Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Βάσεων Γνώσεων και Δεδομένων

Μάθημα: Προχωρημένα Θέματα Βάσεων Δεδομένων

Ακαδημαϊκό Έτος 2022-2023

Διδάσκοντες: Β. Καντερέ, Δ. Τσουμάκος

Εξαμηνιαία Εργασία

GitHub Link: https://github.com/ElliT42/Advanced_db_project

Ομάδα 11

Ανδρώνη Άρτεμις	**03118117**
Τσέρτου Ελένη	**03118165**

Ζητούμενο 1ο:

Αρχικά, εγκαθιστούμε το κατανεμημένο storing και processing framework **Apache Hadoop** έκδοσης 3.3.4 προκειμένου να χρησιμοποιήσουμε το **HDFS** από το οποίο διαβάζουν και γράφουν τα Spark Jobs**.** Στη συνέχεια, εγκαθιστούμε το open source framework **Apache Spark** έκδοσης 3.3.1 για την επεξεργασία των δεδομένων.

Δημιουργούμε δύο DataFrames και δύο RDDs από τα δεδομένα μας, ένα για τα taxi trips και ένα για τα zone lookups.

Ζητούμενο 2ο:

Q1: Για το πρώτο ερώτημα θέλουμε να βρούμε τη διαδρομή με το μεγαλύτερο φιλοδώρημα (tip) τον Μάρτιο και σημείο άφιξης το "Battery Park".

Οι χρόνοι εκτέλεσης του Q1 είναι οι εξής:

Number of Workers	Runtime (sec)
1	14.636770248413086
2	14.584970951080322

Τα αποτελέσματα του Q1 φαίνονται στον παρακάτω πίνακα:

Attribute	Value (trip_distance = 0.0)	Value (trip_distance ≠ 0.0)
VendorID	2	1
tpep_pickup_datetime	17/3/2022 12:27:47	19/3/2022 08:37:46
tpep_dropoff_datetime	17/3/2022 12:27:58	19/3/2022 08:52:34
passenger_count	1.0	2.0
trip_distance	0.0	6.6
RatecodeID	1.0	1.0
store_and_fwd_flag	N	N
PULocationID	12	230
DOLocationID	12	12
payment_type	1	1
fare_amount	2.5	20.5
extra	0.0	2.5
mta_tax	0.5	0.5
tip_amount	40.0	15.0
tolls_amount	0.0	0.0
improvement_surcharge	0.3	0.3
total_amount	45.8	38.8
congestion_surcharge	2.5	2.5
airport_fee	0.0	0.0

Σημείωση: Το αποτέλεσμα που λάβαμε για το μεγαλύτερο φιλοδώρημα αφορούσε μηδενική απόσταση, επομένως αποφασίσαμε να παρουσιάσουμε και το αποτέλεσμα του αντίστοιχου query για μη μηδενικές αποστάσεις.

Q2: Για το δεύτερο ερώτημα θέλουμε να βρούμε, για κάθε μήνα, τη διαδρομή με το υψηλότερο ποσό στα διόδια, αγνοώντας τα μηδενικά ποσά.

Οι χρόνοι εκτέλεσης του Q2 είναι οι εξής:

Number of Workers	Runtime (sec)
1	36.84924030303955
2	34.405351400375366

Τα αποτελέσματα του Q2 φαίνονται στον παρακάτω πίνακα:

Attribute	January	February	March	April	May	June
VendorID	1	1	1	1	1	1
tpep_pickup_datetime	22/1/2022 11:39:07	18/2/2022 02:33:30	11/3/2022 20:08:32	29/4/2022 04:31:21	21/5/2022 16:47:48	12/6/2022 16:51:46
tpep_dropoff_datetime	22/1/2022 12:31:09	18/2/2022 02:35:28	11/3/2022 20:09:45	29/4/2022 04:32:30	21/5/2022 17:05:47	12/6/2022 17:56:48
passenger_count	1.0	1.0	1.0	2.0	1.0	9.0
trip_distance	33.4	1.3	0.0	0.0	2.4	22.0
RatecodeID	1.0	1.0	1.0	1.0	3.0	1.0
store_and_fwd_flag	Y	N	N	N	N	N
PULocationID	70	265	265	249	239	142
DOLocationID	265	265	265	249	246	132
payment_type	4	1	1	3	3	2
fare_amount	88.0	3.0	2.5	3.0	31.5	67.5
extra	0.0	0.5	1.0	3.0	0.0	2.5
mta_tax	0.5	0.5	0.5	0.5	0.0	0.5
tip_amount	0.0	19.85	48.0	0.0	0.0	0.0
tolls_amount	193.3	95.0	235.7	911.87	813.75	800.09
improvement_surcharge	0.3	0.3	0.3	0.3	0.3	0.3
total_amount	282.1	119.15	288.0	918.67	845.55	870.89
congestion_surcharge	0.0	0.0	0.0	2.5	0.0	2.5
airport_fee	0.0	0.0	0.0	0.0	0.0	0.0

Ζητούμενο 3ο:

Q3: Για το τρίτο ερώτημα θέλουμε να βρούμε, ανά 15 ημέρες, τον μέσο όρο της απόστασης και του κόστους για όλες τις διαδρομές με σημείο αναχώρησης διαφορετικό από το σημείο άφιξης.

Οι χρόνοι εκτέλεσης του Q3, χρησιμοποιώντας **DataFrame API**, είναι οι εξής:

Number of Workers	Runtime (sec)
1	19.362547874450684
2	17.153616189956665

Οι χρόνοι εκτέλεσης του Q3, χρησιμοποιώντας **RDD API**, είναι οι εξής:

Number of Workers	Runtime (sec)
1	277.3749454021454
2	266.09274101257324

Τα αποτελέσματα του Q3 φαίνονται στον παρακάτω πίνακα:

Month/Halves	Average Distance	Average Cost
January/1st	5.58	19.90
January/2nd	5.10	19.15
February/1st	6.25	19.49
February/2nd	5.85	20.19
March/1st	6.48	20.65
March/2nd	5.56	21.12
April/1st	5.68	21.51
April/2nd	5.80	21.43
May/1st	6.25	21.92
May/2nd	7.91	22.77
June/1st	6.32	22.47
June/2nd	6.17	22.33

Ζητούμενο 4ο:

Q4: Για το τέταρτο ερώτημα θέλουμε να βρούμε τις 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.

Οι χρόνοι εκτέλεσης του Q4 είναι οι εξής:

Number of Workers	Runtime (sec)
1	14.921222448348999
2	15.300925731658936

Τα αποτελέσματα του Q4 φαίνονται στον παρακάτω πίνακα:

Day of the Week	Time	Total Number of Passengers
1	00:00	228580
1	19:00	226543
1	17:00	226426
2	20:00	247418
2	21:00	238259
2	19:00	23653
3	20:00	276200
3	21:00	268951
3	19:00	257625
4	20:00	281426
4	21:00	276147
4	19:00	258958
5	20:00	285365
5	21:00	283074
5	19:00	268112
6	21:00	289408
6	20:00	282941
6	22:00	255878
7	21:00	274010
7	20:00	272951
7	19:00	261720

Q5: Για το πέμπτο ερώτημα θέλουμε να βρούμε τις κορυφαίες πέντε (top 5) ημέρες ανά μήνα στις οποίες οι κούρσες είχαν το μεγαλύτερο ποσοστό σε tip.

Οι χρόνοι εκτέλεσης του Q5 είναι οι εξής:

Number of Workers	Runtime (sec)
1	14.921222448348999
2	15.300925731658936

Τα αποτελέσματα του Q5 φαίνονται στον παρακάτω πίνακα:

Month	Day	Tips (%)
January	29	21.548
January	15	19.532
January	22	19.337
January	30	19.281
January	1	19.277
February	4	19.557
February	5	19.534
February	6	19.401
February	10	19.355
February	17	19.291
March	9	19.556
March	12	19.392
March	30	19.329
March	24	19.278
March	10	19.274
April	1	19.138
April	7	19.125
April	6	19.091
April	27	19.032
April	28	18.937
May	12	19.214
May	4	19.139
May	11	19.029
May	10	18.972
May	6	18.968
June	16	19.044
June	8	18.967
June	23	18.922
June	9	18.911
June	17	18.830