作业6——寻找共同好友

代码仓库**https://git.nju.edu.cn/Bruce/hw6**

✓ 运行环境

- win10+IDEA+hadoop单机完成编码调试
- bdkit完成集群中任务提交和运行。

✓ 任务分析

1. 仔细分析课程网站提供的输入文件样例，可以得之，输入文件中表示的好友关系是单向的，

   即<,>...只表示friend1,friends2,...,friendn在person的好友列表里，反之**不一定**成立。

2. 为方便表述，以下将问题中的好友关系转化为称**friend是person的学生**或**person是friend的老师**。（只是为了方便表述与理解，因为现实生活中一般认为好友关系是双向的，这使得问题中定义的好友关系理解起来比较困难，因此换成老师与学生的关系便于理解，需要注意，两个人可能互为师生，这在现实中也是可以理解的）

3. 因此，该任务为找出各用户的共有学生。（每个用户都可能是老师也都可能是学生）

4. 考虑使用两个MapReduce的Job完成之。（每个Job有各自的Mapper和Reducer）

   第一个Job得出每个用户是哪些用户的学生，并将该结果输入一个临时（每一行对应一个用户的老师的情况）

   第二个Job逐行读入临时文件，每一用户的老师两两有该用户作为共有学生。之后删除临时文件，并按格式输出最终结果。

✓ 具体设计

- Job1
  - Mapper：解析初始输入，对每一行，定义键值对为<friendi, person>，即学生为key,老师为value。

- Reducer：将同一学生的所有老师记录（对应于mapper输出的同key键值对）为一行。输入临时文件。

  临时文件

- Job2
  - Mapper：读入中间文件，对每一行，定义键值对为<[friendi, friendj],person>，即老师二元组为key（由某学生的所有老师两两组合而成),共有学生为value。
  - Reduce：得到每个老师二元组的所有公共学生，按格式输出并删除临时文件即可。
- 具体的解释见源码中的注释

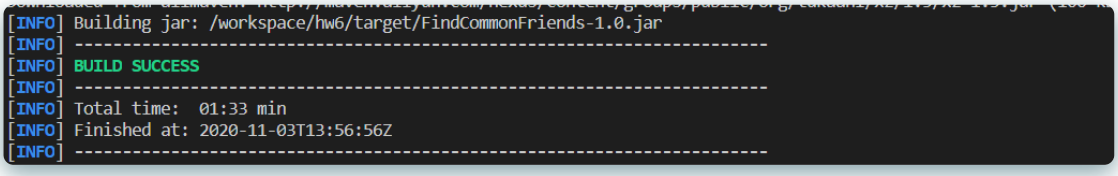
✓ 运行说明

不需要指定主类，可输入若干参数（至少两个，最后一个参数为输出路径，其余所有参数为输入路径。如：

```
hadoop jar target/FindCommonFriends-1.0.jar input_path1 input_path2 output_path
```

✓ 运行截图

源码打包



提交作业完成（两个Job）

- Job 1

```
root@cyj181870013-master:/workspace/hw6# hadoop jar target/FindCommonFriends-1.0.jar /input /output
Java HotSpot(TM) 64-Bit Server VM warning: You have loaded library /usr/local/hadoop/lib/native/libhadoop.so wh
 try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack
20/11/03 14:00:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
20/11/03 14:00:27 INFO client.RMProxy: Connecting to ResourceManager at cyj181870013-master/192.168.219.164:803
20/11/03 14:00:27 INFO input.FileInputFormat: Total input paths to process : 2
20/11/03 14:00:27 INFO mapreduce.JobSubmitter: number of splits:2
20/11/03 14:00:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604411568419_0001
20/11/03 14:00:28 INFO impl.YarnClientImpl: Submitted application application_1604411568419_0001
20/11/03 14:00:28 INFO mapreduce.Job: The url to track the job: http://cyj181870013-master:8088/proxy/applicati
20/11/03 14:00:28 INFO mapreduce.Job: Running job: job_1604411568419_0001
20/11/03 14:00:34 INFO mapreduce.Job: Job job_1604411568419_0001 running in uber mode : false
20/11/03 14:00:34 INFO mapreduce.Job:  map 0% reduce 0%
20/11/03 14:00:39 INFO mapreduce.Job:  map 100% reduce 0%
20/11/03 14:00:44 INFO mapreduce.Job:  map 100% reduce 100%
20/11/03 14:00:44 INFO mapreduce.Job: Job job_1604411568419_0001 completed successfully
```

- Job 2

```
20/11/03 14:00:44 INFO client.RMProxy: Connecting to ResourceManager at cyj181870013-master/192.168.219.164:8032
20/11/03 14:00:44 INFO input.FileInputFormat: Total input paths to process : 1
20/11/03 14:00:44 INFO mapreduce.JobSubmitter: number of splits:1
20/11/03 14:00:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604411568419_0002
20/11/03 14:00:44 INFO impl.YarnClientImpl: Submitted application application_1604411568419_0002
20/11/03 14:00:44 INFO mapreduce.Job: The url to track the job: http://cyj181870013-master:8088/proxy/application_16044115
20/11/03 14:00:44 INFO mapreduce.Job: Running job: job_1604411568419_0002
20/11/03 14:00:53 INFO mapreduce.Job: Job job_1604411568419_0002 running in uber mode : false
20/11/03 14:00:53 INFO mapreduce.Job:  map 0% reduce 0%
20/11/03 14:00:57 INFO mapreduce.Job:  map 100% reduce 0%
20/11/03 14:01:01 INFO mapreduce.Job:  map 100% reduce 100%
20/11/03 14:01:01 INFO mapreduce.Job: Job job_1604411568419_0002 completed successfully
20/11/03 14:01:01 INFO mapreduce.Job: Counters: 49
```

Resource-Manager中显示先后完成两个任务

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1604411568419_0002	root	CommonFriendsStep2	MAPREDUCE	default	Tue Nov 3 22:00:44 +0800 2020	Tue Nov 3 22:01:00 +0800 2020	FINISHED	SUCCEEDED		History	N/A
application_1604411568419_0001	root	CommonFriendsStep1	MAPREDUCE	default	Tue Nov 3 22:00:28 +0800 2020	Tue Nov 3 22:00:42 +0800 2020	FINISHED	SUCCEEDED		History	N/A

输出结果（本路径下output文件夹中也有）

```
root@cyj181870013-master:/workspace/hw6# hadoop fs -cat /output/part-r-00000
Java HotSpot(TM) 64-Bit Server VM warning: You have loaded library /usr/local/hadoop/lib/native/libhadoop.so which might have disabled stack guard. The VM will
 try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
20/11/03 14:02:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
([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])
```

下载到本地的输出文件

part-r-00000 ×

hw6 > output > part-r-00000

```
([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])
```

✓ 遇到的问题

- Job数量开始没有确定好，导致编写困难：

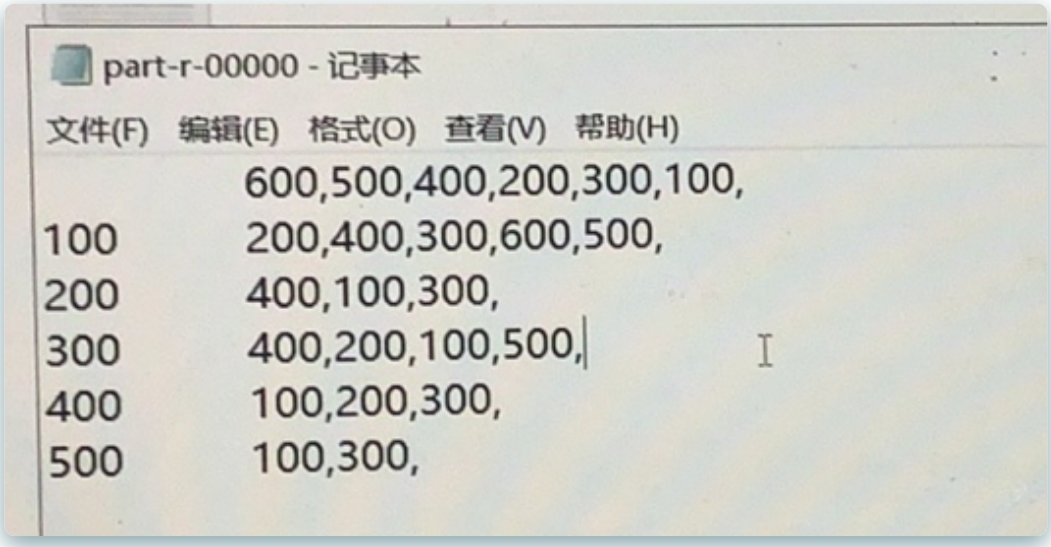
  在编码前，没有先构思好整体框架，一会儿觉得一个Job不够，要两个，一会儿觉得够了。改来改去，十分麻烦。应当在动手前先设计好流程，明确各阶段要实现的目标，而不是边写边改设计。

- 编译运行后一直提示临时路径作为第二个Job的输入路径不存在：

  ```
  Input path does not exist: file:/C:/Users/CYJ/Desktop/hadoop_demo/common_friends/output_temp
  ```

  后来发现，是我在设置Job2的输入路径前，前没有等待Job1完成，因此只要加一个判断Job完成的语句即可。

- 中间文件内容不完全正确：

  原因在于网站上给的输入文件，person和friend之间使用了逗号加空格为分隔符，而不是逗号，发现此问题后，修改字符串处理代码，解决问题。