实验报告

实验名称	实验六 MapReduce 的使用		
实验教室	丹青 922	实验日期	2021 年 6 月 26 日
学　　号	2019210173	姓　　名	刘思远
专业班级	计算机科学与技术 04 班		
指导教师	卢洋		

东北林业大学

信息与计算机科学技术实验中心

一、 实验目的

1. 熟悉 MapReduce 的编程规范；

2. 熟悉 Hadoop 使用的数据类型；

3. 熟悉 Map 阶段的实现方法；

4. 熟悉 Reduce 阶段的实现方法；

5. 熟悉 Driver 的实现方法。

二、 实验环境

（1）计算机的硬件配置 PC 系列微机。

（2）计算机的软件配置 VMware 虚拟机软件及 Ubuntu 虚拟机。

三、 实验内容及结果

基于伪分布式运行模式或完全分布式运行模式，使用 Maven，通过编写 Mapper 、Reducer 和 Driver

进行如下实验：

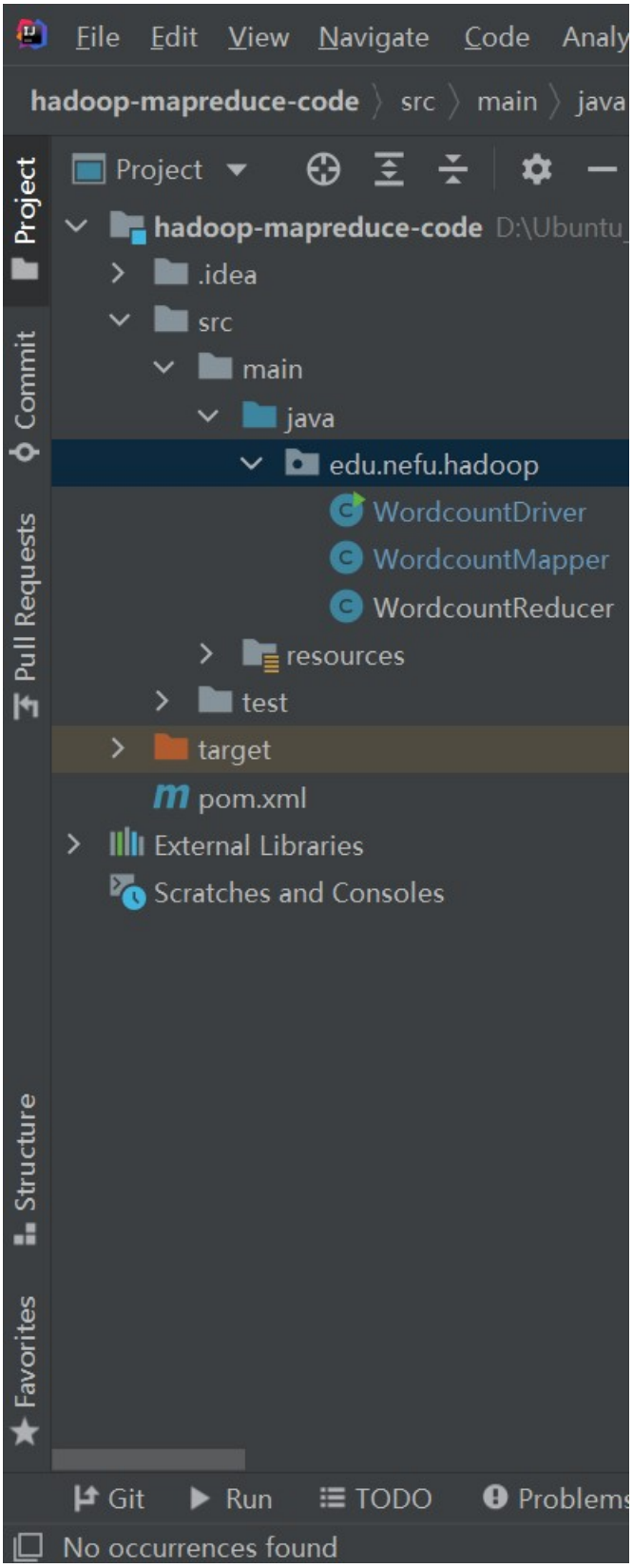
尽量在一个项目下进行实验，不同的功能通过运行时指定不同主类实现。

2.1 WordCount

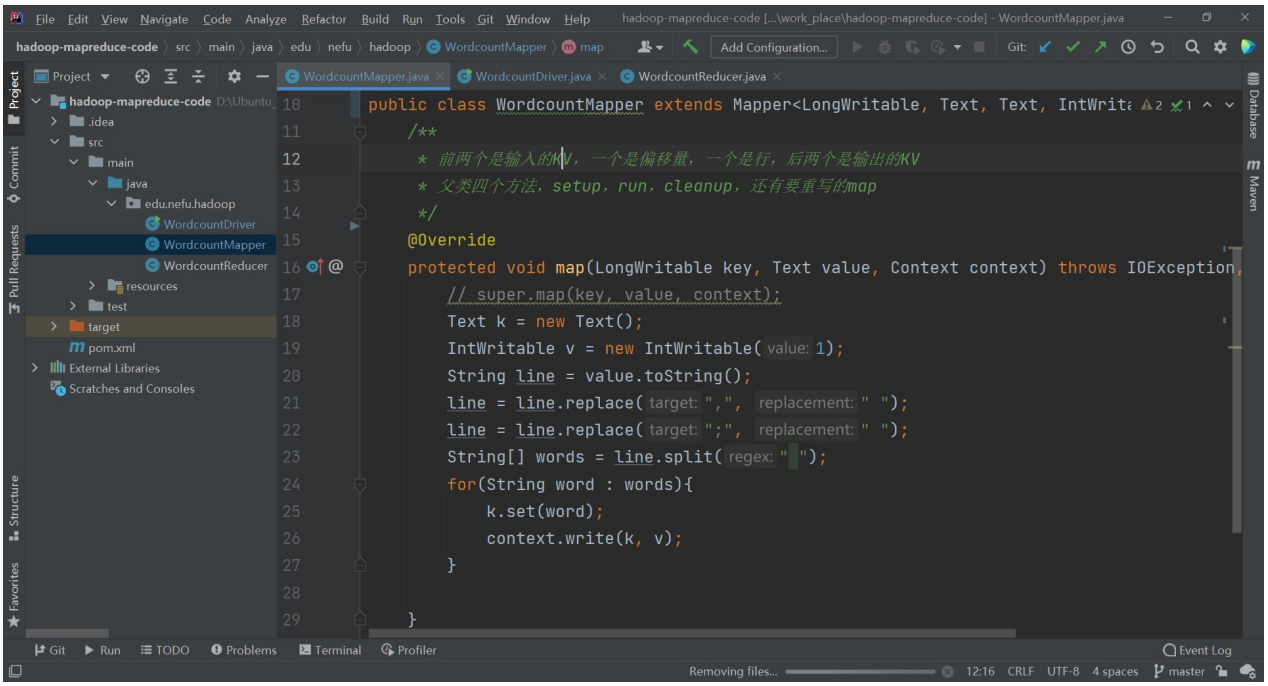
1. 下载 https://gitee.com/lueyoung/hadoop-empirical-data 项目中的 fail2ban.log 文件；


root@DESKTOP-TSQQRSN:/opt/apache-maven-3.0.5/work_place# git clone https://gitee.com/lueyoung/hadoop-empirical-data.git
Cloning into 'hadoop-empirical-data' ...
remote: Enumerating objects: 6, done.
remote: Counting objects: 100% (6/6), done.
remote: Compressing objects: 100% (5/5), done.
remote: Total 6 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (6/6), 34.27 KiB | 177.00 KiB/s, done.

2. 进行需求分析；

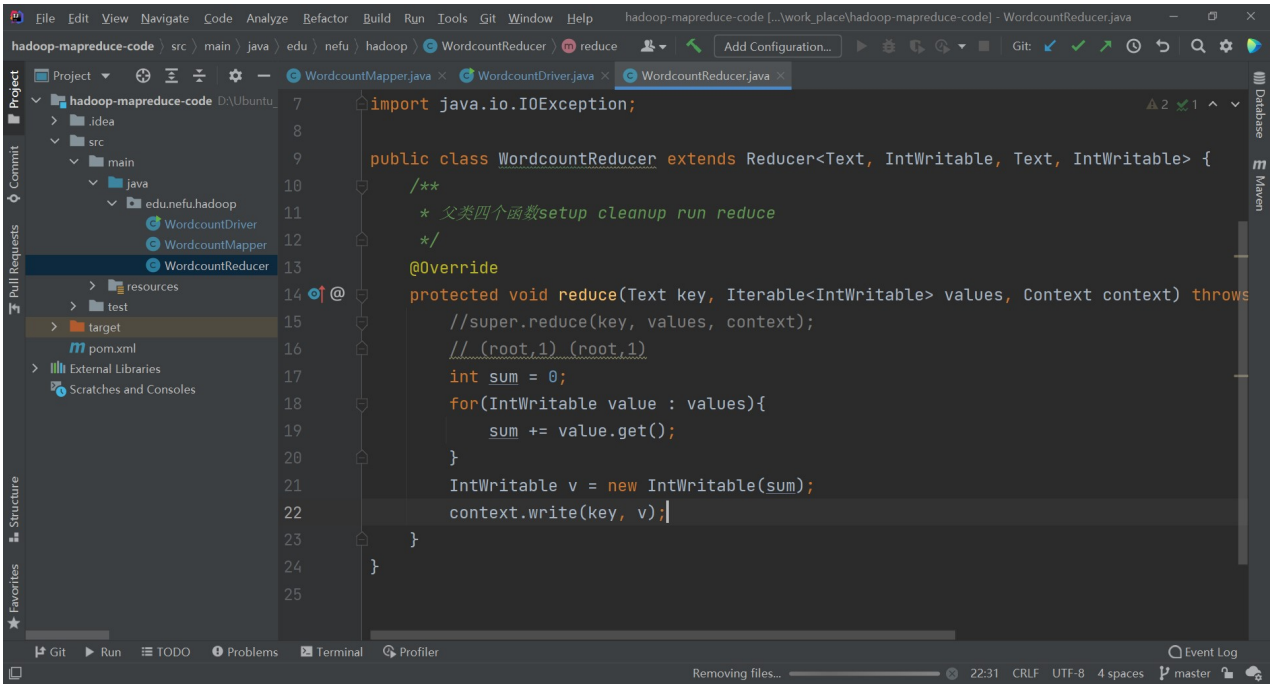

File Edit View Navigate Code Analy
hadoop-mapreduce-code src main java
Project
hadoop-mapreduce-code D:\Ubuntu
.idea
src
main
java
edu.nefu.hadoop
WordcountDriver
WordcountMapper
WordcountReducer
resources
test
target
pom.xml
External Libraries
Scratches and Consoles
Project
Commit
Pull Requests
Structure
Favorites
Git
Run
TODO
Problems
No occurrences found

3. 编写 Mapper ；



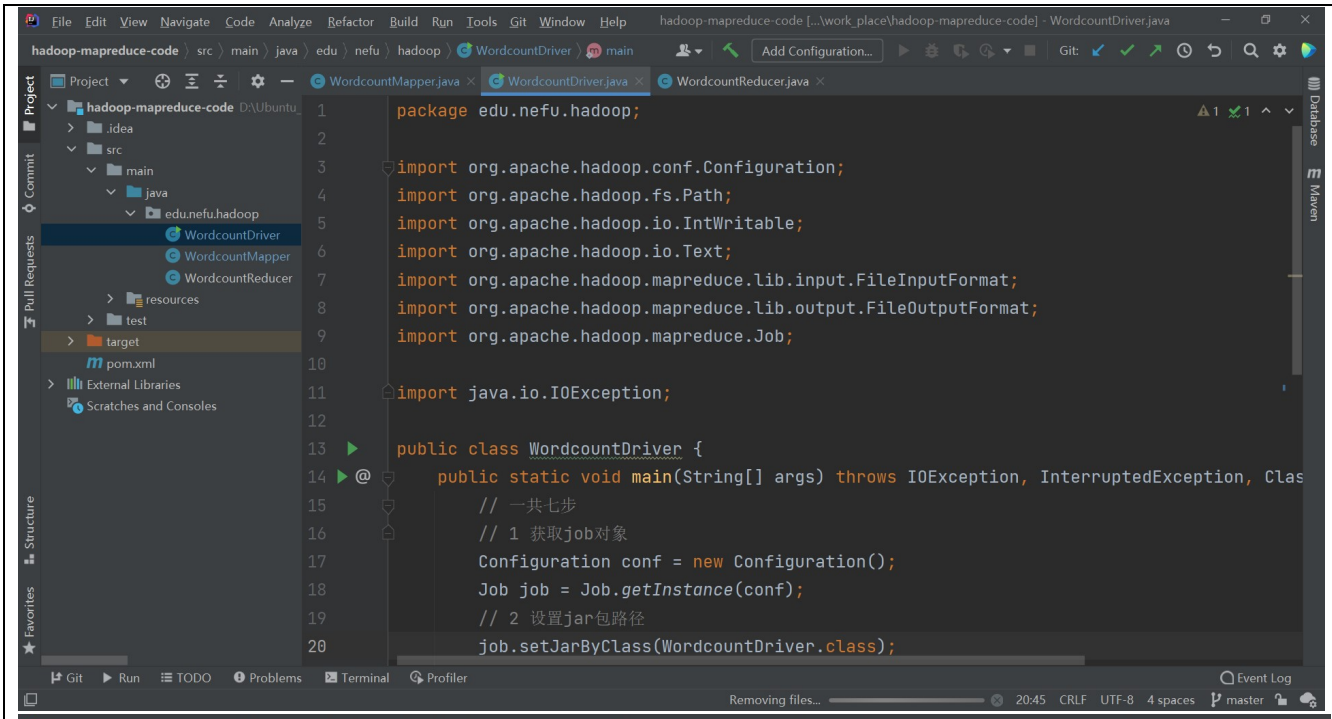
```
public class WordcountMapper extends Mapper<LongWritable, Text, Text, IntWrita
    /**
     * 前两个是输入的KV，一个是偏移量，一个是行，后两个是输出的KV
     * 父类四个方法，setup，run，cleanup，还有要重写的map
     */
    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException,
        // super.map(key, value, context);
        Text k = new Text();
        IntWritable v = new IntWritable( value: 1);
        String line = value.toString();
        line = line.replace( target: ",",  replacement: " ");
        line = line.replace( target: ";",  replacement: " ");
        String[] words = line.split( regex: " ");
        for(String word : words){
            k.set(word);
            context.write(k, v);
        }

    }
```

4. 编写 Reducer ；

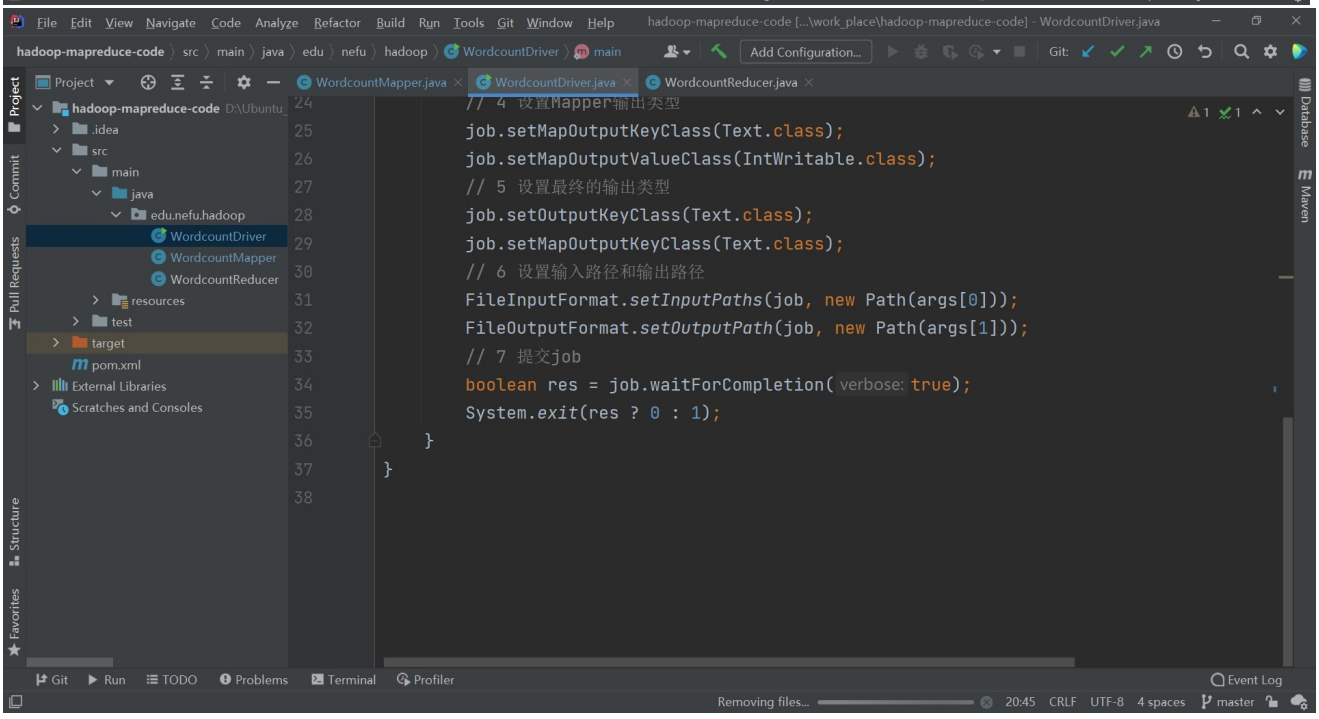


```
import java.io.IOException;

public class WordcountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    /**
     * 父类四个函数setup cleanup run reduce
     */
    @Override
    protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws
        //super.reduce(key, values, context);
        // (root,1) (root,1)
        int sum = 0;
        for(IntWritable value : values){
            sum += value.get();
        }
        IntWritable v = new IntWritable(sum);
        context.write(key, v);
    }
}
```

5. 编写 Driver ；

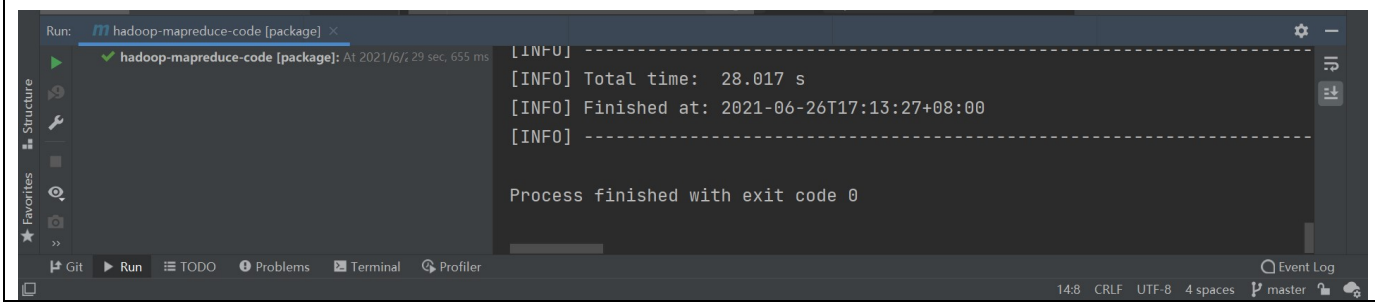


```
package edu.nefu.hadoop;

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

import java.io.IOException;

public class WordcountDriver {
    public static void main(String[] args) throws IOException, InterruptedException, Clas
        // 一共七步
        // 1 获取job对象
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf);
        // 2 设置jar包路径
        job.setJarByClass(WordcountDriver.class);
```



```
        // 4 设置Mapper输出类型
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);
        // 5 设置最终的输出类型
        job.setOutputKeyClass(Text.class);
        job.setMapOutputKeyClass(Text.class);
        // 6 设置输入路径和输出路径
        FileInputFormat.setInputPaths(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        // 7 提交job
        boolean res = job.waitForCompletion( verbose: true);
        System.exit(res ? 0 : 1);
    }
}
```

6. 使用 Maven ，创建 Jar 形式 Package ；



```
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  28.017 s
[INFO] Finished at: 2021-06-26T17:13:27+08:00
[INFO] ------------------------------------------------------------------------

Process finished with exit code 0
```

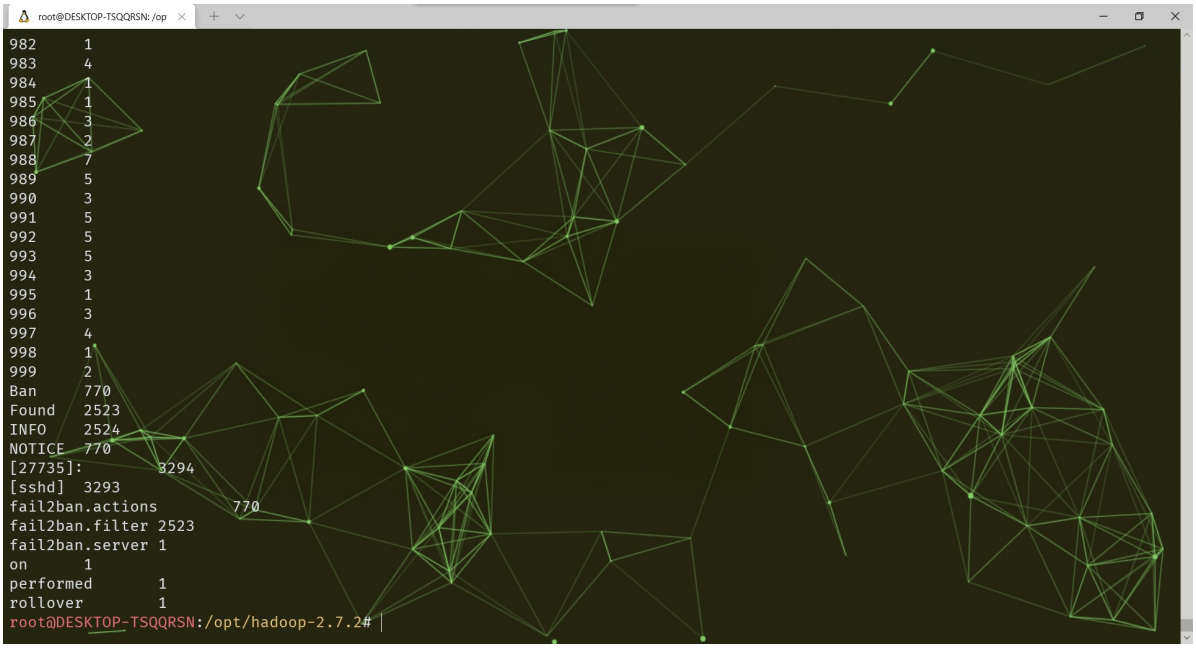
7. 运行项目；

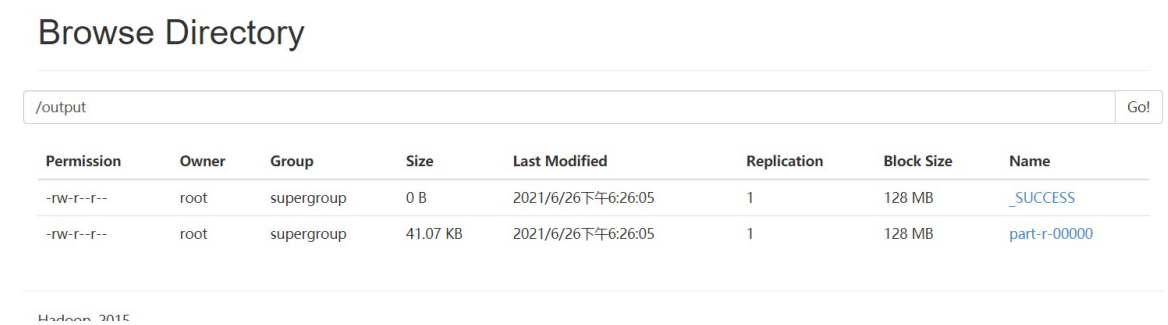
```
root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2# bin/hadoop jar hadoop-mapreduce-code-1.0-SNAPSHOT-jar-with-dependencies.jar edu.nefu.
hadoop.WordcountDriver /input /output
21/06/26 18:25:17 INFO client.RMProxy: Connecting to ResourceManager at DESKTOP-TSQQRSN/10.191.53.85:8032
21/06/26 18:25:17 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
21/06/26 18:25:20 INFO input.FileInputFormat: Total input paths to process : 1
21/06/26 18:25:21 INFO mapreduce.JobSubmitter: number of splits:1
21/06/26 18:25:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624699033715_0002
21/06/26 18:25:24 INFO impl.YarnClientImpl: Submitted application application_1624699033715_0002
21/06/26 18:25:24 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TSQQRSN:8088/proxy/application_1624699033715_0
002/
21/06/26 18:25:24 INFO mapreduce.Job: Running job: job_1624699033715_0002
21/06/26 18:25:48 INFO mapreduce.Job: Job job_1624699033715_0002 running in uber mode : false
21/06/26 18:25:49 INFO mapreduce.Job:  map 0% reduce 0%
21/06/26 18:25:56 INFO mapreduce.Job:  map 100% reduce 0%
21/06/26 18:26:07 INFO mapreduce.Job:  map 100% reduce 100%
21/06/26 18:26:07 INFO mapreduce.Job: Job job_1624699033715_0002 completed successfully
21/06/26 18:26:07 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=784990
                FILE: Number of bytes written=1804465
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

```
root@DESKTOP-TSQQRSN: /op
                Map output records=71140
                Map output bytes=642704
                Map output materialized bytes=784990
                Input split bytes=111
                Combine input records=0
                Combine output records=0
                Reduce input groups=3871
                Reduce shuffle bytes=784990
                Reduce input records=71140
                Reduce output records=3871
                Spilled Records=142280
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=93
                CPU time spent (ms)=3040
                Physical memory (bytes) snapshot=432246784
                Virtual memory (bytes) snapshot=2852217200640
                Total committed heap usage (bytes)=303562752
        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=358144
        File Output Format Counters
                Bytes Written=42059
root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2#
```

8. 观察并分析结果。


root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2# hadoop fs -cat /output/p*
33924
- 2523
/var/log/fail2ban.log 1
000 2
001 4
002 3
003 3
004 4
005 3
006 6
007 3
008 3
009 3
00:01:12 4
00:01:14 2
00:01:15 1
00:02:23 4
00:02:26 3
00:03:39 4
00:03:41 1
00:03:42 2
00:09:50 4
00:09:52 1

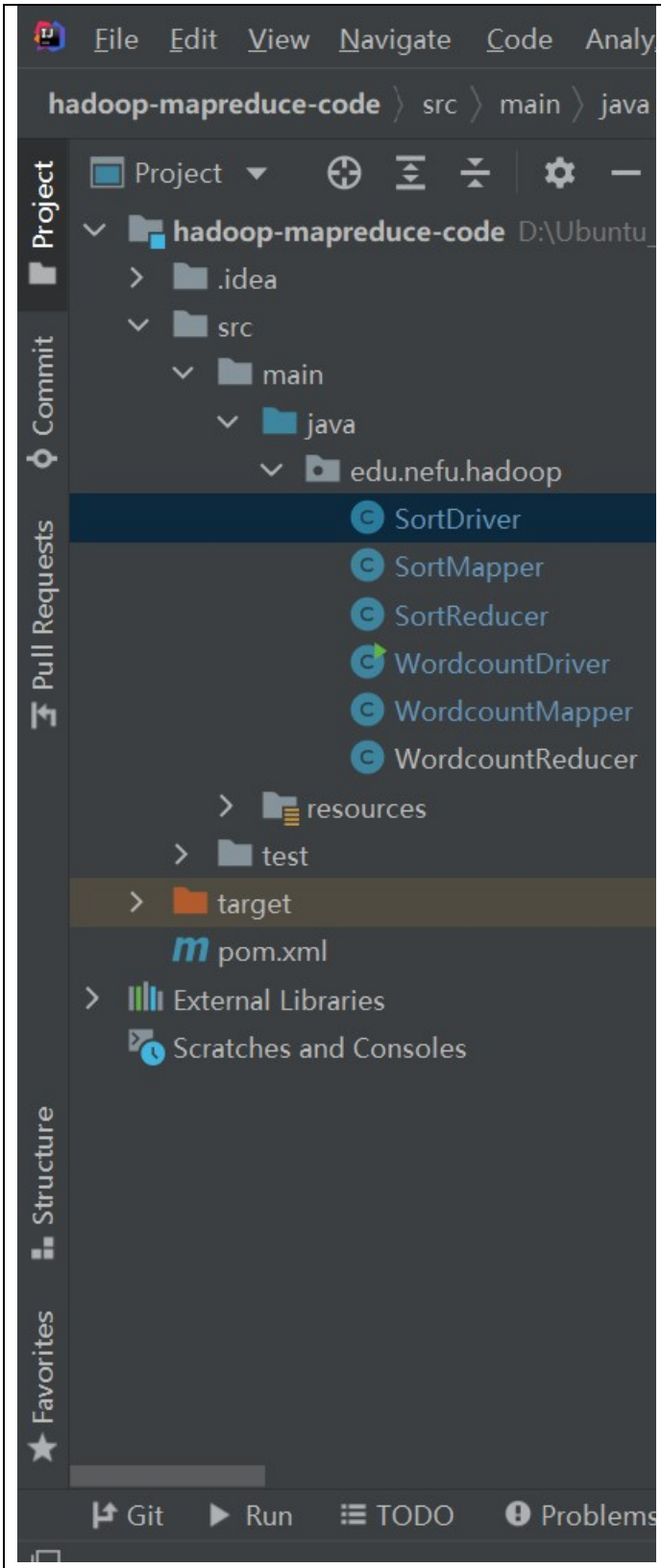

root@DESKTOP-TSQQRSN: /op
982 1
983 4
984 1
985 1
986 3
987 2
988 7
989 5
990 3
991 5
992 5
993 5
994 3
995 1
996 3
997 4
998 1
999 2
Ban 770
Found 2523
INFO 2524
NOTICE 770
[27735]: 3294
[sshd] 3293
fail2ban.actions 770
fail2ban.filter 2523
fail2ban.server 1
on 1
performed 1
rollover 1
root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2#


Browse Directory
/output
Go!
Permission Owner Group Size Last Modified Replication Block Size Name
-rw-r--r-- root supergroup 0 B 2021/6/26下午6:26:05 1 128 MB _SUCCESS
-rw-r--r-- root supergroup 41.07 KB 2021/6/26下午6:26:05 1 128 MB part-r-00000

2.2 排序

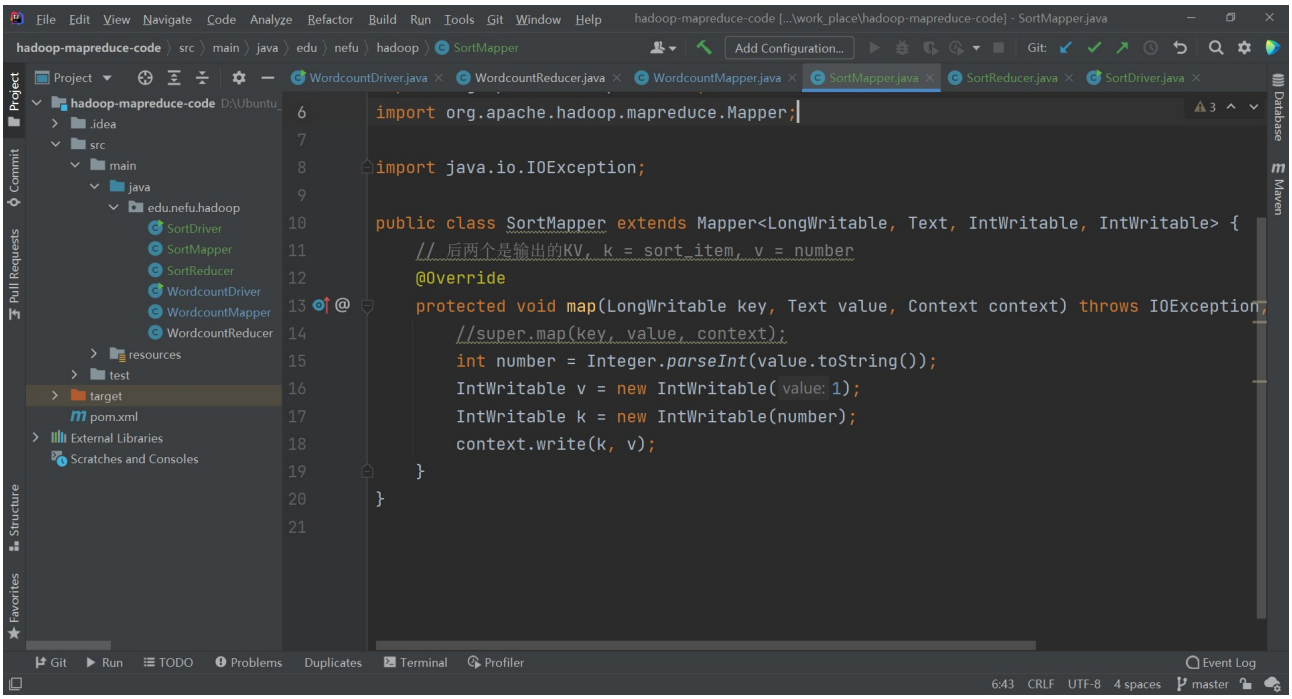
1. 下载 https://gitee.com/lueyoung/hadoop-empirical-data 项目中的 nums.txt 文件；

2. 进行需求分析；



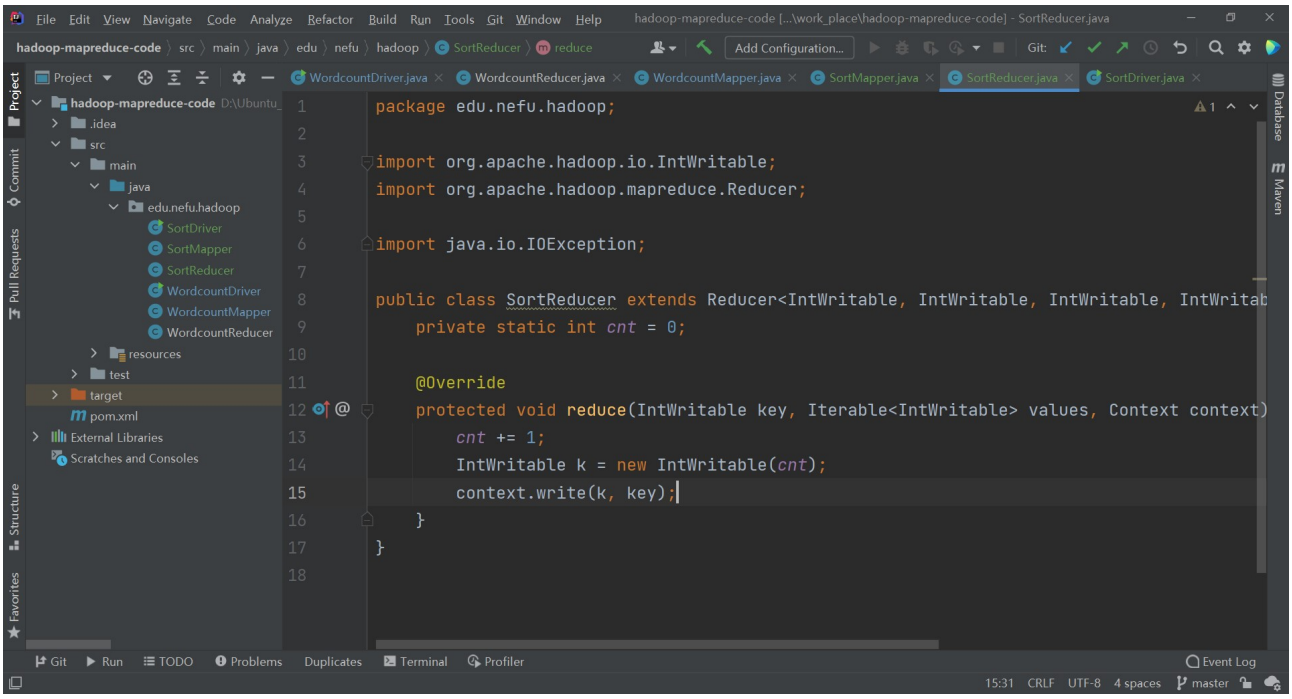
要求输出的形式为 序号 数值 的形式，如：

3. 编写 Mapper ；



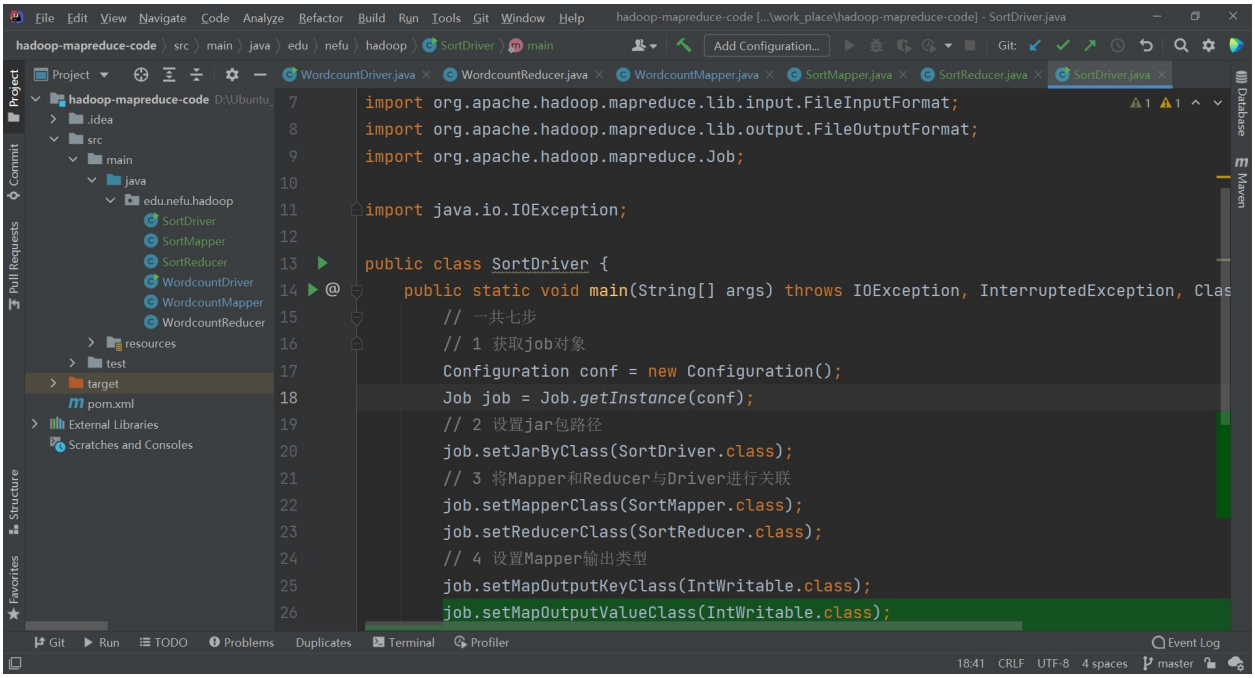
```
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class SortMapper extends Mapper<LongWritable, Text, IntWritable, IntWritable> {
    // 后两个是输出的KV, k = sort_item, v = number
    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException,
        //super.map(key, value, context);
        int number = Integer.parseInt(value.toString());
        IntWritable v = new IntWritable( value: 1);
        IntWritable k = new IntWritable(number);
        context.write(k, v);
    }
}
```

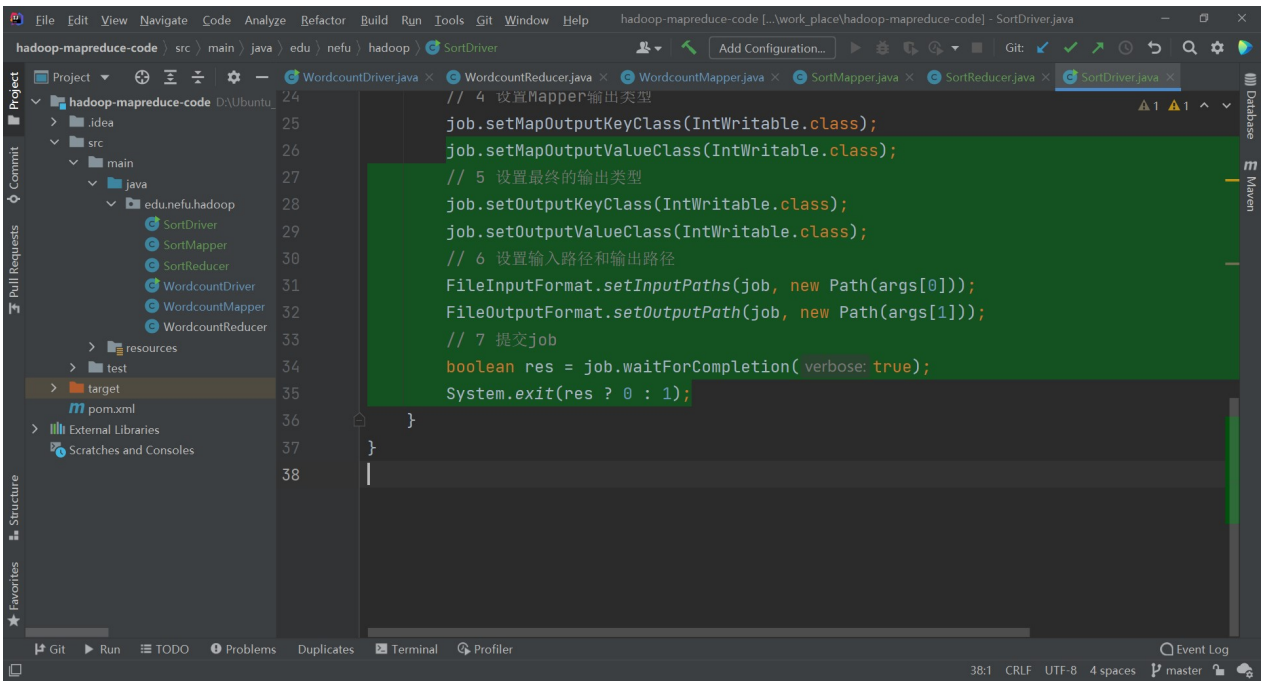
4. 编写 Reducer ；



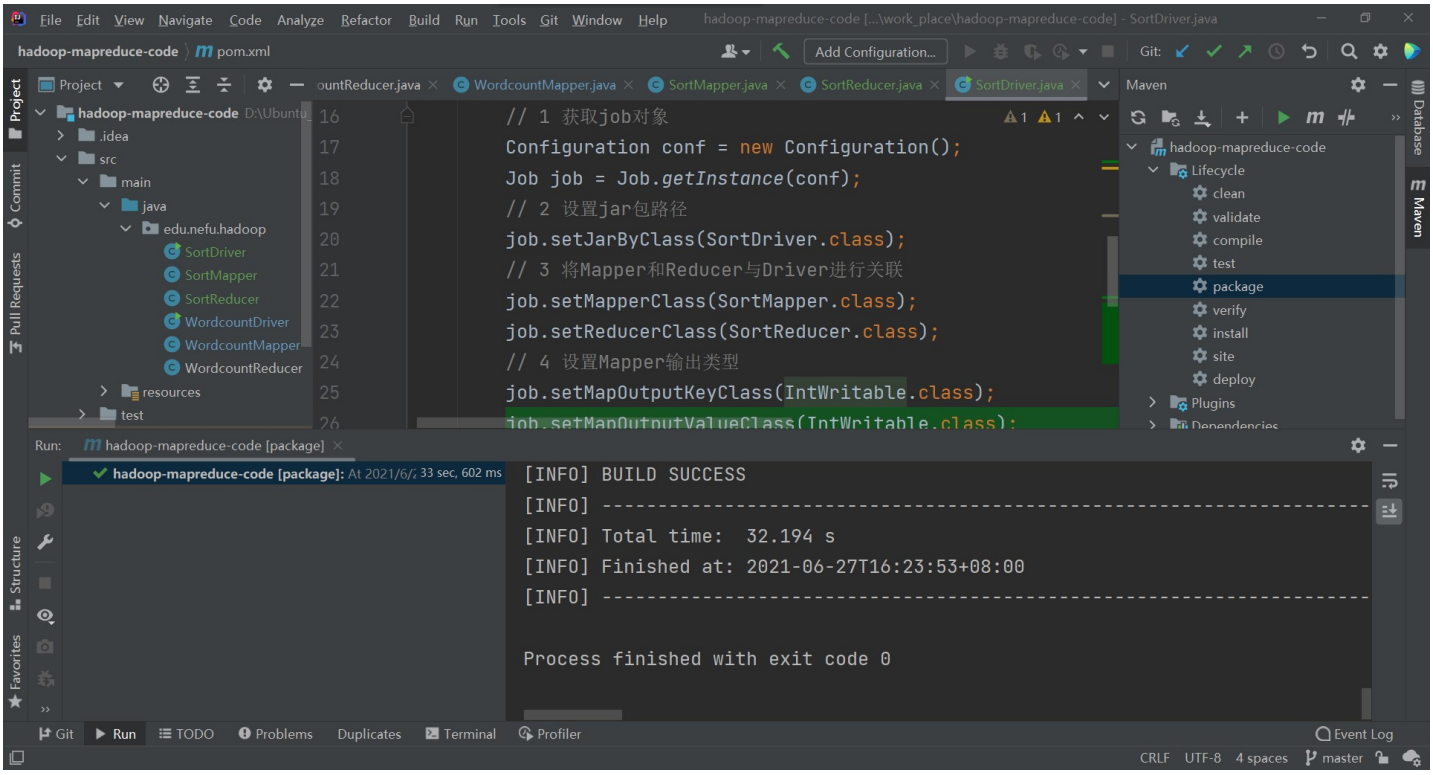
```
package edu.nefu.hadoop;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class SortReducer extends Reducer<IntWritable, IntWritable, IntWritable, IntWritab
    private static int cnt = 0;

    @Override
    protected void reduce(IntWritable key, Iterable<IntWritable> values, Context context)
        cnt += 1;
        IntWritable k = new IntWritable(cnt);
        context.write(k, key);
    }
}
```

5. 编写 Driver ；

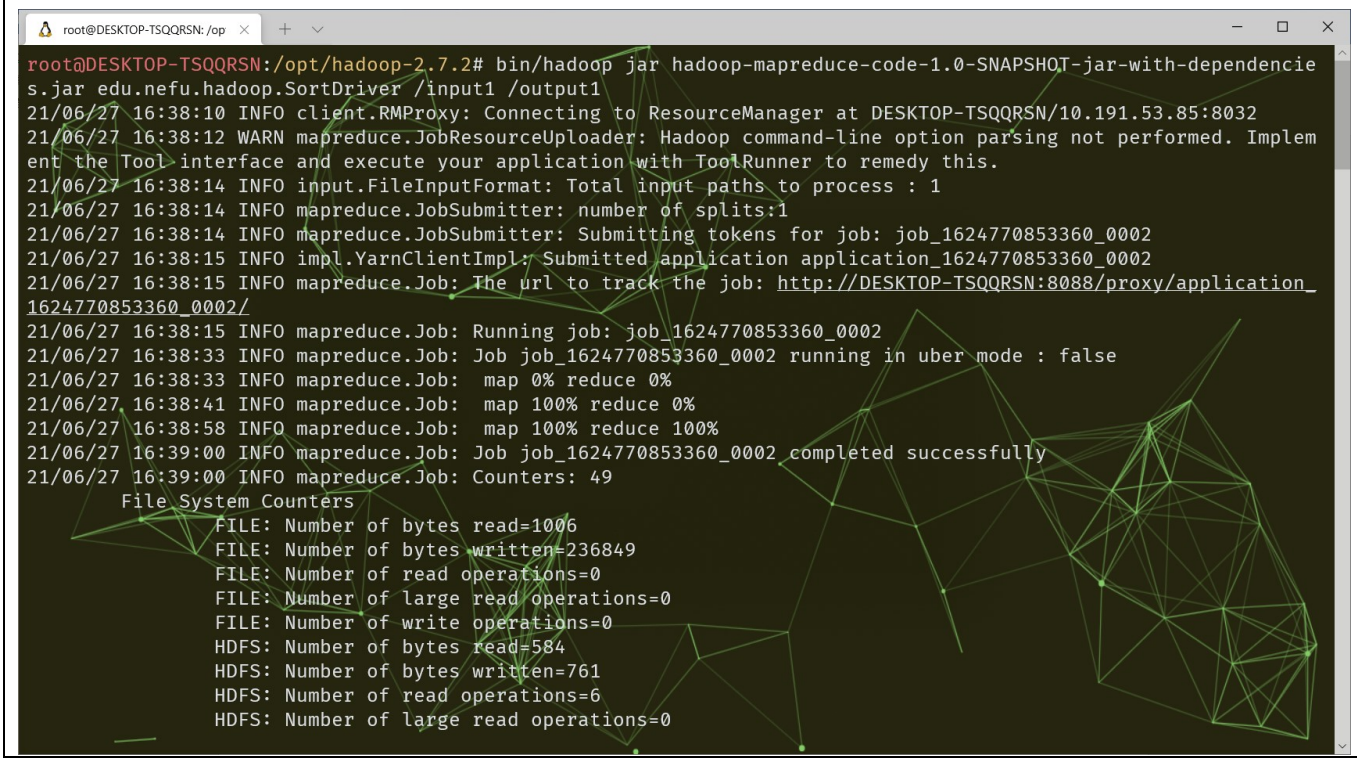


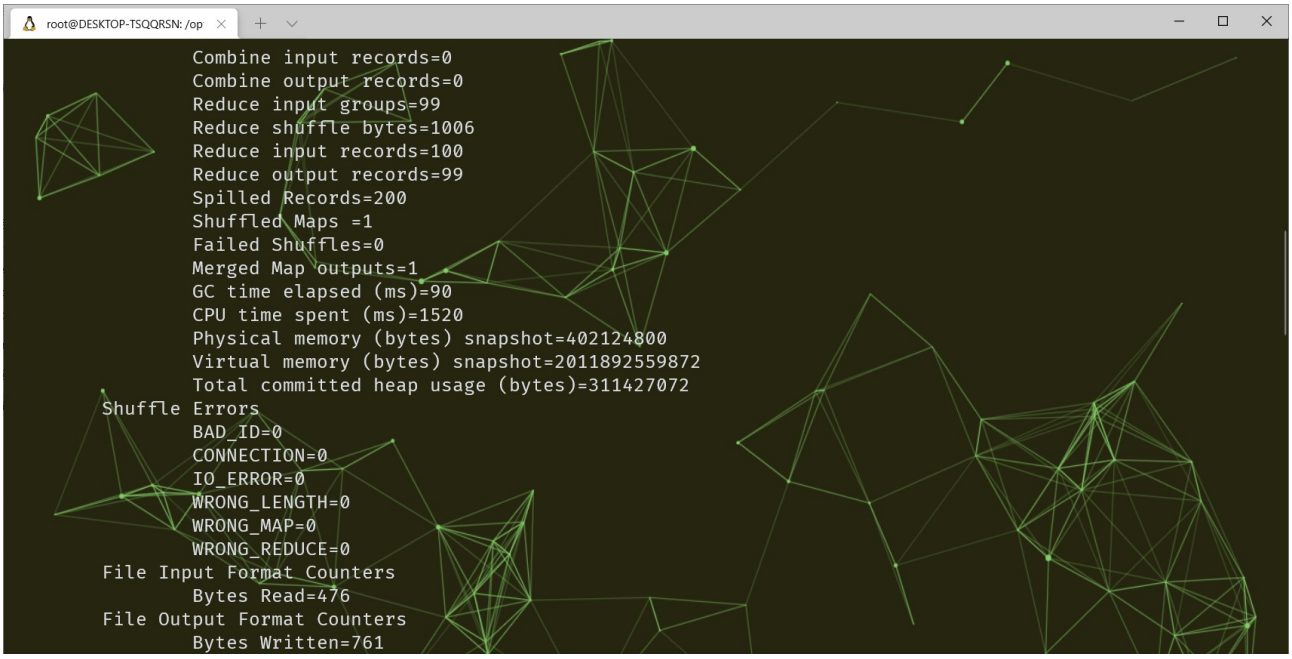


6. 使用 Maven ，创建 Jar 形式 Package ；

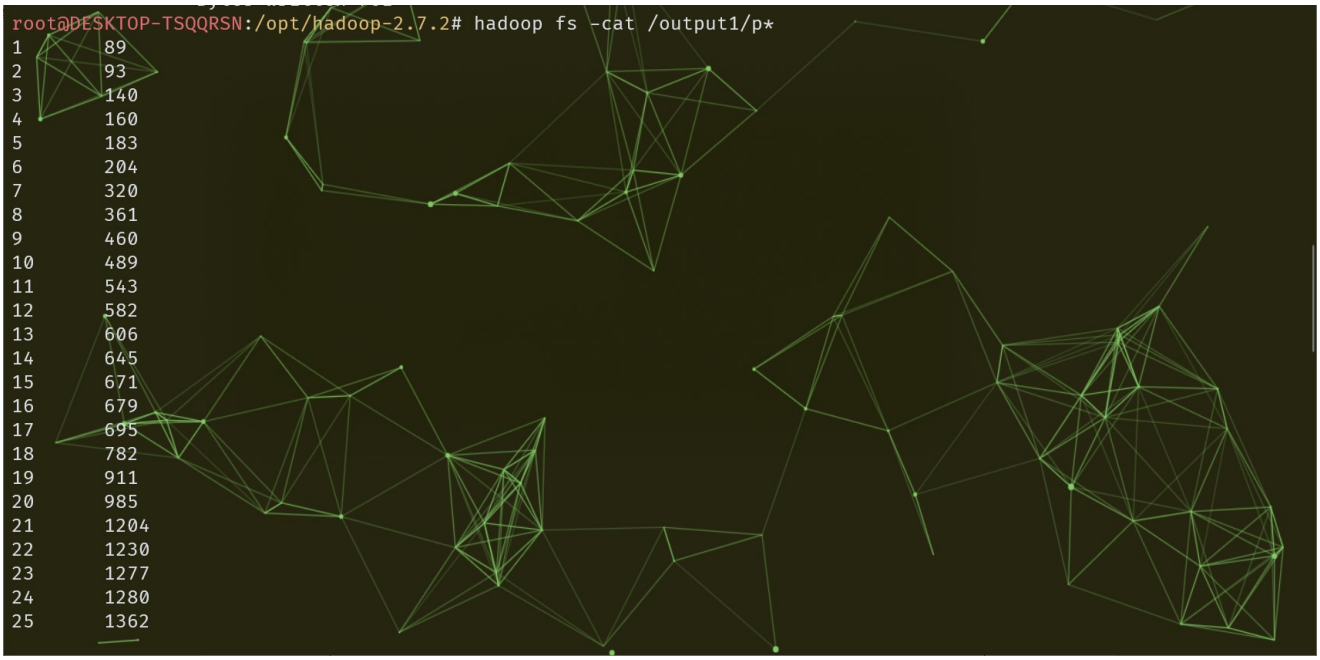

hadoop-mapreduce-code [...\work_place\hadoop-mapreduce-code] - SortDriver.java
hadoop-mapreduce-code
pom.xml
Add Configuration...
Git:
Project
WordcountMapper.java
SortMapper.java
SortReducer.java
SortDriver.java
hadoop-mapreduce-code D:\Ubuntu_
.idea
src
main
java
edu.nefu.hadoop
SortDriver
SortMapper
SortReducer
WordcountDriver
WordcountMapper
WordcountReducer
resources
test
// 1 获取job对象
Configuration conf = new Configuration();
Job job = Job.getInstance(conf);
// 2 设置jar包路径
job.setJarByClass(SortDriver.class);
// 3 将Mapper和Reducer与Driver进行关联
job.setMapperClass(SortMapper.class);
job.setReducerClass(SortReducer.class);
// 4 设置Mapper输出类型
job.setMapOutputKeyClass(IntWritable.class);
job.setMapOutputValueClass(IntWritable.class);
Maven
hadoop-mapreduce-code
Lifecycle
clean
validate
compile
test
package
verify
install
site
deploy
Plugins
Run:
hadoop-mapreduce-code [package]
hadoop-mapreduce-code [package]: At 2021/6/; 33 sec, 602 ms
[INFO] BUILD SUCCESS
[INFO] --
[INFO] Total time: 32.194 s
[INFO] Finished at: 2021-06-27T16:23:53+08:00
[INFO] --
Process finished with exit code 0
Git
Run
TODO
Problems
Duplicates
Terminal
Profiler
Event Log
CRLF
UTF-8
4 spaces
master

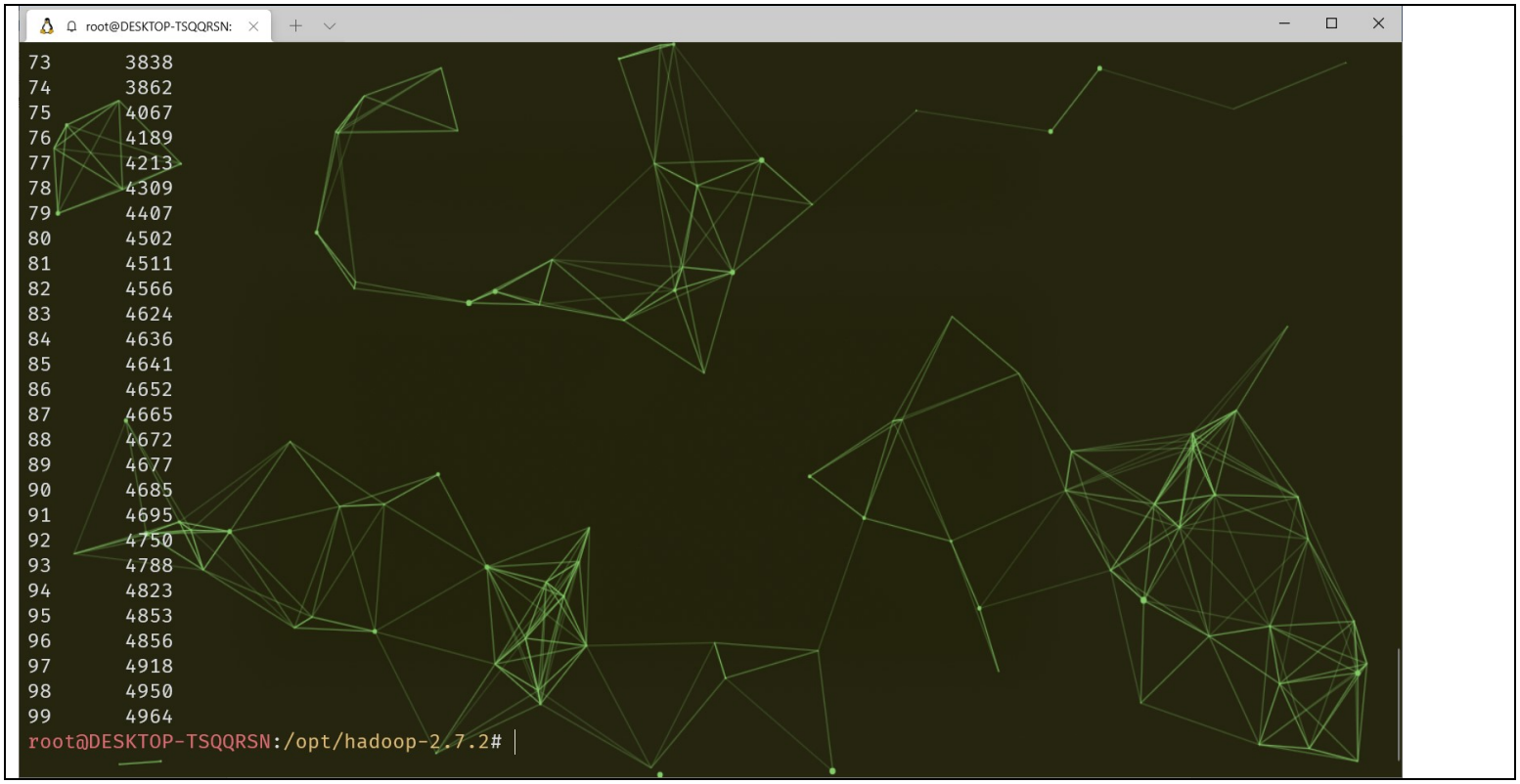
7. 运行项目；


root@DESKTOP-TSQQRSN: /op
root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2# bin/hadoop jar hadoop-mapreduce-code-1.0-SNAPSHOT-jar-with-dependencies.jar edu.nefu.hadoop.SortDriver /input1 /output1
21/06/27 16:38:10 INFO client.RMProxy: Connecting to ResourceManager at DESKTOP-TSQQRSN/10.191.53.85:8032
21/06/27 16:38:12 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
21/06/27 16:38:14 INFO input.FileInputFormat: Total input paths to process : 1
21/06/27 16:38:14 INFO mapreduce.JobSubmitter: number of splits:1
21/06/27 16:38:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624770853360_0002
21/06/27 16:38:15 INFO impl.YarnClientImpl: Submitted application application_1624770853360_0002
21/06/27 16:38:15 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TSQQRSN:8088/proxy/application_1624770853360_0002/
21/06/27 16:38:15 INFO mapreduce.Job: Running job: job_1624770853360_0002
21/06/27 16:38:33 INFO mapreduce.Job: Job job_1624770853360_0002 running in uber mode : false
21/06/27 16:38:33 INFO mapreduce.Job: map 0% reduce 0%
21/06/27 16:38:41 INFO mapreduce.Job: map 100% reduce 0%
21/06/27 16:38:58 INFO mapreduce.Job: map 100% reduce 100%
21/06/27 16:39:00 INFO mapreduce.Job: Job job_1624770853360_0002 completed successfully
21/06/27 16:39:00 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=1006
FILE: Number of bytes written=236849
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=584
HDFS: Number of bytes written=761
HDFS: Number of read operations=6
HDFS: Number of large read operations=0


root@DESKTOP-TSQQRSN: /op
Combine input records=0
Combine output records=0
Reduce input groups=99
Reduce shuffle bytes=1006
Reduce input records=100
Reduce output records=99
Spilled Records=200
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=90
CPU time spent (ms)=1520
Physical memory (bytes) snapshot=402124800
Virtual memory (bytes) snapshot=2011892559872
Total committed heap usage (bytes)=311427072
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=476
File Output Format Counters
Bytes Written=761

8. 观察并分析结果。


root@DESKTOP-TSQQRSN:/opt/hadoop-2.7.2# hadoop fs -cat /output1/p*
1 89
2 93
3 140
4 160
5 183
6 204
7 320
8 361
9 460
10 489
11 543
12 582
13 606
14 645
15 671
16 679
17 695
18 782
19 911
20 985
21 1204
22 1230
23 1277
24 1280
25 1362



四、 实验过程分析与讨论

熟悉 MapReduce 的编程规范；熟悉 Hadoop 使用的数据类型，基本都是由 Java 的基本数据类型重写的；熟悉 Map，Reduce，Driver 阶段的实现方法；知道了如何达成 jar 包，如何自己进行编写一个项目。

五、指导教师意见

指导教师签字：卢洋