Εθνικό Μετσόβιο Πολυτεχνείο Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Προηγμένα Θέματα Βάσεων Δεδομένων Εξαμηνιαία Εργασία, Ακ. Έτος 2022-2023

> Γεώργιος Παπαδούλης: 03118003 Χριστίνα Προεστάκη: 03118877

Απαντήσεις στα ζητούμενα της εργασίας

GitHub Link: https://github.com/G-Papad/adv-db-2022.git

Ζητούμενο 1

Η απάντηση για το Ζητούμενο 1 βρίσκεται στο README στο github.

Συνοπτική παρουσίαση των χρόνων για τα ζητούμενα 2-4.

# of workers Query	1	2
Q1	19.61	20.019
Q2	55.442	57.458
Q3	20.62	22.173
Q3 - RDD	352.227	352.66
Q4	27.372	27.123
Q5	23.586	23.765

Ζητούμενο 2

Εκτέλεση Q1, Q2 χρησιμοποιώντας το DataFrame/SQL API.

1 worker (Q1+Q2)

Zone id Venc Zone id Venc amount improvement_su	dorID tpep_pickup_datetime t urcharge total_amount conges	pep_dropoff_datetime tion_surcharge airpor	passenger_count t_fee month	trip_distance R	atecodeID store	_and_fwd_flag PULo	cationID DOLo	cationID payme	nt_type fare	_amount extra m	ta_tax tip	p_amount tolls
Battery Park 12 0.0	2 2022-03-17 12:27:47 0.3 45.8	2022-03-17 12:27:58 2.5	1.0 1.0 0.0 3	0.0	1.0	N	12	12	1	2.5 0.0	0.5	40.0
	51026120185852			*		-						-

Q1 - 1 worker

	urcharge total_amount conges												
	1 2022-01-22 11:39:07	2022-01-22 12:31:09	1.0	33.4	1.0		70	265		88.0 0.0	0.5	0.0	
	0.3 282.1	0.0	0.0 1										
3 235.7	1 2022-03-11 20:08:32		1.0	0.0	1.0		265	265	1	2.5 1.0	0.5	48.0	
61800.091	0.3 288.0 1 2022-06-12 16:51:46	0.0	0.0 3 9.0	22.0	1.01	N	142	132	21	67.51 2.51	0.51	0.0	
0 000.09	0.3 870.89	2.5	0.01 61	22.0			142	132		07.3 2.3			
4 911.87		2022-04-29 04:32:30	2.0	0.0	1.0	N	249	249		3.0 3.0	0.5	0.0	
	0.3 918.67		0.0 4										
5 813.75	1 2022-05-21 16:47:48		1.0	2.4	3.0		239	246		31.5 0.0	0.0	0.0	
	0.3 845.55	0.0	0.0 5										
2 95.0	1 2022-02-18 02:33:30	2022-02-18 02:35:28	0.0 1.0 0.0 2	1.3	1.0		265	265	1	3.0 0.5	0.5	19.85	

Q2 - 1 worker

2 workers (Q1+Q2)

Zone id Vendo _amount improvement_sur	orID tpep_pickup_datetime t rcharge total_amount conges	tpep_dropoff_datetime pas stion_surcharge airport_f	++ ssenger_count trip fee month	_distance Rat	ecodeID store_ar	nd_fwd_flag PULoca	tionID DOLo	ationID payme	nt_type fare	_amount extra m	ta_tax tip	_amount tolls
Battery Park 12 0.0	2 2022-03-17 12:27:47 0.3 45.8	2022-03-17 12:27:58 2.5	1.0 9.0 3									40.0
												

Q1 - 2 workers

	urcharge total_amount conges													
1 193.3	1 2022-01-22 11:39:07		0.0l	1.0	33.4	1.0		70	265	88.0	0.0	0.5	0.0	19
3 235.7	0.3 282.1 1 2022-03-11 20:08:32	2022-03-11 20:09:45		1.0	0.0	1.0		265	265	2.5	1.0	0.5	48.0	
6 800.09	0.3 288.0 1 2022-06-12 16:51:46		0.0	9.0	22.0	1.0		142	132	67.5	2.5	0.5	0.0	800
09 4 911.87	0.3 870.89 1 2022-04-29 04:31:21		0.0	6 _2.0	0.0	1.0	N	249	249	3.0	3.0	0.5	0.0	911
87 5 813.75	0.3 918.67 1 2022-05-21 16:47:48	2.5 2022-05-21 17:05:47	0.0	4 1.0	2.4	3.0		239	246	31.5	0.0	0.0	0.0	813
75 2 95.0	0.3 845.55 1 2022-02-18 02:33:30	0.0 2022-02-18 02:35:28	0.0	5 1.0	1.3	1.0	N	265	265	3.0	0.5	0.5	19.85	
.0	0.3 119.15	0.0	0.0											

Ζητούμενο 3

Εκτέλεση Q3 χρησιμοποιώντας το DataFrame/SQL API και RDD.

1 worker (Q3 + Q3-rdd)

```
| dekapenthimero avg(Trip_distance) | avg(total_amount) | |
| 0 | 5.576429554927404 | 19.90405084638674 |
| 7 | 5.8003418315344 | 21.428117467105967 |
| 6 | 5.6793230779383554 | 21.515559094570637 |
| 9 | 7.906694182348759 | 22.771948777963715 |
| 5 | 5.556944935850653 | 21.120920554171548 |
| 1 | 5.097880367275346 | 19.14882164234129 |
| 10 | 6.315157336730224 | 22.46630530933543 |
| 3 | 5.849460516243601 | 20.18769180439039 |
| 8 | 6.249697852127271 | 21.92157034889687 |
| 11 | 6.174138574511356 | 22.331380641103525 |
| 2 | 6.248888338463885 | 19.491979067238955 |
| 4 | 6.480485434052913 | 20.652278174141436 |
| DF-query time: 20.62004065513611
```

Q3(DataFrame/SQL API) - 1 worker

```
('4', (648.0485434052913, 20.652278174141436))
('10', (631.5157336730224, 22.46630530933543))
('3', (584.9460516243602, 20.18769180439039))
('6', (567.9323077938355, 21.515559094570637))
('7', (580.03418315344, 21.428117467105967))
('0', (557.6429554927404, 19.90405084638674))
('1', (509.78803672753463, 19.14882164234129))
('9', (790.6694182348759, 22.771948777963715))
('8', (624.9697852127272, 21.92157034889687))
('2', (624.8888338463885, 19.491979067238955))
('5', (555.6944935850653, 21.120920554171548))
('11', (617.4138574511356, 22.331380641103525))
RDD-query time: 352.2271456718445
```

Q3(RDD) - 1 worker

2 workers (Q3 + Q3-rdd)

```
| dekapenthimero|avg(Trip_distance)| avg(total_amount)|
| 0| 5.576429554927404| 19.90405084638674|
| 7| 5.8003418315344| 21.428117467105967|
| 6| 5.6793230779383554| 21.515559094570637|
| 9| 7.906694182348759|22.771948777963715|
| 5| 5.556944935850653|21.120920554171548|
| 1| 5.097880367275346| 19.14882164234129|
| 10| 6.315157336730224| 22.46630530933543|
| 3| 5.849460516243601| 20.18769180439039|
| 8| 6.249697852127271| 21.92157034889687|
| 11| 6.174138574511356| 22.331380641103525|
| 2| 6.248888338463885| 19.491979067238955|
| 4| 6.480485434052913| 20.652278174141436|
| DF-query time: 22.173851013183594
```

```
O3(DataFrame/SQL API) - 2 workers
```

```
DF-query time: 22.173851013183594
('4', (648.0485434052913, 20.652278174141436))
('10', (631.5157336730224, 22.46630530933543))
('3', (584.9460516243602, 20.18769180439039))
('6', (567.9323077938355, 21.515559094570637))
('7', (580.03418315344, 21.428117467105967))
('0', (557.6429554927404, 19.90405084638674))
('1', (509.78803672753463, 19.14882164234129))
('9', (790.6694182348759, 22.771948777963715))
('8', (624.9697852127272, 21.92157034889687))
('2', (624.8888338463885, 19.491979067238955))
('5', (555.6944935850653, 21.120920554171548))
('11', (617.4138574511356, 22.331380641103525))
RDD-query time: 352.6603162288666
```

O3(RDD) - 2 workers

Ζητούμενο 4

Εκτέλεση Q4, Q5 χρησιμοποιώντας το DataFrame/SQL API.

1 worker (Q4+Q5)

++-	++	
days h	our result	
++-	++	
1	0 228580.0	
1	19 226543.0	
1 6 6 3 3	17 226426.0	
6	21 289408.0	
6	20 282941.0	
6	22 255878.0	
3	20 276200.0	
3	21 268951.0	
3	19 257625.0	
5 5 5 5 5 6 6 6 6 6	20 285365.0	
5	21 283074.0	
5	19 268112.0	
4	20 281426.0	
4	21 276147.0	
4	19 258958.0	
7	21 274010.0	
7	20 272951.0	
7 2	19 261720.0	
2	20 247418.0	
2	21 238259.0	
++-	+	
only sh	nowing top 20 rows	
time:	27.372729301452637	

Q4 - 1 worker

++	+	
month	days	result
++		
: :		0.4578674775487535
1		0.4393563580769872
1	1	0.2906301939811919
1	29	0.2405951845436878
1	16	0.23377299918217617
6	13	0.38451369937243063
6	25	0.32913073292653017
6	10	0.27397637812780157
6	16	0.2553497575787421
6	20	0.24242914593518236
3	18	0.29671341612657676
3	21	0.2757992602492
j 3	26	0.22708845953721593
j 3	5	0.22555461372495167
ĺзĺ	12	0.22100859110807622
j si		0.32402658973195914
j 5		0.2603403609036704
j 5		0.23659110789277535
5		0.220524452470084
5		0.2183200616188207
++		
only sh	owine	top 20 rows
3712) 3110	3112	, 100 20 1003
time:	23.58	86294174194336

Q5 - 1 worker

2 workers (Q4+Q5)

+		
ldavs	hour	result
+		++
1	0	228580.0
1		226543.0
1		226426.0
:		289408.0
j 6	20	282941.0
j 6	22	255878.0
jз	20	276200.0
j 3	21	268951.0
j 3	19	257625.0
5	20	285365.0
5	21	283074.0
5	19	268112.0
4	20	281426.0
4	21	276147.0
4	19	258958.0
7	21	274010.0
7	20	272951.0
7	19	261720.0
2	20	247418.0
2		238259.0
+		++
only s	showir	ng top 20 rows
time:	27.1	123685121536255

Q4 - 2 workers

++-	+	+
month d	ays	result
++-	+	+
1	9 0.4	4578674775487535
1	31 0.4	4393563580769872
1	1 0.3	2906301939811919
1	29 0.3	2405951845436878
1	16 0.2	3377299918217617
6	13 0.3	8451369937243063
6	25 0.3	2913073292653017
6	10 0.2	7397637812780157
6	16 0.	2553497575787421
6	20 0.2	4242914593518236
j 3j	18 0.29	9671341612657676
j 3j	21	0.2757992602492
j 3j	26 0.2	2708845953721593
j 3j	5 0.2	2555461372495167
ĺзĺ	12 0.2	2100859110807622
j 5 j	12 0.3	2402658973195914
j 5 j	20 0.	2603403609036704
5	16 0.2	3659110789277535
j si	15 0	. 220524452470084
j 5j		2183200616188207
++-	+	
only sho	wing to	o 20 rows
, , , , , , , , , , , , , , , , , , , ,		
time: 2	3.76535	5825424194

Q5 - 2 workers