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
Διδάσκοντες: Β. Καντερέ, Δ. Τσουμάκος

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Περιγραφή:

Στην παρούσα εργασία καλείστε να επεξεργαστείτε δεδομένα όγκου με χρήση της πλατφόρμας Apache Spark. Τα δεδομένα αφορούν καταγραφές διαδρομών ταξί στην πόλη της Νέας Υόρκης. Οι καταγραφές των ταξιδιών περιλαμβάνουν πεδία που αφορούν τις ημερομηνίες/ώρες παραλαβής και αποβίβασης, τοποθεσίες παραλαβής και αποβίβασης, αποστάσεις ταξιδιού, αναλυτικούς ναύλους, τύπους τιμών, είδη πληρωμών και αριθμό επιβατών που αναφέρονται από τον οδηγό. Τα δεδομένα που χρησιμοποιήθηκαν στα συνημμένα σύνολα δεδομένων συλλέχθηκαν και παρασχέθηκαν στην Επιτροπή Ταξί και Λιμουζίνας της Νέας Υόρκης (TLC) από παρόχους τεχνολογίας. Τα δεδομένα (ανά μήνα) καθώς και περιγραφή τους είναι ανοικτά και βρίσκονται στη σελίδα: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Οι ομάδες καλούνται να χρησιμοποιήσουν έκδοση Spark 3.1 (και άνω) με HDFS ως σύστημα αποθήκευσης αρχείων. Συγκεκριμένα, θα χρησιμοποιήσετε (κυρίως και βασικά) τα DataFrame/SQL API Spark (https://spark.apache.org/docs/latest/sql-programming-guide.html#dataframes) αλλά και το RDD API (https://spark.apache.org/docs/latest/rdd-programming-guide.html).

Τα δεδομένα Yellow Taxi Trip Records που απαιτούνται για την εργασία σας αφορούν τους μήνες Ιανουάριο έως και Ιούνιο 2022 (σε συμπιεσμένο parquet format). Η ακριβής περιγραφή των πεδίων των δεδομένων βρίσκεται εδώ:
https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Επιπλέον, θα σας χρειαστεί και το αρχείο:
https://d37ci6vzurychx.cloudfront.net/misc/taxi+_zone_lookup.csv, που συνδέει το πεδίο LocationID με λεκτική αναπαράσταση της πληροφορίας σε προάστια και ζώνες της πόλης.

Ερωτήματα:

Ερώτημα	Λεκτική Περιγραφή
Q1	Να βρεθεί η διαδρομή με το μεγαλύτερο φιλοδώρημα (tip) τον Μάρτιο και σημείο άφιξης το "Battery Park".
Q2	Να βρεθεί, για κάθε μήνα, η διαδρομή με το υψηλότερο ποσό στα διόδια. Αγνοήστε μηδενικά ποσά.
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.
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.
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%.

Αναλυτικά, τα ζητούμενα της εργασίας και η αντίστοιχη βαθμολόγησή τους είναι η ακόλουθη:

1. Εγκαταστήστε την πλατφόρμα εκτέλεσης Spark & HDFS (5%) και δημιουργήστε δύο RDD (2.5%) και δύο DataFrames (2.5%) από τα αρχικά δεδομένα (taxi trips & zone lookups).
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) (10%+10%).
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. (30%).
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%)

Παραδοτέα:

- Η εργασία να εκπονηθεί σε ομάδες το πολύ των 2 ατόμων.
- Το παραδοτέο της εργασίας θα υποβληθεί στο helios στην σελίδα του μαθήματος σε link που θα ανοίξει αργότερα.
- Η εργασία αποτελεί το 25% του συνολικού βαθμού του μαθήματος. Για να υπολογιστεί ο βαθμός της εργασίας, η κάθε ομάδα θα πρέπει να περάσει επιτυχώς την υποχρεωτική προφορική εξέταση στο αντικείμενο της εργασίας. Η εξέταση θα γίνει μετά την παράδοση της εργασίας και θα αναρτηθεί σχετικό πρόγραμμα αφού ολοκληρωθεί η υποβολή των

εργασιών. Σημειώνεται επίσης ότι η υποβολή παραδοτέου (δείτε παρακάτω) και η εξέταση της εργασίας είναι υποχρεωτικά ώστε να προκύψει βαθμός για την εργασία.

- Απορίες/επεξηγήσεις για την εργασία θα γίνονται μέσω forum στη σελίδα του μαθήματος στο helios. Μη στέλνετε τις απορίες σας στα email των διδασκόντων/βοηθών αλλά να τις υποβάλλετε όπως αναφέρεται.
- Ως παραδοτέο να υποβληθεί ένα pdf αρχείο με όνομα τους ΑΜ των μελών της ομάδας χωρισμένα με κάτω παύλα (ή το ΑΜ του φοιτητή σε περίπτωση μονομελούς ομάδας), π.χ. 03100000_03100001.zip, ή 03100000.zip (ανάλογα με το πλήθος των ατόμων της ομάδας). Το αρχείο θα περιέχει μία αναφορά (αυστηρά με όσα ζητούνται στην εκφώνηση) η οποία θα περιέχει αποκλειστικά της απαντήσεις στις ερωτήσεις που παρατίθενται, καθώς και ένα github (gitlab/bitbucket/...) link που θα περιέχει όλους τους κώδικες που έχετε υλοποιήσει, όπως και πιθανά scripts/howtos για την εκτέλεση του κώδικά σας.
- Η κάθε ομάδα μπορεί να υλοποιήσει τον κώδικά της σε Scala, Java ή Python. Επιπλέον, σας δίνεται η δυνατότητα να χρησιμοποιήσετε δικούς σας πόρους (π.χ. προσωπικούς Η/Υ, VM) ή πόρους από τον Ωκεανό. Σε κάθε περίπτωση, η εξέταση θα απαιτήσει τη ζωντανή επίδειξη του κώδικά σας.