

金融大数据处理技术 实验二（任务一：每日资金流入流出统计）

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 函数主要实现了对于 user_balance_table.csv 中的按照 report_date 相同的条目来统计所有用户每日的 total_purchase_amt 资金流入与 total_redeem_amt 资金流出的总和。

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[] fields = value.toString().split(",");
if (fields.length < 10) return; // Check for valid row

String reportDate = fields[1]; // report_date
BigInteger totalPurchaseAmt = parseBigInteger(fields[4]); //total_purchase_amt
BigInteger totalRedeemAmt = parseBigInteger(fields[8]); // total_redeem_amt

dateKey.set(reportDate);
balancewritable.set(totalPurchaseAmt, totalRedeemAmt);
context.write(dateKey, balancewritable);
```

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 包含有一个 reduce 函数。

reduce 函数就是实现对输入的每一个键值对统计相同键（日期）下的 `total_purchase_amt` 资金流入与 `total_redeem_amt` 资金流出的总和，也就是实现对 `value` 的 `BalanceWritable` 类型数据的求和，最后把按照 < 日期 > TAB < 资金流入量 >,< 资金流出量 >格式进行输出。

以下是 reduce 函数的主要功能语句：

```
BigInteger totalInflow = BigInteger.ZERO;
BigInteger totalOutflow = BigInteger.ZERO;

for (BalanceWritable val : values) {
    totalInflow = totalInflow.add(val.getInflow());
    totalOutflow = totalOutflow.add(val.getOutflow());
}

result.set(totalInflow.toString() + "," + totalOutflow.toString());
context.write(key, result);
```

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/target/shiyan2-1.0-SNAPSHOT.jar
userBalanceAnalysis /input /output
```

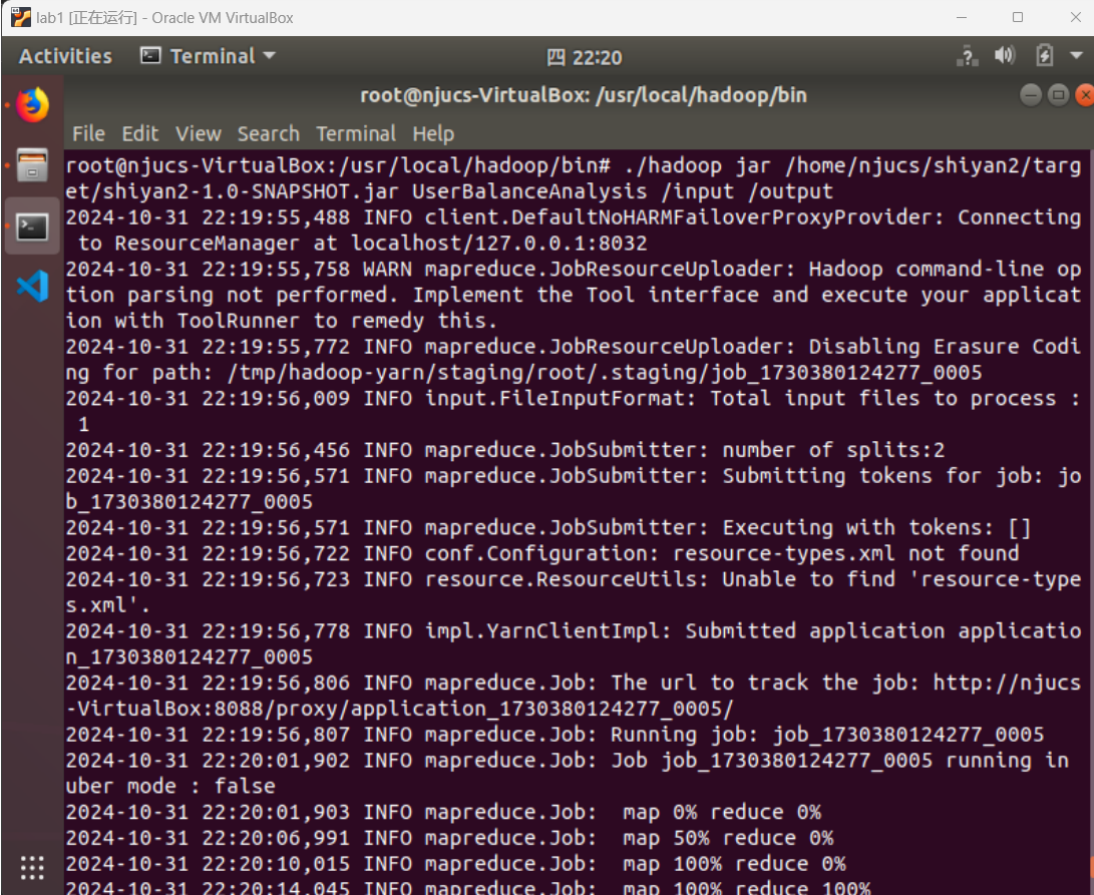
注意要把导出出来的 `part-r-00000` 解锁，以实现普通用户可以打开，命令如下：

```
sudo chown $USER part-r-00000
```

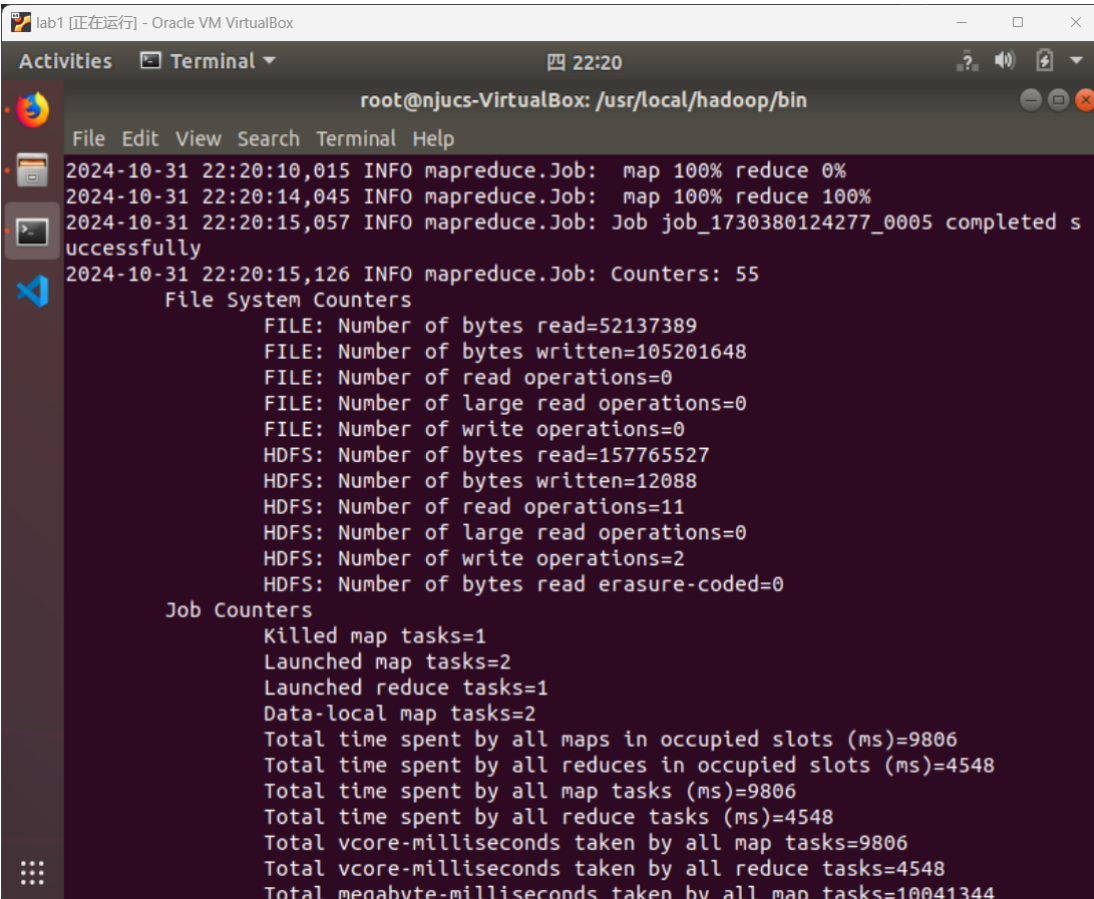
2、程序运行结果

以下即为 UserBalanceAnalysis.java 程序执行的任务一（根据 user_balance_table 表中的数据，统计所有用户每日的资金流入与流出情况。资金流入意味着申购行为，资金流出为赎回行为。输出格式为"< 日期 > TAB < 资金流入量 >,< 资金流出量 >") 的运行结果：

- 程序运行结果图：



```
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/target/shiyan2-1.0-SNAPSHOT.jar UserBalanceAnalysis /input /output
2024-10-31 22:19:55,488 INFO client.DefaultNoHARMFaloverProxyProvider: Connecting to ResourceManager at localhost/127.0.0.1:8032
2024-10-31 22:19:55,758 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-10-31 22:19:55,772 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1730380124277_0005
2024-10-31 22:19:56,009 INFO input.FileInputFormat: Total input files to process : 1
2024-10-31 22:19:56,456 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-31 22:19:56,571 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1730380124277_0005
2024-10-31 22:19:56,571 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-31 22:19:56,722 INFO conf.Configuration: resource-types.xml not found
2024-10-31 22:19:56,723 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-31 22:19:56,778 INFO impl.YarnClientImpl: Submitted application application_1730380124277_0005
2024-10-31 22:19:56,806 INFO mapreduce.Job: The url to track the job: http://njucs-VirtualBox:8088/proxy/application_1730380124277_0005/
2024-10-31 22:19:56,807 INFO mapreduce.Job: Running job: job_1730380124277_0005
2024-10-31 22:20:01,902 INFO mapreduce.Job: Job job_1730380124277_0005 running in uber mode : false
2024-10-31 22:20:01,903 INFO mapreduce.Job: map 0% reduce 0%
2024-10-31 22:20:06,991 INFO mapreduce.Job: map 50% reduce 0%
2024-10-31 22:20:10,015 INFO mapreduce.Job: map 100% reduce 0%
2024-10-31 22:20:14,045 INFO mapreduce.Job: map 100% reduce 100%
```



```
2024-10-31 22:20:10,015 INFO mapreduce.Job: map 100% reduce 0%
2024-10-31 22:20:14,045 INFO mapreduce.Job: map 100% reduce 100%
2024-10-31 22:20:15,057 INFO mapreduce.Job: Job job_1730380124277_0005 completed successfully
2024-10-31 22:20:15,126 INFO mapreduce.Job: Counters: 55
    File System Counters
      FILE: Number of bytes read=52137389
      FILE: Number of bytes written=105201648
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=157765527
      HDFS: Number of bytes written=12088
      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)=9806
      Total time spent by all reduces in occupied slots (ms)=4548
      Total time spent by all map tasks (ms)=9806
      Total time spent by all reduce tasks (ms)=4548
      Total vcore-milliseconds taken by all map tasks=9806
      Total vcore-milliseconds taken by all reduce tasks=4548
      Total megabyte-milliseconds taken by all map tasks=10041344
```

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 22:20
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

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)=9806
Total time spent by all reduces in occupied slots (ms)=4548
Total time spent by all map tasks (ms)=9806
Total time spent by all reduce tasks (ms)=4548
Total vcore-milliseconds taken by all map tasks=9806
Total vcore-milliseconds taken by all reduce tasks=4548
Total megabyte-milliseconds taken by all map tasks=10041344
Total megabyte-milliseconds taken by all reduce tasks=4657152

Map-Reduce Framework
Map input records=2840422
Map output records=2840422
Map output bytes=46456539
Map output materialized bytes=52137395
Input split bytes=230
Combine input records=0
Combine output records=0
Reduce input groups=428
Reduce shuffle bytes=52137395
Reduce input records=2840422
Reduce output records=428
Spilled Records=5680844
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=229
```

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 22:20
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

Reduce shuffle bytes=52137395
Reduce input records=2840422
Reduce output records=428
Spilled Records=5680844
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=229
CPU time spent (ms)=10150
Physical memory (bytes) snapshot=1358491648
Virtual memory (bytes) snapshot=7851220992
Total committed heap usage (bytes)=1175453696
Peak Map Physical memory (bytes)=494202880
Peak Map Virtual memory (bytes)=2616692736
Peak Reduce Physical memory (bytes)=378003456
Peak Reduce Virtual memory (bytes)=2621677568

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=157765297
File Output Format Counters
Bytes Written=12088
root@njucs-VirtualBox: /usr/local/hadoop/bin#
```

- part-r-00000输出结果图:

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 22:21
root@njucs-VirtualBox: /usr/local/hadoop/bin
File Edit View Search Terminal Help
File Output Format Counters
Bytes Written=12088
root@njucs-VirtualBox: /usr/local/hadoop/bin# ./hadoop fs -cat /output/part-r-00000
20130701      32488348,5525022
20130702      29037390,2554548
20130703      27270770,5953867
20130704      18321185,6410729
20130705      11648749,2763587
20130706      36751272,1616635
20130707      8962232,3982735
20130708      57258266,8347729
20130709      26798941,3473059
20130710      30696506,2597169
20130711      44075197,3508800
20130712      34183904,8492573
20130713      15164717,3482829
20130714      22615303,2784107
20130715      48128555,13107943
20130716      50622847,11864981
20130717      29015682,10911513
20130718      24234505,11765356
20130719      33680124,9244769
20130720      20439079,4601143
20130721      21142394,2681331
20130722      40448896,19144267
20130723      58136147,24404051
20130724      48422518,36258592
20130725      57433418,38212836
```

3、WEB页面截图

因为我第一次的程序代码有bug，运行失败了，于是我又修改了一下代码，第二次就 SUCCEEDED 了，同时输出结果也符合预期。

lab1 [正在运行] - Oracle VM VirtualBox

Activities Firefox Web Browser 21:44

All Applications - Mozilla Firefox

All Applications

localhost:8088/cluster/apps 50%

Hadoop All Applications

Cluster Metrics						
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	
2	0	0	2	0	<memory:0 B, vCores:0>	<mem...

Cluster Nodes Metrics			
Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost N...
1	0	0	0

Scheduler Metrics				
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Appl...
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
application_1730380124277_0002	root	User Balance Analysis	MAPREDUCE		root.default	0	Thu Oct 31 21:32:16 +0800 2024	Thu Oct 31 21:32:16 +0800 2024	Thu Oct 31 21:32:33 +0800 2024	FINISHED	SUCCEEDED
application_1730380124277_0001	root	User Balance Analysis	MAPREDUCE		root.default	0	Thu Oct 31 21:21:56 +0800 2024	Thu Oct 31 21:21:57 +0800 2024	Thu Oct 31 21:22:17 +0800 2024	FINISHED	FAILED

Showing 1 to 2 of 2 entries

