

# 金融大数据处理技术 实验二（任务三： 用户活跃度分析）

221275010 屈航

## 1、设计思路

### 1.1、Mapper的设计思路

Mapper 重构了一个 map 函数。

map 函数主要实现了对 user\_balance\_table.csv 表中的每一个用户的活跃天数进行统计，有直接购买（direct\_purchase\_amt 字段大于 0）或赎回行为（total\_redeem\_amt 字段大于 0）时，则该用户当天活跃。实现的过程就是对满足条件 `directPurchaseAmt.compareTo(BigInteger.ZERO) > 0` || `totalRedeemAmt.compareTo(BigInteger.ZERO) > 0` 的条目进行键值对的写入 `context.write(userId, one)` 表示找到该用户有一天活跃；否则就写入 `context.write(userId, new LongWritable(0))` 表示该用户在这一天不活跃。

因为考虑到求和得到的数据量可能很大，导致会超出 int 类型的表示范围，故实现的过程中的资金流入流出量的数据类型都是 BigInteger。

因为要传输两个值，即每日的 total\_purchase\_amt 资金流入与 total\_redeem\_amt 资金流出，故可以把这两个数据作为一个 BalanceWritable 类来作为 value 进行键值对的传输。

以下是 BalanceWritable 类的部分构造语句：

```
public static class BalanceWritable implements Writable {
    private BigInteger inflow;
    private BigInteger outflow;

    public BalanceWritable() {
        this.inflow = BigInteger.ZERO;
        this.outflow = BigInteger.ZERO;
    }

    public void set(BigInteger inflow, BigInteger outflow) {
        this.inflow = inflow;
        this.outflow = outflow;
    }

    @Override
    public void write(DataOutput out) throws IOException {
        out.writeUTF(inflow.toString());
        out.writeUTF(outflow.toString());
    }

    @Override
    public void readFields(DataInput in) throws IOException {
        inflow = new BigInteger(in.readUTF());
        outflow = new BigInteger(in.readUTF());
    }
}
```

以下是 `map` 函数的主要功能语句：

```
String userIdField = fields[0];
BigInteger directPurchaseAmt = new BigInteger(fields[5]);
BigInteger totalRedeemAmt = new BigInteger(fields[8]);

userId.set(userIdField);
if (directPurchaseAmt.compareTo(BigInteger.ZERO) > 0 ||
    totalRedeemAmt.compareTo(BigInteger.ZERO) > 0) {
    context.write(userId, one);
} else {
    context.write(userId, new LongWritable(0)); // Count as active day even if
    amounts are zero
}
```

**注意：实验开始前要对 `user_balance_table.csv` 中的第一行删去，第一行并不是需要统计的内容。**

## 1.2、Reducer的设计思路

Reducer 包含有 `reduce` 函数和 `cleanup` 函数。

1. `reduce` 函数就是实现对于同一个key的 `Iterable<LongWritable> values` 中的每一个元素按照 `userID` 作为键，对于活跃天数进行求和操作。最后把统计完成的键值对，存入 `Map<Long, List<Text>> activeDaysMap = new TreeMap<>(Comparator.reverseOrder())`，这个 `activeDaysMap` 变量存入的元素是一个键值对为（活跃天数，`userID`列表），可以实现按照 key 即活跃天数进行倒序排列，以待 `cleanup` 函数实现对输出格式的处理。

以下是 `reduce` 函数的主要功能语句：

```
private LongWritable result = new LongWritable();
private Map<Long, List<Text>> activeDaysMap = new TreeMap<>
(Comparator.reverseOrder());

@Override
protected void reduce(Text key, Iterable<LongWritable> values, Context
context) throws IOException, InterruptedException {
    long activeDays = 0;
    for (LongWritable val : values) {
        activeDays += val.get();
    }
    result.set(activeDays);

    activeDaysMap.computeIfAbsent(activeDays, k -> new ArrayList<>
()).add(new Text(key));
}
```

2. `cleanup` 函数就是实现对输出结果的格式控制。对于 `reduce` 函数中实现的键为活跃天数，值为 `userID`列表，这已经按照活跃天数完成了倒序排序了，所以只需要把键值对的 `userID`列表展开依次输出即可。

以下是 `cleanup` 函数的主要功能语句：

```
for (Map.Entry<Long, List<Text>> entry : activeDaysMap.entrySet()) {  
    for (Text userId : entry.getValue()) {  
        context.write(userId, new LongWritable(entry.getKey()));  
    }  
}
```

## 1.3、项目运行的配置设计

- 此次项目主要使用 Maven 进行项目管理，通过编辑 pom.xml 文件对该项目进行配置。pom.xml 文件的配置信息包含有该项目需要哪些库文件需要下载，该项目的项目文件有哪些。
- 依次使用 mvn clean install 进行配置，同时还可以使用 mvn compile 对 .class 文件进行生成，mvn package 实现对项目文件的 .class 文件打包成 jar 文件。
- 将 user\_balance\_table.csv 上传至 HDFS 的 /input 文件夹里面，最后运行该项目的 jar 文件，运行命令为：

```
./hadoop jar /home/njucs/shiyan2_3/target/shiyan2_3-1.0-SNAPSHOT.jar  
ActiveDaysCount /input /output_3
```

**注意要把导出出来的 part-r-000000 解锁，以实现普通用户可以打开，命令如下：**

```
sudo chown $USER part-r-000000
```

## 2、程序运行结果

以下即为 ActiveDaysCount.java 程序执行的任务三（根据 user\_balance\_table.csv 表中的数据，统计每个用户的活跃天数，并按照活跃天数降序排列。输出格式为“< 用户 ID> TAB < 活跃天数>”) 的运行结果：

- **程序运行结果图：**

```
lab1 [正在运行] - Oracle VM VirtualBox
五 23:42
root@njucs-VirtualBox: /usr/local/hadoop/bin

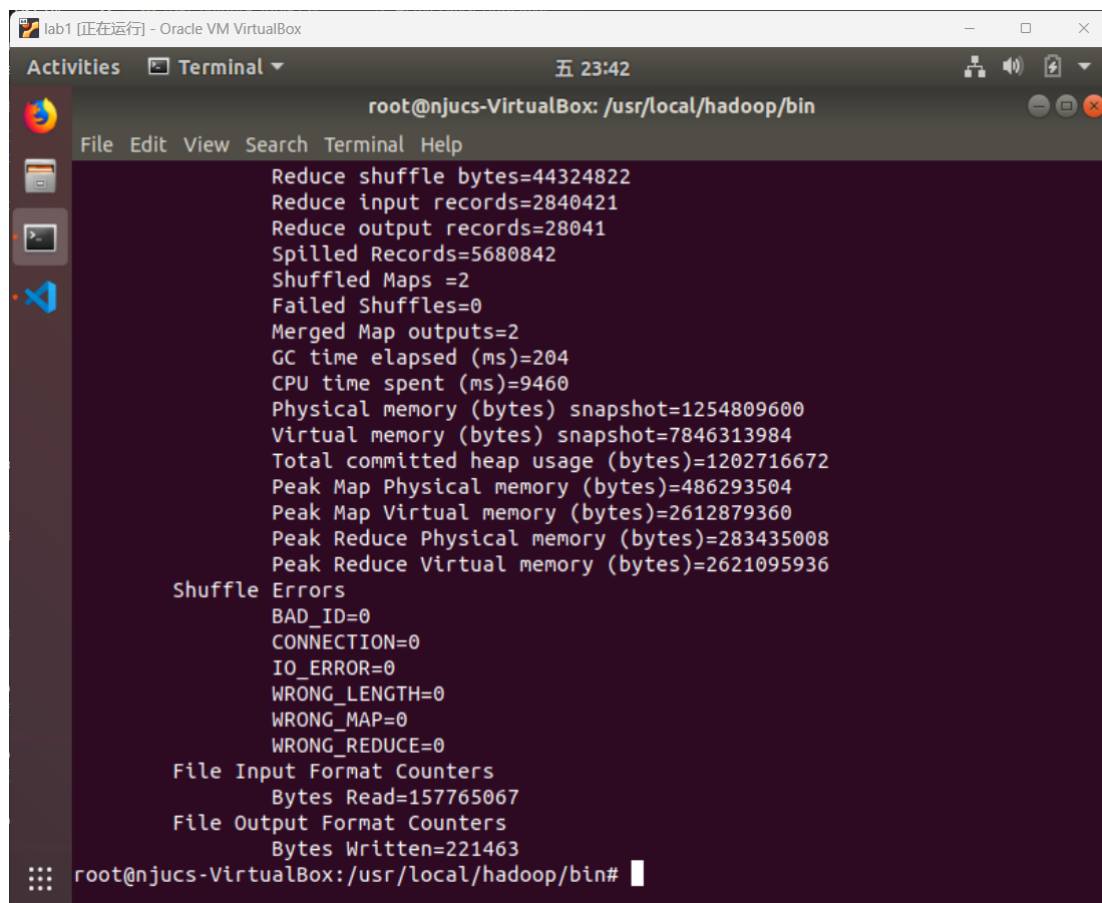
File Edit View Search Terminal Help

root@njucs-VirtualBox:/usr/local/hadoop/bin# ./hadoop jar /home/njucs/shiyan2_3/target/shiyan2_3-1.0-SNAPSHOT.jar ActiveDaysCount /input /output_3
2024-11-01 23:41:33,720 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceManager at localhost/127.0.0.1:8032
2024-11-01 23:41:33,987 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-11-01 23:41:34,022 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1730472846024_0006
2024-11-01 23:41:34,242 INFO input.FileInputFormat: Total input files to process : 1
2024-11-01 23:41:34,698 INFO mapreduce.JobSubmitter: number of splits:2
2024-11-01 23:41:34,807 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1730472846024_0006
2024-11-01 23:41:34,807 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-01 23:41:34,949 INFO conf.Configuration: resource-types.xml not found
2024-11-01 23:41:34,949 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-11-01 23:41:35,003 INFO impl.YarnClientImpl: Submitted application application_1730472846024_0006
2024-11-01 23:41:35,042 INFO mapreduce.Job: The url to track the job: http://njucs-VirtualBox:8088/proxy/application_1730472846024_0006/
2024-11-01 23:41:35,043 INFO mapreduce.Job: Running job: job_1730472846024_0006
2024-11-01 23:41:40,132 INFO mapreduce.Job: Job job_1730472846024_0006 running in uber mode : false
2024-11-01 23:41:40,133 INFO mapreduce.Job: map 0% reduce 0%
2024-11-01 23:41:46,200 INFO mapreduce.Job: map 50% reduce 0%
2024-11-01 23:41:49,219 INFO mapreduce.Job: map 100% reduce 0%
2024-11-01 23:41:52,236 INFO mapreduce.Job: map 100% reduce 100%
```

```
lab1 [正在运行] - Oracle VM VirtualBox
五 23:42
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

successfully
2024-11-01 23:41:53,315 INFO mapreduce.Job: Counters: 55
  File System Counters
    FILE: Number of bytes read=44324816
    FILE: Number of bytes written=89576484
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=157765297
    HDFS: Number of bytes written=221463
    HDFS: Number of read operations=11
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Killed map tasks=1
    Launched map tasks=2
    Launched reduce tasks=1
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=9315
    Total time spent by all reduces in occupied slots (ms)=3587
    Total time spent by all map tasks (ms)=9315
    Total time spent by all reduce tasks (ms)=3587
    Total vcore-milliseconds taken by all map tasks=9315
    Total vcore-milliseconds taken by all reduce tasks=3587
    Total megabyte-milliseconds taken by all map tasks=9538560
    Total megabyte-milliseconds taken by all reduce tasks=3673088
  Map-Reduce Framework
    Map input records=2840421
```



lab1 [正在运行] - Oracle VM VirtualBox

五 23:42

root@njucs-VirtualBox: /usr/local/hadoop/bin

```
File Edit View Search Terminal Help

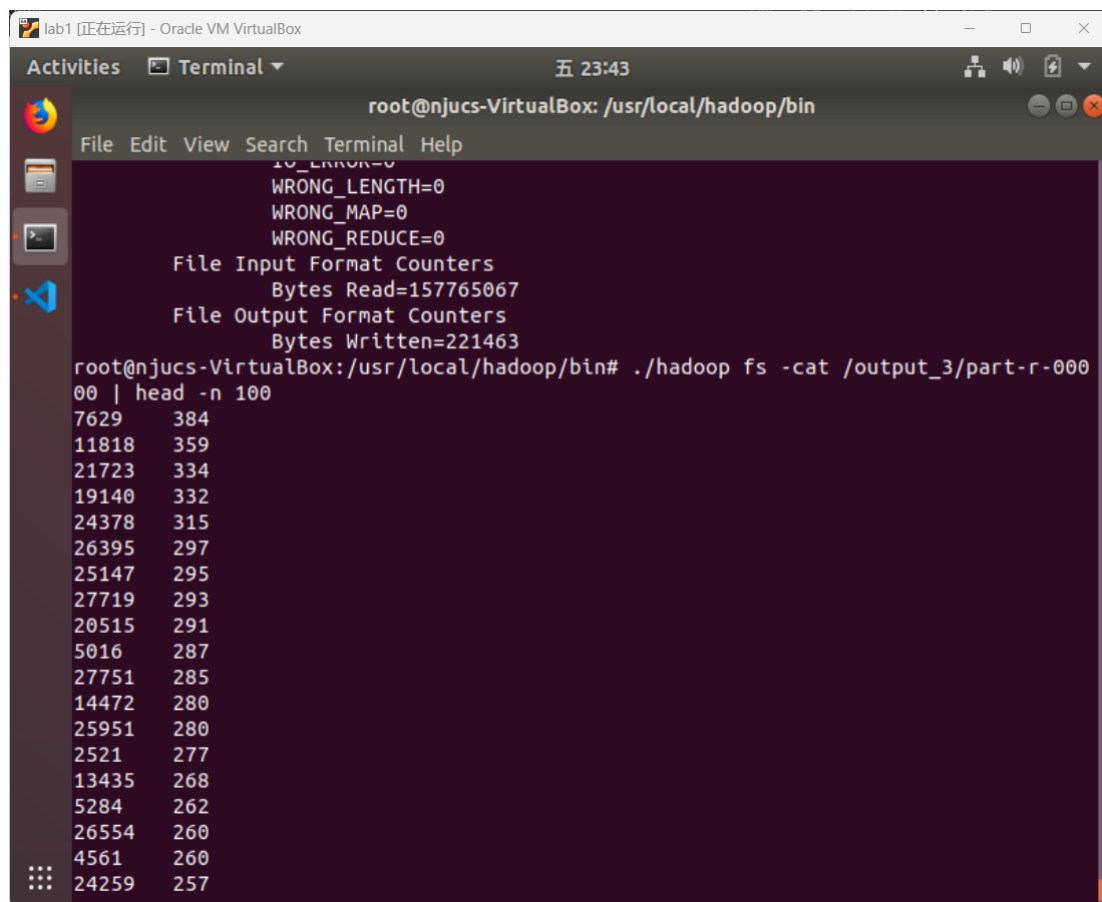
Reduce shuffle bytes=44324822
Reduce input records=2840421
Reduce output records=28041
Spilled Records=5680842
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=204
CPU time spent (ms)=9460
Physical memory (bytes) snapshot=1254809600
Virtual memory (bytes) snapshot=7846313984
Total committed heap usage (bytes)=1202716672
Peak Map Physical memory (bytes)=486293504
Peak Map Virtual memory (bytes)=2612879360
Peak Reduce Physical memory (bytes)=283435008
Peak Reduce Virtual memory (bytes)=2621095936

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=157765067
File Output Format Counters
  Bytes Written=221463

root@njucs-VirtualBox: /usr/local/hadoop/bin#
```

- part-r-00000输出结果图：



lab1 [正在运行] - Oracle VM VirtualBox

五 23:43

root@njucs-VirtualBox: /usr/local/hadoop/bin

```
File Edit View Search Terminal Help

IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=157765067
File Output Format Counters
  Bytes Written=221463

root@njucs-VirtualBox: /usr/local/hadoop/bin# ./hadoop fs -cat /output_3/part-r-000
00 | head -n 100
7629    384
11818   359
21723   334
19140   332
24378   315
26395   297
25147   295
27719   293
20515   291
5016    287
27751   285
14472   280
25951   280
2521    277
13435   268
5284    262
26554   260
4561    260
24259   257
7040    254
```

### 3、WEB页面截图

因为我的代码一开始写的有bug，所有修改了4次，也就运行了5次，第5次程序运行结果符合预期，任务三完成。

lab1 [正在运行] - Oracle VM VirtualBox

Activities

Firefox Web Browser

五 23:43

All Applications - Mozilla Firefox

All Applications

localhost:8088/cluster/apps

50%

doop

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
6	0	0	6	0	<memory:0 B, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Last No
1	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Appli
Capacity Scheduler	[memory-mb (unit-Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	F Cr
<a href="#">application_1730472846024_0005</a>	root	Active Days Count	MAPREDUCE		root.default	0	Fri Nov 1 23:41:34 +0800 2024	Fri Nov 1 23:41:35 +0800 2024	Fri Nov 1 23:41:51 +0800 2024	FINISHED	SUCCEEDED	N/A
<a href="#">application_1730472846024_0005</a>	root	Active Days Count	MAPREDUCE		root.default	0	Fri Nov 1 23:38:37 +0800 2024	Fri Nov 1 23:38:37 +0800 2024	Fri Nov 1 23:38:54 +0800 2024	FINISHED	SUCCEEDED	N/A
<a href="#">application_1730472846024_0004</a>	root	Active Days Count	MAPREDUCE		root.default	0	Fri Nov 1 23:29:38 +0800 2024	Fri Nov 1 23:29:39 +0800 2024	Fri Nov 1 23:29:55 +0800 2024	FINISHED	SUCCEEDED	N/A
<a href="#">application_1730472846024_0003</a>	root	Active Days Count	MAPREDUCE		root.default	0	Fri Nov 1 23:17:17 +0800 2024	Fri Nov 1 23:17:18 +0800 2024	Fri Nov 1 23:17:35 +0800 2024	FINISHED	SUCCEEDED	N/A
<a href="#">application_1730472846024_0002</a>	root	Active Days	MAPREDUCE		root.default	0	Fri Nov 1 23:12:04	Fri Nov 1 23:12:05 +0800	Fri Nov 1 23:12:21	FINISHED	SUCCEEDED	N/A