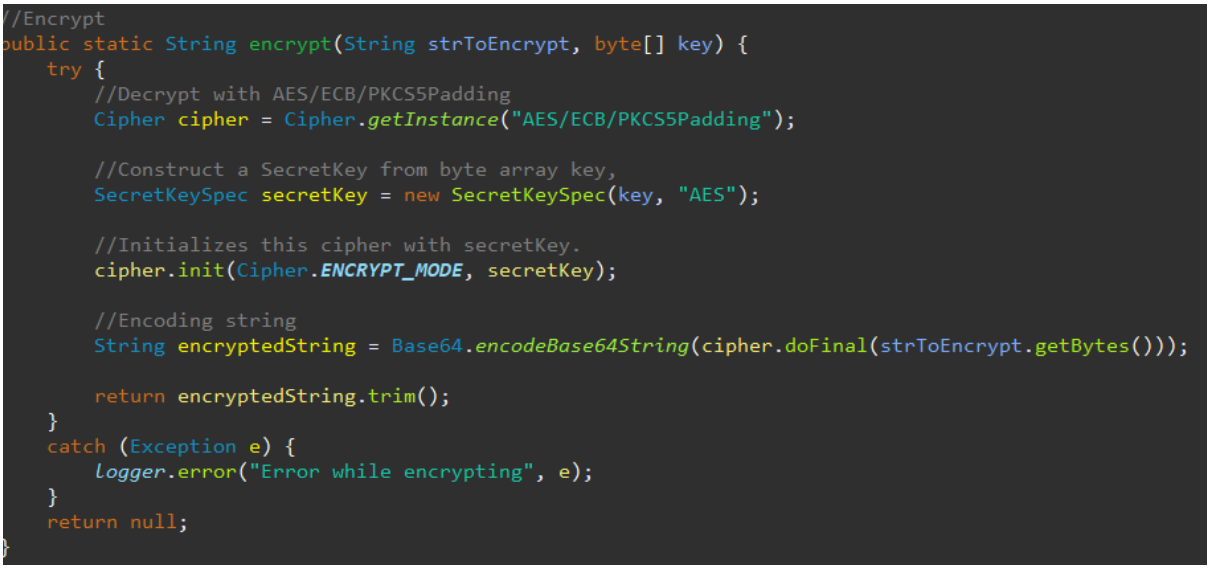
```
StringTokenizer tokenizer = new StringTokenizer(value.toString(), "[|]");

//Create Arraylist and add encryptCol to list, then list=2,3,4,5,6,8
List < Integer > list = new ArrayList < Integer > ();
Collections.addAll(list, encryptCol);
//System.out.println("Mapper one " + value);

String newStr = "";
int counter = 1;
//iterating through all the words available in that line.
while (tokenizer.hasMoreTokens()) {
    String token = tokenizer.nextToken();

    //get the list and token with key_07 and save to variable newStr
    if (list.contains(counter)) {
        if (newStr.length() > 0)
            newStr += ",";
        newStr += encrypt(token, key_07);
    } else {
        if (newStr.length() > 0)
            newStr += ",";
        newStr += token;
    }
    counter++;
}
context.write(NullWritable.get(), new Text(newStr.toString()));
```

- Một số thiết lập đầu Mapper.



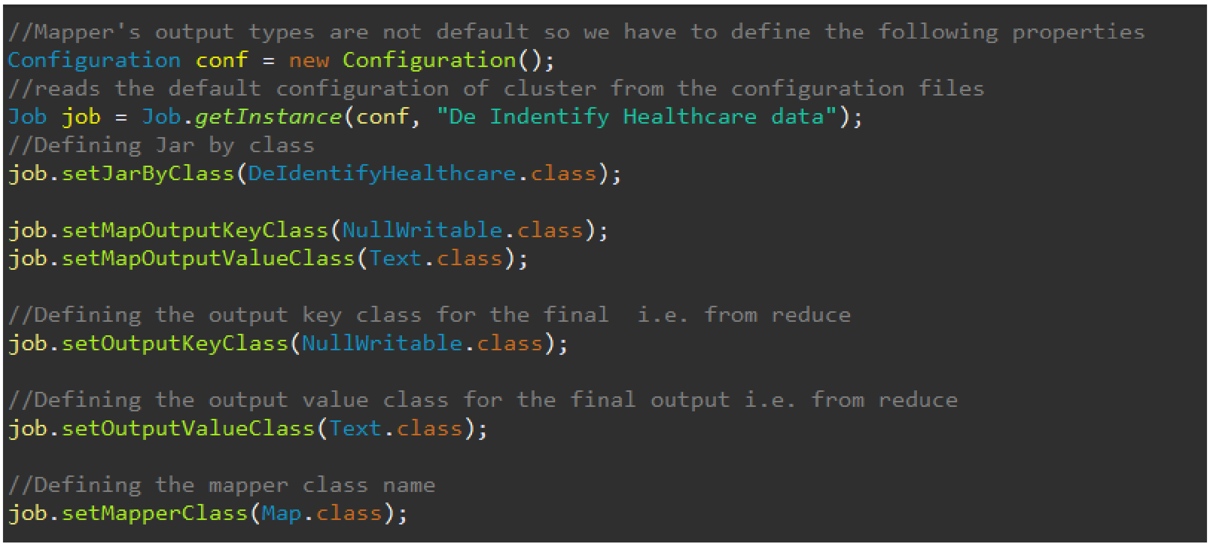
```java
static Logger logger = Logger.getLogger(DeIdentifyHealthcare.class.getName());

// Sử dụng các cột 2, 3, 4, 5, 6, 8
// 11111|bbb1|3/29/1995|4494428023|bbb1@xxx.com|90922865|F|HIV|84
public static Integer[] encryptCol = {2, 3, 4, 5, 6, 8};

// tạo key để encrypt
private static byte[] key_07 = new String("key07").getBytes();
```

- Hàm encrypt



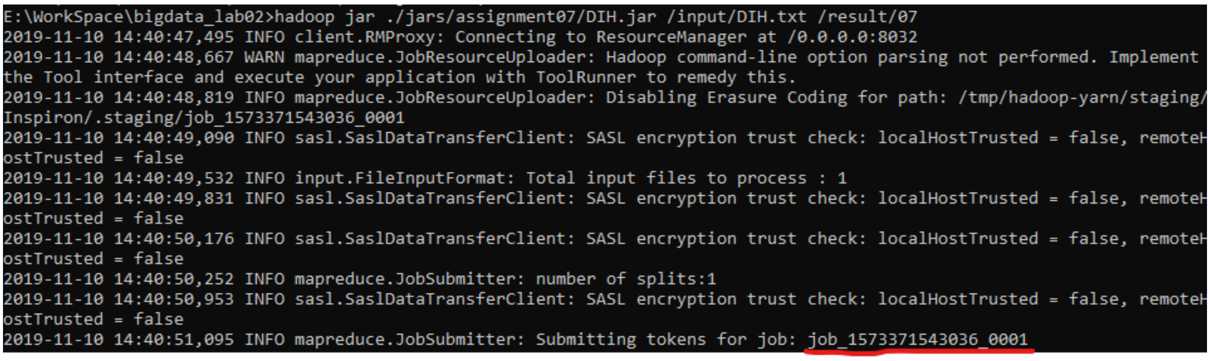
```java
/Encrypt
ublic static String encrypt(String strToEncrypt, byte[] key) {
    try {
        //Decrypt with AES/ECB/PKCS5Padding
        Cipher cipher = Cipher.getInstance("AES/ECB/PKCS5Padding");

        //Construct a SecretKey from byte array key,
        SecretKeySpec secretKey = new SecretKeySpec(key, "AES");

        //Initializes this cipher with secretKey.
        cipher.init(Cipher.ENCRYPT_MODE, secretKey);

        //Encoding string
        String encryptedString = Base64.encodeBase64String(cipher.doFinal(strToEncrypt.getBytes()));

        return encryptedString.trim();
    }
    catch (Exception e) {
        logger.error("Error while encrypting", e);
    }
    return null;
}
```

- Hàm main
  - Thiết lập configuration và job.
  - Thiết lập MapOutputKeyClass là NullWritable, MapOutputValueClass là Text.
  - Thiết lập OutputKeyClass là NullWritable, OutputValueClass là Text.

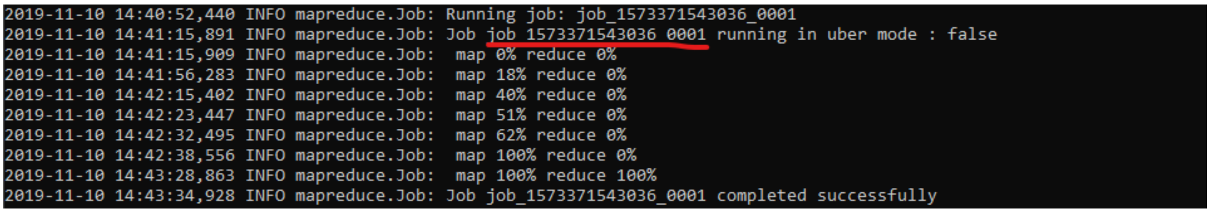


```java
//Mapper's output types are not default so we have to define the following properties
Configuration conf = new Configuration();
//reads the default configuration of cluster from the configuration files
Job job = Job.getInstance(conf, "De Indentify Healthcare data");
//Defining Jar by class
job.setJarByClass(DeIdentifyHealthcare.class);

job.setMapOutputKeyClass(NullWritable.class);
job.setMapOutputValueClass(Text.class);

//Defining the output key class for the final  i.e. from reduce
job.setOutputKeyClass(NullWritable.class);

//Defining the output value class for the final output i.e. from reduce
job.setOutputValueClass(Text.class);

//Defining the mapper class name
job.setMapperClass(Map.class);
```

- Thực thi
  - Copy dataset vào hdfs:
  - ➔ hadoop fs -copyFromLocal ./Dataset/DeIdentifyHealthcare.txt /input/DIH.txt
  - Lệnh thực thi jar file
  - ➔ Hadoop jar ./jars/assignment07/DIH.jar /input/DIH.txt /result/07

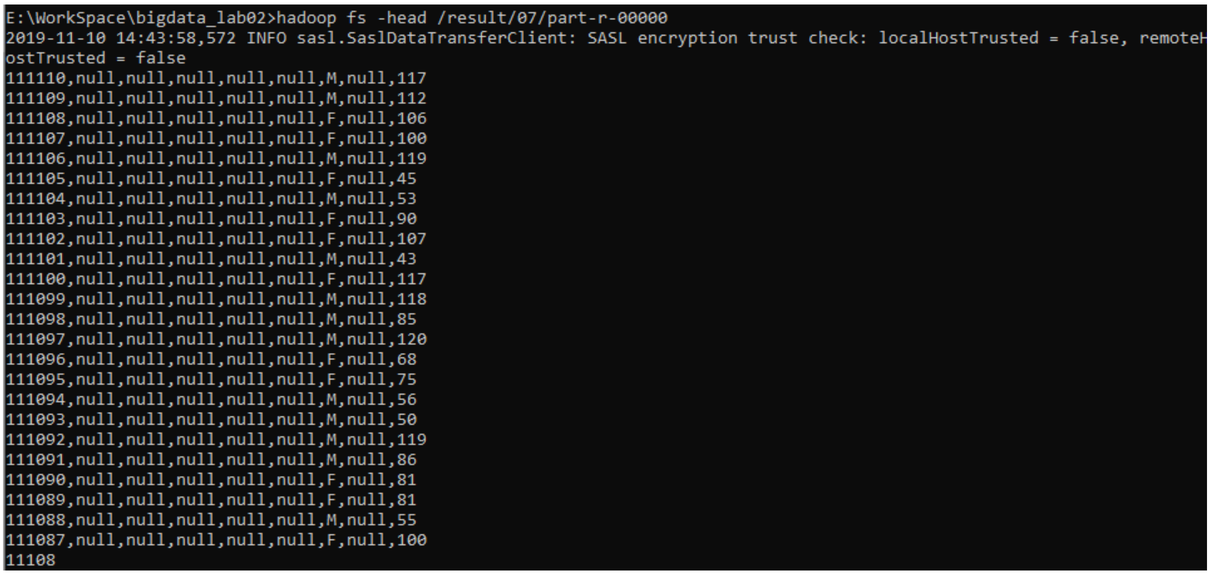


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment07/DIH.jar /input/DIH.txt /result/07
2019-11-10 14:40:47,495 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-10 14:40:48,667 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2019-11-10 14:40:48,819 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Inspiron/.staging/job_1573371543036_0001
2019-11-10 14:40:49,090 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 14:40:49,532 INFO input.FileInputFormat: Total input files to process : 1
2019-11-10 14:40:49,831 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 14:40:50,176 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 14:40:50,252 INFO mapreduce.JobSubmitter: number of splits:1
2019-11-10 14:40:50,953 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2019-11-10 14:40:51,095 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1573371543036_0001
```



```
2019-11-10 14:40:52,440 INFO mapreduce.Job: Running job: job_1573371543036_0001
2019-11-10 14:41:15,891 INFO mapreduce.Job: Job job_1573371543036_0001 running in uber mode : false
2019-11-10 14:41:15,909 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-10 14:41:56,283 INFO mapreduce.Job:  map 18% reduce 0%
2019-11-10 14:42:15,402 INFO mapreduce.Job:  map 40% reduce 0%
2019-11-10 14:42:23,447 INFO mapreduce.Job:  map 51% reduce 0%
2019-11-10 14:42:32,495 INFO mapreduce.Job:  map 62% reduce 0%
2019-11-10 14:42:38,556 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-10 14:43:28,863 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-10 14:43:34,928 INFO mapreduce.Job: Job job_1573371543036_0001 completed successfully
```

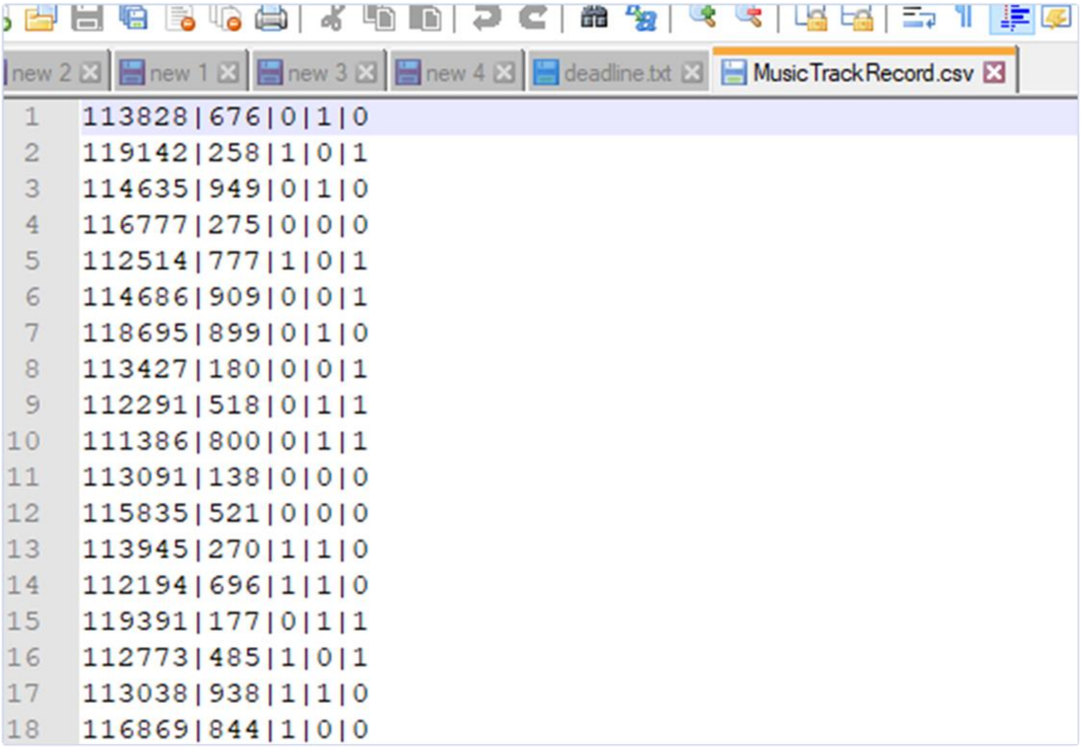
- Kết quả



```
E:\WorkSpace\bigdata_lab02>hadoop fs -head /result/07/part-r-00000
2019-11-10 14:43:58,572 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
111110,null,null,null,null,null,M,null,117
111109,null,null,null,null,null,M,null,112
111108,null,null,null,null,null,F,null,106
111107,null,null,null,null,null,F,null,100
111106,null,null,null,null,null,M,null,119
111105,null,null,null,null,null,F,null,45
111104,null,null,null,null,null,M,null,53
111103,null,null,null,null,null,F,null,90
111102,null,null,null,null,null,F,null,107
111101,null,null,null,null,null,M,null,43
111100,null,null,null,null,null,F,null,117
111099,null,null,null,null,null,M,null,118
111098,null,null,null,null,null,M,null,85
111097,null,null,null,null,null,M,null,120
111096,null,null,null,null,null,F,null,68
111095,null,null,null,null,null,F,null,75
111094,null,null,null,null,null,M,null,56
111093,null,null,null,null,null,M,null,50
111092,null,null,null,null,null,M,null,119
111091,null,null,null,null,null,M,null,86
111090,null,null,null,null,null,F,null,81
111089,null,null,null,null,null,F,null,81
111088,null,null,null,null,null,M,null,55
111087,null,null,null,null,null,F,null,100
11108
```

8. Assignment 08 – Music Track

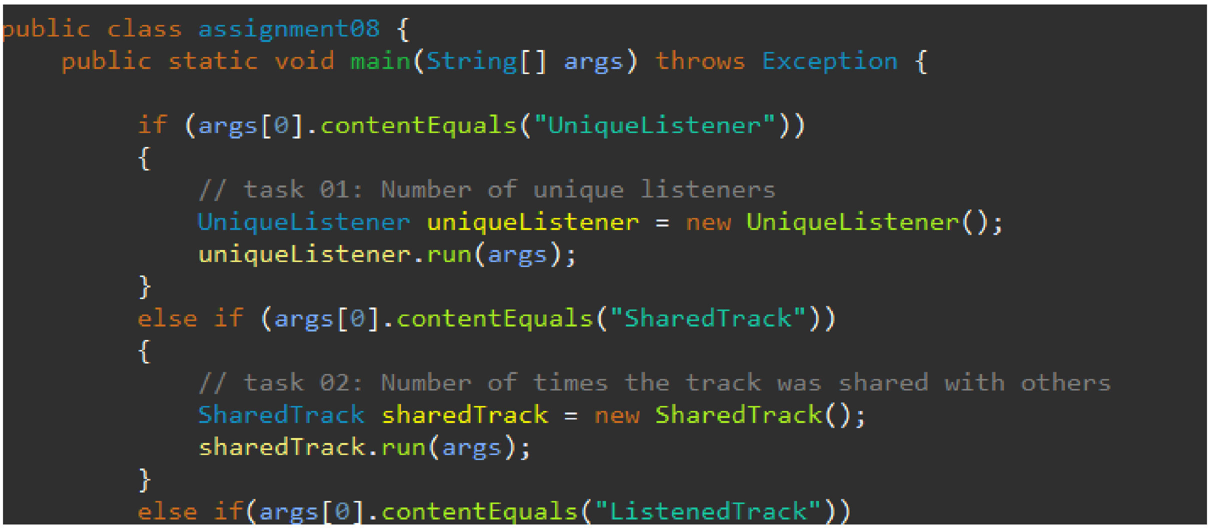
- Dataset tự tạo *MusicTrackRecords.csv*



```
113828|676|0|1|0
119142|258|1|0|1
114635|949|0|1|0
116777|275|0|0|0
112514|777|1|0|1
114686|909|0|0|1
118695|899|0|1|0
113427|180|0|0|1
112291|518|0|1|1
111386|800|0|1|1
113091|138|0|0|0
115835|521|0|0|0
113945|270|1|1|0
112194|696|1|1|0
119391|177|0|1|1
112773|485|1|0|1
113038|938|1|1|0
116869|844|1|0|0
```

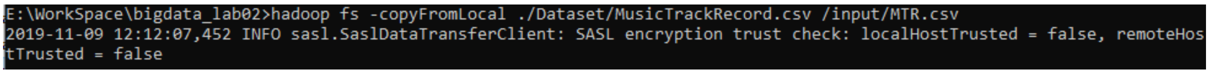
- Có 5 lớp tương ứng với 5 task cần thực thi
  - UniqueListener: Number of unique listeners per track.
  - SharedTrack: Number of times the track was shared with others.
  - ListenedTrack: Number of times the track was listened to on the radio.
  - ListenedTotalTrack: Number of times the track was listened to in total. (Tổng số lần track được nghe – không bị skip).
  - SkippedOnRadioTrack: Number of times the track was skipped on the radio.
- Trong lớp assignment08, nhóm viết hàm main để thực thi chương trình.
  - Lệnh thực thi chung:
  - Args[0]: tên task.
  - Args[1]: đường dẫn input.
  - args[2]: đường dẫn output.

➔ hadoop jar <đường dẫn file jar> <tên task> <input> <output>



```
public class assignment08 {
    public static void main(String[] args) throws Exception {

        if (args[0].contentEquals("UniqueListener"))
        {
            // task 01: Number of unique listeners
            UniqueListener uniqueListener = new UniqueListener();
            uniqueListener.run(args);
        }
        else if (args[0].contentEquals("SharedTrack"))
        {
            // task 02: Number of times the track was shared with others
            SharedTrack sharedTrack = new SharedTrack();
            sharedTrack.run(args);
        }
        else if(args[0].contentEquals("ListenedTrack"))
```

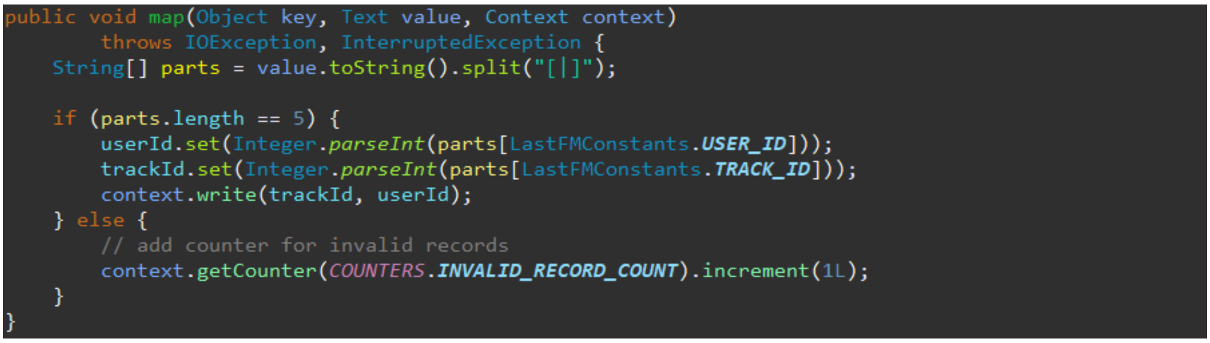
- Copy dataset vào hdfs.



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyFromLocal ./Dataset/MusicTrackRecord.csv /input/MTR.csv
2019-11-09 12:12:07,452 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHos
tTrusted = false
```

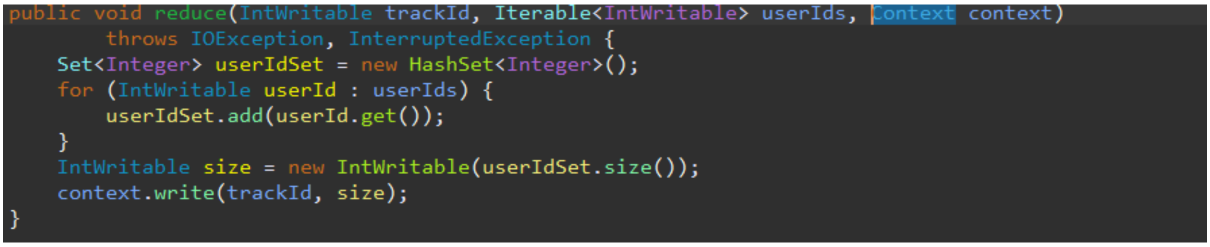
(1) Task 01: UniqueListener

- Phương thức map:



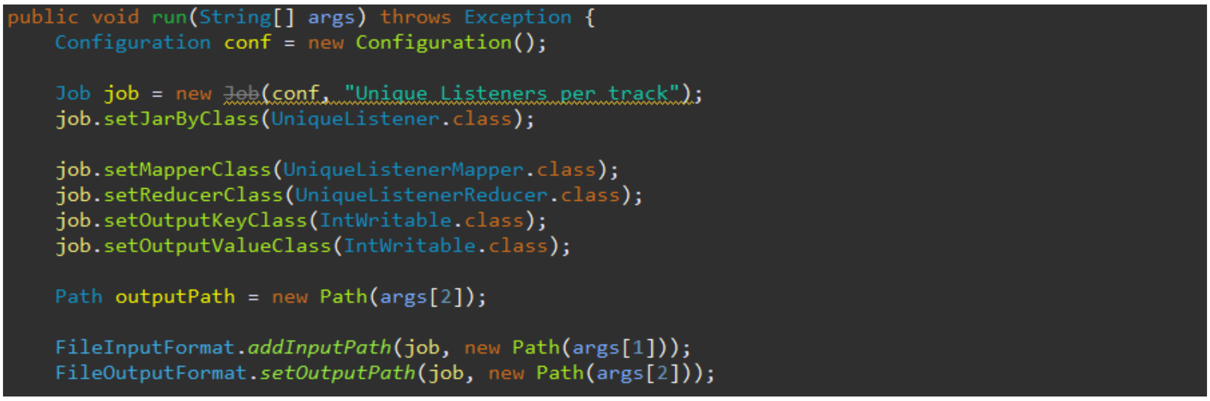
```
public void map(Object key, Text value, Context context)
        throws IOException, InterruptedException {
    String[] parts = value.toString().split("[|]");

    if (parts.length == 5) {
        userId.set(Integer.parseInt(parts[LastFMConstants.USER_ID]));
        trackId.set(Integer.parseInt(parts[LastFMConstants.TRACK_ID]));
        context.write(trackId, userId);
    } else {
        // add counter for invalid records
        context.getCounter(COUNTERS.INVALID_RECORD_COUNT).increment(1L);
    }
}
```

- Phương thức reduce:



```
public void reduce(IntWritable trackId, Iterable<IntWritable> userIds, Context context)
        throws IOException, InterruptedException {
    Set<Integer> userIdSet = new HashSet<Integer>();
    for (IntWritable userId : userIds) {
        userIdSet.add(userId.get());
    }
    IntWritable size = new IntWritable(userIdSet.size());
    context.write(trackId, size);
}
```

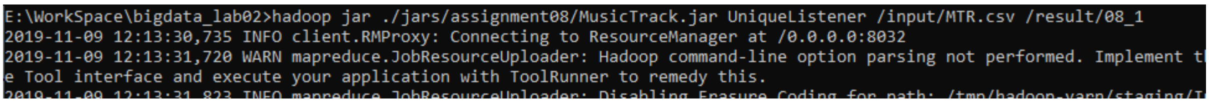
- Hàm run



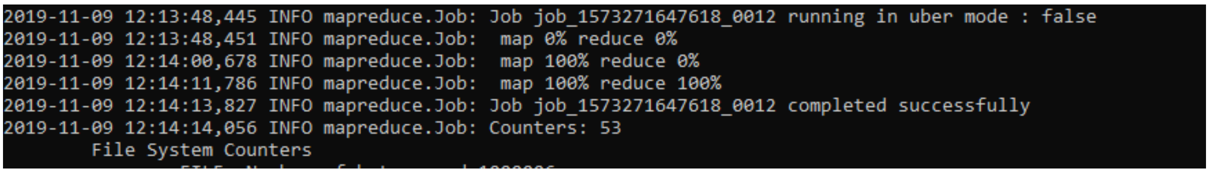
```
public void run(String[] args) throws Exception {
    Configuration conf = new Configuration();

    Job job = new Job(conf, "Unique Listeners per track");
    job.setJarByClass(UniqueListener.class);

    job.setMapperClass(UniqueListenerMapper.class);
    job.setReducerClass(UniqueListenerReducer.class);
    job.setOutputKeyClass(IntWritable.class);
    job.setOutputValueClass(IntWritable.class);

    Path outputPath = new Path(args[2]);

    FileInputFormat.addInputPath(job, new Path(args[1]));
    FileOutputFormat.setOutputPath(job, new Path(args[2]));
```

- Thực thi:

➔ hadoop ./jars/assignment08/MusicTrack.jar UniqueListener /input/MTR.csv /result/08_1.

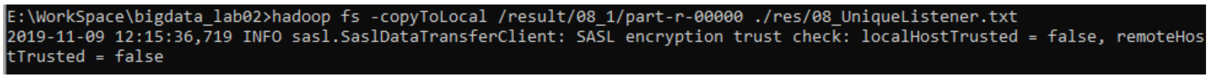


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment08/MusicTrack.jar UniqueListener /input/MTR.csv /result/08_1
2019-11-09 12:13:30,735 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 12:13:31,720 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement th
e Tool interface and execute your application with ToolRunner to remedy this.
2019-11-09 12:13:31,823 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/I
```



```
2019-11-09 12:13:48,445 INFO mapreduce.Job: Job job_1573271647618_0012 running in uber mode : false
2019-11-09 12:13:48,451 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 12:14:00,678 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 12:14:11,786 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 12:14:13,827 INFO mapreduce.Job: Job job_1573271647618_0012 completed successfully
2019-11-09 12:14:14,056 INFO mapreduce.Job: Counters: 53
        File System Counters
```

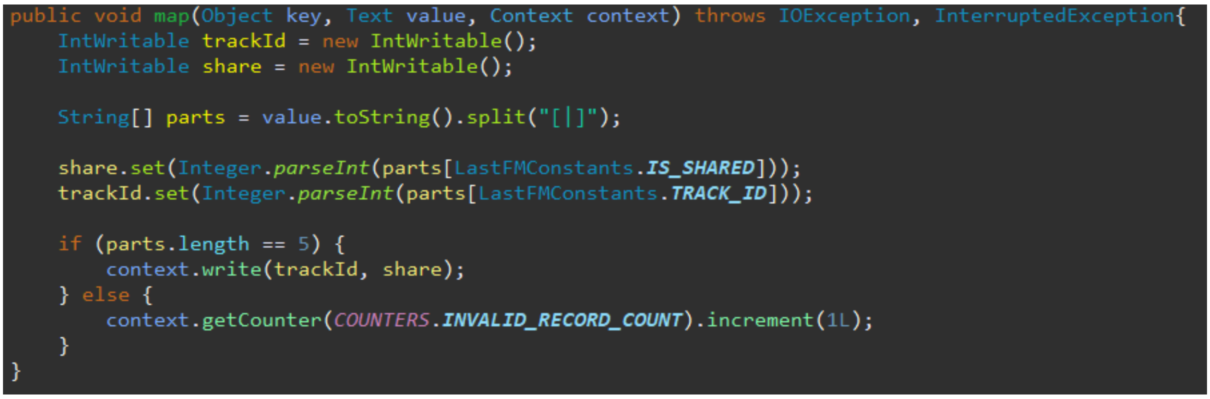
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/08_1/part-r-00000 ./res/08_UniqueListener.txt
2019-11-09 12:15:36,719 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHos
tTrusted = false
```

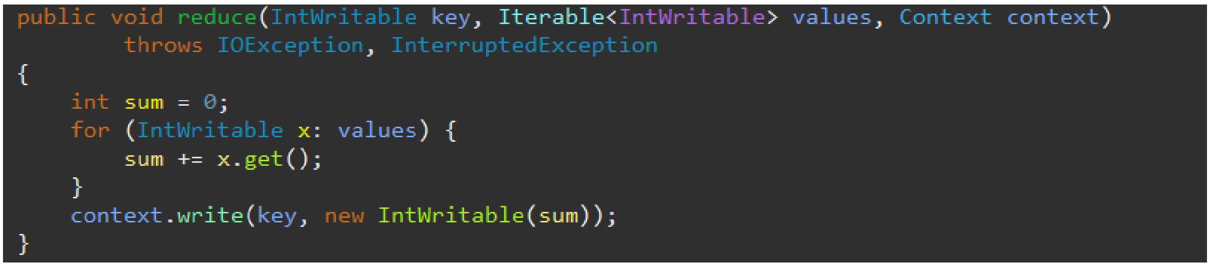
(2) Task 02: SharedTrack

- Phương thức map



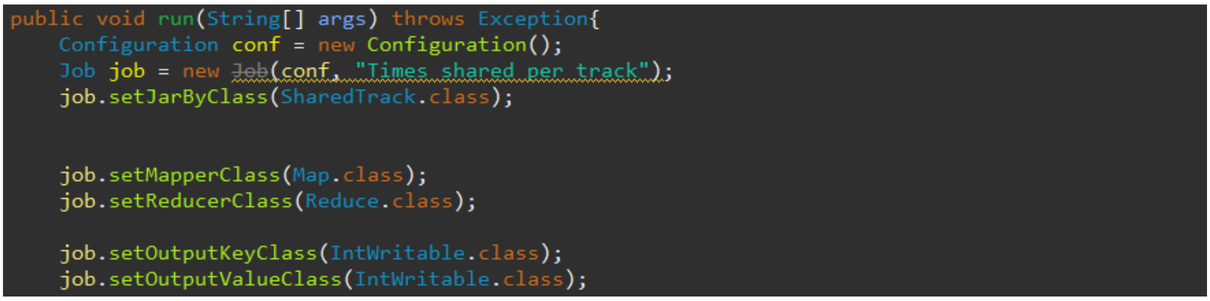
```
public void map(Object key, Text value, Context context) throws IOException, InterruptedException{
    IntWritable trackId = new IntWritable();
    IntWritable share = new IntWritable();

    String[] parts = value.toString().split("[|]");

    share.set(Integer.parseInt(parts[LastFMConstants.IS_SHARED]));
    trackId.set(Integer.parseInt(parts[LastFMConstants.TRACK_ID]));

    if (parts.length == 5) {
        context.write(trackId, share);
    } else {
        context.getCounter(COUNTERS.INVALID_RECORD_COUNT).increment(1L);
    }
}
```

- Phương thức reduce



```
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException
{
    int sum = 0;
    for (IntWritable x: values) {
        sum += x.get();
    }
    context.write(key, new IntWritable(sum));
}
```

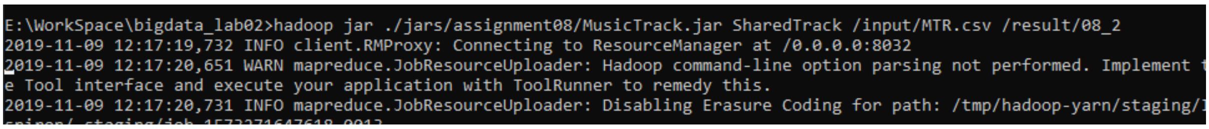
- Hàm run



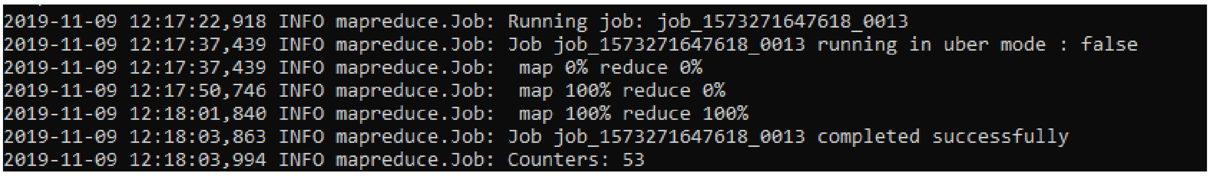
```
public void run(String[] args) throws Exception{
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Times shared per track");
    job.setJarByClass(SharedTrack.class);


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

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

- Thực thi:

➔ hadoop ./jars/assignment08/MusicTrack.jar SharedTrack /input/MTR.csv /result/08_2.

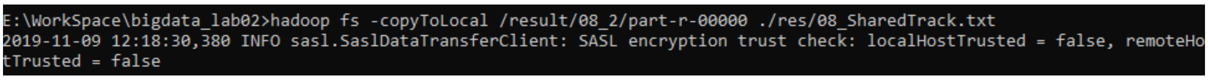


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment08/MusicTrack.jar SharedTrack /input/MTR.csv /result/08_2
2019-11-09 12:17:19,732 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 12:17:20,651 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
e Tool interface and execute your application with ToolRunner to remedy this.
2019-11-09 12:17:20,731 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/I
```



```
2019-11-09 12:17:22,918 INFO mapreduce.Job: Running job: job_1573271647618_0013
2019-11-09 12:17:37,439 INFO mapreduce.Job: Job job_1573271647618_0013 running in uber mode : false
2019-11-09 12:17:37,439 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 12:17:50,746 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 12:18:01,840 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 12:18:03,863 INFO mapreduce.Job: Job job_1573271647618_0013 completed successfully
2019-11-09 12:18:03,994 INFO mapreduce.Job: Counters: 53
```

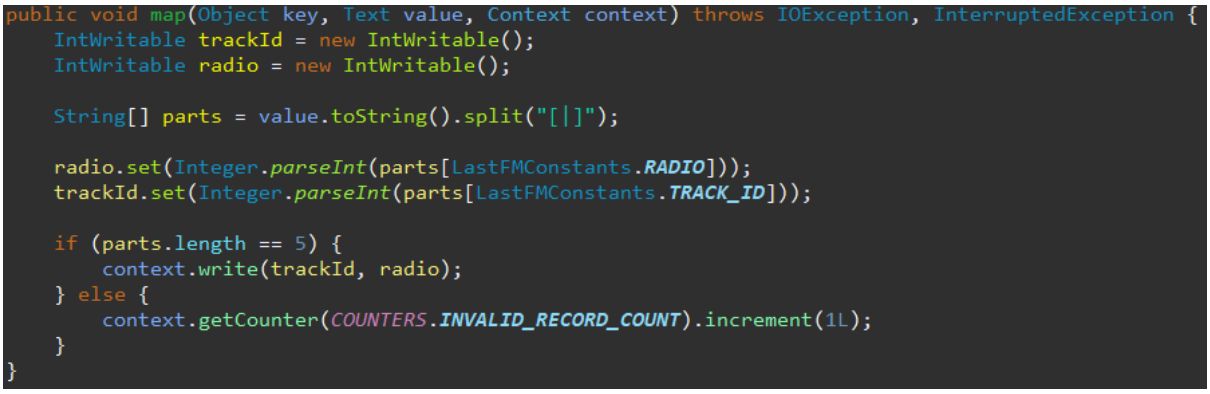
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/08_2/part-r-00000 ./res/08_SharedTrack.txt
2019-11-09 12:18:30,380 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHo
tTrusted = false
```

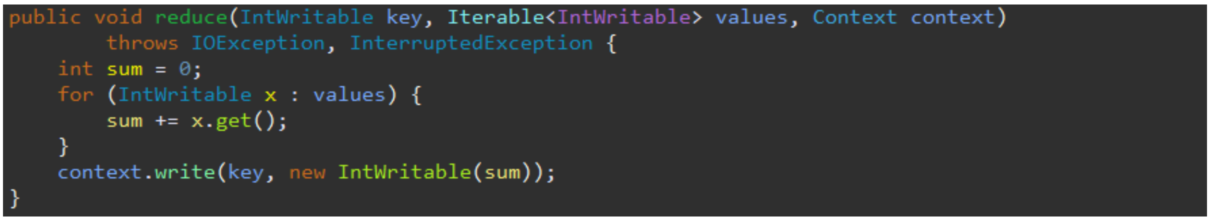
(3) Task 03: ListenedTrack (on Radio)

- Phương thức map



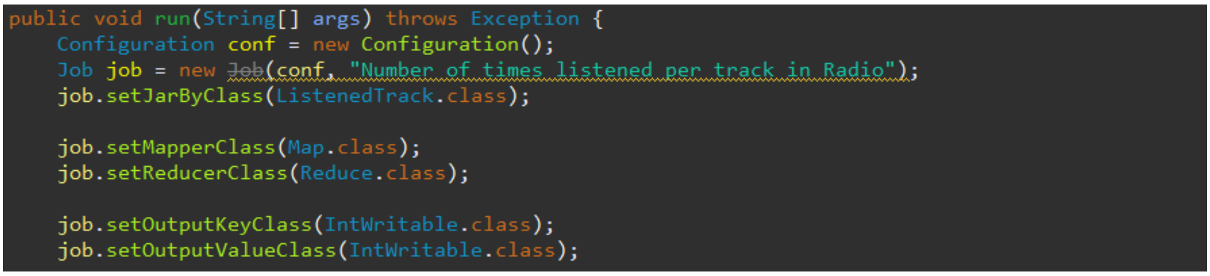
```
public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
    IntWritable trackId = new IntWritable();
    IntWritable radio = new IntWritable();

    String[] parts = value.toString().split("[|]");

    radio.set(Integer.parseInt(parts[LastFMConstants.RADIO]));
    trackId.set(Integer.parseInt(parts[LastFMConstants.TRACK_ID]));

    if (parts.length == 5) {
        context.write(trackId, radio);
    } else {
        context.getCounter(COUNTERS.INVALID_RECORD_COUNT).increment(1L);
    }
}
```

- Phương thức reduce



```
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable x : values) {
        sum += x.get();
    }
    context.write(key, new IntWritable(sum));
}
```

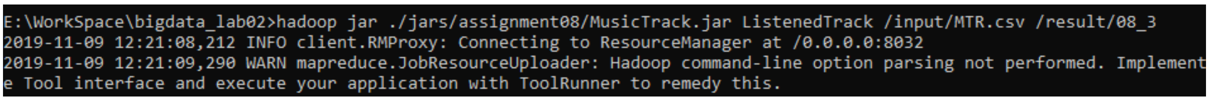
- Phương thức run



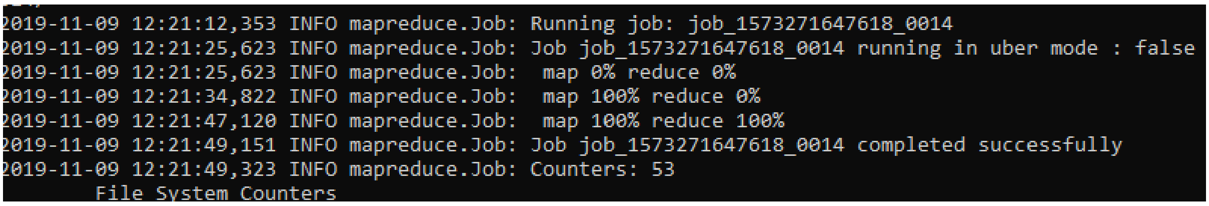
```
public void run(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Number of times listened per track in Radio");
    job.setJarByClass(ListenedTrack.class);

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

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

- Thực thi:

➔ hadoop ./jars/assignment08/MusicTrack.jar ListenedTrack /input/MTR.csv /result/08_3.

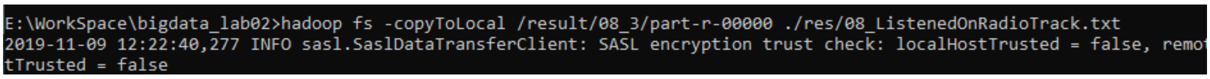


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment08/MusicTrack.jar ListenedTrack /input/MTR.csv /result/08_3
2019-11-09 12:21:08,212 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 12:21:09,290 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
e Tool interface and execute your application with ToolRunner to remedy this.
```



```
2019-11-09 12:21:12,353 INFO mapreduce.Job: Running job: job_1573271647618_0014
2019-11-09 12:21:25,623 INFO mapreduce.Job: Job job_1573271647618_0014 running in uber mode : false
2019-11-09 12:21:25,623 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 12:21:34,822 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 12:21:47,120 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 12:21:49,151 INFO mapreduce.Job: Job job_1573271647618_0014 completed successfully
2019-11-09 12:21:49,323 INFO mapreduce.Job: Counters: 53
        File System Counters
```

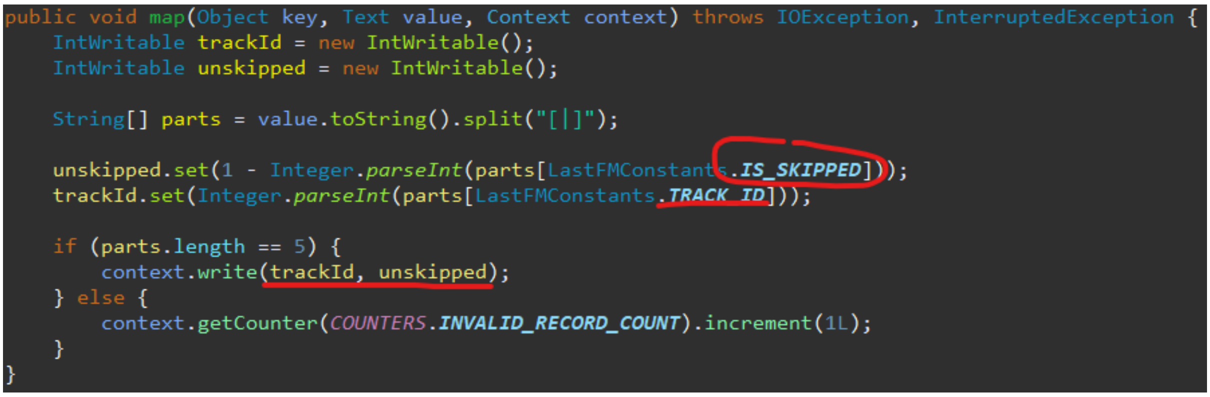
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/08_3/part-r-00000 ./res/08_ListenedOnRadioTrack.txt
2019-11-09 12:22:40,277 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remot
tTrusted = false
```

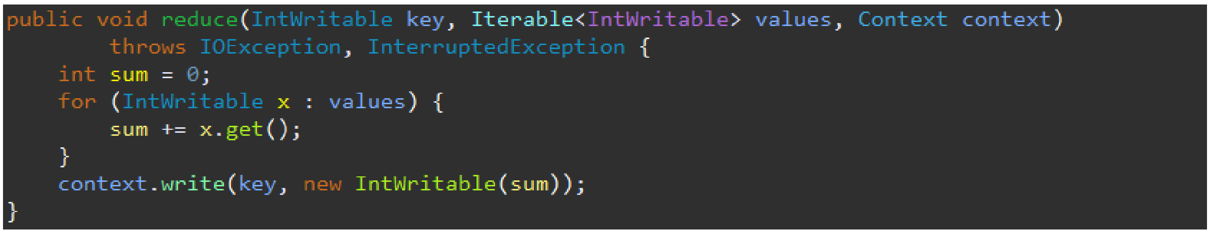
(4) Task 04: ListenedTrack (in total)

- Phương thức map



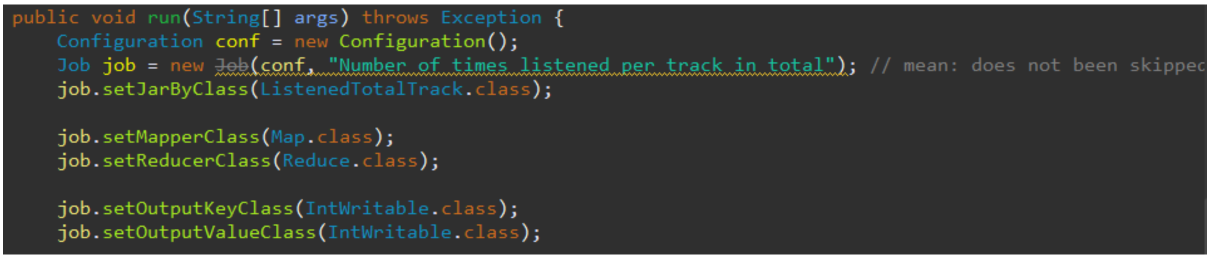
```
public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
    IntWritable trackId = new IntWritable();
    IntWritable unskipped = new IntWritable();

    String[] parts = value.toString().split("[|]");

    unskipped.set(1 - Integer.parseInt(parts[LastFMConstants.IS_SKIPPED]));
    trackId.set(Integer.parseInt(parts[LastFMConstants.TRACK_ID]));

    if (parts.length == 5) {
        context.write(trackId, unskipped);
    } else {
        context.getCounter(COUNTERS.INVALID_RECORD_COUNT).increment(1L);
    }
}
```

- Phương thức reduce



```
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable x : values) {
        sum += x.get();
    }
    context.write(key, new IntWritable(sum));
}
```

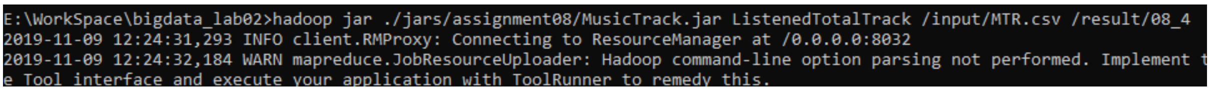
- Hàm run



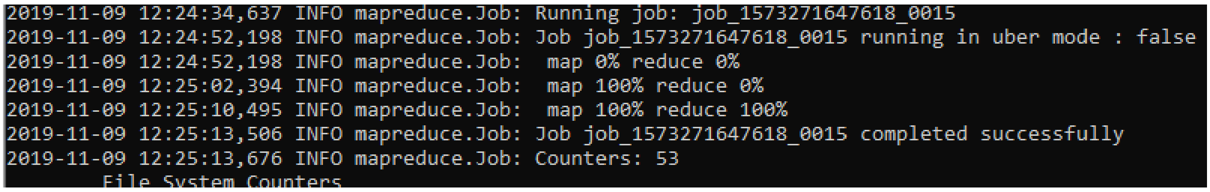
```
public void run(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Number of times listened per track in total"); // mean: does not been skippec
    job.setJarByClass(ListenedTotalTrack.class);

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

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

- Thực thi:

➔ hadoop ./jars/assignment08/MusicTrack.jar ListenedTotalTrack /input/MTR.csv /result/08_4.

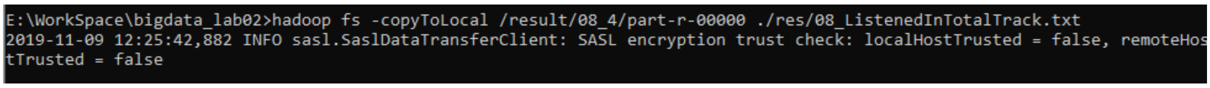


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment08/MusicTrack.jar ListenedTotalTrack /input/MTR.csv /result/08_4
2019-11-09 12:24:31,293 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 12:24:32,184 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
e Tool interface and execute your application with ToolRunner to remedy this.
```



```
2019-11-09 12:24:34,637 INFO mapreduce.Job: Running job: job_1573271647618_0015
2019-11-09 12:24:52,198 INFO mapreduce.Job: Job job_1573271647618_0015 running in uber mode : false
2019-11-09 12:24:52,198 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 12:25:02,394 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 12:25:10,495 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 12:25:13,506 INFO mapreduce.Job: Job job_1573271647618_0015 completed successfully
2019-11-09 12:25:13,676 INFO mapreduce.Job: Counters: 53
        File System Counters
```

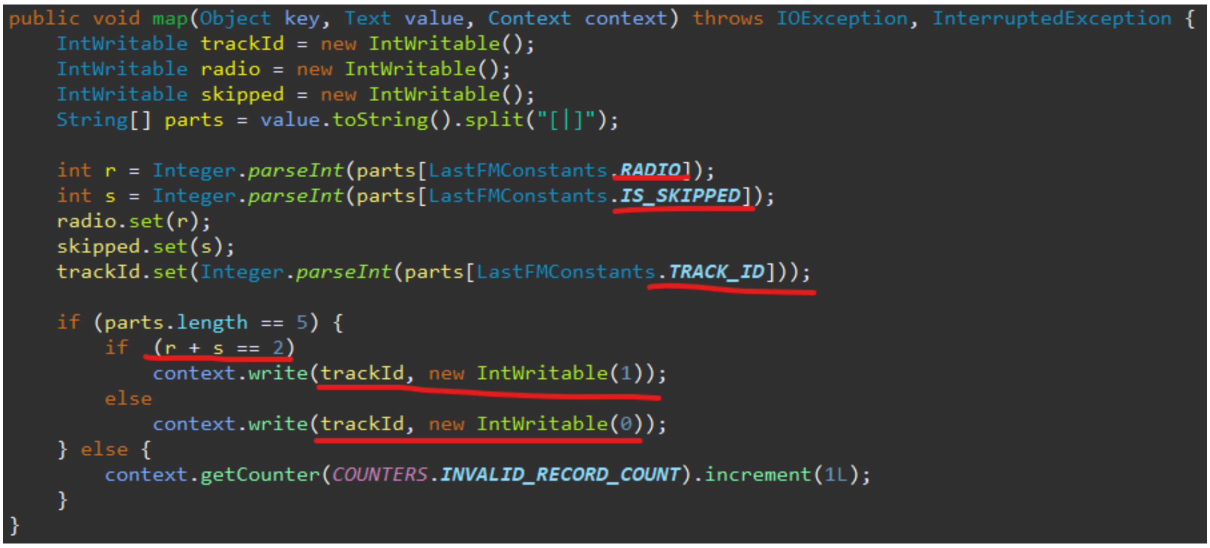
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/08_4/part-r-00000 ./res/08_ListenedInTotalTrack.txt
2019-11-09 12:25:42,882 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHos
tTrusted = false
```

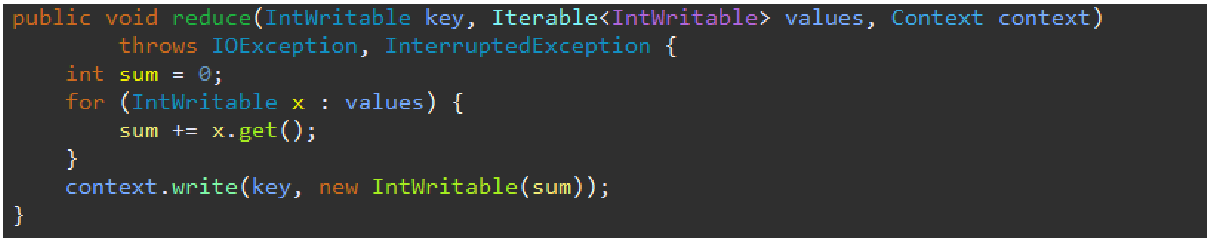
(5) Task 05: SkippedOnRadio

- Phương thức map



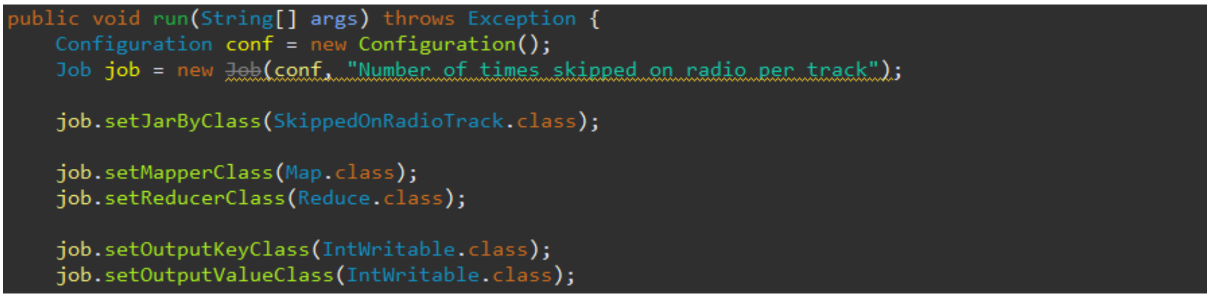
```
public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
    IntWritable trackId = new IntWritable();
    IntWritable radio = new IntWritable();
    IntWritable skipped = new IntWritable();
    String[] parts = value.toString().split("[|]");

    int r = Integer.parseInt(parts[LastFMConstants.RADIO]);
    int s = Integer.parseInt(parts[LastFMConstants.IS_SKIPPED]);
    radio.set(r);
    skipped.set(s);
    trackId.set(Integer.parseInt(parts[LastFMConstants.TRACK_ID]));

    if (parts.length == 5) {
        if  (r + s == 2)
            context.write(trackId, new IntWritable(1));
        else
            context.write(trackId, new IntWritable(0));
    } else {
        context.getCounter(COUNTERS.INVALID_RECORD_COUNT).increment(1L);
    }
}
```

- Phương thức reduce



```
public void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
        throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable x : values) {
        sum += x.get();
    }
    context.write(key, new IntWritable(sum));
}
```

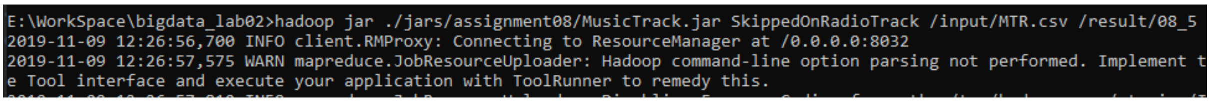
- Hàm run



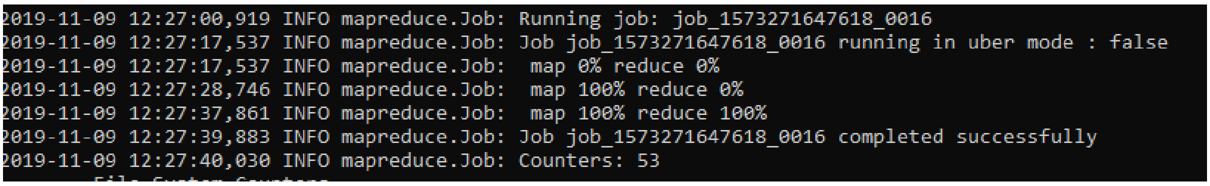
```
public void run(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new ~~Job~~(conf, "Number of times skipped on radio per track");

    job.setJarByClass(SkippedOnRadioTrack.class);

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

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

- Thực thi:

➔ hadoop ./jars/assignment08/MusicTrack.jar SkippedOnRadioTrack /input/MTR.csv /result/08_5.

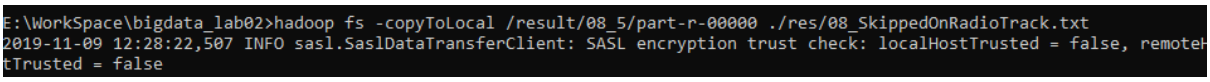


```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment08/MusicTrack.jar SkippedOnRadioTrack /input/MTR.csv /result/08_5
2019-11-09 12:26:56,700 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 12:26:57,575 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
e Tool interface and execute your application with ToolRunner to remedy this.
```



```
2019-11-09 12:27:00,919 INFO mapreduce.Job: Running job: job_1573271647618_0016
2019-11-09 12:27:17,537 INFO mapreduce.Job: Job job_1573271647618_0016 running in uber mode : false
2019-11-09 12:27:17,537 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 12:27:28,746 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 12:27:37,861 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 12:27:39,883 INFO mapreduce.Job: Job job_1573271647618_0016 completed successfully
2019-11-09 12:27:40,030 INFO mapreduce.Job: Counters: 53
```

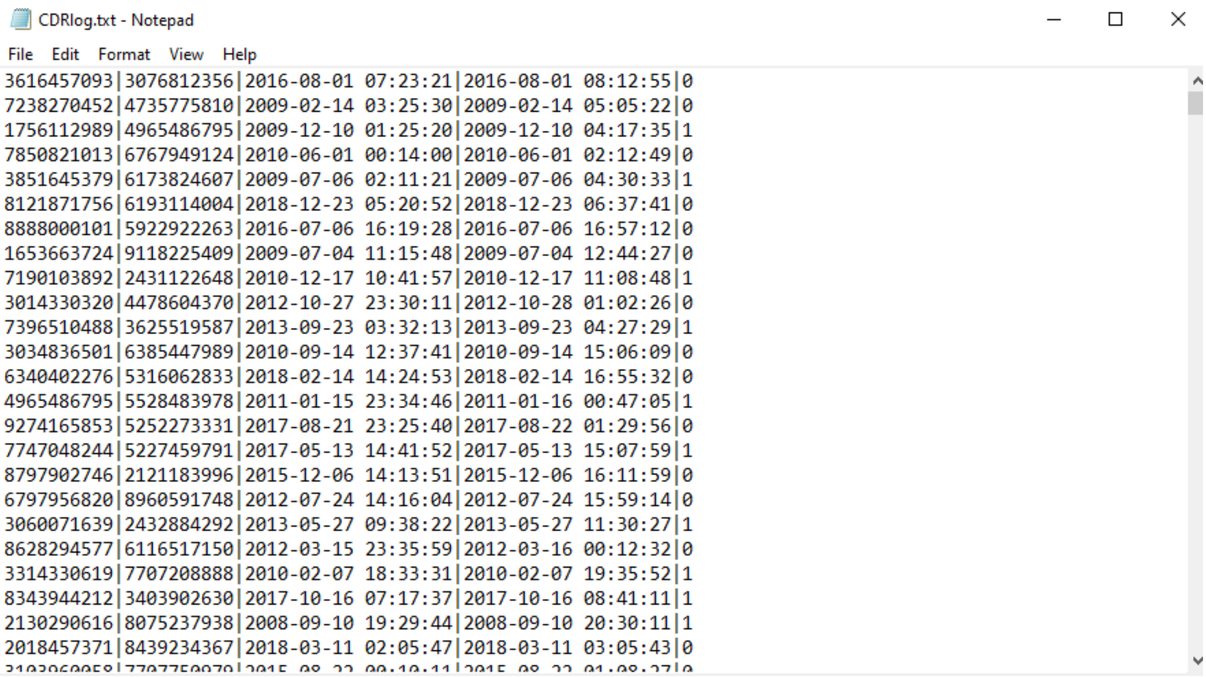
- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyToLocal /result/08_5/part-r-00000 ./res/08_SkippedOnRadioTrack.txt
2019-11-09 12:28:22,507 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
tTrusted = false
```

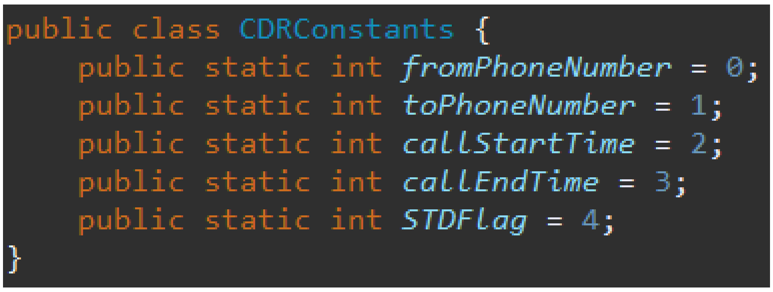
9. Assignment 09 – Telecom CDR

- Dataset



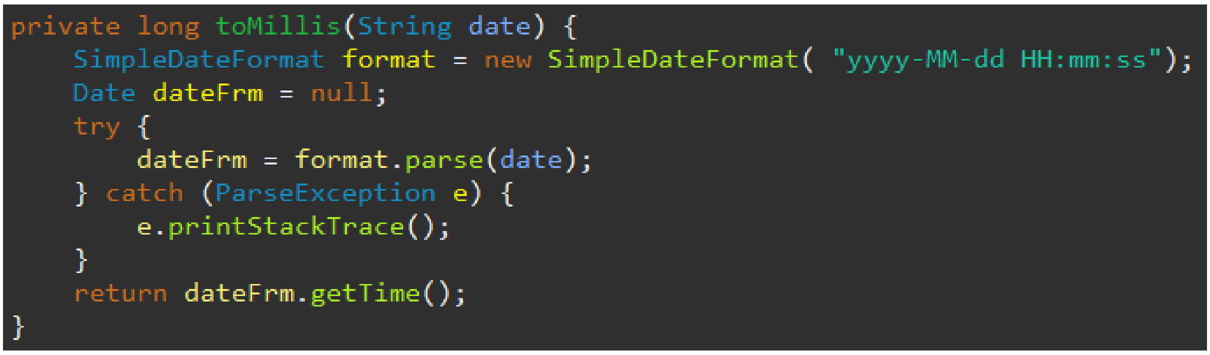
```
3616457093|3076812356|2016-08-01 07:23:21|2016-08-01 08:12:55|0
7238270452|4735775810|2009-02-14 03:25:30|2009-02-14 05:05:22|0
1756112989|4965486795|2009-12-10 01:25:20|2009-12-10 04:17:35|1
7850821013|6767949124|2010-06-01 00:14:00|2010-06-01 02:12:49|0
3851645379|6173824607|2009-07-06 02:11:21|2009-07-06 04:30:33|1
8121871756|6193114004|2018-12-23 05:20:52|2018-12-23 06:37:41|0
8888000101|5922922263|2016-07-06 16:19:28|2016-07-06 16:57:12|0
1653663724|9118225409|2009-07-04 11:15:48|2009-07-04 12:44:27|0
7190103892|2431122648|2010-12-17 10:41:57|2010-12-17 11:08:48|1
3014330320|4478604370|2012-10-27 23:30:11|2012-10-28 01:02:26|0
7396510488|3625519587|2013-09-23 03:32:13|2013-09-23 04:27:29|1
3034836501|6385447989|2010-09-14 12:37:41|2010-09-14 15:06:09|0
6340402276|5316062833|2018-02-14 14:24:53|2018-02-14 16:55:32|0
4965486795|5528483978|2011-01-15 23:34:46|2011-01-16 00:47:05|1
9274165853|5252273331|2017-08-21 23:25:40|2017-08-22 01:29:56|0
7747048244|5227459791|2017-05-13 14:41:52|2017-05-13 15:07:59|1
8797902746|2121183996|2015-12-06 14:13:51|2015-12-06 16:11:59|0
6797956820|8960591748|2012-07-24 14:16:04|2012-07-24 15:59:14|0
3060071639|2432884292|2013-05-27 09:38:22|2013-05-27 11:30:27|1
8628294577|6116517150|2012-03-15 23:35:59|2012-03-16 00:12:32|0
3314330619|7707208888|2010-02-07 18:33:31|2010-02-07 19:35:52|1
8343944212|3403902630|2017-10-16 07:17:37|2017-10-16 08:41:11|1
2130290616|8075237938|2008-09-10 19:29:44|2008-09-10 20:30:11|1
2018457371|8439234367|2018-03-11 02:05:47|2018-03-11 03:05:43|0
```

- Lớp CDRConstants để lưu vị trí tương ứng các giá trị trong mỗi record.



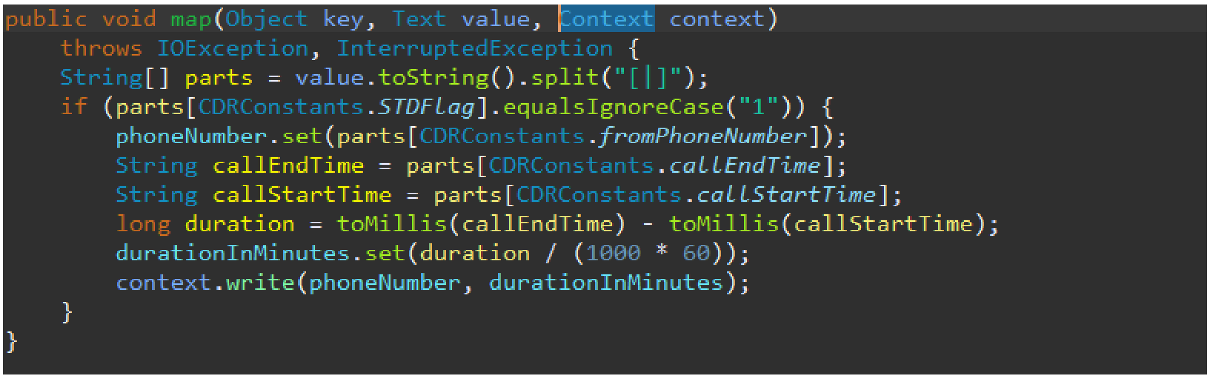
```java
public class CDRConstants {
    public static int fromPhoneNumber = 0;
    public static int toPhoneNumber = 1;
    public static int callStartTime = 2;
    public static int callEndTime = 3;
    public static int STDFlag = 4;
}
```

- Hàm toMillis chuyển kiểu dữ liệu String sang Date (đơn vị là mi li giây).



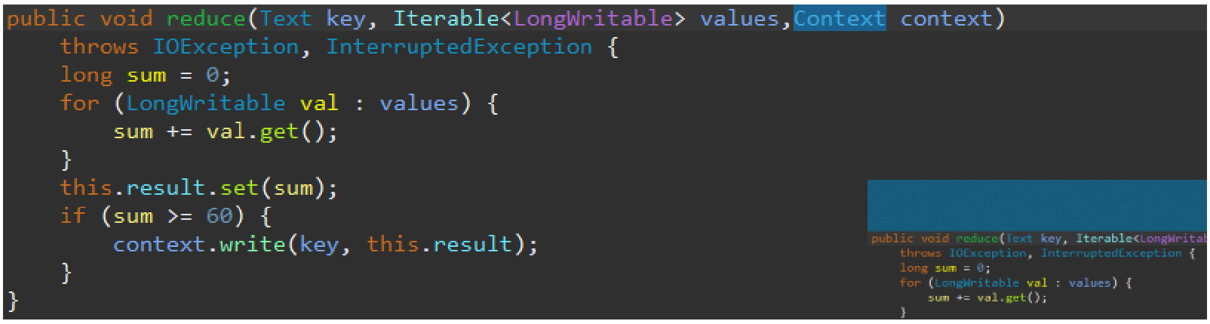
```
private long toMillis(String date) {
    SimpleDateFormat format = new SimpleDateFormat( "yyyy-MM-dd HH:mm:ss");
    Date dateFrm = null;
    try {
        dateFrm = format.parse(date);
    } catch (ParseException e) {
        e.printStackTrace();
    }
    return dateFrm.getTime();
}
```

- Phương thức map nhận đầu vào value ứng với mỗi record. Context ghi ra cặp <phone_number, minutes>
    - Chuyển record từ sang các chuỗi được split bởi "|".
    - Chỉ xét các record có flag (CDRConstants.STDFlag) là 1.
    - Với mỗi record thỏa điều kiện lấy được fromPhoneNumber, callStartTime và callEndTime.
    - Tính minutes từ start và end time ở trên.



```
public void map(Object key, Text value, Context context)
    throws IOException, InterruptedException {
    String[] parts = value.toString().split("[|]");
    if (parts[CDRConstants.STDFlag].equalsIgnoreCase("1")) {
        phoneNumber.set(parts[CDRConstants.fromPhoneNumber]);
        String callEndTime = parts[CDRConstants.callEndTime];
        String callStartTime = parts[CDRConstants.callStartTime];
        long duration = toMillis(callEndTime) - toMillis(callStartTime);
        durationInMinutes.set(duration / (1000 * 60));
        context.write(phoneNumber, durationInMinutes);
    }
}
```

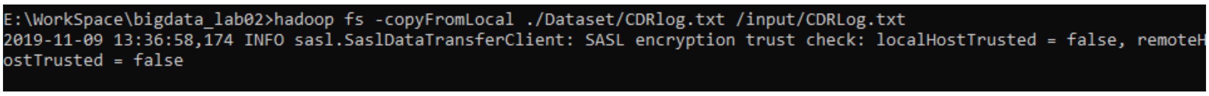
- Phương thức reduce đầu vào từ pha map key là phoneNumber, values là minutes. Context ghi ra cặp <phoneNumber, sum_minutes> thỏa sum_minutes lớn hơn 60.



```
public void reduce(Text key, Iterable<LongWritable> values,Context context)
    throws IOException, InterruptedException {
    long sum = 0;
    for (LongWritable val : values) {
        sum += val.get();
    }
    this.result.set(sum);
    if (sum >= 60) {
        context.write(key, this.result);
    }
}
```

- Thực thi:
    - Sao chép dữ liệu (dataset) từ local sang hdfs.

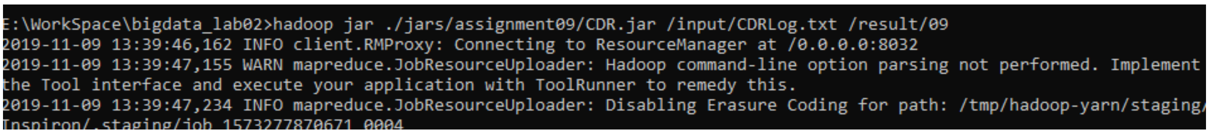
➔ hadoop fs -copyFromLocal ./Dataset/CDRLog.txt /input/CDRLog.txt



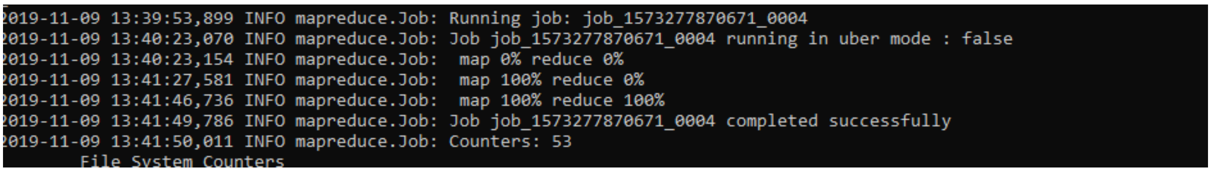
```
E:\WorkSpace\bigdata_lab02>hadoop fs -copyFromLocal ./Dataset/CDRlog.txt /input/CDRLog.txt
2019-11-09 13:36:58,174 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
```

- o Gọi lệnh jar trong Hadoop thực thi mapreduce với file jar tương ứng.

➔ hadoop jar ./jars/assignment09/CDR.jar /input/CDRLog.txt /result/09



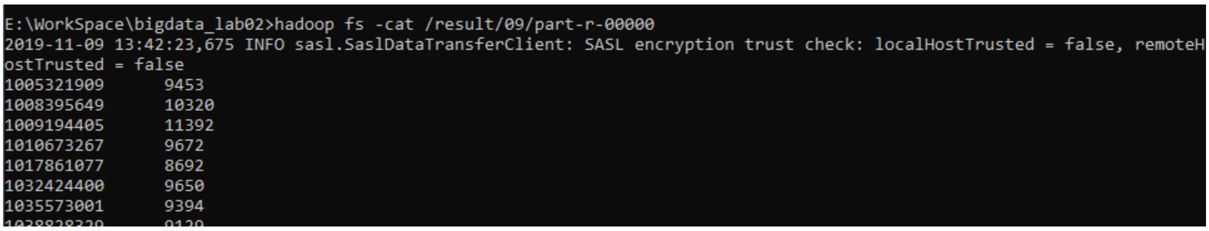
```
E:\WorkSpace\bigdata_lab02>hadoop jar ./jars/assignment09/CDR.jar /input/CDRLog.txt /result/09
2019-11-09 13:39:46,162 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2019-11-09 13:39:47,155 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement
the Tool interface and execute your application with ToolRunner to remedy this.
2019-11-09 13:39:47,234 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Inspiron/.staging/job_1573277870671_0004
```



```
2019-11-09 13:39:53,899 INFO mapreduce.Job: Running job: job_1573277870671_0004
2019-11-09 13:40:23,070 INFO mapreduce.Job: Job job_1573277870671_0004 running in uber mode : false
2019-11-09 13:40:23,154 INFO mapreduce.Job:  map 0% reduce 0%
2019-11-09 13:41:27,581 INFO mapreduce.Job:  map 100% reduce 0%
2019-11-09 13:41:46,736 INFO mapreduce.Job:  map 100% reduce 100%
2019-11-09 13:41:49,786 INFO mapreduce.Job: Job job_1573277870671_0004 completed successfully
2019-11-09 13:41:50,011 INFO mapreduce.Job: Counters: 53
        File System Counters
```

Quá trình chạy mapreduce job

- Kết quả:



```
E:\WorkSpace\bigdata_lab02>hadoop fs -cat /result/09/part-r-00000
2019-11-09 13:42:23,675 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
1005321909      9453
1008395649      10320
1009194405      11392
1010673267      9672
1017861077      8692
1032424400      9650
1035573001      9394
1038828320      9129
```

IV. THAM KHẢO:

- Lab02_Map Reduce.pdf