第5章MapReduce进阶编程实训

(源自:https://biglab.site)

(版本:Ver1.2-20230828)

原因二：编译器内部错误

实训1统计全球每年月的最高气温和最低气温

实训目的

1. 掌握MapReduce编程中Combiner的使用
2. 掌握自定义数据类型
3. 掌握自定义计数器
4. 掌握MapReduce 参数的传递
5. 掌握Toolrunner的使用和 Eclipse 提交MapReduce任务

训练要点

1. 掌握Combiner的使用
2. 掌握自定义数据类型

需求说明

获取ncdc.noaa.gov上的全球气候数据，进行数据处理后生成data.txt文件，将文件上传至 hdfs，然后统计每年的最高温和最低温

实现思路及步骤

1. 准备测试数据
2. 编写自定义一个数据类型YearMaxTAndMinT,定义字符串类型year,double类型的maxTemp和minTemp
3. 创建MaxTAndMinTMapper,实现获取年份和气温，并将年月作为key，将气温作为value输出
4. 创建一个MaxTAndMinTCombiner,实现年份最高气温和最低气温的获取，将月份作为key,将气温作为value输出
5. 创建一个MaxTAndMinTReducer，实现获取年月最高气温和最低气温获取，并创建YearMaxTAndMinT对象存放，将该对象作为value,将NummWritable.get()作为key输出
6. 编译成jar，然后上传到集群，使用 hadoop jar执行

作业要求

1. 环境说明:本小组主机:,本小组成员机:,本成员机:
2. 在http://master:9870上拍照截取本小组集群中本成员目录下/user/myname中上传的文件,需包含temp目录和文件
3. 在eclipse中，分别截图 map类，reduce类等，main方法等的源码图
4. 在eclipse中，运行，截取运行console内容图
5. 查集群linux本成员虚拟下运行hadoop.jar程序， 截图
6. 在http://master:9870的文件系统中，打开运行输出结果:/user/myname/output_tempcount/下的文件内容，截图

实现参考

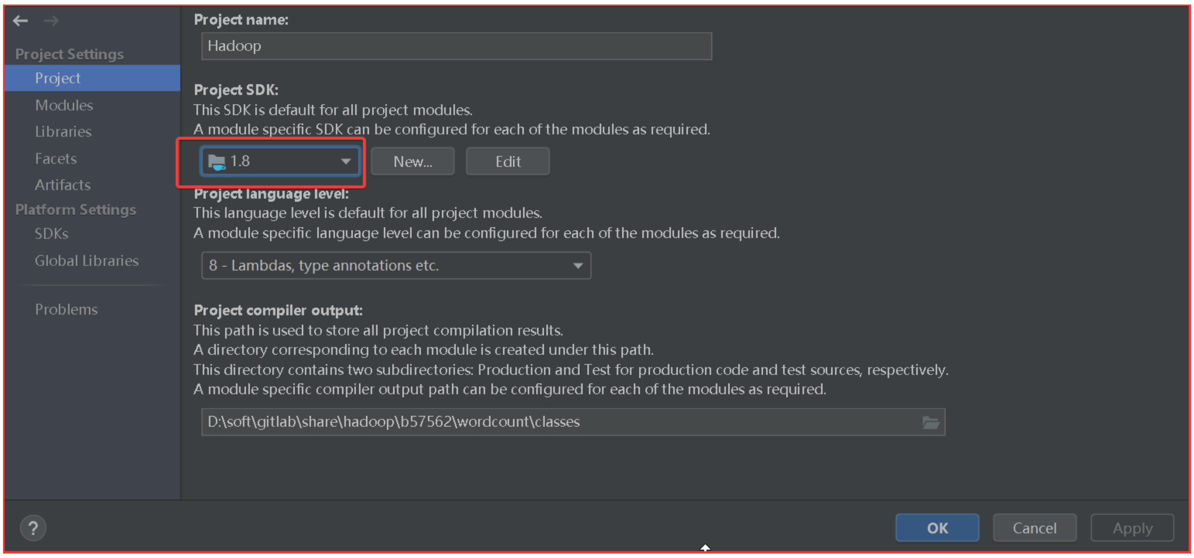
准备测试数据

```
cd /root/hadoop
wget https://biglab.site/b37066/file/temp.tar
tar -xvf ./temp.tar
hdfs dfs -mkdir -p /user/myname/temp
hdfs dfs -put ./temp2021.txt /user/myname/temp
hdfs dfs -ls /user/myname/temp/
hdfs dfs -chmod -R 777 /
```

注意事项

设置项目为jdk1.8

菜单->File->Project Structure:

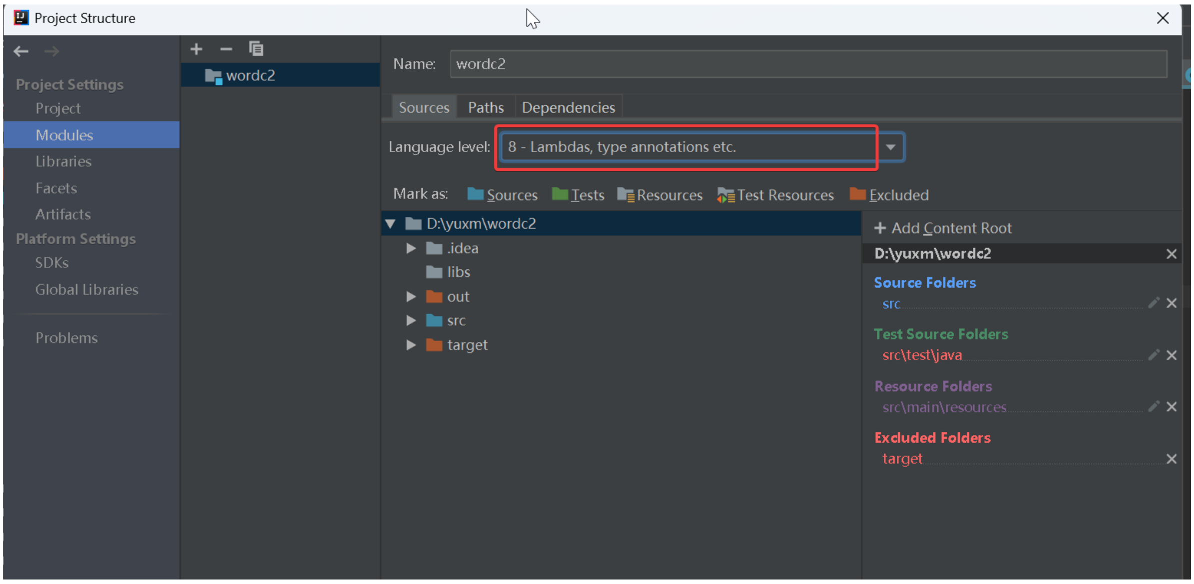


Project name:
Hadoop

Project SDK:
This SDK is default for all project modules.
A module specific SDK can be configured for each of the modules as required.
1.8

Project language level:
This language level is default for all project modules.
A module specific language level can be configured for each of the modules as required.
8 - Lambdas, type annotations etc.

Project compiler output:
This path is used to store all project compilation results.
A directory corresponding to each module is created under this path.
This directory contains two subdirectories: Production and Test for production code and test sources, respectively.
A module specific compiler output path can be configured for each of the modules as required.
D:\soft\gitlab\share\hadoop\b57562\wordcount\classes



Name: wordc2

Language level: 8 - Lambdas, type annotations etc.

D:\yuxm\wordc2
Source Folders: src
Test Source Folders: src\test\java
Resource Folders: src\main\resources
Excluded Folders: target

编写代码

源码参考：https://jihulab.com/biglab-share/hadoop/-/tree/main/b57562/wordcount/src/chap5_tempcount?ref_type=heads

自定义YearMaxTAndMinT

```java
package chap5_tempcount;

import java.io.DataInput;
import java.io.DataOutput;
import java.io.IOException;

import org.apache.hadoop.io.WritableComparable;

public class YearMaxTAndMinT implements WritableComparable<YearMaxTAndMinT>{
    private String year;
    private Double maxTemp;
    private Double mintemp;
    public String getYear() {
        return year;
    }

    public void setYear(String year) {
        this.year = year;
    }

    public Double getMaxTemp() {
        return maxTemp;
    }

    public void setMaxTemp(Double maxTemp) {
        this.maxTemp = maxTemp;
    }

    public Double getMintemp() {
        return mintemp;
    }

    public void setMintemp(Double mintemp) {
        this.mintemp = mintemp;
    }


    public YearMaxTAndMinT() {

    }

    @Override
    public void readFields(DataInput in) throws IOException {
        this.year=in.readUTF();
        this.maxTemp=in.readDouble();
        this.mintemp=in.readDouble();
    }

    @Override
    public void write(DataOutput out) throws IOException {
        out.writeUTF(year);
        out.writeDouble(maxTemp);
        out.writeDouble(mintemp);
    }
    @Override
    public int compareTo(YearMaxTAndMinT o) {
//      return this.getYear().compareTo(o.getYear());
        return this.getMaxTemp().compareTo(o.getMaxTemp());
    }
    @Override
    public String toString() {
        return this.year+"\t"+this.maxTemp.toString()+"\t"+this.mintemp.toString();
    }
}

```

MaxTAndMinTMapper

```
package chap5_tempcount;

import java.io.IOException;

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

public class MaxTAndMinTMapper extends Mapper<Object, Text, Text, DoubleWritable> {
    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

        try {
            String line = value.toString();
//872220 99999  20210221    82.0 10    65.0 10   1007.8  6   955.2 10   12.4 10   13.5 10   23.9  999.9    91.0*   65.5*  0.00I 999.9  000000
            String year = line.substring(14, 20).trim();
            double airTemperature;
            airTemperature = Double.parseDouble(line.substring(23, 30).trim());

            context.write(new Text(year), new DoubleWritable(airTemperature));

        } catch (NumberFormatException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (InterruptedException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
    }
}

```

MaxTAndMinTCombiner

```java
package chap5_tempcount;

import java.io.IOException;

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

public class MaxTAndMinTCombiner extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {
    @Override
    protected void reduce(Text key, Iterable<DoubleWritable> value,
            Context context)
                    throws IOException, InterruptedException {
        double maxtemp=0;
        double mintemp=0;
        for (DoubleWritable val : value) {
            if (val.get()>maxtemp)
            {
                maxtemp=val.get();
            }
            if (val.get()<mintemp)
            {
                mintemp=val.get();
            }
        }
        context.write(key, new DoubleWritable(maxtemp));
        context.write(key, new DoubleWritable(mintemp));
    }
}


```

MaxTAndMinTReducer

```java
package chap5_tempcount;

import java.io.IOException;

import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class MaxTAndMinTReducer extends Reducer<Text, DoubleWritable, NullWritable, YearMaxTAndMinT> {
    private YearMaxTAndMinT result = new YearMaxTAndMinT();
    @Override
    protected void reduce(Text key, Iterable<DoubleWritable> value, Context context)                     {
        double maxtemp=0;
        double mintemp=0;
        for (DoubleWritable val : value) {
            if (val.get()>maxtemp)
            {
                maxtemp=val.get();
            }
            if (val.get()<mintemp)
            {
                mintemp=val.get();
            }
        }
        result.setYear(key.toString());
        result.setMaxTemp(maxtemp);
        result.setMintemp(mintemp);


        try {
            context.write(NullWritable.get(), result);
        } catch (IOException | InterruptedException e) {
            e.printStackTrace();
        }
    }
}

```

驱动类MaxTAndMinT

```
package chap5_tempcount;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

//import com.sun.jersey.core.impl.provider.entity.XMLJAXBElementProvider.Text;
import org.apache.hadoop.io.Text;

import utils.ConfUtil;

public class MaxTAndMinT {

    public static void main(String[] args) throws Exception {
        Configuration conf = ConfUtil.GetConf(MaxTAndMinT.class);
        String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
        if (otherArgs.length < 2) {
            otherArgs = new String[] { "/user/myname/temp/temp2021.txt", "/user/myname/output_tempcount" };
        }
        Job job = Job.getInstance(conf, "maxtandmint");
        job.setJarByClass(MaxTAndMinT.class);
        job.setMapperClass(MaxTAndMinTMapper.class);

        job.setReducerClass(MaxTAndMinTReducer.class);
        job.setCombinerClass(MaxTAndMinTCombiner.class);
        job.setNumReduceTasks(1);// 设置Reducer任务数为0

        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(DoubleWritable.class);
        job.setOutputKeyClass(NullWritable.class);
        job.setOutputValueClass(YearMaxTAndMinT.class);

        FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
        FileSystem.get(conf).delete(new Path(otherArgs[1]), true);
        FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));

        System.err.println(job.waitForCompletion(true) ? -1 : 1);
    }

}

```

编译与导出jar

编译生成jar包

菜单-》Build -》 Build Artifacts -》hadoop-》Build， 参考4.3节编译生成jar包图例

上传jar包到master

1. 在Hadoop项目，左侧树图中-》out-》artifacts -》hadoop -》 hadoop.jar ，右击hadoop.jar，菜单中选择复制
2. 打开xftp，进入master主机，进入root-》hadoop目录，右击选择粘贴

运行MR程序

在master主机上

```
cd /root/hadoop
hadoop jar /root/hadoop/hadoop.jar chap5_tempcount.MaxTAndMinT \
-D mapreduce.ifile.readahead=false \
/user/myname/temp/temp2021.txt \
/user/myname/output_tempcount
```

运行结果如：

```
[root@master hadoop]# cd /root/hadoop
[root@master hadoop]# hadoop jar /root/hadoop/hadoop.jar chap5_tempcount.MaxTAndMinT \
> -D mapreduce.ifile.readahead=false \
> /user/myname/temp/temp2021.txt \
> /user/myname/output_tempcount
SLF4J: Class path contains multiple SLF4J bindings.
```

```
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
class name:chap5_tempcount.MaxTAndMinT
2023-11-02 22:27:57,075 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2.properties
2023-11-02 22:27:57,240 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-11-02 22:27:57,240 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-11-02 22:27:57,830 INFO input.FileInputFormat: Total input files to process : 1
2023-11-02 22:27:57,934 INFO mapreduce.JobSubmitter: number of splits:1
2023-11-02 22:27:58,165 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1509502677_0001
2023-11-02 22:27:58,168 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-02 22:27:58,424 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-11-02 22:27:58,425 INFO mapreduce.Job: Running job: job_local1509502677_0001
2023-11-02 22:27:58,434 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-11-02 22:27:58,446 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 22:27:58,447 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-02 22:27:58,448 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-11-02 22:27:58,520 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-11-02 22:27:58,521 INFO mapred.LocalJobRunner: Starting task: attempt_local1509502677_0001_m_000000_0
2023-11-02 22:27:58,567 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 22:27:58,567 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-02 22:27:58,610 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-11-02 22:27:58,615 INFO mapred.MapTask: Processing split: hdfs://master:8020/user/myname/temp/temp2021.txt:0+109023121
2023-11-02 22:27:58,834 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-11-02 22:27:58,834 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-11-02 22:27:58,834 INFO mapred.MapTask: soft limit at 83886080
2023-11-02 22:27:58,834 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
```

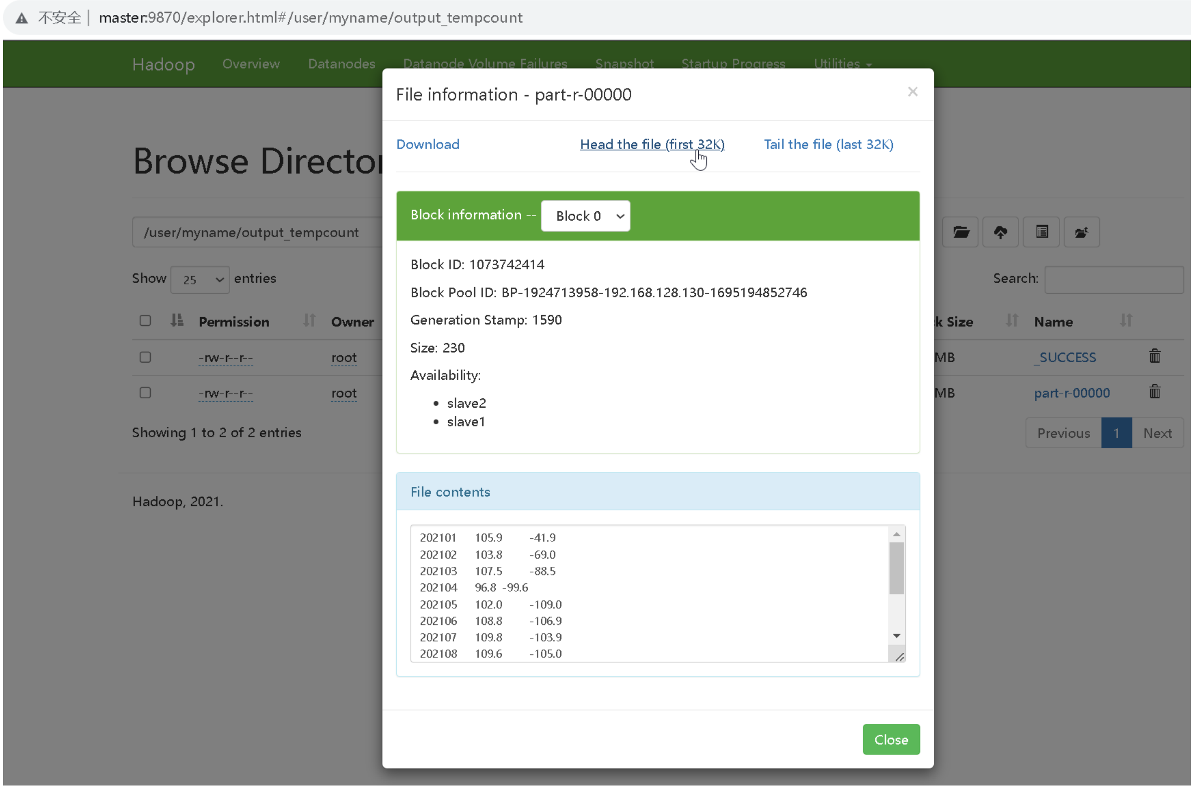
```
2023-11-02 22:27:58,834 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-11-02 22:27:58,859 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2023-11-02 22:27:59,433 INFO mapreduce.Job: Job job_local1509502677_0001 running in uber mode : false
2023-11-02 22:27:59,435 INFO mapreduce.Job:  map 0% reduce 0%
2023-11-02 22:28:01,484 INFO mapred.LocalJobRunner:
2023-11-02 22:28:01,488 INFO mapred.MapTask: Starting flush of map output
2023-11-02 22:28:01,488 INFO mapred.MapTask: Spilling map output
2023-11-02 22:28:01,488 INFO mapred.MapTask: bufstart = 0; bufend = 11765085; bufvoid = 104857600
2023-11-02 22:28:01,488 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23077044(92308176); length = 3137353/6553600
2023-11-02 22:28:02,282 INFO mapred.MapTask: Finished spill 0
2023-11-02 22:28:02,297 INFO mapred.Task: Task:attempt_local1509502677_0001_m_000000_0 is done. And is in the process of committing
2023-11-02 22:28:02,308 INFO mapred.LocalJobRunner: map
2023-11-02 22:28:02,308 INFO mapred.Task: Task 'attempt_local1509502677_0001_m_000000_0' done.
2023-11-02 22:28:02,321 INFO mapred.Task: Final Counters for attempt_local1509502677_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=81613
                FILE: Number of bytes written=597053
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=109023121
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Map-Reduce Framework
                Map input records=784339
                Map output records=784339
                Map output bytes=11765085
                Map output materialized bytes=414
                Input split bytes=113
                Combine input records=784339
                Combine output records=24
                Spilled Records=24
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=299
                Total committed heap usage (bytes)=126791680
        File Input Format Counters
                Bytes Read=109023121
2023-11-02 22:28:02,321 INFO mapred.LocalJobRunner: Finishing task: attempt_local1509502677_0001_m_000000_0
2023-11-02 22:28:02,322 INFO mapred.LocalJobRunner: map task executor complete.
2023-11-02 22:28:02,327 INFO mapred.LocalJobRunner: Waiting for reduce tasks
```

```
2023-11-02 22:28:02,328 INFO mapred.LocalJobRunner: Starting task: attempt_local1509502677_0001_r_000000_0
2023-11-02 22:28:02,346 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 22:28:02,346 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-02 22:28:02,347 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-11-02 22:28:02,358 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@772261fa
2023-11-02 22:28:02,367 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-11-02 22:28:02,405 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=173133008, maxSingleShuffleLimit=43283252, mergeThreshold=114267792, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2023-11-02 22:28:02,414 INFO reduce.EventFetcher: attempt_local1509502677_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2023-11-02 22:28:02,460 INFO mapreduce.Job:  map 100% reduce 0%
2023-11-02 22:28:02,470 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1509502677_0001_m_000000_0 decomp: 410 len: 414 to MEMORY
2023-11-02 22:28:02,474 INFO reduce.InMemoryMapOutput: Read 410 bytes from map-output for attempt_local1509502677_0001_m_000000_0
2023-11-02 22:28:02,476 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 410, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->410
2023-11-02 22:28:02,482 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2023-11-02 22:28:02,484 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:28:02,485 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2023-11-02 22:28:02,497 INFO mapred.Merger: Merging 1 sorted segments
2023-11-02 22:28:02,497 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 401 bytes
2023-11-02 22:28:02,502 INFO reduce.MergeManagerImpl: Merged 1 segments, 410 bytes to disk to satisfy reduce memory limit
2023-11-02 22:28:02,503 INFO reduce.MergeManagerImpl: Merging 1 files, 414 bytes from disk
2023-11-02 22:28:02,504 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2023-11-02 22:28:02,504 INFO mapred.Merger: Merging 1 sorted segments
2023-11-02 22:28:02,504 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 401 bytes
2023-11-02 22:28:02,505 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:28:02,568 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2023-11-02 22:28:02,701 INFO mapred.Task: Task:attempt_local1509502677_0001_r_000000_0 is done. And is in the process of committing
2023-11-02 22:28:02,708 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:28:02,708 INFO mapred.Task: Task attempt_local1509502677_0001_r_000000_0 is allowed to commit now
```

```
2023-11-02 22:28:02,755 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1509502677_0001_r_000000_0' to hdfs://master:8020/user/myname/output_tempcount
2023-11-02 22:28:02,757 INFO mapred.LocalJobRunner: reduce > reduce
2023-11-02 22:28:02,757 INFO mapred.Task: Task 'attempt_local1509502677_0001_r_000000_0' done.
2023-11-02 22:28:02,758 INFO mapred.Task: Final Counters for attempt_local1509502677_0001_r_000000_0: Counters: 29
        File System Counters
                FILE: Number of bytes read=82473
                FILE: Number of bytes written=597467
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=109023121
                HDFS: Number of bytes written=230
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=414
                Reduce input records=24
                Reduce output records=12
                Spilled Records=24
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=12
                Total committed heap usage (bytes)=126791680
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=230
2023-11-02 22:28:02,758 INFO mapred.LocalJobRunner: Finishing task: attempt_local1509502677_0001_r_000000_0
2023-11-02 22:28:02,759 INFO mapred.LocalJobRunner: reduce task executor complete.
2023-11-02 22:28:03,461 INFO mapreduce.Job:  map 100% reduce 100%
2023-11-02 22:28:03,462 INFO mapreduce.Job: Job job_local1509502677_0001 completed successfully
2023-11-02 22:28:03,481 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=164086
                FILE: Number of bytes written=1194520
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=218046242
                HDFS: Number of bytes written=230
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
        Map-Reduce Framework
                Map input records=784339
                Map output records=784339
                Map output bytes=11765085
                Map output materialized bytes=414
                Input split bytes=113
                Combine input records=784339
                Combine output records=24
                Reduce input groups=12
                Reduce shuffle bytes=414
                Reduce input records=24
                Reduce output records=12
                Spilled Records=48
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=311
                Total committed heap usage (bytes)=253583360
        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=109023121
        File Output Format Counters
                Bytes Written=230
-1
[root@master hadoop]#
```

HDFS上运行结果

结果目录：/user/myname/output_tempcount



File information - part-r-00000

Block ID: 1073742414

Block Pool ID: BP-1924713958-192.168.128.130-1695194852746

Generation Stamp: 1590

Size: 230

Availability:

- slave2
- slave1

```
202101	105.9	-41.9
202102	103.8	-69.0
202103	107.5	-88.5
202104	96.8	-99.6
202105	102.0	-109.0
202106	108.8	-106.9
202107	109.8	-103.9
202108	109.6	-105.0
```

实训2筛选气温在15~25C之间的数据

实训目的

1. 掌握MapReduce编程中Combiner的使用
2. 掌握自定义数据类型
3. 掌握自定义计数器
4. 掌握MapReduce 参数的传递
5. 掌握Toolrunner的使用和 Eclipse 提交MapReduce任务

训练要点

1. 掌握Combiner的使用
2. 掌握自定义数据类型

需求说明

获取ncdc.noaa.gov上的全球气候数据，进行数据处理后生成data.txt文件，将文件上传至 hdfs，然后统计每年的最高温和最低温

实现思路及步骤

1. 准备测试数据
2. 创建TempSelectMapper,实现温度数据筛选， 将记录作为value输出，NullWritable作为key输出
3. 创建TempSelectRun继承自 Tool,实现参数的设置和ToolRunner的run调用
4. 编译成jar，然后上传到集群，使用 hadoop jar执行

作业要求

1. 环境说明:本小组主机:,本小组成员机:,本成员机:
2. 在http://master:9870上拍照截取本小组集群中本成员目录下/user/myname中上传的文件,需包含temp目录和文件
3. 在eclipse中，分别截图 map类，main方法的源码图
4. 在eclipse中，运行，截取运行console内容图
5. 查集群linux本成员虚拟下运行程序tempselect.jar， 截图
6. 在http://master:9870的文件系统中，打开运行输出结果:/user/myname/output_tempselectrun/下的文件内容，截图

实现参考

准备测试数据

```
cd /root/hadoop
wget http://bigdata.hddly.cn/b37066/file/temp.tar
tar -xvf ./temp.tar
hdfs dfs -mkdir -p /user/myname/temp
hdfs dfs -put ./temp2021.txt /user/myname/temp
hdfs dfs -ls /user/myname/temp/
```

编写代码

源码参考:https://jihulab.com/biglab-share/hadoop/-/tree/main/b57562/wordcount/src/chap5_tempselect?ref_type=heads

TempSelectMapper

```java
package chap5_tempselect;

import java.io.IOException;

import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import enums.EnumSumCounter;

public class TempSelectMapper extends Mapper<Object, Text, NullWritable, Text> {

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

        try {
            String line = value.toString();
//872220 99999  20210221    82.0 10    65.0 10   1007.8  6    955.2 10    12.4 10    13.5 10    23.9   999.9    91.0*    65.5*   0.00I 999.9  000000
            String year = line.substring(14, 20).trim();
            Float airTemperature;
            airTemperature = Float.parseFloat(line.substring(23,
30).trim());
            Float
maxtemp=context.getConfiguration().getFloat("maxtemp",25.0f);
            Float
mintemp=context.getConfiguration().getFloat("mintemp",15.0f);
            if (mintemp<= airTemperature && airTemperature<=maxtemp)
            {
//              context.write(new Text(year), new
DoubleWritable(airTemperature));

context.getCounter(EnumSumCounter.TempNormalCount).increment(1);
                context.write(NullWritable.get(), value);
            }
            else
            {

context.getCounter(EnumSumCounter.TempOverCount).increment(1);
            }


        } catch (NumberFormatException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (InterruptedException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
    }
}

```

TempSelectRun

```
package chap5_tempselect;


import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.NullWritable;
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.input.SequenceFileAsTextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

```

```java
import utils.ConfUtil;
import utils.FinalUtil;
public class TempSelectRun extends Configured implements Tool{
    public static void main(String[] args){
        String[] myArgs={
                "/user/myname/temp", "/user/myname/output_tempselectrun"
        };

        try {
            ToolRunner.run( ConfUtil.GetConf(TempSelect.class), new TempSelectRun(), myArgs);
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
    @Override
    public int run(String[] args) throws Exception {
        Configuration conf = ConfUtil.GetConf(TempSelect.class);
        conf.setFloat("maxtemp",FinalUtil.MaxTemp);
        conf.setFloat("mintemp",FinalUtil.MinTemp);
        Job job = Job.getInstance(conf, "tempselectrun");
        job.setJarByClass(TempSelectRun.class);
        job.setMapperClass(TempSelectMapper.class);

        job.setNumReduceTasks(0);// 锟斤拷锟斤拷Reducer锟斤拷锟斤拷锟斤拷为0

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

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileSystem.get(conf).delete(new Path(args[1]), true); //锟斤拷删锟斤拷目锟斤拷路锟斤拷
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        return job.waitForCompletion(true)?-1:1;
    }
}

```

编译与导出jar

编译生成jar包

菜单-》Build -》 Build Artifacts -》hadoop-》Build， 参考4.3节编译生成jar包图例

上传jar包到master

1. 在Hadoop项目，左侧树图中-》out-》artifacts -》hadoop -》 hadoop.jar， 右击hadoop.jar，菜单中选择复制
2. 打开xftp，进入master主机，进入root-》hadoop目录，右击选择粘贴

运行MR程序

在master主机上运行

```
cd /root/hadoop
hadoop jar /root/hadoop/hadoop.jar chap5_tempselect.TempSelectRun
```

运行结果

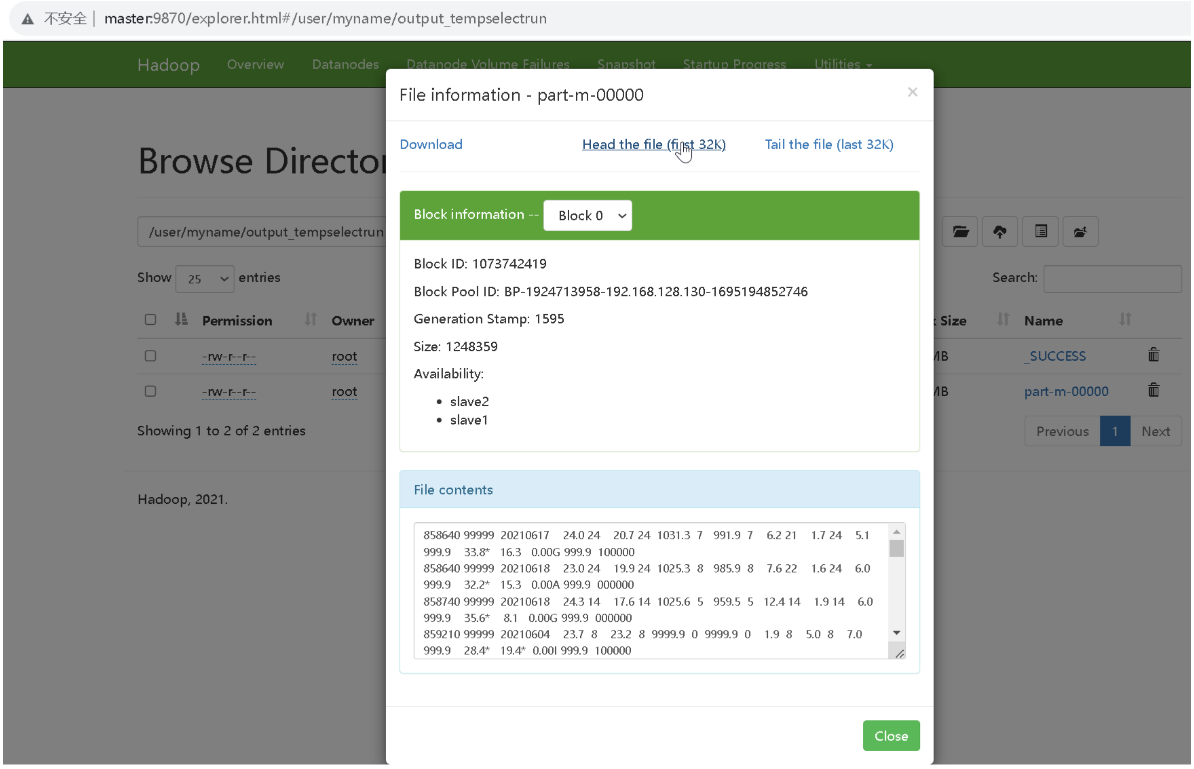
```
[root@master hadoop]# hadoop jar /root/hadoop/hadoop.jar chap5_tempselect.TempSelectRun
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
class name:chap5_tempselect.TempSelect
class name:chap5_tempselect.TempSelect
2023-11-02 23:10:57,171 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2.properties
2023-11-02 23:10:57,330 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-11-02 23:10:57,330 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-11-02 23:10:58,060 INFO input.FileInputFormat: Total input files to process : 1
2023-11-02 23:10:58,115 INFO mapreduce.JobSubmitter: number of splits:1
2023-11-02 23:10:58,371 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local826601610_0001
2023-11-02 23:10:58,374 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-02 23:10:58,619 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-11-02 23:10:58,620 INFO mapreduce.Job: Running job: job_local826601610_0001
2023-11-02 23:10:58,630 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-11-02 23:10:58,640 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 23:10:58,640 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-02 23:10:58,641 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-11-02 23:10:58,713 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-11-02 23:10:58,714 INFO mapred.LocalJobRunner: Starting task: attempt_local826601610_0001_m_000000_0
2023-11-02 23:10:58,758 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 23:10:58,758 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
```

```
2023-11-02 23:10:58,804 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-11-02 23:10:58,809 INFO mapred.MapTask: Processing split: hdfs://master:8020/user/myname/temp/temp2021.txt:0+109023121
2023-11-02 23:10:59,644 INFO mapreduce.Job: Job job_local826601610_0001 running in uber mode : false
2023-11-02 23:10:59,646 INFO mapreduce.Job:  map 0% reduce 0%
2023-11-02 23:11:01,950 INFO mapred.LocalJobRunner: 
2023-11-02 23:11:02,023 INFO mapred.Task: Task:attempt_local826601610_0001_m_000000_0 is done. And is in the process of committing
2023-11-02 23:11:02,029 INFO mapred.LocalJobRunner: 
2023-11-02 23:11:02,030 INFO mapred.Task: Task attempt_local826601610_0001_m_000000_0 is allowed to commit now
2023-11-02 23:11:02,064 INFO output.FileOutputCommitter: Saved output of task 'attempt_local826601610_0001_m_000000_0' to hdfs://master:8020/user/myname/output_tempselectrun
2023-11-02 23:11:02,066 INFO mapred.LocalJobRunner: map
2023-11-02 23:11:02,066 INFO mapred.Task: Task 'attempt_local826601610_0001_m_000000_0' done.
2023-11-02 23:11:02,080 INFO mapred.Task: Final Counters for attempt_local826601610_0001_m_000000_0: Counters: 22
        File System Counters
                FILE: Number of bytes read=151904
                FILE: Number of bytes written=662807
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=109023121
                HDFS: Number of bytes written=1248359
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=784339
                Map output records=8981
                Input split bytes=113
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=269
                Total committed heap usage (bytes)=21561344
        enums.EnumSumCounter
                TempNormalCount=8981
                TempOverCount=775358
        File Input Format Counters 
                Bytes Read=109023121
        File Output Format Counters 
                Bytes Written=1248359
2023-11-02 23:11:02,080 INFO mapred.LocalJobRunner: Finishing task: attempt_local826601610_0001_m_000000_0
2023-11-02 23:11:02,081 INFO mapred.LocalJobRunner: map task executor complete.
2023-11-02 23:11:02,664 INFO mapreduce.Job:  map 100% reduce 0%
```

```
2023-11-02 23:11:02,666 INFO mapreduce.Job: Job job_local826601610_0001 completed successfully
2023-11-02 23:11:02,683 INFO mapreduce.Job: Counters: 22
        File System Counters
                FILE: Number of bytes read=151904
                FILE: Number of bytes written=662807
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=109023121
                HDFS: Number of bytes written=1248359
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=784339
                Map output records=8981
                Input split bytes=113
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=269
                Total committed heap usage (bytes)=21561344
        enums.EnumSumCounter
                TempNormalCount=8981
                TempOverCount=775358
        File Input Format Counters
                Bytes Read=109023121
        File Output Format Counters
                Bytes Written=1248359
[root@master hadoop]#
```

HDFS上运行结果

查看hdfs上/user/myname/output_tempselectrun路径



	Permission	Owner	Size	Name
□	-rw-r--r--	root	B	_SUCCESS
□	-rw-r--r--	root	B	part-m-00000

Showing 1 to 2 of 2 entries

File information - part-m-00000

Block ID: 1073742419

Block Pool ID: BP-1924713958-192.168.128.130-1695194852746

Generation Stamp: 1595

Size: 1248359

Availability:

- slave2
- slave1

File contents

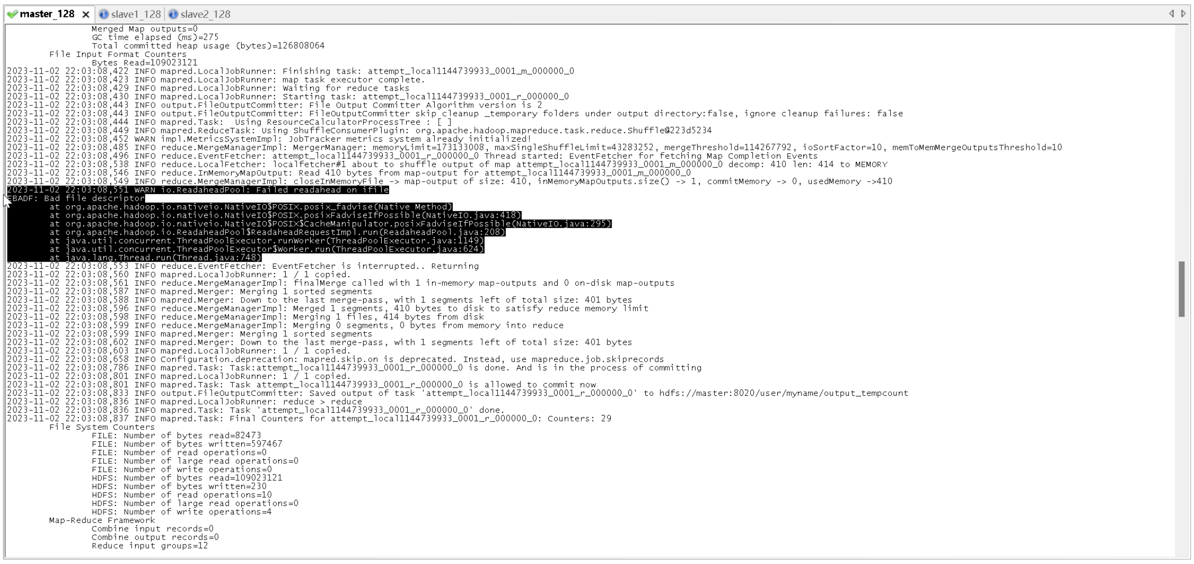
```
858640 99999  20210617    24.0 24    20.7 24  1031.3  7   991.9  7    6.2 21    1.7 24    5.1
999.9    33.8*   16.3   0.00G 999.9  100000
858640 99999  20210618    23.0 24    19.9 24  1025.3  8   985.9  8    7.6 22    1.6 24    6.0
999.9    32.2*   15.3   0.00A 999.9  000000
858740 99999  20210618    24.3 14    17.6 14  1025.6  5   959.5  5   12.4 14    1.9 14    6.0
999.9    35.6*    8.1   0.00G 999.9  000000
859210 99999  20210604    23.7  8    23.2  8  9999.9  0  9999.9  0    1.9  8    5.0  8    7.0
999.9    28.4*   19.4*  0.00I 999.9  100000
```

常见问题

问题一：EBADF: Bad file descriptor

WARN io.ReadaheadPool: Failed readahead on ifile
EBADF: Bad file descriptor

如图：



```
                Merged Map outputs=0
                GC time elapsed (ms)=275
                Total committed heap usage (bytes)=126808064
        File Input Format Counters
                Bytes Read=109023121
2023-11-02 22:03:08,422 INFO mapred.LocalJobRunner: Finishing task: attempt_local1144739933_0001_m_000000_0
2023-11-02 22:03:08,423 INFO mapred.LocalJobRunner: map task executor complete.
2023-11-02 22:03:08,429 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2023-11-02 22:03:08,430 INFO mapred.LocalJobRunner: Starting task: attempt_local1144739933_0001_r_000000_0
2023-11-02 22:03:08,443 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-02 22:03:08,443 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-02 22:03:08,444 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2023-11-02 22:03:08,449 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@223d5234
2023-11-02 22:03:08,452 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-11-02 22:03:08,485 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=173133008, maxSingleShuffleLimit=43283252, mergeThreshold=114267792, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2023-11-02 22:03:08,496 INFO reduce.EventFetcher: attempt_local1144739933_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2023-11-02 22:03:08,538 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1144739933_0001_m_000000_0 decomp: 410 len: 414 to MEMORY
2023-11-02 22:03:08,546 INFO reduce.InMemoryMapOutput: Read 410 bytes from map-output for attempt_local1144739933_0001_m_000000_0
2023-11-02 22:03:08,549 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 410, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->410
2023-11-02 22:03:08,551 WARN io.ReadaheadPool: Failed readahead on ifile
EBADF: Bad file descriptor
        at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posix_fadvise(Native Method)
        at org.apache.hadoop.io.nativeio.NativeIO$POSIX.posixFadviseIfPossible(NativeIO.java:418)
        at org.apache.hadoop.io.nativeio.NativeIO$POSIX$CacheManipulator.posixFadviseIfPossible(NativeIO.java:295)
        at org.apache.hadoop.io.ReadaheadPool$ReadaheadRequestImpl.run(ReadaheadPool.java:208)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
2023-11-02 22:03:08,553 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2023-11-02 22:03:08,560 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:03:08,561 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2023-11-02 22:03:08,587 INFO mapred.Merger: Merging 1 sorted segments
2023-11-02 22:03:08,588 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 401 bytes
2023-11-02 22:03:08,596 INFO reduce.MergeManagerImpl: Merged 1 segments, 410 bytes to disk to satisfy reduce memory limit
2023-11-02 22:03:08,598 INFO reduce.MergeManagerImpl: Merging 1 files, 414 bytes from disk
2023-11-02 22:03:08,599 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2023-11-02 22:03:08,599 INFO mapred.Merger: Merging 1 sorted segments
2023-11-02 22:03:08,602 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 401 bytes
2023-11-02 22:03:08,603 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:03:08,658 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2023-11-02 22:03:08,786 INFO mapred.Task: Task:attempt_local1144739933_0001_r_000000_0 is done. And is in the process of committing
2023-11-02 22:03:08,801 INFO mapred.LocalJobRunner: 1 / 1 copied.
2023-11-02 22:03:08,801 INFO mapred.Task: Task attempt_local1144739933_0001_r_000000_0 is allowed to commit now
2023-11-02 22:03:08,833 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1144739933_0001_r_000000_0' to hdfs://master:8020/user/myname/output_tempcount
2023-11-02 22:03:08,836 INFO mapred.LocalJobRunner: reduce > reduce
2023-11-02 22:03:08,836 INFO mapred.Task: Task 'attempt_local1144739933_0001_r_000000_0' done.
2023-11-02 22:03:08,837 INFO mapred.Task: Final Counters for attempt_local1144739933_0001_r_000000_0: Counters: 29
        File System Counters
                FILE: Number of bytes read=82473
                FILE: Number of bytes written=597467
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=109023121
                HDFS: Number of bytes written=230
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
```

查阅信息后，说由于在快速读取文件的时候，文件被关闭引起，也可能是其他bug导致，此处忽略。
也可以 mapreduce.ifile.readahead = false 临时禁掉

```
cd /root/hadoop
hadoop jar /root/hadoop/hadoop.jar chap5_tempcount.MaxTAndMinT \
-D mapreduce.ifile.readahead=false \
/user/myname/temp/temp2021.txt \
/user/myname/output_tempcount
```

问题二：IDEA Compilation failed internal java compiler error

IDEA在编译项目时报错

如图：

问题原因分两种：

原因一：多处的JDK的版本不匹配

导致这个错误的原因主要是因为jdk版本问题，此处有两个因素，一个是编译版本不匹配，一个是当前项目jdk版本不支持。

查看项目的jdk

File ->Project Structure->Project Settings ->Project或使用快捷键Ctrl+Alt+shift+S打开项目的jdk配置
要求：1，project jdk版本要求：1.8 ; 2，project language level要求:8

查看工程的jdk

点击上例中Modules（File ->Project Structure->Project Settings ->Modules）查看对应jdk版本，其中Language level要求:8

查看java编译器版本：

File ->Settings->Build,Execution,Deployment->Compiler->Java Compiler

要求：1，Project bytecode version是：8； 2，module的target bytecode version也是：8

原因二：编译器内部错误

真的就是编译器内部错误，此时就得去查看 错误日志，
我之前遇到的错误：java.lang.OutOfMemoryError：GC overhead limit exceeded：
加载太多资源到内存，导致GC耗时较多
GC overhead limt exceed检查是Hotspot VM 1.6定义的一个策略，通过统计GC时间来预测是否要OOM了，提前抛出异常，防止OOM发生。Sun 官方对此的定义是：“并行/并发回收器在GC回收时间过长时会抛出OutOfMemroyError。过长的定义是，超过98%的时间用来做GC并且回收了不到2%的堆内存。用来避免内存过小造成应用不能正常工作。“

听起来没啥用...预测OOM有啥用？起初开来这玩意只能用来Catch住释放内存资源，避免应用挂掉。后来发现一般情况下这个策略不能拯救你的应用，但是可以在应用挂掉之前做最后的挣扎，比如数据保存或者保存现场（Heap Dump）。
解决办法
1.增加参数-XX:-UseGCOverheadLimit，关闭这个特性，同时增加heap大小-Xmx1024m -Xms512m，系统环境变量 新增两行

```
_JAVA_OPTIONS
-Xms512m Xmx512m
```

每个人的电脑配置不一样，上面那个适合电脑配置低的，配置完点击ok关闭环境变量窗口，重启idea！编译完成后，将新加的环境变量参数删除、再重启，点击debug 或者run ok

将jdk升级到jdk1.8及以上版本，就能完全解决 java.lang.OutOfMemoryError：GC overhead limit exceeded 的问题