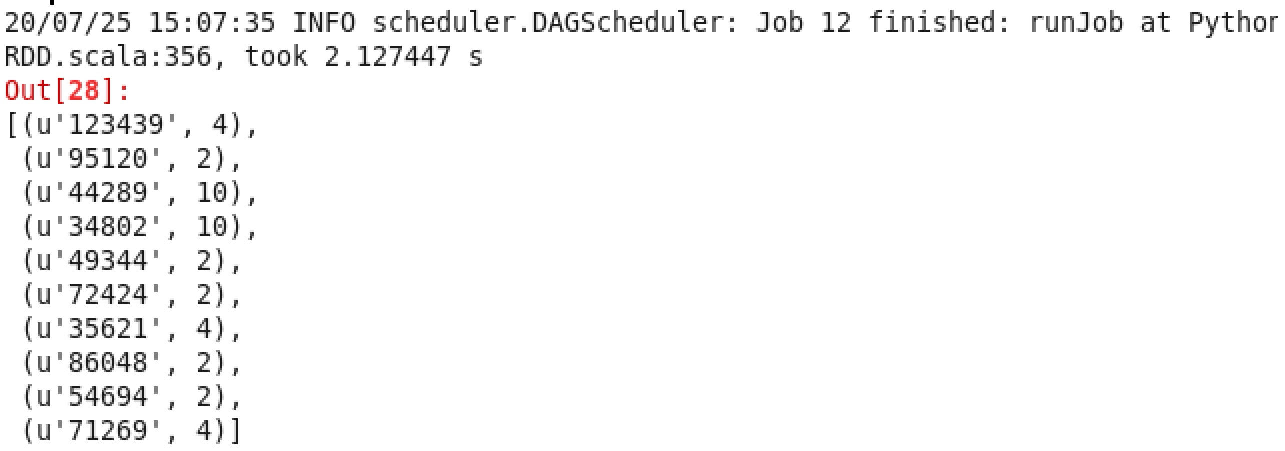
BUAN 6346 Big Data

Homework 12

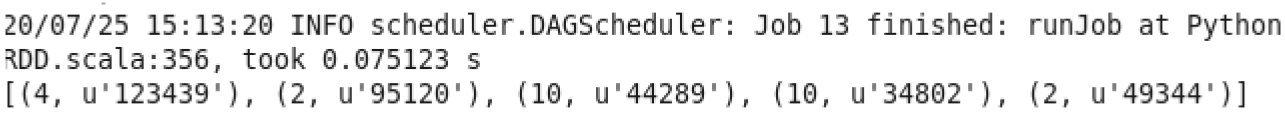
Use pair RDDs to join two datasets

Step1:



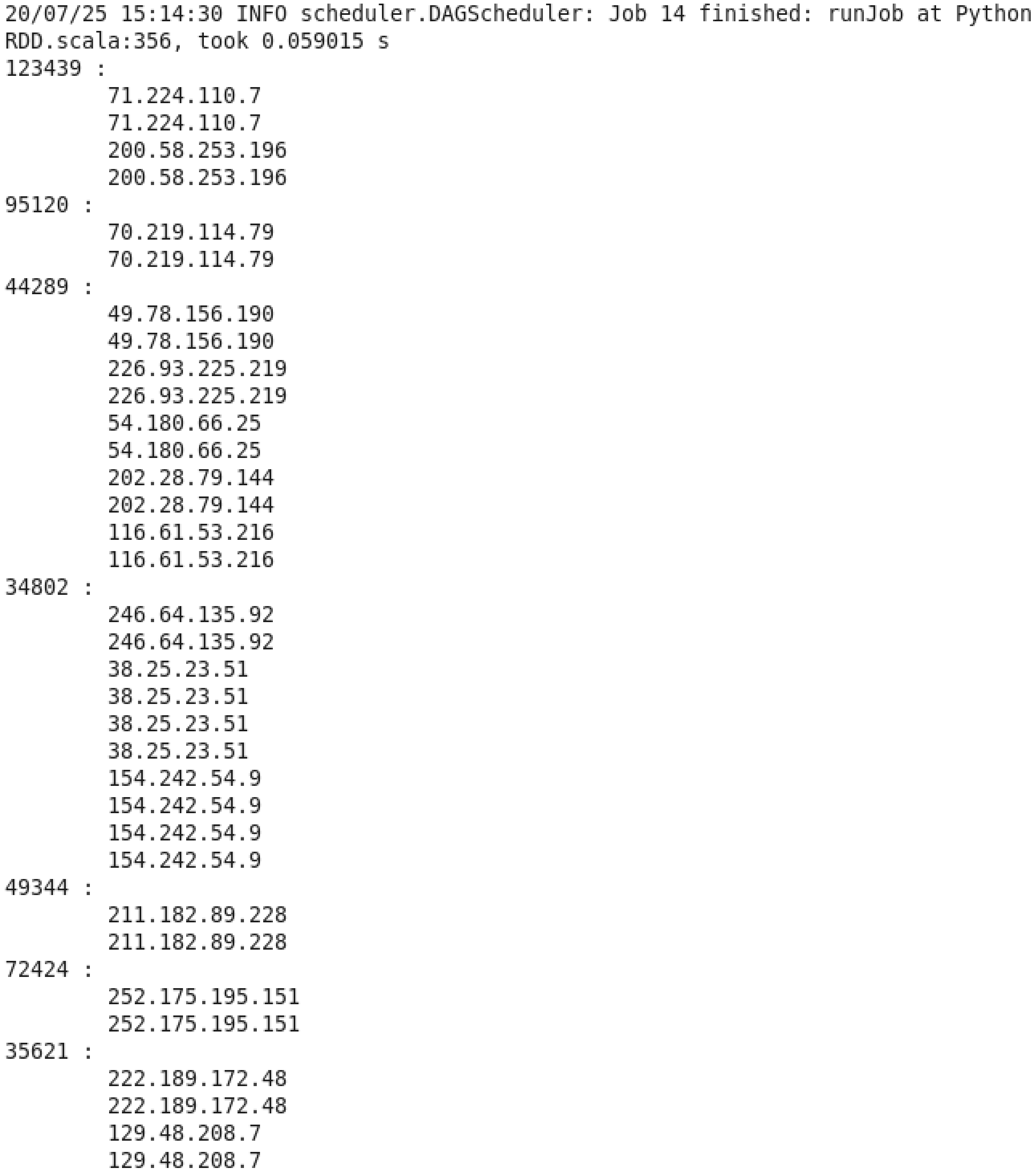
```
20/07/25 15:07:35 INFO scheduler.DAGScheduler: Job 12 finished: runJob at Pythor
RDD.scala:356, took 2.127447 s
Out[28]:
[(u'123439', 4),
 (u'95120', 2),
 (u'44289', 10),
 (u'34802', 10),
 (u'49344', 2),
 (u'72424', 2),
 (u'35621', 4),
 (u'86048', 2),
 (u'54694', 2),
 (u'71269', 4)]
```

Step2:



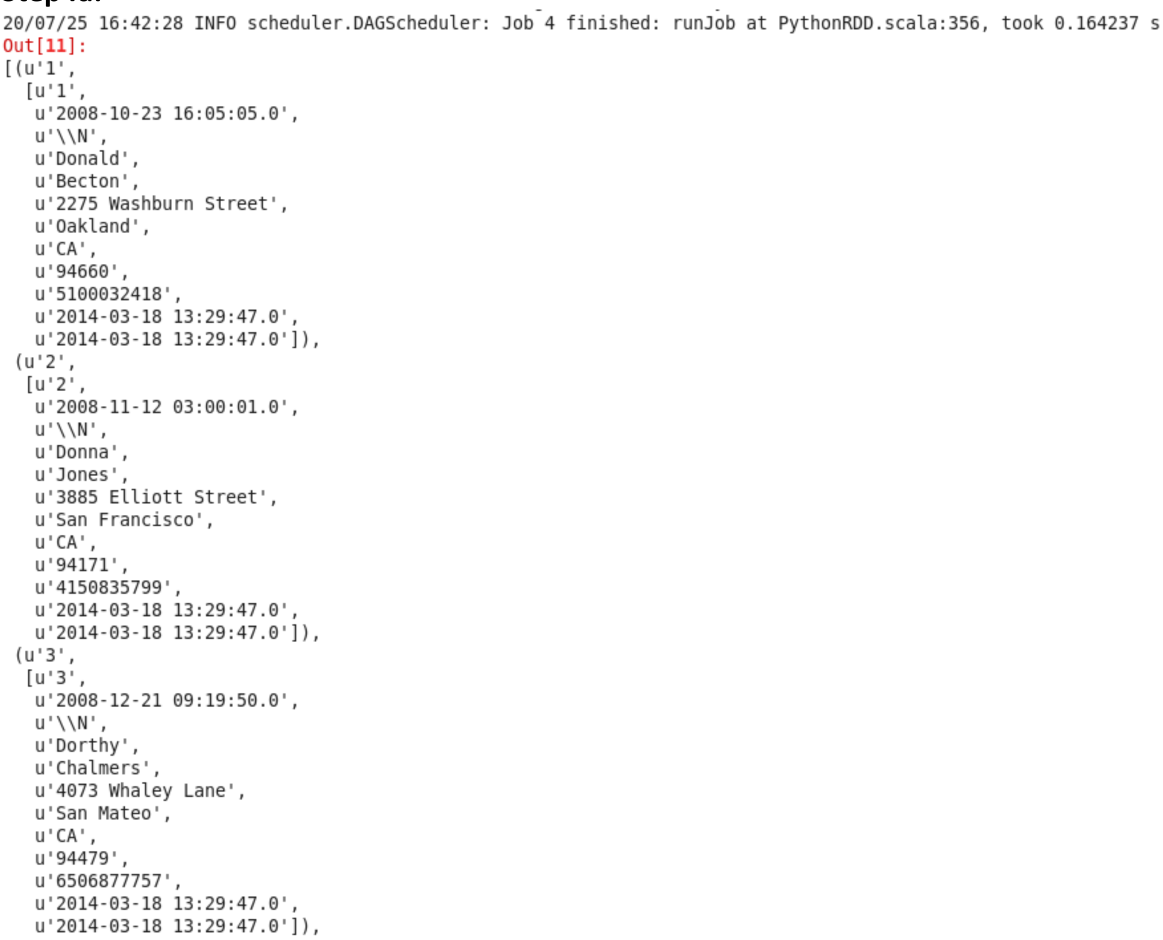
```
20/07/25 15:13:20 INFO scheduler.DAGScheduler: Job 13 finished: runJob at Python
RDD.scala:356, took 0.075123 s
[(4, u'123439'), (2, u'95120'), (10, u'44289'), (10, u'34802'), (2, u'49344')]
```

Step3:



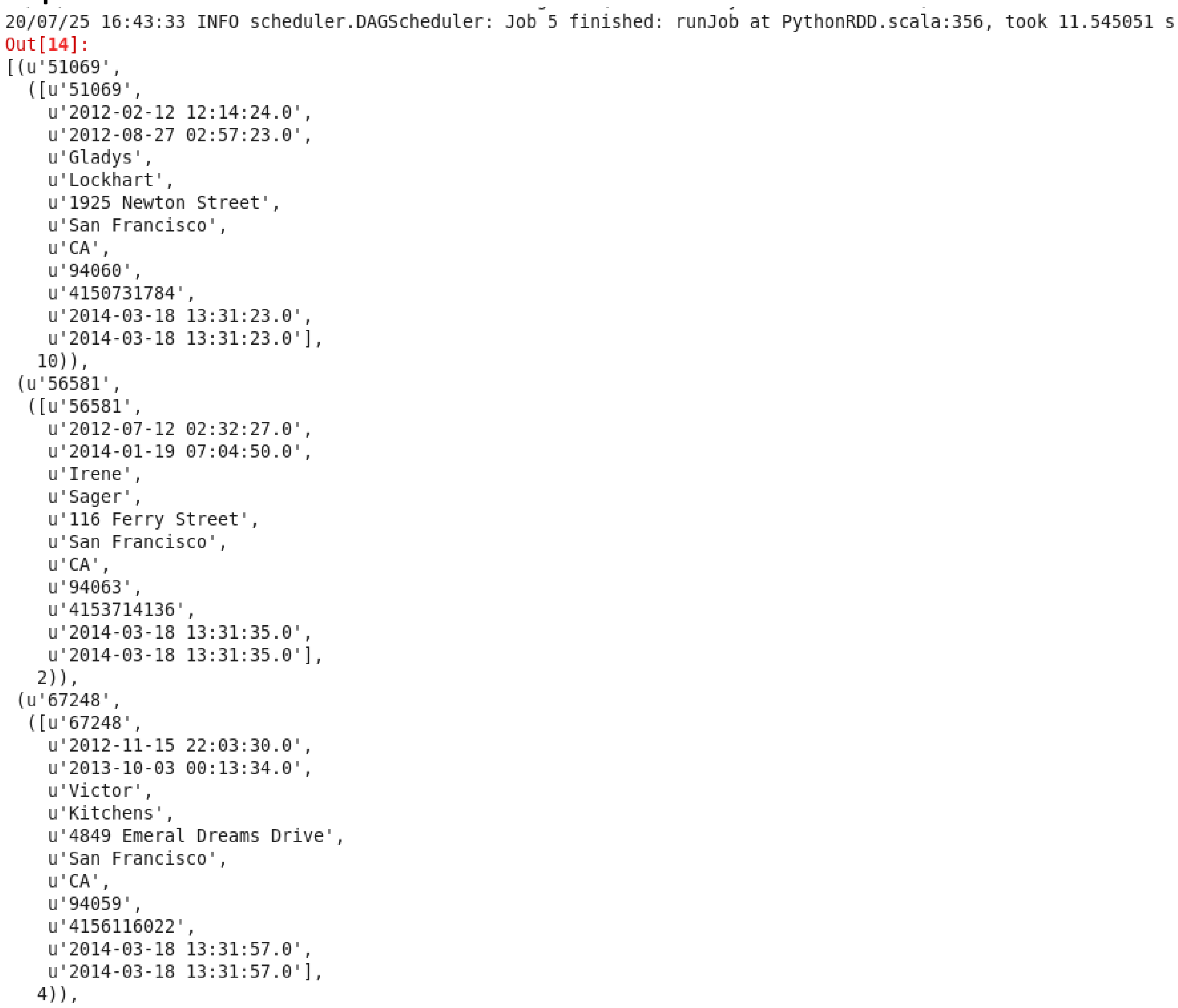
```
20/07/25 15:14:30 INFO scheduler.DAGScheduler: Job 14 finished: runJob at Python
RDD.scala:356, took 0.059015 s
123439 :
        71.224.110.7
        71.224.110.7
        200.58.253.196
        200.58.253.196
95120 :
        70.219.114.79
        70.219.114.79
44289 :
        49.78.156.190
        49.78.156.190
        226.93.225.219
        226.93.225.219
        54.180.66.25
        54.180.66.25
        202.28.79.144
        202.28.79.144
        116.61.53.216
        116.61.53.216
34802 :
        246.64.135.92
        246.64.135.92
        38.25.23.51
        38.25.23.51
        38.25.23.51
        38.25.23.51
        154.242.54.9
        154.242.54.9
        154.242.54.9
        154.242.54.9
49344 :
        211.182.89.228
        211.182.89.228
72424 :
        252.175.195.151
        252.175.195.151
35621 :
        222.189.172.48
        222.189.172.48
        129.48.208.7
        129.48.208.7
```

Step4a:



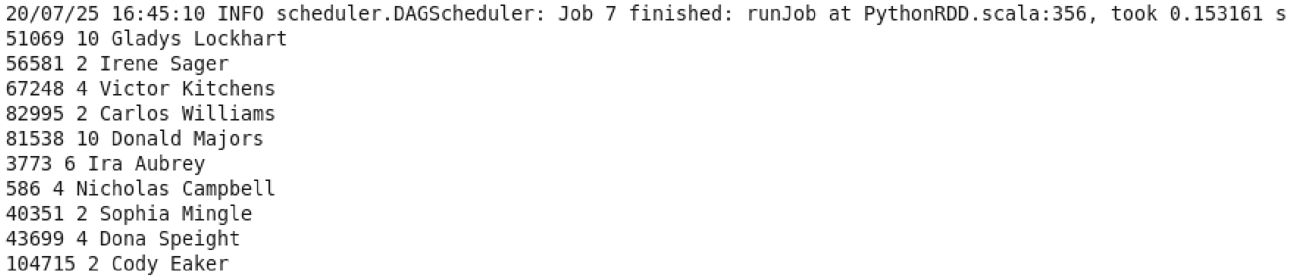
```
20/07/25 16:42:28 INFO scheduler.DAGScheduler: Job 4 finished: runJob at PythonRDD.scala:356, took 0.164237 s
Out[11]:
[(u'1',
  [u'1',
   u'2008-10-23 16:05:05.0',
   u'\\N',
   u'Donald',
   u'Becton',
   u'2275 Washburn Street',
   u'Oakland',
   u'CA',
   u'94660',
   u'5100032418',
   u'2014-03-18 13:29:47.0',
   u'2014-03-18 13:29:47.0']),
 (u'2',
  [u'2',
   u'2008-11-12 03:00:01.0',
   u'\\N',
   u'Donna',
   u'Jones',
   u'3885 Elliott Street',
   u'San Francisco',
   u'CA',
   u'94171',
   u'4150835799',
   u'2014-03-18 13:29:47.0',
   u'2014-03-18 13:29:47.0']),
 (u'3',
  [u'3',
   u'2008-12-21 09:19:50.0',
   u'\\N',
   u'Dorthy',
   u'Chalmers',
   u'4073 Whaley Lane',
   u'San Mateo',
   u'CA',
   u'94479',
   u'6506877757',
   u'2014-03-18 13:29:47.0',
   u'2014-03-18 13:29:47.0']),
```

Step4b:



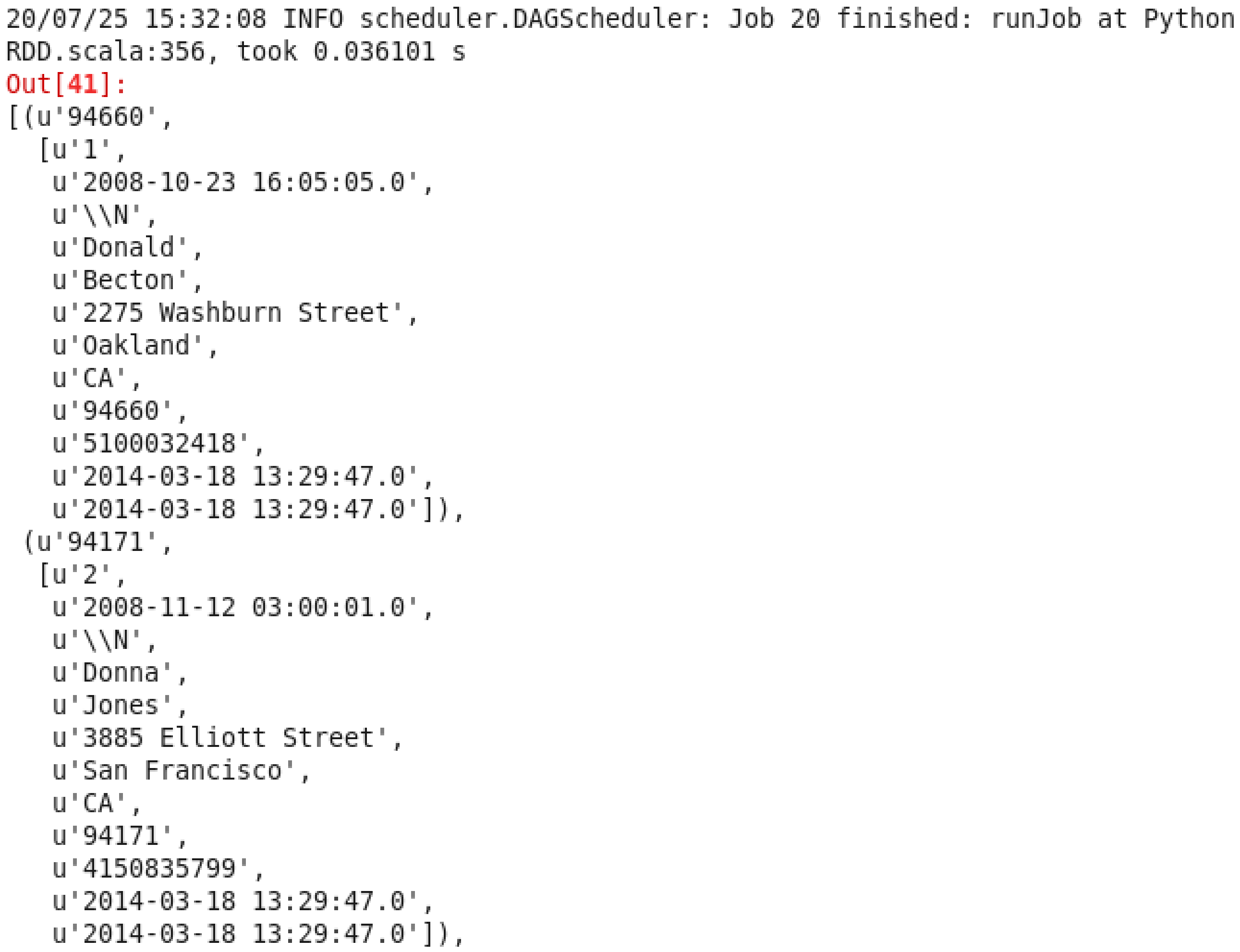
```
20/07/25 16:43:33 INFO scheduler.DAGScheduler: Job 5 finished: runJob at PythonRDD.scala:356, took 11.545051 s
Out[14]:
[(u'51069',
  ([u'51069',
    u'2012-02-12 12:14:24.0',
    u'2012-08-27 02:57:23.0',
    u'Gladys',
    u'Lockhart',
    u'1925 Newton Street',
    u'San Francisco',
    u'CA',
    u'94060',
    u'4150731784',
    u'2014-03-18 13:31:23.0',
    u'2014-03-18 13:31:23.0'],
   10)),
 (u'56581',
  ([u'56581',
    u'2012-07-12 02:32:27.0',
    u'2014-01-19 07:04:50.0',
    u'Irene',
    u'Sager',
    u'116 Ferry Street',
    u'San Francisco',
    u'CA',
    u'94063',
    u'4153714136',
    u'2014-03-18 13:31:35.0',
    u'2014-03-18 13:31:35.0'],
   2)),
 (u'67248',
  ([u'67248',
    u'2012-11-15 22:03:30.0',
    u'2013-10-03 00:13:34.0',
    u'Victor',
    u'Kitchens',
    u'4849 Emeral Dreams Drive',
    u'San Francisco',
    u'CA',
    u'94059',
    u'4156116022',
    u'2014-03-18 13:31:57.0',
    u'2014-03-18 13:31:57.0'],
   4)),
```

Step4c:



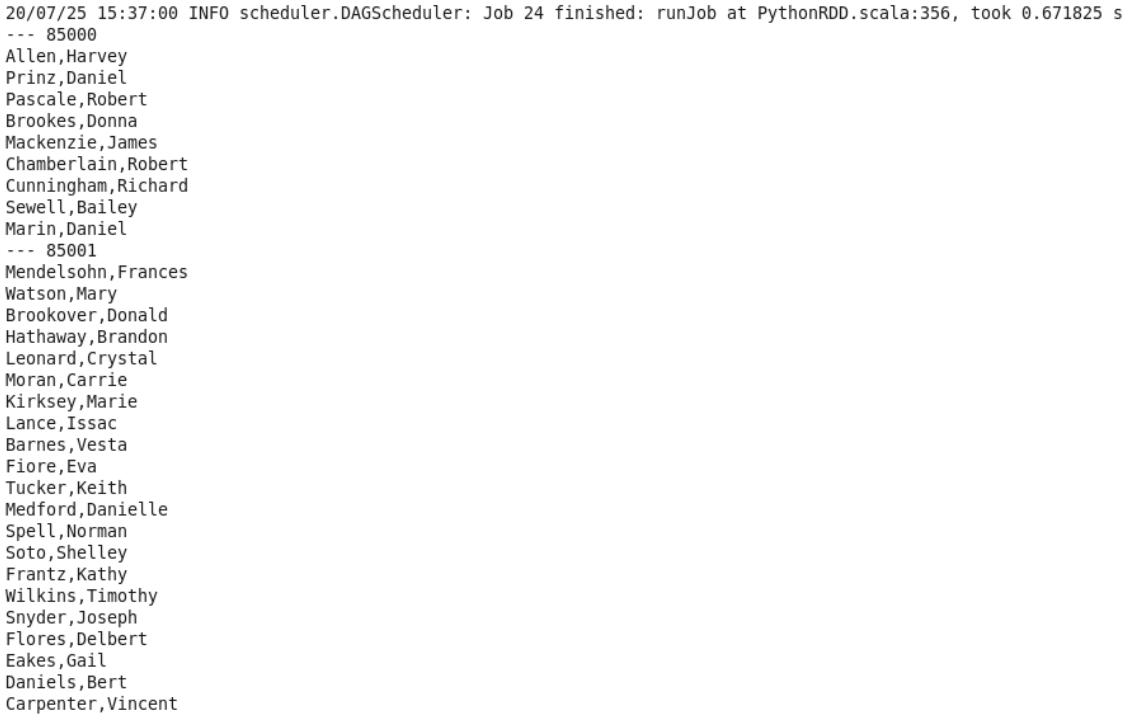
```
20/07/25 16:45:10 INFO scheduler.DAGScheduler: Job 7 finished: runJob at PythonRDD.scala:356, took 0.153161 s
51069 10 Gladys Lockhart
56581 2 Irene Sager
67248 4 Victor Kitchens
82995 2 Carlos Williams
81538 10 Donald Majors
3773 6 Ira Aubrey
586 4 Nicholas Campbell
40351 2 Sophia Mingle
43699 4 Dona Speight
104715 2 Cody Eaker
```

Challenge1:



```
20/07/25 15:32:08 INFO scheduler.DAGScheduler: Job 20 finished: runJob at Python
RDD.scala:356, took 0.036101 s
Out[41]:
[(u'94660',
  [u'1',
   u'2008-10-23 16:05:05.0',
   u'\\N',
   u'Donald',
   u'Becton',
   u'2275 Washburn Street',
   u'Oakland',
   u'CA',
   u'94660',
   u'5100032418',
   u'2014-03-18 13:29:47.0',
   u'2014-03-18 13:29:47.0']),
 (u'94171',
  [u'2',
   u'2008-11-12 03:00:01.0',
   u'\\N',
   u'Donna',
   u'Jones',
   u'3885 Elliott Street',
   u'San Francisco',
   u'CA',
   u'94171',
   u'4150835799',
   u'2014-03-18 13:29:47.0',
   u'2014-03-18 13:29:47.0']),
```

Challenge2 & 3:



```
20/07/25 15:37:00 INFO scheduler.DAGScheduler: Job 24 finished: runJob at PythonRDD.scala:356, took 0.671825 s
--- 85000
Allen,Harvey
Prinz,Daniel
Pascale,Robert
Brookes,Donna
Mackenzie,James
Chamberlain,Robert
Cunningham,Richard
Sewell,Bailey
Marin,Daniel
--- 85001
Mendelsohn,Frances
Watson,Mary
Brookover,Donald
Hathaway,Brandon
Leonard,Crystal
Moran,Carrie
Kirksey,Marie
Lance,Issac
Barnes,Vesta
Fiore,Eva
Tucker,Keith
Medford,Danielle
Spell,Norman
Soto,Shelley
Frantz,Kathy
Wilkins,Timothy
Snyder,Joseph
Flores,Delbert
Eakes,Gail
Daniels,Bert
Carpenter,Vincent
```