

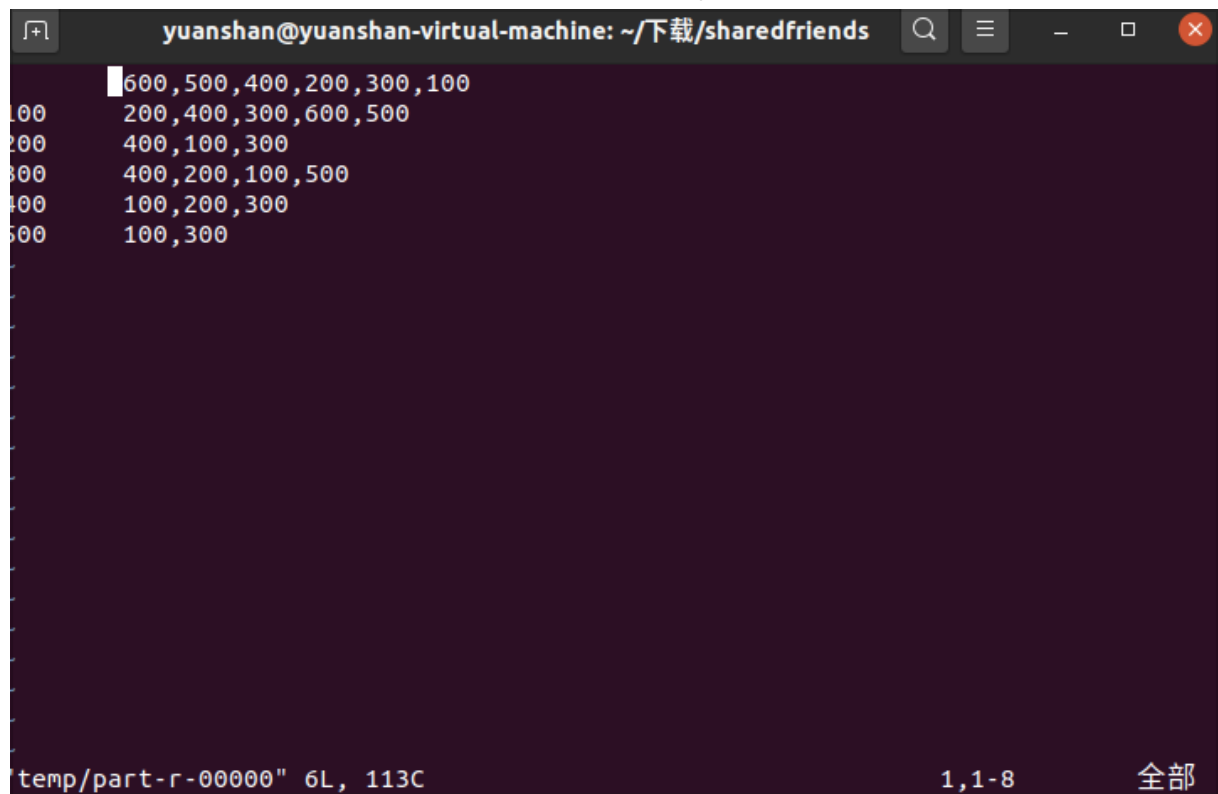
FBDP作业六

设计思路

1. 目前拥有的是人：好友1，好友2，...，好友k的形式，要找到共同好友，在mapreduce的算法设计思想下，应该分为两步
2. 首先，map1将数据格式改为，key：好友；value：存储过好友的人；这样一来，reduce1执行后可以得到对于一个好友，所有存储过其好友的名单，将结果输出保存
3. 第二次执行mapreduce，map2读取存储信息，得到存储过好友的人之间两两匹配，key：存储过好友的人1，存储过好友的人2；value：被存储的共同好友。reduce2将存储过好友的人1，存储过好友的人2所有的value结合在一起，最后实现输出

难点

1. 设计中间文件的读取等问题
2. 发现输出格式不对，保存了中间文件打开看第一次map-reduce结果



```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
100 600,500,400,200,300,100
200 200,400,300,600,500
300 400,100,300
400 400,200,100,500
500 100,200,300
600 100,300

temp/part-r-00000" 6L, 113C 1,1-8 全部
```

```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
100,200]      [,300,400]
100,300]      [,200]
100,400]      [300,]
100,500]      []
200,300]      [,100]
200,400]      [,100,300]
200,500]      [100,]
300,400]      [100,]
300,500]      [,100]
400,500]      [100,]

output/part-r-000000" 10L, 174C 1,9 全部
```

发现是，运用 `String[] friends = record[1].split("\\s+");` 后，数组中保存多了一个null值，重新去除所有null值后排列，看中间产出文件已正确

```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
100      600,500,300,400,200
200      400,100,300
300      400,200,100,500
400      100,300,200
500      300,100

temp/part-r-000000" 5L, 88C 2,15-19 全部
```

```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
100,200]      [400,300]
100,300]      [200]
100,400]      [300]
200,300]      [100]
200,400]      [300,100]
200,500]      [100]
300,400]      [100]
300,500]      [100]
400,500]      [100]

output/part-r-000000" 9L, 152C 1,9 全部
```

3. 随后发现均有缺少值，检查发现在map使两两组合时，没有定位完全，应到length-1的位置，修改后结果正确，如下

```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
Reduce shuffle bytes=374
Reduce input records=23
Reduce output records=14
Spilled Records=46
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=969932800
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=88
File Output Format Counters
Bytes Written=260
yuanshan@yuanshan-virtual-machine:~/下载/sharedfriends$ hdfs dfs -get /sharedfri
ends/output ./output
2020-10-31 18:30:47,726 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
```

```
yuanshan@yuanshan-virtual-machine: ~/下载/sharedfriends
100,200]      [300,400]
100,300]      [200,400,500]
100,400]      [200,300]
100,500]      [300]
200,300]      [400,100]
200,400]      [300,100]
200,500]      [300,100]
200,600]      [100]
300,400]      [200,100]
300,500]      [100]
300,600]      [100]
400,500]      [100,300]
400,600]      [100]
500,600]      [100]

output/part-r-000000" 14L, 260C 1,9 全部
```

本次作业收获

1. 进一步熟悉了map-reduce 的用法以及Hadoop相关操作
2. 对文件输入流的了解更深厚
3. 通过对中间输出文件的检查进行debug的方式比较有效