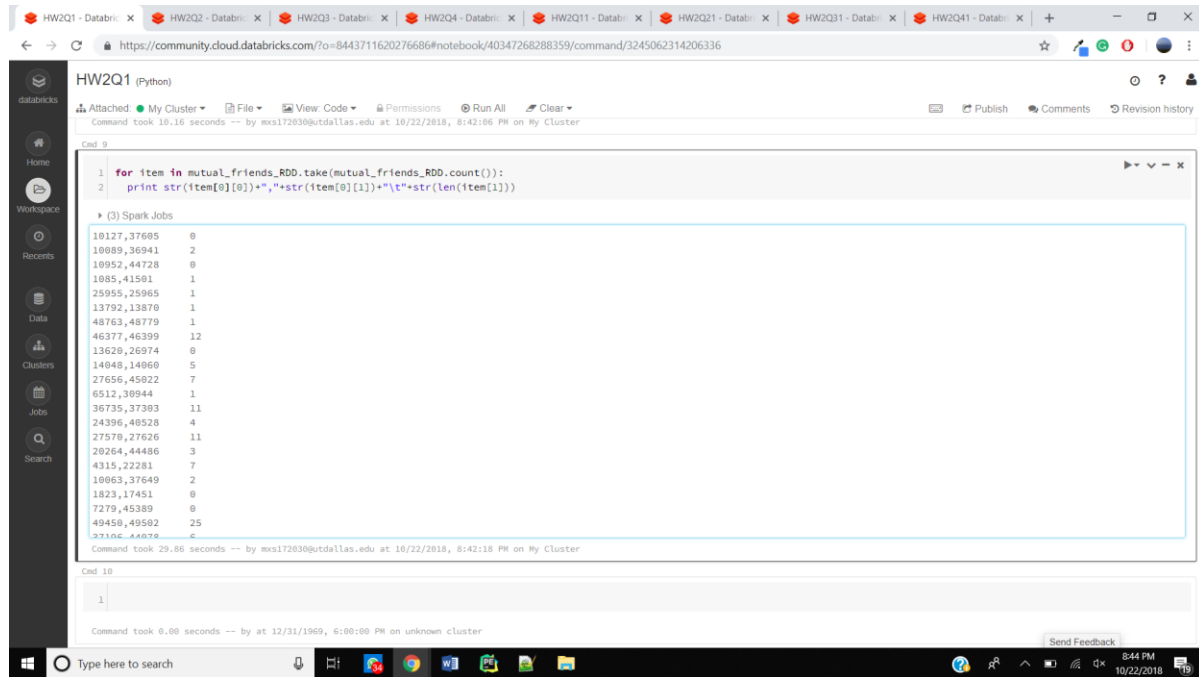


Homework 2
Name: Maitri Dharmendrakumar Shah
NetID: mxs172030

Q1:
Spark Code:
Command:
Python HW2Q1.py
Answer snapshot:



The screenshot shows a Databricks notebook interface with a sidebar on the left containing navigation icons for Home, Workspace, Recent, Data, Clusters, Jobs, and Search. The main area displays a Python notebook titled 'HW2Q1 (Python)'. The code in the notebook is as follows:

```
1 for item in mutual_friends_RDD.take(mutual_friends_RDD.count()):
2   print str(item[0][0])+", "+str(item[0][1])+"\t"+str(len(item[1]))
```

The output of the code is displayed below the code cell, showing 25 rows of data. Each row contains three fields: a pair of IDs separated by a comma, a tab character, and a count. The data is as follows:

mutual_friends_RDD.take(mutual_friends_RDD.count())	len(item[1])
10127,37695	0
10089,36941	2
10952,44728	0
1085,41501	1
25955,25965	1
13792,13870	1
48763,48779	1
46377,46399	12
13620,26974	0
14048,14060	5
27656,45022	7
6512,30944	1
36735,37303	11
24396,40528	4
27570,27626	11
29264,44486	3
4315,22201	7
10063,37649	2
1823,17451	0
7279,45389	0
49450,49502	25
37306,44679	6

The notebook interface also shows a command bar at the bottom with the text 'Command took 29.06 seconds -- by mxs172030@utdallas.edu at 10/22/2018, 8:42:18 PM on My Cluster'.

Spark SQL:
Command:
Python HW2Q11.py
Answer Snapshot:

The screenshot shows a Databricks community notebook titled 'HW2Q11 (Python)'. The command executed is `common_friend_pair.show()`. The output displays a list of friend pairs and their friend counts, truncated to the top 20 rows. The command took 19.02 seconds to execute.

```

common_friend_pair.show()

+-----+
| friend_pair | friends_list2 |
+-----+
| [18, 12562] | 0 |
| [79, 364] | 0 |
| [106, 15860] | 0 |
| [123, 2591] | 0 |
| [123, 12672] | 0 |
| [127, 26366] | 0 |
| [146, 25722] | 4 |
| [146, 25738] | 0 |
| [161, 15130] | 3 |
| [170, 17037] | 0 |
| [220, 2591] | 0 |
| [222, 21915] | 1 |
| [239, 43190] | 3 |
| [239, 43492] | 0 |
| [246, 249] | 0 |
| [311, 41391] | 0 |
| [336, 374] | 0 |
| [338, 7684] | 0 |
| [340, 30866] | 2 |
| [343, 36624] | 0 |
+-----+

only showing top 20 rows

Command took 19.02 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 8:55:34 PM on My Cluster

```

Q2:
Spark Code:
Command:
Python HW2Q2.py
Answer Snapshot:

The screenshot shows a Databricks community notebook titled 'HW2Q2 (Python)'. The command executed is `for item in final_output_RDD.take(final_output_RDD.count()):`. The output displays a list of user details, truncated to the top 20 rows. The command took 0.42 seconds to execute.

```

for item in final_output_RDD.take(final_output_RDD.count()):
    user1_details=item[1].split(":")
    user2_details=item[2].split(":")
    print
    str(item[0])+"\t"+str(user1_details[0])+"\t"+str(user1_details[1])+"\t"+str(user1_details[2])+"\t"+str(user2_details[0])+"\t"+str(user2_details[1])+"\t"+str(user2_details[2])

+-----+
| 99 | William Carey | 91 School Street,Beltsville,Washington DC,20705,US | Jane | Irish | 582 Dogwood Road,Phoenix,Arizona,85016,US | |
| 99 | Gregory Won | 1146 Meadow Drive,Missoula,Montana,59801,US | Gerald | Wisner | 4414 Cambridge Place,Bel Air,Maryland,21014,US |
| 99 | Caroline | Knight | 2486 School House Road,Jackson,Mississippi,39201,US | Martha | Stott | 468 Twin House Lane,Springfield,Massouri,65804,US |
| 99 | Arron Batz | 4301 Doe Meadow Drive,Baltimore,Maryland,21202,US | Roger | Salls | 1144 Kinney Street,Pittsfield,Massachusetts,1201,US |
| 99 | William Carey | 91 School Street,Beltsville,Washington DC,20705,US | Nola | Joyner | 1872 Lyndon Street,Nazareth,Pennsylvania,18064,US |
| 99 | Thomas Hook | 1153 Sycamore Circle,Fort Worth,Texas,76102,US | Amanda | Degroot | 1705 Upton Avenue,South China,Maine,4358,US |
| 99 | Gerald Wisner | 4414 Cambridge Place,Bel Air,Maryland,21014,US | Anthony | Manley | 258 Eastland Avenue,Jackson,Mississippi,39206,US |
| 99 | Gregory Won | 1146 Meadow Drive,Missoula,Montana,59801,US | Anthony | Manley | 258 Eastland Avenue,Jackson,Mississippi,39206,US |
| 99 | Arthur Stephenson | 3405 Jewell Road,Minneapolis,Minnesota,55406,US | Raymond | Norman | 2937 Black Stallion Road,Fort Thomas,Kentucky,41075,US |
| 99 | Paul Mackenzie | 4092 Marigold Lane,Coral Gables,Florida,33134,US | Paul | Nguyen | 471 Spring Street,Springfield,Illinois,62701,US |
+-----+

Command took 0.42 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 8:50:08 PM on My Cluster

```

Spark SQL:
Command:
Python HW2Q21.py
Answer Snapshot:

HW2Q21 (Python)

Attached My Cluster File View Code Permissions Run All Clear

(5) Spark Jobs

friends_number	user1_firstname	user1_lastname	user1_full_address	user2_firstname	user2_lastname	user2_full_address
99	Paul	Mackenzie	4092 Marigold Lane...	Paul	Nguyen	471 Spring Street...
99	Kenneth	Miller	2590 Westfall Ave...	Adrian	Wells	138 Kooter Lane, C...
99	Bambi	Villegas	1329 Grove Avenue...	Amanda	Turner	1164 Hewes Avenue...
99	Nola	Joyner	1872 Lyndon Stree...	Jane	Irish	582 Dogwood Road...
99	Arthur	Stephenson	3405 Jewell Road...	Raymond	Norman	2937 Black Stalli...
99	Alan	Hiltz	3441 Peck Street...	Titus	Beach	379 Weekley Stree...
99	William	Carey	91 School Street...	Nola	Joyner	1872 Lyndon Stree...
99	Charles	Davis	2359 West Fork Dr...	Anthony	Manley	258 Eastland Aven...
99	Gregory	Won	1146 Meadow Drive...	Anthony	Manley	258 Eastland Aven...
99	Gerald	Wisner	4414 Cambridge Pl...	Anthony	Manley	258 Eastland Aven...

Command took 23.73 seconds -- by mxs172030@utdallas.edu at 10/22/2018, 9:08:46 PM on My Cluster

Cell 29

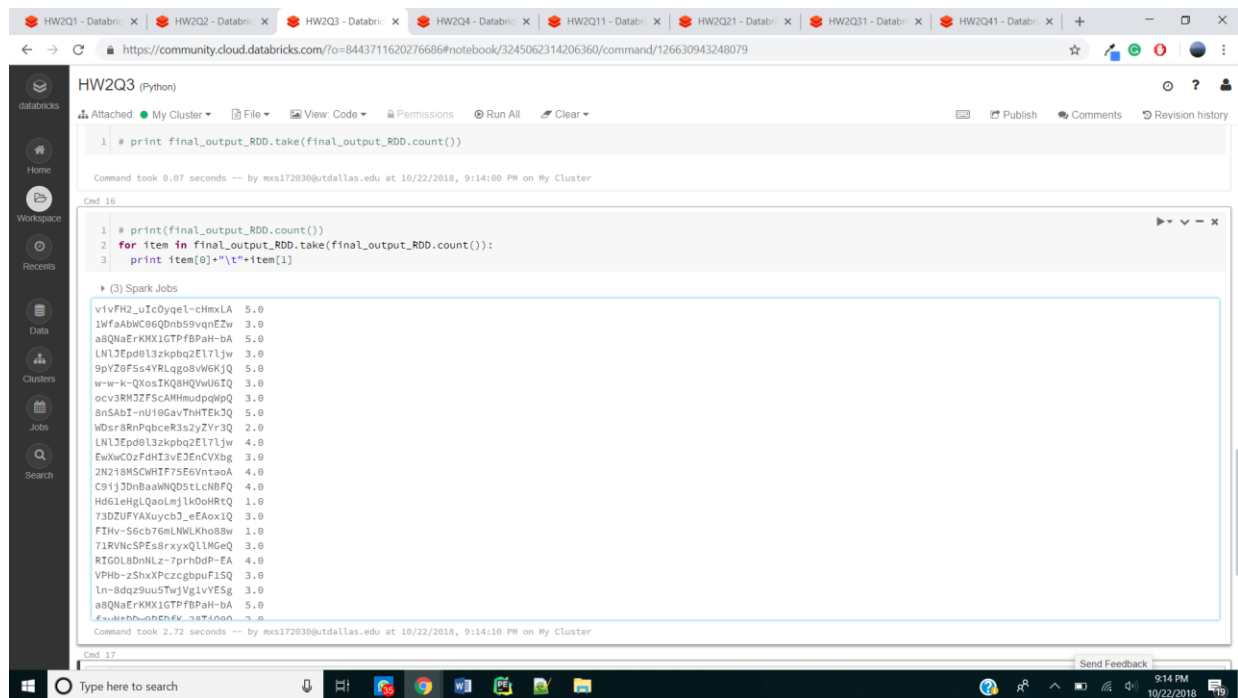
```
1 for item in second_user_info.take(second_user_info.count()):
2     print str(item[0])+"\t"+item[1]+"\t"+item[2]+"\t"+item[3]+"\t"+item[4]+"\t"+item[5]+"\t"+item[6]
```

(6) Spark Jobs

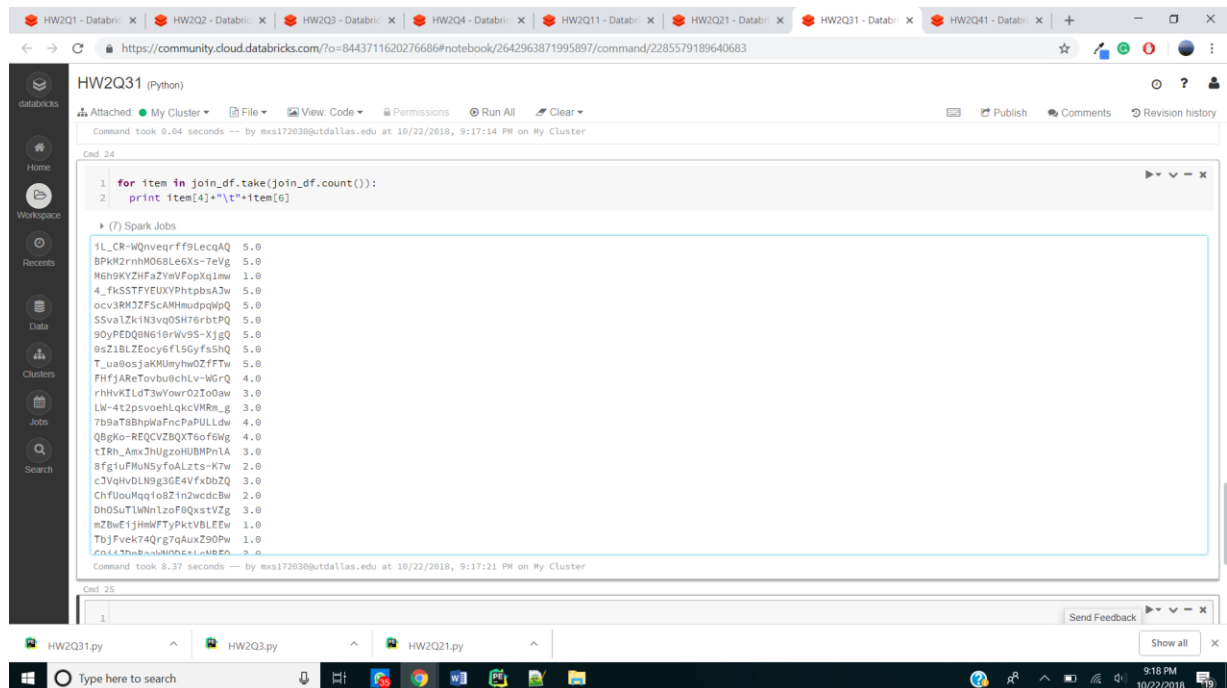
```
99 Paul Mackenzie 4092 Marigold Lane,Coral Gables,Florida,33134,US Paul Nguyen 471 Spring Street,Springfield,Illinois,62701,US
99 Kenneth Miller 2590 Westfall Avenue,Lafayette,Minnesota,56054,US Adrian Wells 138 Kooter Lane,Charlotte,North Carolina,28202,US
99 Bambi Villegas 1329 Grove Avenue,Goodwell,Oklahoma,73939,US Amanda Turner 1164 Hewes Avenue,Baltimore,Maryland,21202,US
99 Nola Joyner 1872 Lyndon Street,Nazareth,Pennsylvania,18064,US Jane Irish 582 Dogwood Road,Phoenix,Arizona,85016,US
99 Arthur Stephenson 3405 Jewell Road,Minneapolis,Minnesota,55406,US Raymond Norman 2937 Black Stallion Road,Fort Thomas,Kentucky,41075,US
99 Alan Hiltz 3441 Peck Street,Nashua,New Hampshire,3061,US Titus Beach 379 Weekley Street,San Antonio,Texas,78238,US
99 William Carey 91 School Street,Beltsville,Washington DC,20705,US Nola Joyner 1872 Lyndon Street,Nazareth,Pennsylvania,18064,US
99 Charles Davis 2359 West Fork Drive,Fort Lauderdale,Florida,33308,US Anthony Manley 258 Eastland Avenue,Jackson,Mississippi,39206,US
99 Gregory Won 1146 Meadow Drive,Missoula,Montana,59801,US Anthony Manley 258 Eastland Avenue,Jackson,Mississippi,39206,US
99 Gerald Wisner 4414 Cambridge Place,Bel Air,Maryland,21014,US Anthony Manley 258 Eastland Avenue,Jackson,Mississippi,39206,US
```

Command took 55.49 seconds -- by mxs172030@utdallas.edu at 10/22/2018, 9:08:21 PM on My Cluster

Q3:
Spark Code:
Command:
Python HW2Q3.py
Answer Snapshot:



Spark SQL:
Command:
Python HW2Q31.py
Answer Snapshot:



Q4:
Spark Code:
Command:
Python HW2Q4.py
Answer Snapshot:

The screenshot shows the Databricks community interface with a notebook titled "HW2Q4 (Python)". The notebook contains three code cells. The first cell prints the count of the review_detail_RDD. The second cell joins the review_detail_RDD with the business_detail_RDD into a final_output_RDD. The third cell prints the first 10 items of the final_output_RDD. The output of the third cell is a list of 10 items, each containing a business name and its count.

```
1 # print review_detail_RDD.take(review_detail_RDD.count())
```

Command took 0.04 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 9:23:03 PM on My Cluster

```
1 final_output_RDD=review_detail_RDD.join(business_detail_RDD)
```

Command took 0.12 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 9:23:04 PM on My Cluster

```
1 # print final_output_RDD.take(10)
```

Command took 0.04 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 9:23:07 PM on My Cluster

```
1 for item in final_output_RDD.take(final_output_RDD.count()):
2     print item[0]+"\t"+item[1][1].split(":")[0]+"\t"+item[1][1].split(":")[1]+"\t"+str(item[1][0])
```

Command took 1.41 seconds -- by mxi172030@utdallas.edu at 10/22/2018, 9:23:10 PM on My Cluster

Business Name	Count
List(Banks & Credit Unions, Financial Services)	5.0
List(Keys & Locksmiths, Home Services)	5.0
List(Professional Sports Teams, Arts & Entertainment)	5.0
List(Photographers, Event Planning & Services)	5.0
List(Barbers, Beauty and Spas)	5.0
List(Hair Salons, Beauty and Spas)	5.0
List(Shopping, Jewelry)	5.0
List(Bars, Sports Bars, Nightlife)	5.0
List(Doctors, Health and Medical, Internal Medicine, Pediatricians)	5.0
List(Food, Grocery)	5.0

Spark SQL:
Command:
Python HW2Q41.py
Answer Snapshot:

HW2Q1 - Databricks | HW2Q2 - Databricks | HW2Q3 - Databricks | HW2Q4 - Databricks | HW2Q11 - Databricks | HW2Q12 - Databricks | HW2Q13 - Databricks

→ → ↺ https://community.cloud.databricks.com/?o=8443711620276686#notebook/4498925691907815/command/307237484121725 ☆ ⚙ 📄 🔍 🌐

HW2Q41 (Python)

Attached • My Cluster View Code Permissions Run All Clear

Cnd 13

```
1 review_df_top=review_df_top.sort(desc("avg_rating")).limit(10)
```

▶ review_df_top pyspark.sql.dataframe.DataFrame = [business_id string avg_rating double]

Command took 0.07 seconds -- by mxsl17203@utdallas.edu at 10/22/2018, 9:24:38 PM on My Cluster

Cnd 14

```
1 join_df = review_df_top.join(business_df, business_df.b_id == review_df_top.business_id)
```

▶ join_df pyspark.sql.dataframe.DataFrame = [business_id string avg_rating double ... 3 more fields]

Command took 0.12 seconds -- by mxsl17203@utdallas.edu at 10/22/2018, 9:24:33 PM on My Cluster

Cnd 15

```
1 for item in join_df.take(join_df.count()):
2     print item[0]+"\\t"+item[3]+"\\t"+item[4]+"\\t"+str(item[1])
```

▶ (6) Spark Jobs

CJ1N7Zw_1B3ue4ebTm3dw	1176 E Colorado BlvdPasadenaPasadena, CA 91106	List(Investing, Financial Services, Shopping, Antiques)	5.0
VtL0X3FTT79T79UD3mfKw	3636 Nobel DrSte 100University CitySan Diego, CA 92122	List(Real Estate Services, Home Services, Real Estate)	5.0
AzqK7Dr-9zMoXbpsZNlkoA	1212 W Springfield AveUrbana, IL 61801	List(Middle Schools & High Schools, Education)	5.0
ZCELOqe_TRJ3TSIdIXalw	707 S 16th StLafayette, IN 47905	List(Automotive, Body Shops)	5.0
pBkkIzqn2jSfPG0-Jj1Wug	4150 Regents Park RowSte 230La Jolla, CA 92037	List(Dentists, Health and Medical)	5.0
VBBHdfPpzxKTGVAo8jBmw	1712 W Jefferson AveSouth Los AngelesLos Angeles, CA 90018	List(Elementary Schools, Education, Preschools)	5.0
SF3nqkSolZnmQWnf75vuw	4225 Executive SquareSte 600La Jolla, CA 92037	List(Professional Services, Lawyers)	5.0
wUVZsnLRWpQy73or3W	103 Carnegie CtrSte 300Princeton, NJ 08540	List(Shopping, Computers, Local Services, IT Services & Computer Repair, Professional Services, Web Design)	5.0
ANQK7eQLoPU2z_nG03pDnQ	731 Peachtree Street NEMidtownAtlanta, GA 30308	List(Religious Organizations)	5.0
oBsblz5sK7oxZeGFvF4w	200 Lothrop StOaklandPittsburgh, PA 15213	List(Doctors, Hospitals, Health and Medical, Medical Centers)	5.0

Command took 0.58 seconds -- by mxsl17203@utdallas.edu at 10/22/2018, 9:24:50 PM on My Cluster

Cnd 16

Send Feedback

HW2Q4.py

Show all