ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Προχωρημένα Θέματα Βάσεων Δεδομένων Ακ. έτος 2022-23, 9ο Εξάμηνο

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Παυλανά Λουκία — 03118711

Χατζησάββας Ανδρέας — 03118701

Ομάδα 57

Github repo: https://github.com/AndreasHadjisavvas99/Advanced-Databases.git

Ερωτήματα:

2. Εκτελέστε τα Q1, Q2 χρησιμοποιώντας το DataFrame/SQL API. Θέλουμε τα αποτελέσματα και τους χρόνους εκτέλεσης του ερωτήματος με χρήση 1 και 2 workers (και όλες τις διαθέσιμες CPUs). Για να λάβετε σωστά τους χρόνους εκτέλεσης, φροντίστε να κάνετε collect το αποτέλεσμα του κάθε query (ή γράψιμο στο hdfs-δίσκο).

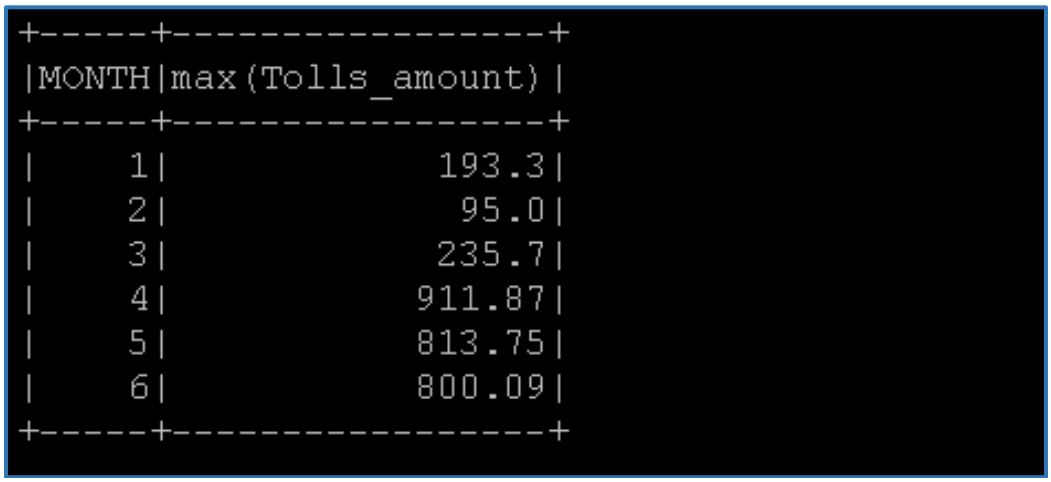
Για την εκτέλεση του παραπάνω ερωτήματος εκτελέσαμε το περιεχόμενο των προγραμμάτων q1-team-57.py και q2-team-57.py με 1 και 2 executors. Οι εντολές και τα αποτελέσματα φαίνονται παρακάτω:
spark-submit q1-team-57.py –deploy_mode cluster
spark-submit q2-team-57.py –deploy_mode cluster

Αποτελέσματα Q1:

```
+--------+--------------------+--------------------+---------------+-------------+----------+------------+------------+------------+-----------+-----+-------+----------+----------+---------------------+-----------+--------------------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_amount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee|
+--------+--------------------+--------------------+---------------+-------------+----------+------------+------------+------------+-----------+-----+-------+----------+----------+---------------------+-----------+--------------------+
|       2| 2022-03-17 12:27:47| 2022-03-17 12:27:58|            1.0|          0.0|       1.0|                N|         132|         132|          1|        2.5|  0.0|    0.5|       0.0|         0.0|                  0.3|       45.8|                 2.5|        0.0|
+--------+--------------------+--------------------+---------------+-------------+----------+------------+------------+------------+-----------+-----+-------+----------+----------+---------------------+-----------+--------------------+
```

Χρόνος Q1 με 1 worker: 33.6841 sec.
Χρόνος Q1 με 2 workers: 20.9451 sec.

Αποτελέσματα Q2:



```
+-----+-----------------+
|MONTH|max(Tolls_amount)|
+-----+-----------------+
|    1|            193.3|
|    2|             95.0|
|    3|            235.7|
|    4|           911.87|
|    5|           813.75|
|    6|           800.09|
+-----+-----------------+
```

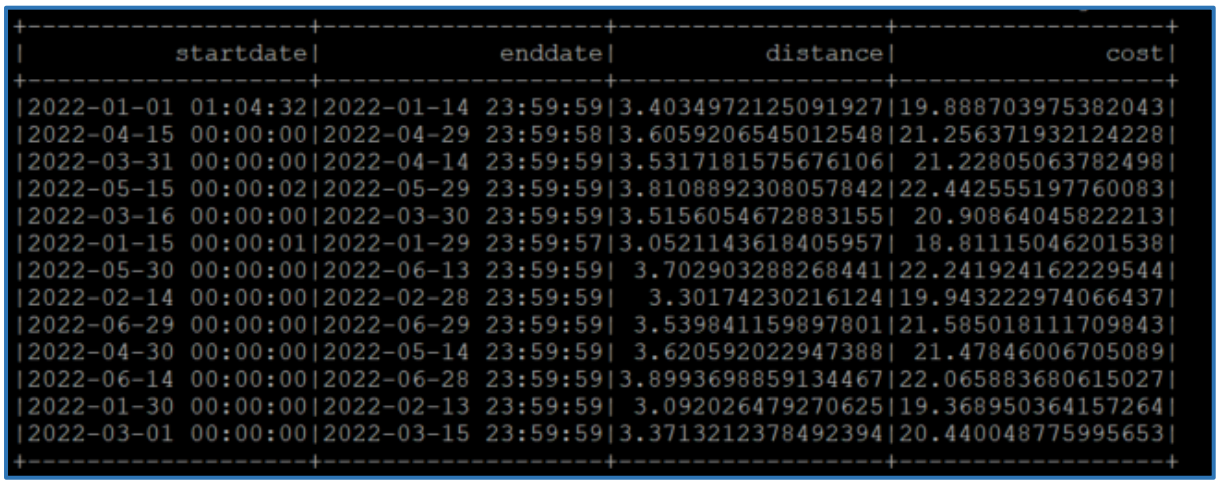
Χρόνος Q2 με 1 worker: 31.0249 sec.
Χρόνος Q2 με 2 workers: 19.5321 sec.

3. Εκτελέστε το Q3 χρησιμοποιώντας το DataFrame/SQL API και το RDD API. Θέλουμε τα αποτελέσματα και τους χρόνους εκτέλεσης του ερωτήματος με χρήση 1 και 2 workers.

Για την εκτέλεση του παραπάνω ερωτήματος εκτελέσαμε το περιεχόμενο των προγραμμάτων q3-team-57.py και q3-rdd-team57.py με 1 και 2 executors. Οι εντολές και τα αποτελέσματα φαίνονται παρακάτω:

spark-submit q3-team-57.py –deploy_mode cluster
spark-submit q3-rdd-team57.py –deploy_mode cluster

Αποτελέσματα Q3 με DataFrame:



```
+-------------------+-------------------+------------------+------------------+
|          startdate|            enddate|          distance|              cost|
+-------------------+-------------------+------------------+------------------+
|2022-01-01 01:04:32|2022-01-14 23:59:59|3.4034972125091927|19.888703975382043|
|2022-04-15 00:00:00|2022-04-29 23:59:58|3.6059206545012548|21.256371932124228|
|2022-03-31 00:00:00|2022-04-14 23:59:59|3.5317181575676106| 21.22805063782498|
|2022-05-15 00:00:02|2022-05-29 23:59:59|3.8108892308057842|22.442555197760083|
|2022-03-16 00:00:00|2022-03-30 23:59:59|3.5156054672883155| 20.90864045822213|
|2022-01-15 00:00:01|2022-01-29 23:59:57|3.0521143618405957| 18.81115046201538|
|2022-05-30 00:00:00|2022-06-13 23:59:59| 3.702903288268441|22.241924162229544|
|2022-02-14 00:00:00|2022-02-28 23:59:59|  3.30174230216124|19.943222974066437|
|2022-06-29 00:00:00|2022-06-29 23:59:59| 3.539841159897801|21.585018111709843|
|2022-04-30 00:00:00|2022-05-14 23:59:59| 3.620592022947388| 21.47846006705089|
|2022-06-14 00:00:00|2022-06-28 23:59:59|3.8993698859134467|22.065883680615027|
|2022-01-30 00:00:00|2022-02-13 23:59:59| 3.092026479270625|19.368950364157264|
|2022-03-01 00:00:00|2022-03-15 23:59:59|3.3713212378492394|20.440048775995653|
+-------------------+-------------------+------------------+------------------+
```

Χρόνος Q3 με 1 worker: 29.4612 sec.
Χρόνος Q3 με 2 workers: 17.5810 sec.

Αποτελέσματα Q3 με RDD:

```
(0, 3.3682419971726216, 19.788390697674604)
(1, 3.0606004056439473, 18.896049553565256)
(2, 3.074041257534215, 19.38250672787478)
(3, 3.304107410980879, 20.023875664365608)
(4, 3.3674038711623266, 20.502171614730525)
(5, 3.508210987800328, 20.950830366111834)
(6, 3.539083550567074, 21.274477188968717)
(7, 3.5860544966450116, 21.24403430304125)
(8, 3.657246868790049, 21.648890803824244)
(9, 3.809222735639307, 22.506562082036915)
(10, 3.715318110735683, 22.24297554458931)
(11, 3.81958054670484, 22.095514053946587)
```

Χρόνος Q2 με 1 worker: 239.2116 sec.
Χρόνος Q2 με 2 workers: 148.4248 sec.

4. Εκτελέστε τα Q4, Q5 χρησιμοποιώντας το DataFrame/SQL API. Θέλουμε τα αποτελέσματα και τους χρόνους εκτέλεσης του ερωτήματος με χρήση 1 και 2 workers. (20%+20%)

Για την εκτέλεση του παραπάνω ερωτήματος εκτελέσαμε το περιεχόμενο των προγραμμάτων q4-team-57.py και q5-team-57.py με 1 και 2 executors. Οι εντολές και τα αποτελέσματα φαίνονται παρακάτω:

spark-submit q4-team-57.py –deploy_mode cluster
spark-submit q5-team-57.py –deploy_mode cluster

Αποτελέσματα Q4:

```
+-----------+-----+----------+---+
|Day of week|Hours|passengers| rn|
+-----------+-----+----------+---+
|        Sun|    0|  228580.0|  1|
|        Sun|   19|  226543.0|  2|
|        Sun|   17|  226426.0|  3|
|        Mon|   20|  247418.0|  1|
|        Mon|   21|  238259.0|  2|
|        Mon|   19|  236534.0|  3|
|        Thu|   20|  275631.0|  1|
|        Thu|   21|  272210.0|  2|
|        Thu|   19|  258972.0|  3|
|        Sat|   21|  274010.0|  1|
|        Sat|   20|  272951.0|  2|
|        Sat|   19|  261720.0|  3|
|        Wed|   20|  281426.0|  1|
|        Wed|   21|  276147.0|  2|
|        Wed|   19|  258958.0|  3|
|        Fri|   21|  289408.0|  1|
|        Fri|   20|  282941.0|  2|
|        Fri|   22|  255878.0|  3|
|        Tue|   20|  276200.0|  1|
|        Tue|   21|  268951.0|  2|
+-----------+-----+----------+---+
only showing top 20 rows
```

Χρόνος Q4 με 1 worker: 42.1210 sec.
Χρόνος Q4 με 2 workers: 25.7447 sec.

Αποτελέσματα Q5:

```
+-----+---+--------------------------+---+
|Month|Day|avg tip percentage per day| rn|
+-----+---+--------------------------+---+
|    1|  9|         0.4633978928950156|  1|
|    1| 31|        0.45033640549183185|  2|
|    1|  1|        0.293686280101199 07|  3|
|    1| 29|        0.24405057978033298|  4|
|    1|  3|         0.2346723604537071|  5|
|    6| 13|         0.3905783100048706|  1|
|    6| 25|         0.3342221571012584|  2|
|    6| 10|        0.27711431354994775|  3|
|    6| 16|         0.2576058599413441|  4|
|    6| 20|        0.24386840861361336|  5|
|    3| 18|        0.30060154641527953|  1|
|    3| 21|        0.27816095192492624|  2|
|    3| 26|        0.22757314434025755|  3|
|    3|  5|        0.22593650242474292|  4|
|    3| 12|        0.22119149413891917|  5|
|    5| 12|        0.32831795026646854|  1|
|    5| 20|        0.26275972974605066|  2|
|    5| 16|        0.23785072190672488|  3|
|    5| 15|        0.22115316873811383|  4|
|    5|  6|        0.21923903947696885|  5|
+-----+---+--------------------------+---+
only showing top 20 rows
```

Χρόνος Q5 με 1 worker: 32.2719 sec
Χρόνος Q5 με 2 workers: 23.8169 sec.

Χρόνοι με ένα worker όπως αποθηκευτήκαν στο αρχείο times-sql_1worker.txt :

```
Time elapsed for q1 is 33.6841 sec.
Time elapsed for q2 is 31.0249 sec.
Time elapsed for q3 is 29.4612 sec.
Time elapsed for q4 is 42.1210 sec.
Time elapsed for q5 is 32.2719 sec.
Time elapsed for q3 with RDD is 239.2116 sec.
~
```

Χρόνοι με δύο workers όπως αποθηκευτήκαν στο αρχείο times-sql_2worker.txt:

```
Time elapsed for q1 is 20.9451 sec.
Time elapsed for q2 is 19.5321 sec.
Time elapsed for q3 is 17.5810 sec.
Time elapsed for q3 with RDD is 148.4248 sec.
Time elapsed for q4 is 25.7447 sec.
Time elapsed for q5 is 23.8169 sec.
~
```