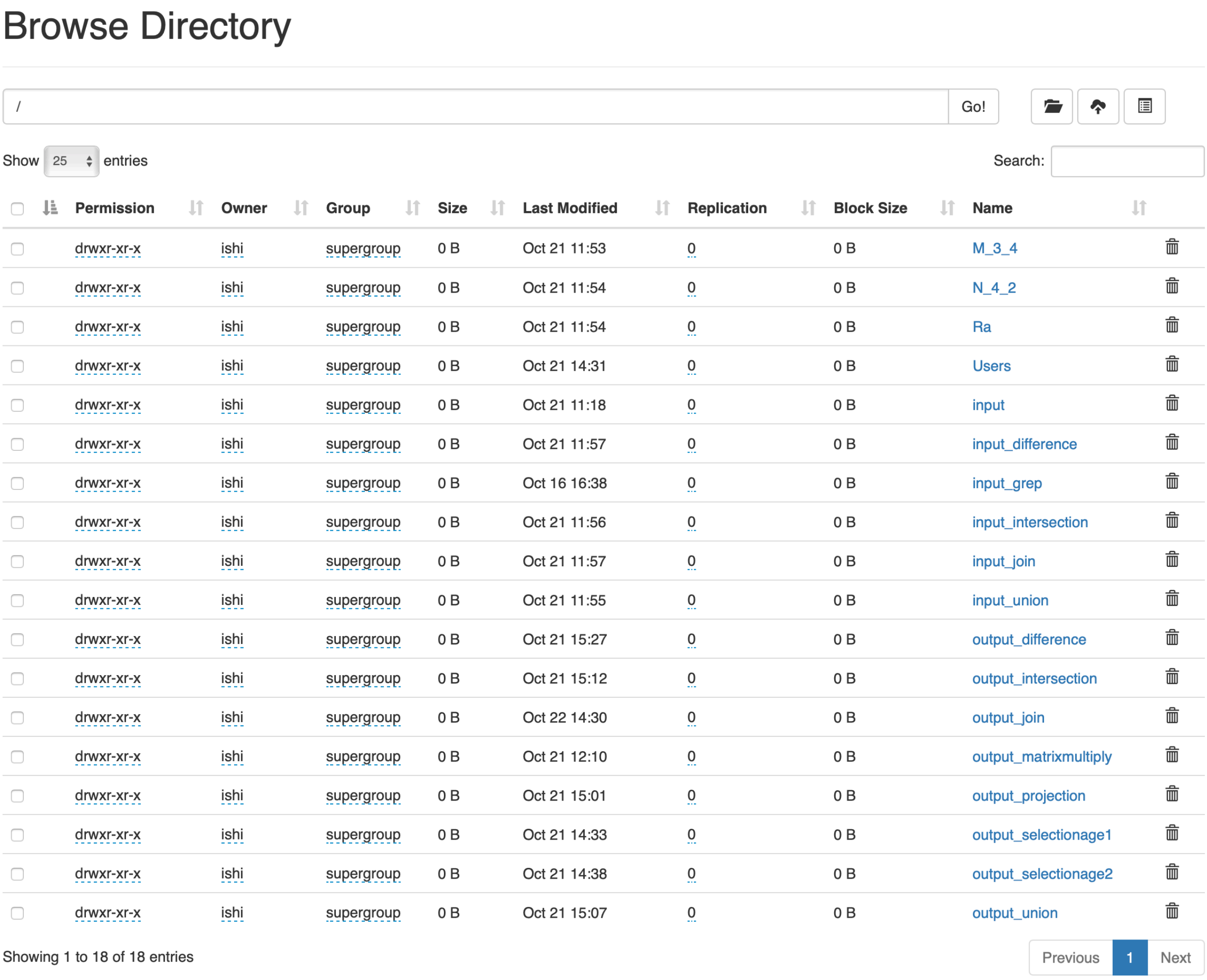
金融大数据处理技术_作业4

171860015-计金-石霭青

本次作业内容为MapReduce基础编程，运行环境为mac os系统+hadoop3.2.1伪分布式+Intellij IDEA。

全部程序伪分布运行完后HDFS文件系统中的输入输出文件目录展示如下。



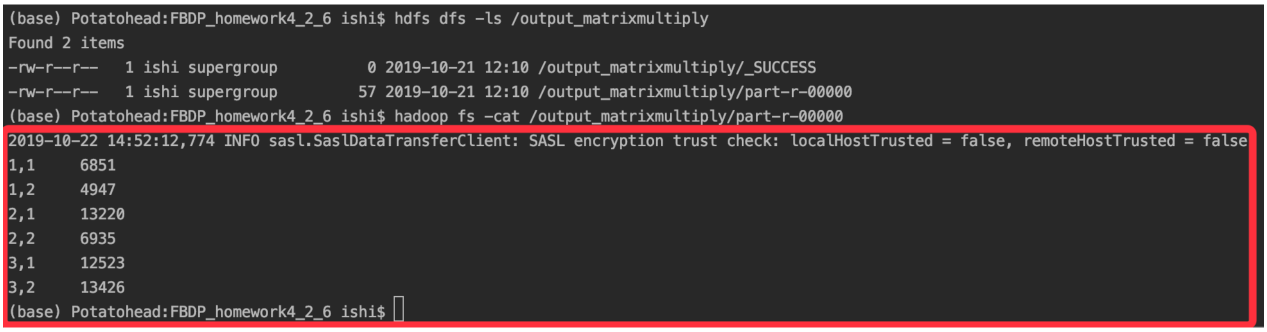
Browse Directory

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:53	0	0 B	M_3_4	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:54	0	0 B	N_4_2	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:54	0	0 B	Ra	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 14:31	0	0 B	Users	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:18	0	0 B	input	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:57	0	0 B	input_difference	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 16 16:38	0	0 B	input_grep	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:56	0	0 B	input_intersection	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:57	0	0 B	input_join	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 11:55	0	0 B	input_union	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 15:27	0	0 B	output_difference	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 15:12	0	0 B	output_intersection	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 22 14:30	0	0 B	output_join	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 12:10	0	0 B	output_matrixmultiply	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 15:01	0	0 B	output_projection	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 14:33	0	0 B	output_selectionage1	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 14:38	0	0 B	output_selectionage2	🗑
□	drwxr-xr-x	ishi	supergroup	0 B	Oct 21 15:07	0	0 B	output_union	🗑

Showing 1 to 18 of 18 entries

以下为各题运行截图。

1 给出矩阵乘法的MapReduce实现，以M_3_4和N_4_2作为输入进行测试。

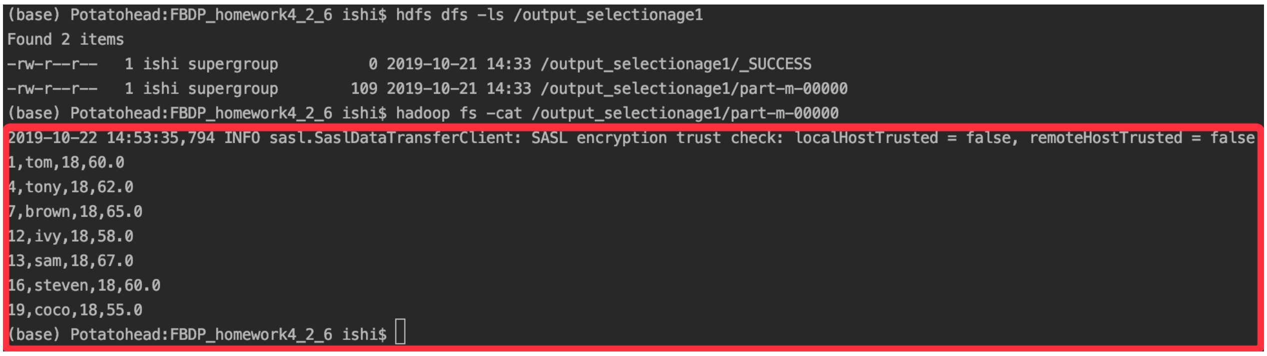


```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_matrixmultiply
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 12:10 /output_matrixmultiply/_SUCCESS
-rw-r--r--   1 ishi supergroup         57 2019-10-21 12:10 /output_matrixmultiply/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_matrixmultiply/part-r-00000
2019-10-22 14:52:12,774 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1,1	6851
1,2	4947
2,1	13220
2,2	6935
3,1	12523
3,2	13426
(base) Potatohead:FBDP_homework4_2_6 ishi$ 
```

2 给出关系代数的选择、投影、并集、交集、差集及自然连接的MapReduce实现。测试集如下：

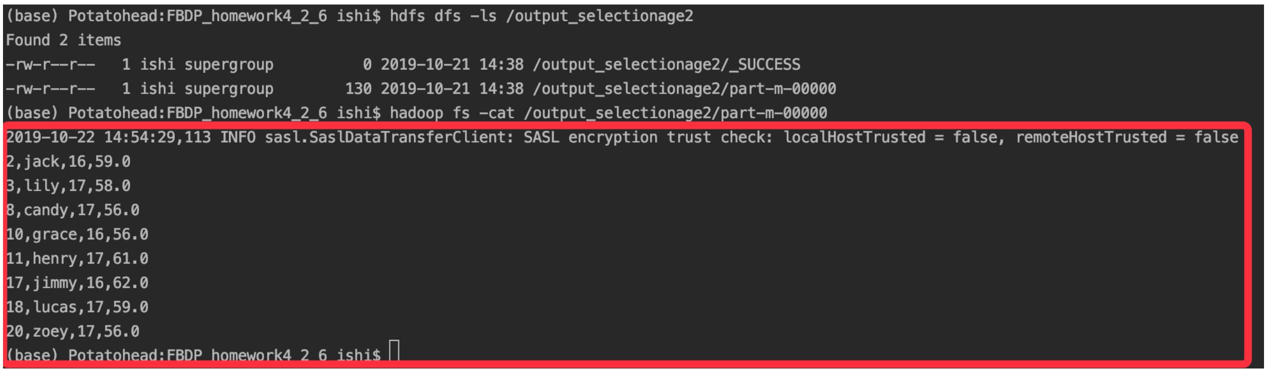
- 关系Ra：(id, name, age, weight)
- 关系Rb：(id, gender, height)

2.1 在Ra.txt上选择age=18的记录



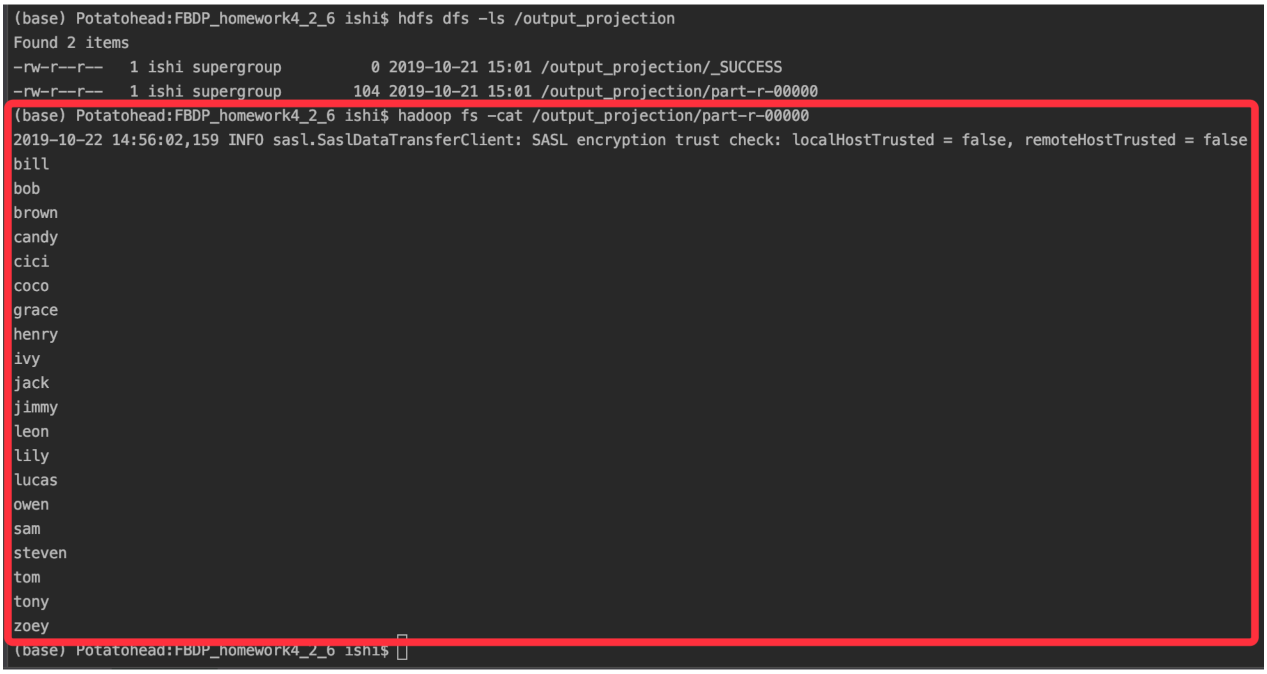
```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_selectionage1
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 14:33 /output_selectionage1/_SUCCESS
-rw-r--r--   1 ishi supergroup        109 2019-10-21 14:33 /output_selectionage1/part-m-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_selectionage1/part-m-00000
2019-10-22 14:53:35,794 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1,tom,18,60.0
4,tony,18,62.0
7,brown,18,65.0
12,ivy,18,58.0
13,sam,18,67.0
16,steven,18,60.0
19,coco,18,55.0
(base) Potatohead:FBDP_homework4_2_6 ishi$ 
```

在Ra.txt上选择age<18的记录



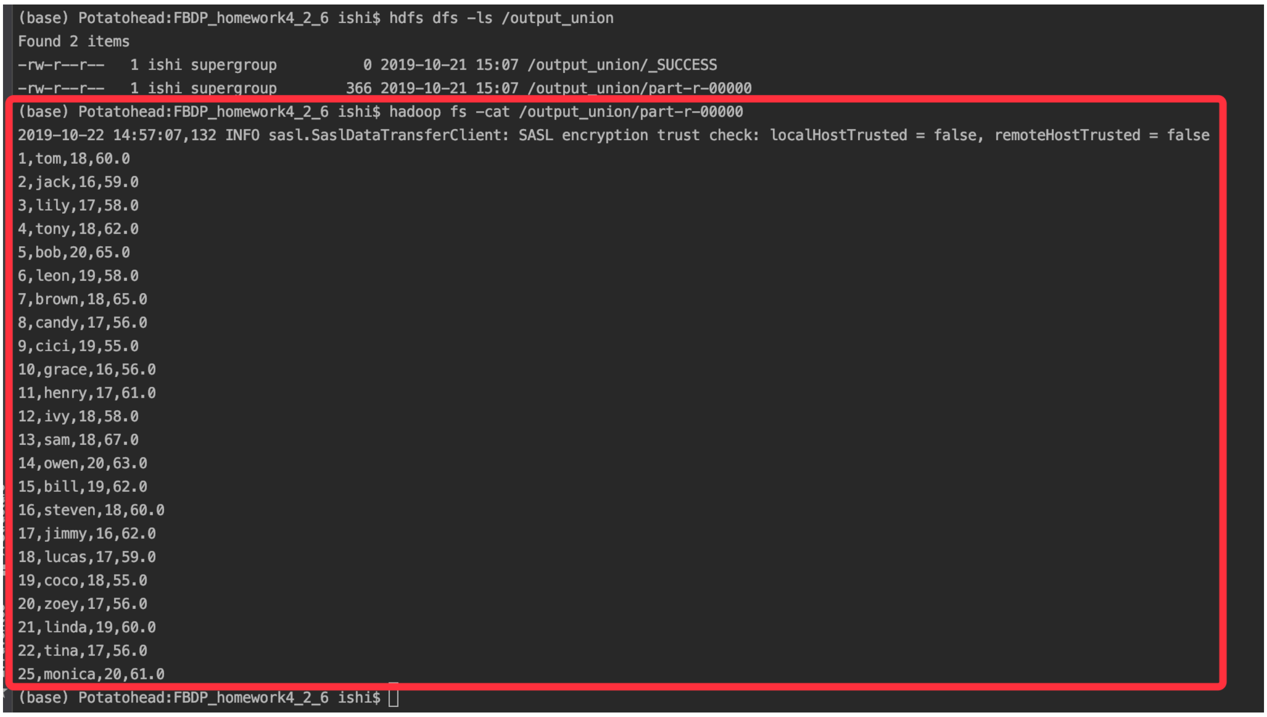
```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_selectionage2
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 14:38 /output_selectionage2/_SUCCESS
-rw-r--r--   1 ishi supergroup        130 2019-10-21 14:38 /output_selectionage2/part-m-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_selectionage2/part-m-00000
2019-10-22 14:54:29,113 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2,jack,16,59.0
3,lily,17,58.0
8,candy,17,56.0
10,grace,16,56.0
11,henry,17,61.0
17,jimmy,16,62.0
18,lucas,17,59.0
20,zoey,17,56.0
(base) Potatohead:FBDP_homework4_2_6 ishi$ 
```

2.2 在Ra.txt上对属性name进行投影



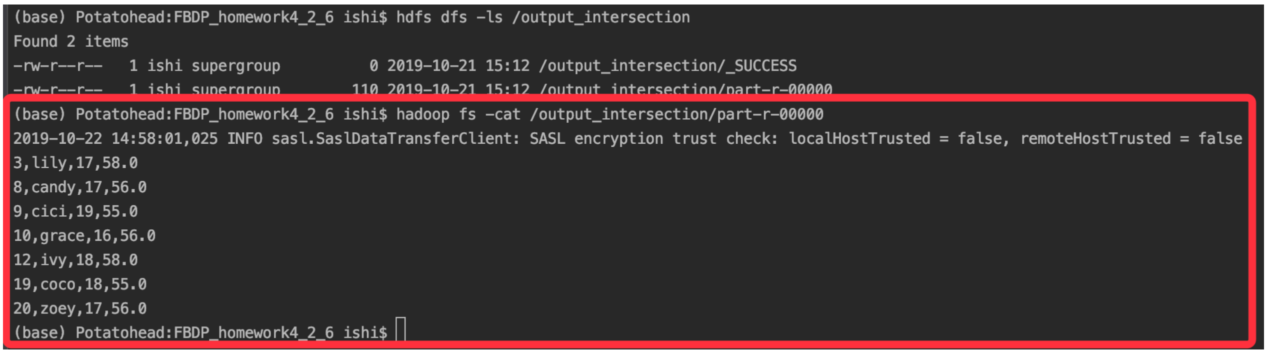
```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_projection
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 15:01 /output_projection/_SUCCESS
-rw-r--r--   1 ishi supergroup        104 2019-10-21 15:01 /output_projection/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_projection/part-r-00000
2019-10-22 14:56:02,159 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
bill
bob
brown
candy
cici
coco
grace
henry
ivy
jack
jimmy
leon
lily
lucas
owen
sam
steven
tom
tony
zoey
(base) Potatohead:FBDP_homework4_2_6 ishi$
```

2.3 求Ra1和Ra2的并集



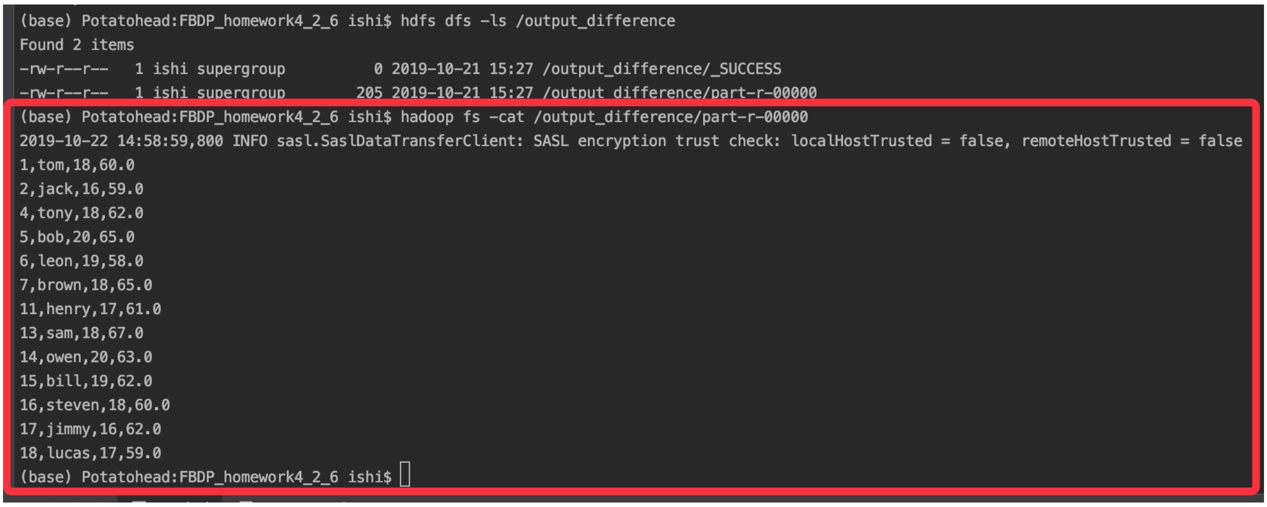
```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_union
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 15:07 /output_union/_SUCCESS
-rw-r--r--   1 ishi supergroup        366 2019-10-21 15:07 /output_union/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_union/part-r-00000
2019-10-22 14:57:07,132 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1,tom,18,60.0
2,jack,16,59.0
3,lily,17,58.0
4,tony,18,62.0
5,bob,20,65.0
6,leon,19,58.0
7,brown,18,65.0
8,candy,17,56.0
9,cici,19,55.0
10,grace,16,56.0
11,henry,17,61.0
12,ivy,18,58.0
13,sam,18,67.0
14,owen,20,63.0
15,bill,19,62.0
16,steven,18,60.0
17,jimmy,16,62.0
18,lucas,17,59.0
19,coco,18,55.0
20,zoey,17,56.0
21,linda,19,60.0
22,tina,17,56.0
25,monica,20,61.0
(base) Potatohead:FBDP_homework4_2_6 ishi$
```

2.4 求Ra1和Ra2的交集



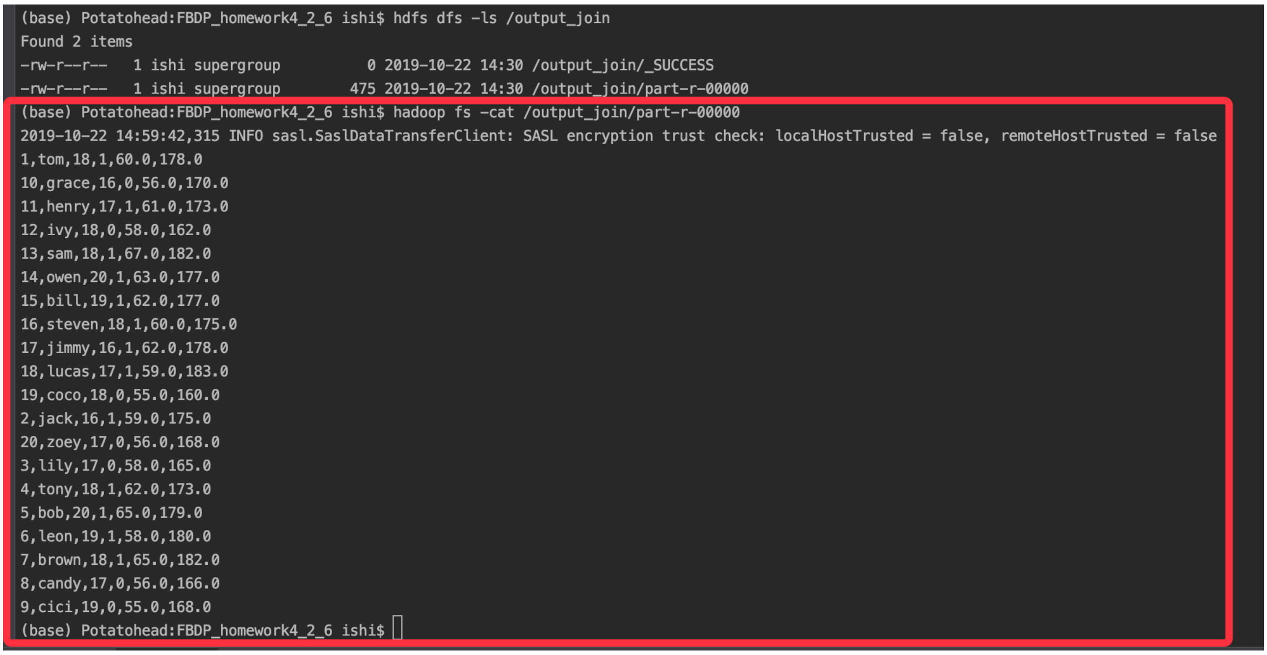
```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_intersection
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 15:12 /output_intersection/_SUCCESS
-rw-r--r--   1 ishi supergroup        110 2019-10-21 15:12 /output_intersection/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_intersection/part-r-00000
2019-10-22 14:58:01,025 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
3,lily,17,58.0
8,candy,17,56.0
9,cici,19,55.0
10,grace,16,56.0
12,ivy,18,58.0
19,coco,18,55.0
20,zoey,17,56.0
(base) Potatohead:FBDP_homework4_2_6 ishi$
```

2.5 求Ra2 - Ra1



```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_difference
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-21 15:27 /output_difference/_SUCCESS
-rw-r--r--   1 ishi supergroup        205 2019-10-21 15:27 /output_difference/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_difference/part-r-00000
2019-10-22 14:58:59,800 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1,tom,18,60.0
2,jack,16,59.0
4,tony,18,62.0
5,bob,20,65.0
6,leon,19,58.0
7,brown,18,65.0
11,henry,17,61.0
13,sam,18,67.0
14,owen,20,63.0
15,bill,19,62.0
16,steven,18,60.0
17,jimmy,16,62.0
18,lucas,17,59.0
(base) Potatohead:FBDP_homework4_2_6 ishi$ 
```

2.6 Ra和Rb在属性id上进行自然连接，要求最后的输出格式为(id, name, age, gender, weight, height)



```
(base) Potatohead:FBDP_homework4_2_6 ishi$ hdfs dfs -ls /output_join
Found 2 items
-rw-r--r--   1 ishi supergroup          0 2019-10-22 14:30 /output_join/_SUCCESS
-rw-r--r--   1 ishi supergroup        475 2019-10-22 14:30 /output_join/part-r-00000
(base) Potatohead:FBDP_homework4_2_6 ishi$ hadoop fs -cat /output_join/part-r-00000
2019-10-22 14:59:42,315 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1,tom,18,1,60.0,178.0
10,grace,16,0,56.0,170.0
11,henry,17,1,61.0,173.0
12,ivy,18,0,58.0,162.0
13,sam,18,1,67.0,182.0
14,owen,20,1,63.0,177.0
15,bill,19,1,62.0,177.0
16,steven,18,1,60.0,175.0
17,jimmy,16,1,62.0,178.0
18,lucas,17,1,59.0,183.0
19,coco,18,0,55.0,160.0
2,jack,16,1,59.0,175.0
20,zoey,17,0,56.0,168.0
3,lily,17,0,58.0,165.0
4,tony,18,1,62.0,173.0
5,bob,20,1,65.0,179.0
6,leon,19,1,58.0,180.0
7,brown,18,1,65.0,182.0
8,candy,17,0,56.0,166.0
9,cici,19,0,55.0,168.0
(base) Potatohead:FBDP_homework4_2_6 ishi$ 
```