Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών

Υπολογιστών

Προχωρημένα Θέματα Βάσεων Δεδομένων

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Δημήτρης Λυμπερόπουλος, 03118208

Email: el18208@mail.ntua.gr

Νικόλας Τασιόπουλος , 03118858

Email: el18858@mail.ntua.gr

Εγκατάσταση:

Σχετικά με την εγκατάσταση ΟΛΩΝ των απαραίτητων εργαλείων φτιάξαμε αναλυτικό οδηγό με reference από πηγές στο github μας εντός του φακέλου setup.

Γενικά σχόλια:

Για να υπολογίσουμε τον χρόνο εκτέλεσης των ερωτημάτων, τρέξαμε το κάθε ερώτημα 10 φορές και πήραμε τον μέσο όρο των χρόνων εκτέλεσης κάθε φοράς.

Γενικά, παρατηρήσαμε ότι με 2 workers τα ερωτήματα χρειάζονται αρκετά μικρότερο χρόνο εκτέλεσης σε σχέση με την περίπτωση που έχουμε 1 worker, κάτι το οποίο είναι λογικό. Ωστόσο, ο διπλασιασμός του αριθμού των workers (από 1 σε 2) δεν συνεπάγεται σε όλες τις περιπτώσεις και υποδιπλασιασμό του χρόνου εκτέλεσης. Επιπλέον, το ερώτημα Q3 με την χρήση του RDD API, ήταν το πιο αργό από όλα τα ερωτήματα, με τον χρόνο εκτέλεσής του, τόσο με έναν όσο και με δύο workers, να είναι σημαντικά μεγαλύτερος από τους χρόνους εκτέλεσης των υπόλοιπων ερωτημάτων (κατά 1-2 τάξεις μεγέθους).

Τέλος, να σημειώσουμε ότι στα δεδομένα parquet που πήραμε για όλα τα ερωτήματα κάναμε παραδοχή ότι είναι σωστά αν είναι εντός του 2022 (ακόμα και αν ο μήνας είναι μετά τον Ιούνιο).

Github repository: https://github.com/nikolastas/advancedDb

Έχουμε προσκαλέσει ως collaborator τον λογαριασμό dtsouma όπως αναφέρθηκε στο forum. Σε περίπτωση που αντιμετωπίζετε θέμα παρακαλώ στείλτε μας μήνυμα στα email που αναγράφονται στην πρώτη σελίδα.

Ερώτημα 1.

Να βρεθεί η διαδρομή με το μεγαλύτερο φιλοδώρημα (tip) τον Μάρτιο και σημείο άφιξης το "Battery Park".

Αποτελέσματα:

Όνομα στήλης	Τιμή
VendorID	2
tpep_pickup_datetime	datetime.datetime(2022, 3, 17, 12, 27, 47)
tpep_dropoff_datetime	datetime.datetime(2022, 3, 17, 12, 27, 58)
passenger_count	1.0
trip_distance	0.0
RatecodeID	1.0
store_and_fwd_flag	N
PULocationID	12

DOLocationID	12
payment_type	1
fare_amount	2.5
extra	0.0
mta_tax	0.5
tip_amount	40.0
tolls_amount	0.0
improvement_surcharge	0.3
total_amount	45.8
congestion_surcharge	2.5
airport_fee	0.0
LocationID	12
Borough	Manhattan
Zone	Battery Park
service_zone	Yellow Zone

Χρόνοι εκτέλεσης:

	2 Workers	1 Worker
Average execution time:	5 sec	32 sec

Ερώτημα 2.

Να βρεθεί, για κάθε μήνα, η διαδρομή με το υψηλότερο ποσό στα διόδια. Αγνοήστε μηδενικά ποσά.

Αποτελέσματα:

max_toll_fees_per_month	month	max_date
193.3	1	datetime.datetime(2022, 1, 31, 23, 59, 48)
800.09	6	datetime.datetime(2022, 6, 30, 23, 59, 52)
235.7	3	datetime.datetime(2022, 3, 31, 23, 59, 50)
813.75	5	datetime.datetime(2022, 5, 31, 23, 59, 55)

911.87	4	datetime.datetime(2022, 4, 30, 23, 59, 39)
29.55	7	datetime.datetime(2022, 7, 1, 2, 59, 49)
95.0	2	datetime.datetime(2022, 2, 28, 23, 59, 55)

Χρόνοι εκτέλεσης:

	2 Workers	1 Worker
Average execution time:	3.4 sec	10 sec

Ερώτημα 3.

Να βρεθεί, ανά 15 ημέρες, ο μέσος όρος της απόστασης και του κόστους για όλες τις διαδρομές με σημείο αναχώρησης διαφορετικό από το σημείο άφιξης.

Αποτελέσματα:

Στον παρακάτω πίνακα η στήλη period αντιστοιχεί στην ακόλουθη μορφή περιόδου:

$$2022 - Υ - XX \rightarrow \begin{cases} 2022 - Υ - 01, εαν\ XX \leq 15\ αρα\ 15\ πρώτες\ μερες \\ 2022 - Υ - 16\ εαν\ XX > 15\ , αρα\ 15\ τελευταίες\ μέρες \end{cases}$$

avg_trip_distance	avg_trip_cost	period
3.924531002558172	21.8054026069	2022-07-01 00:00:00
6.174138574511356	22.3313806411	2022-06-16 00:00:00
6.315157336730177	22.4663053093	2022-06-01 00:00:00
7.906694182348757	22.7719487780	2022-05-16 00:00:00
6.249697852127242	21.9215703489	2022-05-01 00:00:00
5.800344707645977	21.4280883762	2022-04-16 00:00:00
5.679323077938295	21.5155590946	2022-04-01 00:00:00
5.556944935850653	21.1209205542	2022-03-16 00:00:00
6.480485434052824	20.6522781742	2022-03-01 00:00:00
5.849460516243601	20.1876918043	2022-02-16 00:00:00
6.248888338463885	19.4919790672	2022-02-01 00:00:00

5.097880367275346	19.1488216423	2022-01-16 00:00:00
5.576410377852007	19.9037026378	2022-01-01 00:00:00

Χρόνοι εκτέλεσης:

	2 Workers	1 Worker
Average execution time SQL	15 sec	22 sec
Average execution time RDD (μόνο 3 φορές εκτέλεσης)	233 sec	420 sec

Ερώτημα 4.

Να βρεθούν οι τρεις μεγαλύτερες (top 3) ώρες αιχμής ανά ημέρα της εβδομάδος, εννοώντας τις ώρες (π.χ., 7-8πμ, 3-4μμ, κλπ) της ημέρας με τον μεγαλύτερο αριθμό επιβατών σε μια κούρσα ταξί. Ο υπολογισμός αφορά όλους τους μήνες.

Αποτελέσματα:

Στον παρακάτω πίνακα έχουμε η στήλη hour_of_day αναφέρετε στην περίοδο που προκύπτει :

$$hour\ of\ day : X \rightarrow [X, X+1]$$

day_of_week	hour_of_day	num_clients
1	0	1.5299456507188562
1	1	1.527838567375201
1	2	1.5080726185191242
2	0	1.4679887711672552
2	1	1.4442867916810471
2	2	1.4231993989051486
3	0	1.4200313882151518
3	1	1.4175124740006593
3	2	1.4104520814693964
4	1	1.4088480212656305
4	0	1.4012291857176276
4	2	1.4011489645958584

5	23	1.4053823152498932
5	1	1.402590728520038
5	0	1.4010382527988254
6	23	1.475576918073731
6	22	1.444813976205668
6	2	1.4236394379113688
7	23	1.522606766277207
7	22	1.5068176194011382
7	0	1.4993154284898547

Χρόνοι εκτέλεσης:

	2 Workers	1 Worker
Average execution time:	11 sec	16 sec

Ερώτημα 5.

Να βρεθούν οι κορυφαίες πέντε (top 5) ημέρες ανά μήνα στις οποίες οι κούρσες είχαν το μεγαλύτερο ποσοστό σε tip. Για παράδειγμα, εάν η κούρσα κόστισε 10$ (fare_amount) και το tip ήταν 5$, το ποσοστό είναι 50%.

Αποτελέσματα:

trip_month	**trip_day**	**tip_percentage**
1	9	45.78674775487207
1	31	43.93563580770273
1	1	29.07803686136836
1	29	24.059518454370057
1	16	23.377299918220096
2	21	25.981657452766274

2	13	24.572068389402546
2	9	23.904535643412483
2	10	23.33961589934868
2	27	23.3006799515465
3	18	29.671341612659685
3	21	27.57992602492248
3	26	22.70884595372165
3	5	22.55546137249565
3	12	22.100859110808635
4	12	48.36884410450339
4	2	31.175092883998968
4	21	30.44861250236277
4	3	24.46372770475391
4	30	21.99676965994668
5	12	32.402658973198044
5	20	26.034036090366385
5	16	23.659110789279985
5	15	22.05244524700949
5	6	21.832006161884486
6	13	38.45136993724611
6	25	32.91307329265353
6	10	27.397637812780694
6	16	25.534975757875227
6	20	24.242914593519107
7	1	21.26102932306057

Χρόνοι εκτέλεσης:

	2 Workers	1 Worker
Average execution time:	16 sec	20 sec