Ví dụ

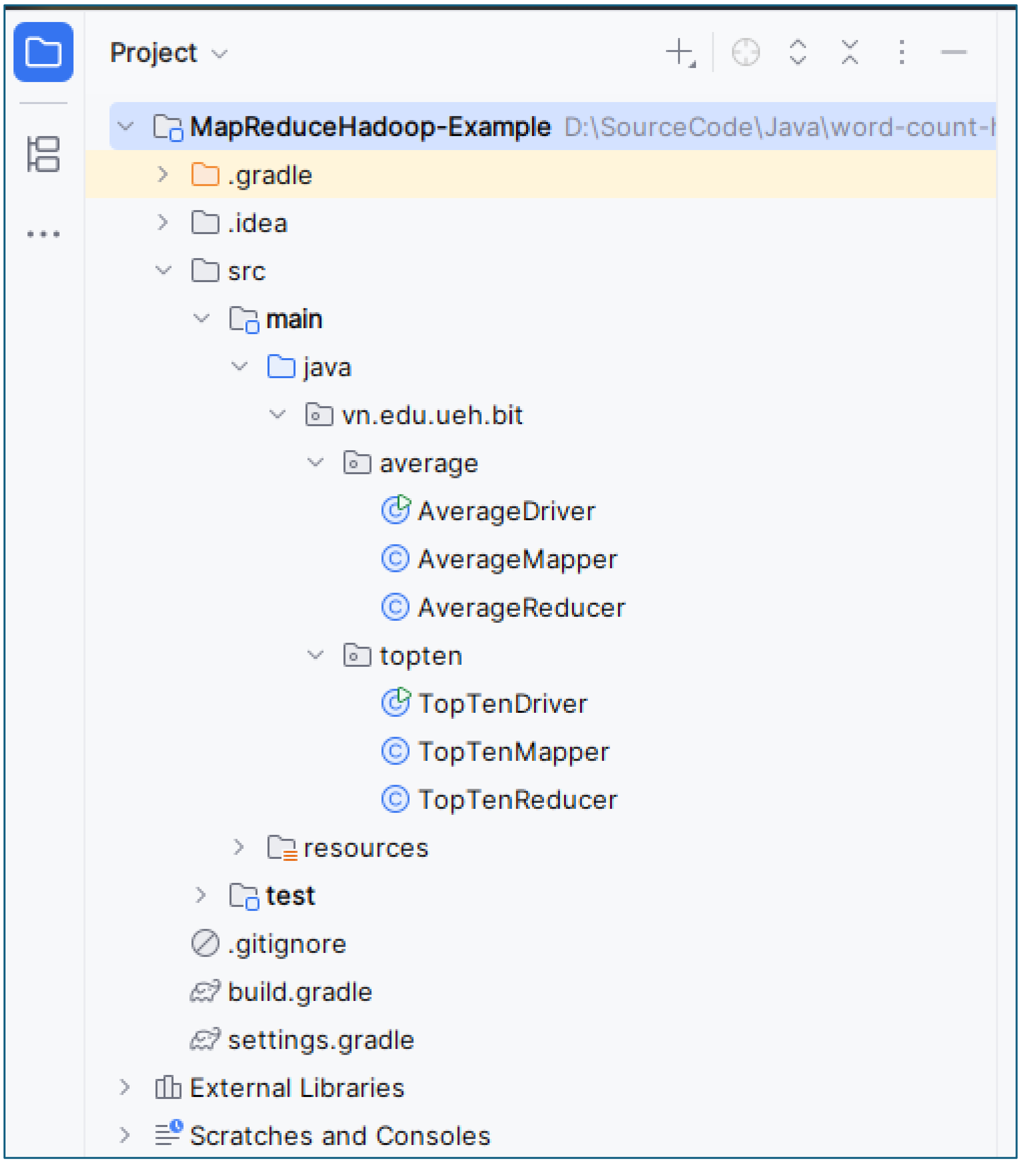
Trong phần này, chúng ta sẽ tiếp tục với một ví dụ trong đó hai công việc cùng đóng gói vào 1 file jar.

Dự án sau bao gồm 2 Chương trình

1. Average, tính giá trị giao dịch trung bình từ mọi ID (mỗi ID có thể có nhiều giao dịch).
2. TopTen, tìm kiếm mười giao dịch hàng đầu có giá trị cao nhất.

Coding

Tổ chức code



```
MapReduceHadoop-Example D:\SourceCode\Java\word-count-l
  .gradle
  .idea
  src
    main
      java
        vn.edu.ueh.bit
          average
            AverageDriver
            AverageMapper
            AverageReducer
          topten
            TopTenDriver
            TopTenMapper
            TopTenReducer
      resources
    test
  .gitignore
  build.gradle
  settings.gradle
External Libraries
Scratches and Consoles
```

Thêm dependencies cho application

```
implementation("org.apache.hadoop:hadoop-common:3.4.1")
implementation("org.apache.hadoop:hadoop-hdfs:3.4.1")
implementation("org.apache.hadoop:hadoop-mapreduce-client-core:3.4.1")
```

Code cho chương trình Average

```java
package vn.edu.ueh.bit.average;

import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class AverageMapper extends Mapper<Object, Text, Text, LongWritable> {
  private final LongWritable result = new LongWritable();

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

    String name = tokens[0];
    long val = Long.parseLong(tokens[1]);

    result.set(val);
    context.write(new Text(name), result);
  }
}
```

```java
package vn.edu.ueh.bit.average;

import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class AverageReducer extends Reducer<Text,
    LongWritable, Text, LongWritable> {
  private final LongWritable result = new LongWritable();

  @Override
  public void reduce(Text key, Iterable<LongWritable> values,
           Context context) throws IOException, InterruptedException {
    String name = key.toString();
    long sum = 0, count = 0;

    for (LongWritable val : values) {
      sum += val.get();
      count += 1;
    }

    result.set(sum / count);
    context.write(key, result);
  }
}
```

```java
package vn.edu.ueh.bit.average;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

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

        Job job = Job.getInstance(conf, "Average");

        job.setJarByClass(AverageDriver.class);
        job.setMapperClass(AverageMapper.class);
        job.setReducerClass(AverageReducer.class);

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

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

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

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

Code cho phần Topten

```
package vn.edu.ueh.bit.topten;

import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;
import java.util.Map;
import java.util.TreeMap;

public class TopTenMapper extends Mapper<Object, Text, Text, LongWritable> {

    private TreeMap<Long, String> record;

    @Override
    public void setup(Context context) throws IOException,
            InterruptedException {
        record = new TreeMap<Long, String>();
    }

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

        String name = tokens[0];
        long count = Long.parseLong(tokens[1]);

        record.put(count, name);
        if (record.size() > 10) {
            record.remove(record.firstKey());
        }
    }

    @Override
    public void cleanup(Context context) throws IOException,
            InterruptedException {
        for (Map.Entry<Long, String> entry : record.entrySet()) {

            long count = entry.getKey();
            String name = entry.getValue();

            context.write(new Text(name), new LongWritable(count));
        }
    }
}
```

```
package vn.edu.ueh.bit.topten;

import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;
import java.util.Map;
import java.util.TreeMap;

public class TopTenReducer extends Reducer<Text, LongWritable, Text, LongWritable> {
    private TreeMap<Long, String> record;

    @Override
    public void setup(Context context) throws IOException, InterruptedException {
        record = new TreeMap<Long, String>();
    }

    @Override
    public void reduce(Text key, Iterable<LongWritable> values,
                       Context context) throws IOException, InterruptedException {

        String name = key.toString();
        long count = 0;

        for (LongWritable val : values) {
            count = val.get();
        }

        record.put(count, name);

        if (record.size() > 10) {
            record.remove(record.firstKey());
        }
    }

    @Override
    public void cleanup(Context context) throws IOException, InterruptedException {
        for (Map.Entry<Long, String> entry : record.entrySet()) {

            long count = entry.getKey();
            String name = entry.getValue();
            context.write(new Text(name), new LongWritable(count));
        }
    }
}
```

```
package vn.edu.ueh.bit.topten;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;


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

        Job job = Job.getInstance(conf, "topten");

        job.setJarByClass(TopTenDriver.class);
        job.setMapperClass(TopTenMapper.class);
        job.setReducerClass(TopTenReducer.class);

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

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

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

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

Triển khai

Bước 1

Start HADOOP. Chạy 2 lệnh

```
start-dfs
start-yarn
```

Bước 2

Chuẩn bị dữ liệu

Tạo hai file text average.txt và topten.txt trong thư mục D:\data. Nội dung như sau

average.txt	topten.txt
transaction10 10 transaction20 20 transaction30 10 transaction40 10 transaction50 50 transaction40 10 transaction30 20 transaction10 10 transaction10 40	transaction10 13 transaction12 23 transaction13 1222 transaction14 123 transaction15 1230 transaction16 1 transaction17 223 transaction18 1023 transaction19 1213 transaction20 11 transaction21 13 transaction22 123 transaction23 12220 transaction24 12113 transaction25 1230 transaction26 1 transaction27 23 transaction28 1023 transaction29 11213 transaction30 10

Đặt thư mục hiện hành tại D:\data

Tạo thư mục textfiles trong hdfs

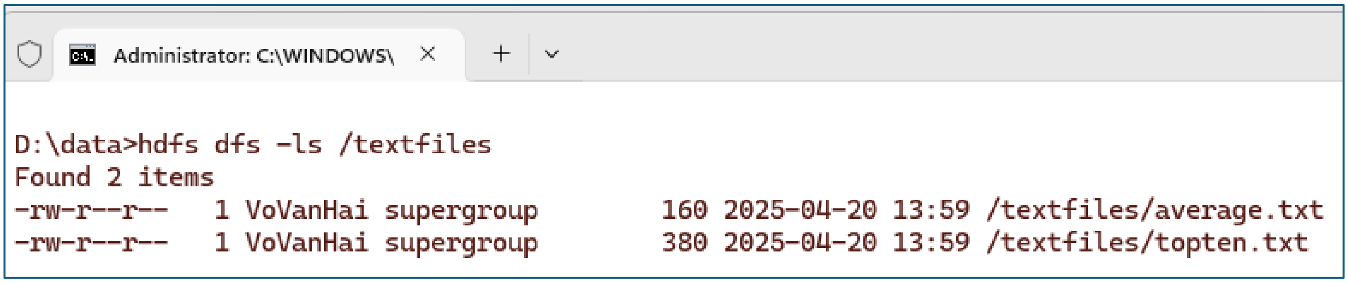
```
hdfs dfs -mkdir /textfiles
```

Tải hai file text lên

```
hdfs dfs -put topten.txt /textfiles/topten.txt
hdfs dfs -put average.txt /textfiles/average.txt
```

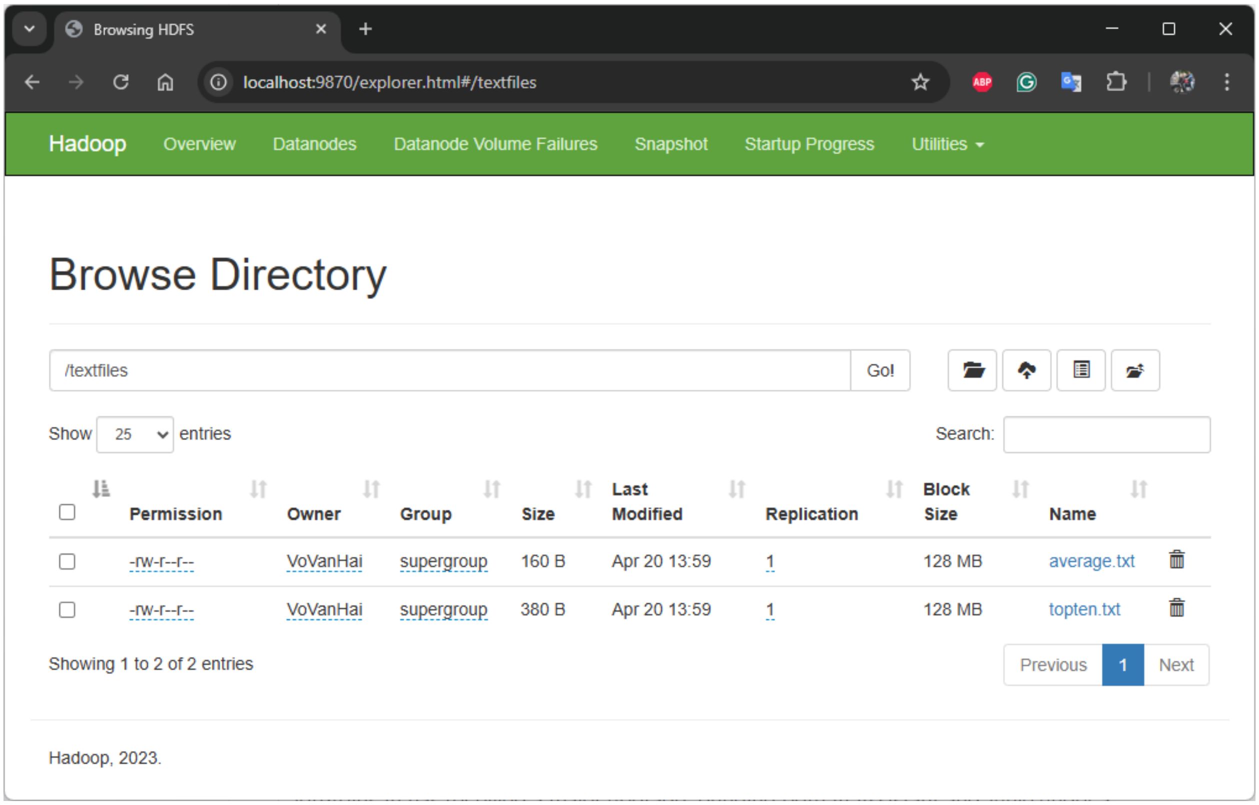
Xem nội dung thư mục thử có upload lên đầy đủ chưa

```
hadoop fs -ls /textfiles
```



```
D:\data>hdfs dfs -ls /textfiles
Found 2 items
-rw-r--r--   1 VoVanHai supergroup        160 2025-04-20 13:59 /textfiles/average.txt
-rw-r--r--   1 VoVanHai supergroup        380 2025-04-20 13:59 /textfiles/topten.txt
```

Có thể xem ở chế độ web ở địa chỉ: http://localhost:9870/explorer.html#/textfiles



Browse Directory

/textfiles

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	VoVanHai	supergroup	160 B	Apr 20 13:59	1	128 MB	average.txt
□	-rw-r--r--	VoVanHai	supergroup	380 B	Apr 20 13:59	1	128 MB	topten.txt

Showing 1 to 2 of 2 entries

Hadoop, 2023.

Bước 3

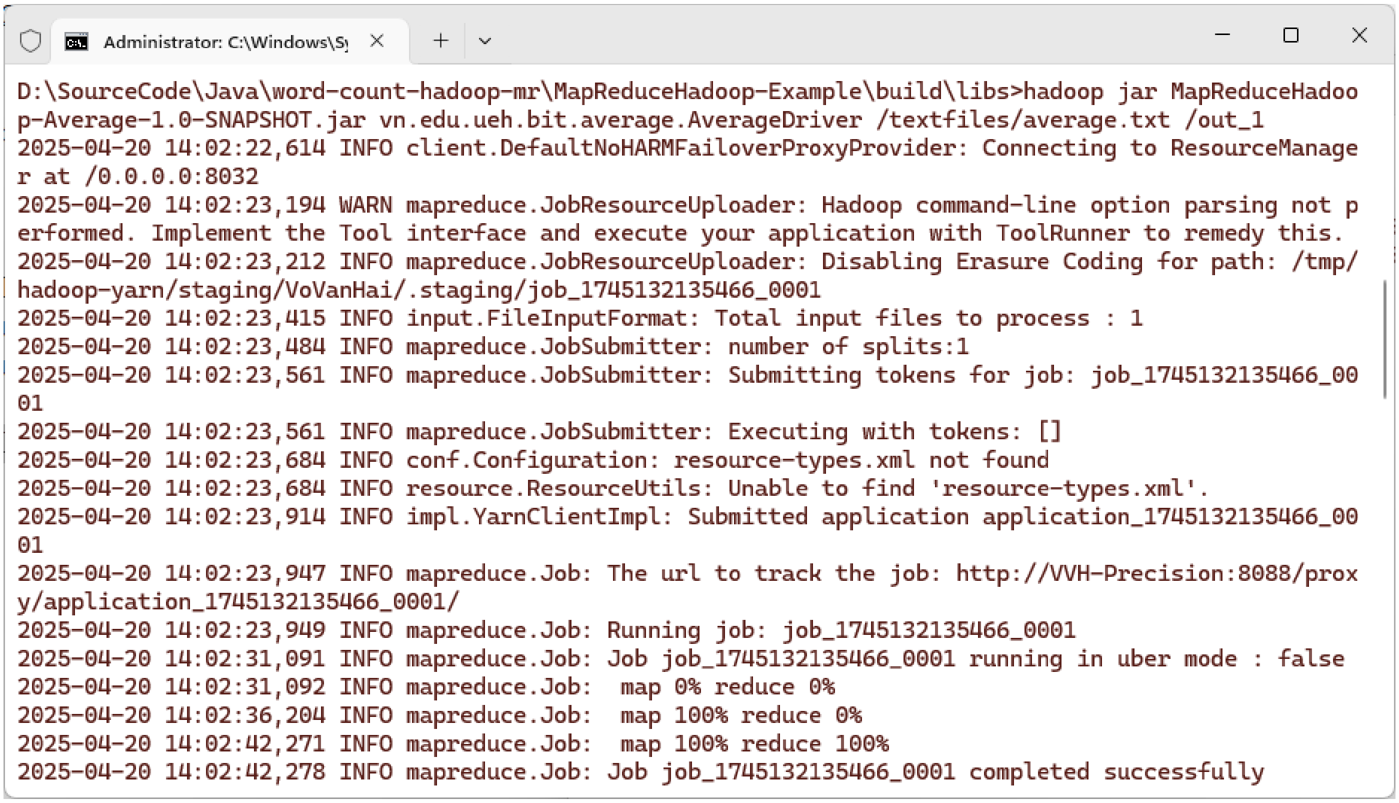
Thực thi chương trình

Giả sử thư mục jar file sau khi build project ở D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs.

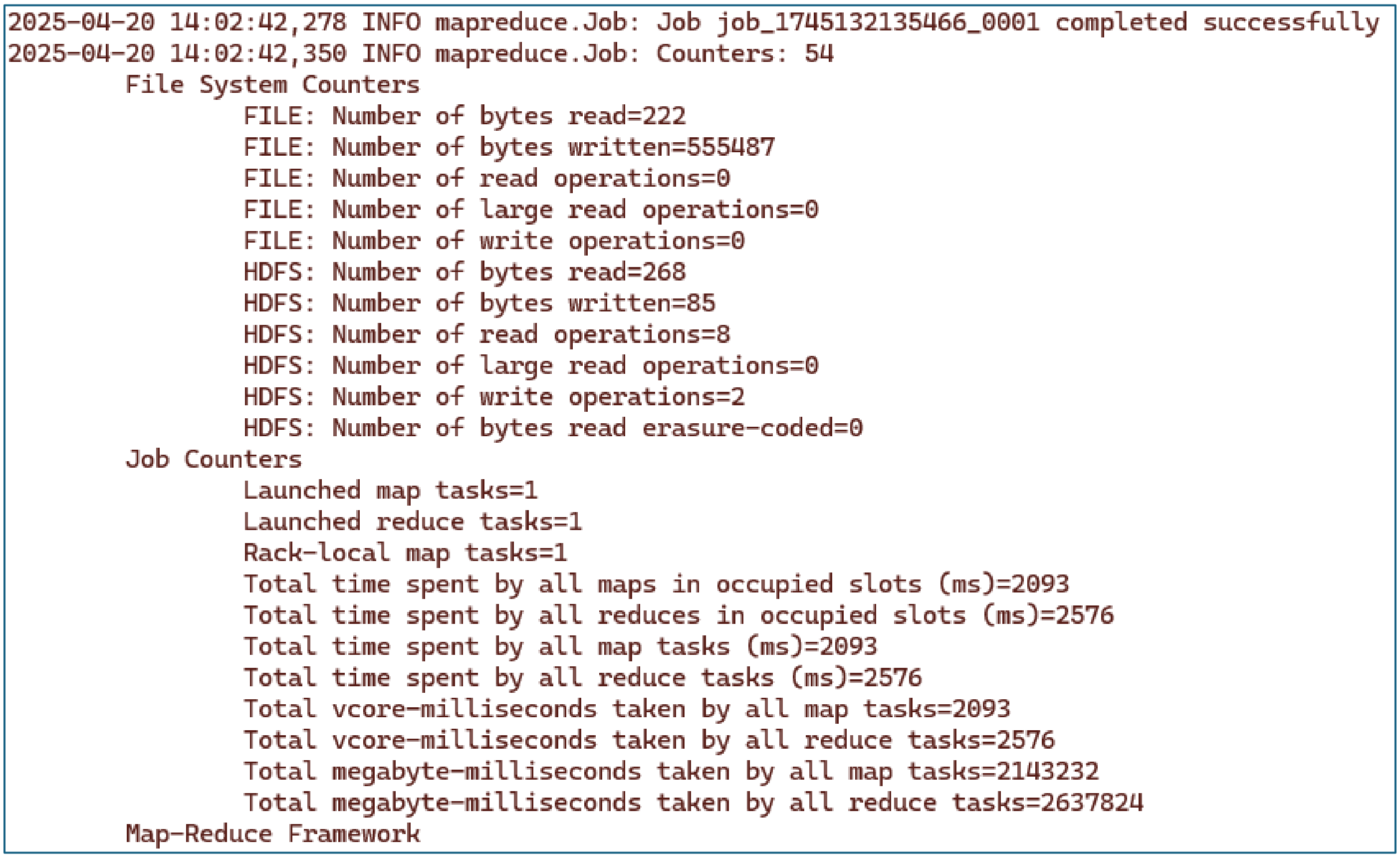
Ta có thể đặt thư mục hiện hành ở đây và tiến hành chạy lệnh Hadoop. Điều này tránh bớt việc copy file jar sau mỗi lần build.

Chạy lệnh sau để thự thi việc tính trung bình:

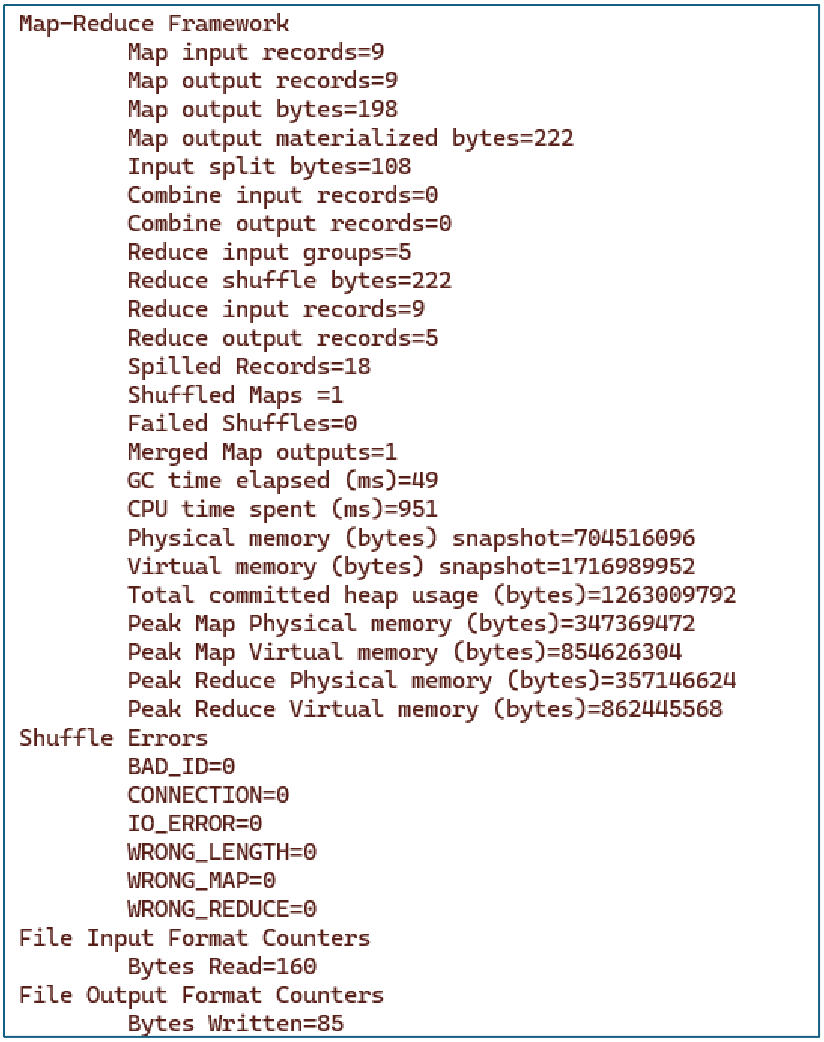
```
hadoop jar MapReduceHadoop-Average-1.0-SNAPSHOT.jar vn.edu.ueh.bit.average.AverageDriver
/textfiles/average.txt /out_1
```



```
D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop jar MapReduceHadoo
p-Average-1.0-SNAPSHOT.jar vn.edu.ueh.bit.average.AverageDriver /textfiles/average.txt /out_1
2025-04-20 14:02:22,614 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManage
r at /0.0.0.0:8032
2025-04-20 14:02:23,194 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not p
erformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-20 14:02:23,212 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/
hadoop-yarn/staging/VoVanHai/.staging/job_1745132135466_0001
2025-04-20 14:02:23,415 INFO input.FileInputFormat: Total input files to process : 1
2025-04-20 14:02:23,484 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-20 14:02:23,561 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745132135466_00
01
2025-04-20 14:02:23,561 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-20 14:02:23,684 INFO conf.Configuration: resource-types.xml not found
2025-04-20 14:02:23,684 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-20 14:02:23,914 INFO impl.YarnClientImpl: Submitted application application_1745132135466_00
01
2025-04-20 14:02:23,947 INFO mapreduce.Job: The url to track the job: http://VVH-Precision:8088/prox
y/application_1745132135466_0001/
2025-04-20 14:02:23,949 INFO mapreduce.Job: Running job: job_1745132135466_0001
2025-04-20 14:02:31,091 INFO mapreduce.Job: Job job_1745132135466_0001 running in uber mode : false
2025-04-20 14:02:31,092 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-20 14:02:36,204 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-20 14:02:42,271 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-20 14:02:42,278 INFO mapreduce.Job: Job job_1745132135466_0001 completed successfully
```



```
2025-04-20 14:02:42,278 INFO mapreduce.Job: Job job_1745132135466_0001 completed successfully
2025-04-20 14:02:42,350 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=222
                FILE: Number of bytes written=555487
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=268
                HDFS: Number of bytes written=85
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2093
                Total time spent by all reduces in occupied slots (ms)=2576
                Total time spent by all map tasks (ms)=2093
                Total time spent by all reduce tasks (ms)=2576
                Total vcore-milliseconds taken by all map tasks=2093
                Total vcore-milliseconds taken by all reduce tasks=2576
                Total megabyte-milliseconds taken by all map tasks=2143232
                Total megabyte-milliseconds taken by all reduce tasks=2637824
        Map-Reduce Framework
```



```
Map-Reduce Framework
        Map input records=9
        Map output records=9
        Map output bytes=198
        Map output materialized bytes=222
        Input split bytes=108
        Combine input records=0
        Combine output records=0
        Reduce input groups=5
        Reduce shuffle bytes=222
        Reduce input records=9
        Reduce output records=5
        Spilled Records=18
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=49
        CPU time spent (ms)=951
        Physical memory (bytes) snapshot=704516096
        Virtual memory (bytes) snapshot=1716989952
        Total committed heap usage (bytes)=1263009792
        Peak Map Physical memory (bytes)=347369472
        Peak Map Virtual memory (bytes)=854626304
        Peak Reduce Physical memory (bytes)=357146624
        Peak Reduce Virtual memory (bytes)=862445568
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=160
File Output Format Counters
        Bytes Written=85
```

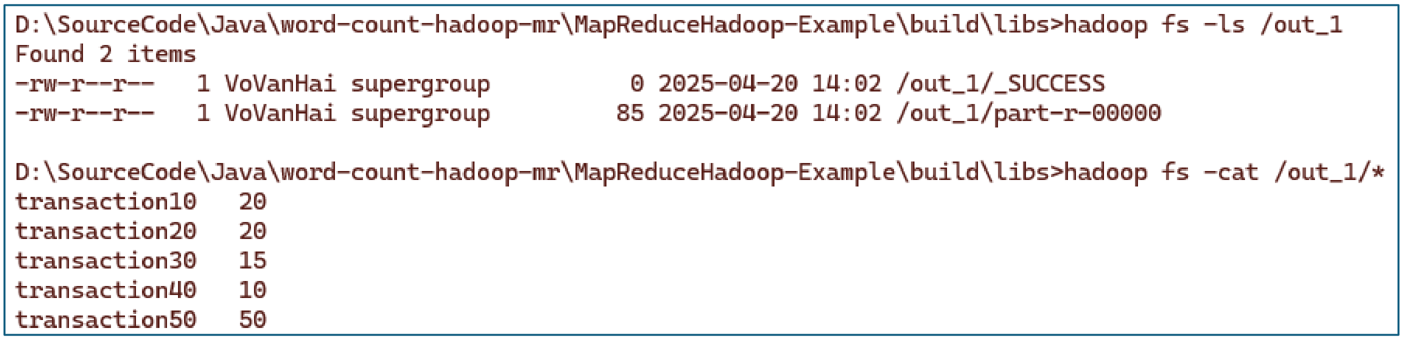
Chạy lệnh sau để xem thư mục kết quả

```
hadoop fs -ls /out_1
```

Sau đó chạy lệnh sau để xem kết quả

```
hadoop fs -cat /out_1/*
```

Kết quả như sau



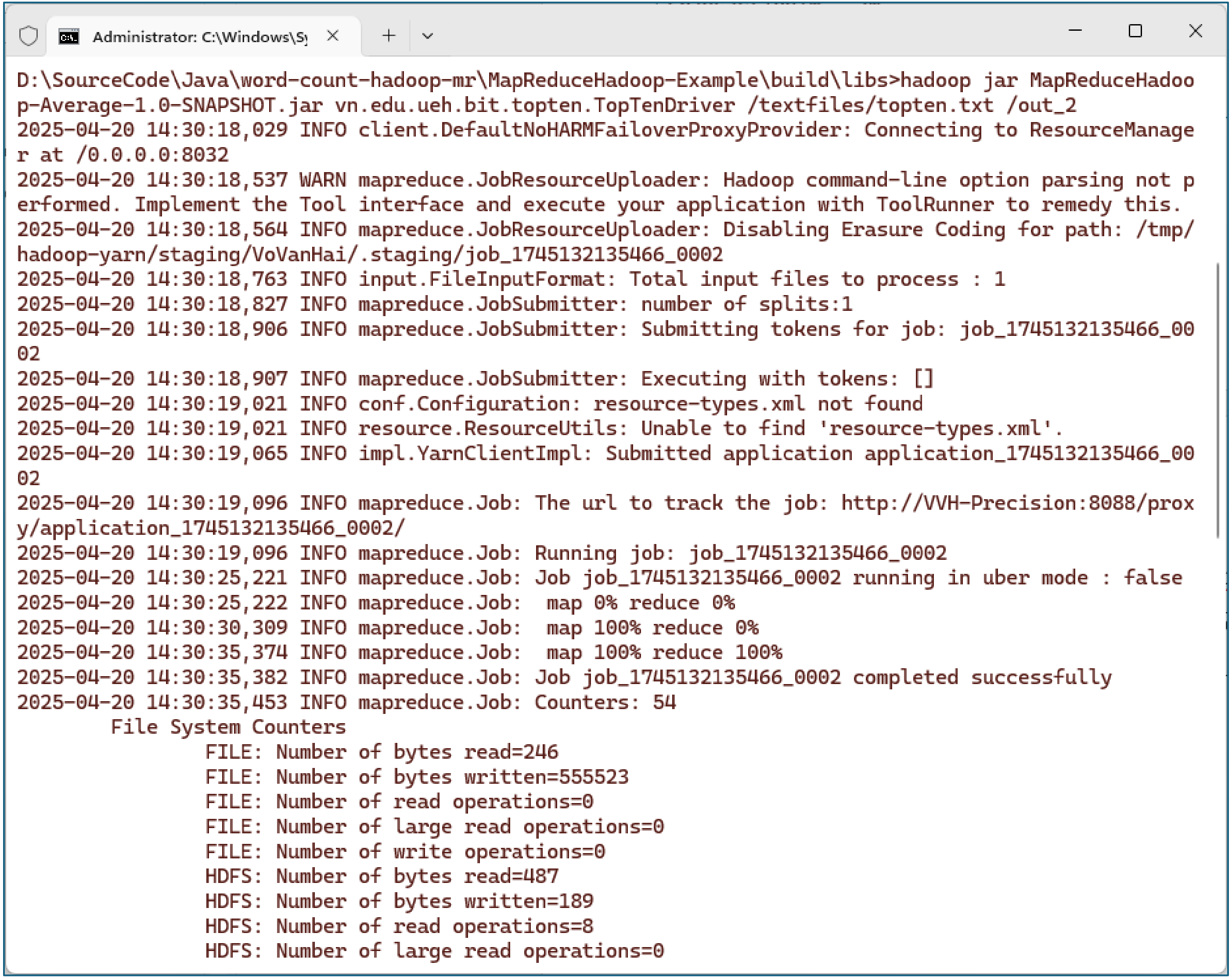
```
D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop fs -ls /out_1
Found 2 items
-rw-r--r--   1 VoVanHai supergroup          0 2025-04-20 14:02 /out_1/_SUCCESS
-rw-r--r--   1 VoVanHai supergroup         85 2025-04-20 14:02 /out_1/part-r-00000

D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop fs -cat /out_1/*
transaction10   20
transaction20   20
transaction30   15
transaction40   10
transaction50   50
```

Top ten

Tương tự, ta chạy câu lệnh sau để thực thi TopTenDriver

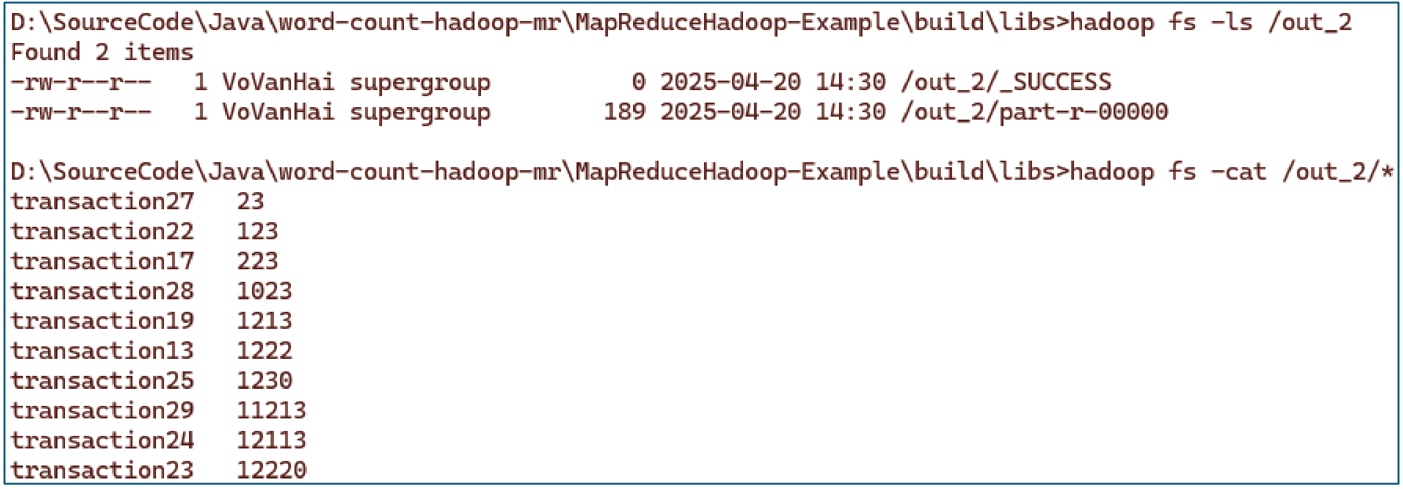
```
hadoop jar MapReduceHadoop-Average-1.0-SNAPSHOT.jar vn.edu.ueh.bit.topten.TopTenDriver
/textfiles/topten.txt /out_2
```



```
D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop jar MapReduceHadoo
p-Average-1.0-SNAPSHOT.jar vn.edu.ueh.bit.topten.TopTenDriver /textfiles/topten.txt /out_2
2025-04-20 14:30:18,029 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManage
r at /0.0.0.0:8032
2025-04-20 14:30:18,537 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not p
erformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-20 14:30:18,564 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/
hadoop-yarn/staging/VoVanHai/.staging/job_1745132135466_0002
2025-04-20 14:30:18,763 INFO input.FileInputFormat: Total input files to process : 1
2025-04-20 14:30:18,827 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-20 14:30:18,906 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1745132135466_00
02
2025-04-20 14:30:18,907 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-20 14:30:19,021 INFO conf.Configuration: resource-types.xml not found
2025-04-20 14:30:19,021 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-20 14:30:19,065 INFO impl.YarnClientImpl: Submitted application application_1745132135466_00
02
2025-04-20 14:30:19,096 INFO mapreduce.Job: The url to track the job: http://VVH-Precision:8088/prox
y/application_1745132135466_0002/
2025-04-20 14:30:19,096 INFO mapreduce.Job: Running job: job_1745132135466_0002
2025-04-20 14:30:25,221 INFO mapreduce.Job: Job job_1745132135466_0002 running in uber mode : false
2025-04-20 14:30:25,222 INFO mapreduce.Job:  map 0% reduce 0%
2025-04-20 14:30:30,309 INFO mapreduce.Job:  map 100% reduce 0%
2025-04-20 14:30:35,374 INFO mapreduce.Job:  map 100% reduce 100%
2025-04-20 14:30:35,382 INFO mapreduce.Job: Job job_1745132135466_0002 completed successfully
2025-04-20 14:30:35,453 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=246
                FILE: Number of bytes written=555523
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=487
                HDFS: Number of bytes written=189
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
```

Xem kết quả

```
Hadoop fs -ls /out_2
```



```
D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop fs -ls /out_2
Found 2 items
-rw-r--r--   1 VoVanHai supergroup          0 2025-04-20 14:30 /out_2/_SUCCESS
-rw-r--r--   1 VoVanHai supergroup        189 2025-04-20 14:30 /out_2/part-r-00000

D:\SourceCode\Java\word-count-hadoop-mr\MapReduceHadoop-Example\build\libs>hadoop fs -cat /out_2/*
transaction27   23
transaction22   123
transaction17   223
transaction28   1023
transaction19   1213
transaction13   1222
transaction25   1230
transaction29   11213
transaction24   12113
transaction23   12220
```