Εθνικό Μετσόβειο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Προχωρημένα Θέματα Βάσεων Δεδομένων – 9ο Εξάμηνο
Εξαμηνιαία Εργασία – Ακ. Έτος 2022-2023

Ομάδα 6
Κυριακόπουλος Γεώργιος – 03118153
Τζελέπης Σεραφείμ – 03118849
GitHub Repository

Ερώτημα 1

Αρχικά, ακολουθήσαμε τις οδηγίες για τη δημιουργία δύο μηχανημάτων στο ~okeanos και για το στήσιμο του τοπικού δικτύου και της σύνδεσης με τον εξωτερικό κόσμο.

Έπειτα, προχωρήσαμε με βάση έναν παλιότερο οδηγό, όπου κάναμε τις απαραίτητες αλλαγές όπου χρειαζόταν, είτε λόγω νέων εκδόσεων είτε λόγω κάποιου deprecation. Με βάση τα αντίστοιχα scripts, ρυθμίσαμε το passwordless ssh μεταξύ master και slave, δημιουργήσαμε το NAT των δύο υπολογιστών, εγκαταστήσαμε τη Java, εγκαταστήσαμε το Hadoop και ρυθμίσαμε τα κατάλληλα environmental variables και όποια config files χρειαζόταν, ενώ, στη συνέχεια, εγκαταστήσαμε το Spark, με αντίστοιχη ρύθμιση σε όσα env variables/configs χρειάζεται.

Έχοντας ένα λειτουργικό multi-node cluster πλέον, ξεκινήσαμε το Hadoop File System και το Apache Spark με τους αντίστοιχους workers (αρχικά 2, 1 στο master και 1 στο slave, με όλους τους διαθέσιμους CPU cores και όλη τη διαθέσιμη RAM). Τέλος, περάσαμε τα data, σε μορφή parquet, στο hdfs ώστε να τα έχουμε διαθέσιμα στο cluster.

Ερώτημα 2

Q1. Να βρεθεί η διαδρομή με το μεγαλύτερο φιλοδώρημα (tip) τον Μάρτιο και σημείο άφιξης το "Battery Park".

Time elapsed for Query 1 with 1 worker: 11.391889333724976 sec.
Time elapsed for Query 1 with 2 workers: 10.844828844070435 sec.

Το αποτέλεσμα του ερωτήματος είναι το ίδιο τόσο για 1 όσο και για 2 workers και είναι το εξής:

VendorID	2
tpep_pickup_datetime	44637.5192939815
tpep_dropoff_datetime	44637.5194212963
passenger_count	1
trip_distance	0
RatecodeID	1
store_and_fwd_flag	N
PULocationID	12
DOLocationID	12
payment_type	1
fare_amount	2.5
extra	0
mta_tax	0.5
tip_amount	40
tolls_amount	0
improvement_surcharge	0.3
total_amount	45.8
congestion_surcharge	2.5
airport_fee	0

Q2. Να βρεθεί, για κάθε μήνα, η διαδρομή με το υψηλότερο ποσό στα διόδια. Αγνοήστε μηδενικά ποσά.

Time elapsed for Query 2 with 1 worker: 27.938428163528442 sec.
Time elapsed for Query 2 with 2 workers: 23.334222555160522 sec.

Το αποτέλεσμα του ερωτήματος είναι το ίδιο τόσο για 1 όσο και για 2 workers και είναι το εξής:

	January	February	March	April	May	June
VendorID	1	1	1	1	1	1
tpep_pickup_datetime	2022-01-22 11:39:07	2022-02-18 02:33:30	2022-03-11 20:08:32	2022-04-29 04:31:21	2022-05-21 16:47:48	2022-06-12 16:51:46
tpep_dropoff_datetime	2022-01-22 12:31:09	2022-02-18 02:35:28	2022-03-11 20:09:45	2022-04-29 04:32:30	2022-05-21 17:05:47	2022-06-12 17:56:48
passenger_count	1	1	1	2	1	9
trip_distance	33.4	1.3	0	0	2.4	22
RatecodeID	1	1	1	1	3	1
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	3	2.5	3	31.5	67.5
extra	0	0.5	1	3	0	2.5
mta_tax	0.5	0.5	0.5	0.5	0	0.5
tip_amount	0	19.85	48	0	0	0
tolls_amount	193.3	95	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	918.67	845.55	870.89
congestion_surcharge	0	0	0	2.5	0	2.5
airport_fee	0	0	0	0	0	0

Ερώτημα 3

Q3. 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.

Για την εκδοχή με το DataFrame/SQL API:

Time elapsed for Query 3 with 1 worker: 13.266789674758911 sec.
Time elapsed for Query 3 with 2 workers: 13.361640930175781 sec.

Το αποτέλεσμα του ερωτήματος είναι το ίδιο τόσο για 1 όσο και για 2 workers και είναι το εξής:

15-day group	Average Trip Distance	Average Total Amount
0	5.576410377852007	19.903702637879007
1	4.804840472309411	19.03660791389491
2	5.950485844928121	19.553891327960553
3	6.1857672125677	20.17207809365826
4	6.606992664131458	20.692371844024798
5	5.533001951325831	21.121666826650515
6	5.679323077938295	21.515559094583587
7	5.800344707645977	21.428088376232783
8	6.249697852127242	21.921570348909114
9	7.9990632224691165	22.806499070460386
10	6.378971191608972	22.452110839872283
11	6.153370128239474	22.352167683521646
12	5.811220970695942	22.16938397436561

Για την εκδοχή με το RDD API:

Ερώτημα 4

Q4. 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.

Time elapsed for Query 4 with 1 worker: 10.996814012527466 sec.
Time elapsed for Query 4 with 2 workers: 13.40391206741333 sec.

Το αποτέλεσμα του ερωτήματος είναι το ίδιο τόσο για 1 όσο και για 2 workers και είναι το εξής:

Day of the Week	Hour of the Day	Sum of Passengers Count
1	0	228580.0
1	19	226543.0
1	17	226426.0
2	20	247418.0
2	21	238259.0
2	19	236534.0
3	20	276200.0
3	21	268951.0
3	19	257625.0
4	20	281426.0
4	21	276147.0
4	19	258958.0
5	20	285365.0
5	21	283074.0
5	19	268112.0
6	21	289408.0
6	20	282941.0
6	22	255878.0
7	21	274010.0
7	20	272951.0
7	19	261720.0

Σημείωση:

- το **Day of the Week** ξεκινάει από την ημέρα Κυριακή για τιμή ίση με 1 και τελειώνει στην ημέρα Σάββατο για τιμή ίση με 7
- το **Hour of the Day** ορίζει την ώρα εκκίνησης της ώρας αιχμής, δηλαδή το 0 σημαίνει 00.00-01.00 και το 19 σημαίνει 19.00-20.00

Q5. 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%.

Time elapsed for Query 5 with 1 worker: 10.121625661849976 sec.
Time elapsed for Query 5 with 2 workers: 12.232147693634033 sec.

Το αποτέλεσμα του ερωτήματος είναι το ίδιο τόσο για 1 όσο και για 2 workers και είναι το εξής:

Month	Day of the Month	Average 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

Σημείωση:

- το **Month** ξεκινάει από το μήνα Ιανουάριο για τιμή ίση με 1 και τελειώνει στο μήνα Ιούνιο για τιμή ίση με 6
- το **Day of the Month** ξεκινάει από την πρώτη μέρα ενός μήνα για τιμή ίση με 1
- το **Average Tip Percentage** ορίζεται ως ο μέσος όρος των ποσοστών % για κάθε ημέρα του μήνα, δηλαδή η τιμή 45.78674775487207 σημαίνει περίπου 45.79% average of tip amount to fare amount percent ratio

Συμπεράσματα

Σχετικά με τους χρόνους εκτέλεσης των ερωτημάτων έχουμε τις παρακάτω παρατηρήσεις:

- Για το Q1 έχουμε 5% ταχύτερη εκτέλεση με 2 workers
- Για το Q2 έχουμε 16% ταχύτερη εκτέλεση με 2 workers
- Για το Q3 έχουμε 1% ταχύτερη εκτέλεση με 1 worker
- Για το Q4 έχουμε 18% ταχύτερη εκτέλεση με 1 worker
- Για το Q5 έχουμε 17% ταχύτερη εκτέλεση με 1 worker

Επομένως, γενικά παρατηρούμε ότι υπάρχουν ερωτήματα τα οποία παρουσίαζουν ταχύτερη εκτέλεση με τη χρήση 2 workers, όπου μοιράζονται το φορτίο υπολογισμού των απαντήσεων, όμως υπάρχουν και ερωτήματα όπου η χρήση ενός μόνο worker είναι αρκετά ταχύτερη. Αυτό ίσως οφείλεται σε πιθανά overhead καθυστέρησης που περιλαμβάνει η χρήση 2 workers, καθώς και το σχετικά μικρό dataset που χρησιμοποιείται στην εργασία, κάτι το οποίο δεν επιτρέπει στην εκτέλεση με 2 workers να επιδείξει την ταχύτητα της.