Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής & Επικοινωνιών
Τμήμα Πληροφορικής

Ακαδ. έτος 2017-18 (χειμ. εξάμηνο)

ΣΥΣΤΗΜΑΤΑ ΔΙΑΧΕΙΡΙΣΗΣ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

5ο εξ., υποχρ. κατευθ. ΠΣΥ / ΤΛΕΣ

Διδάσκοντες:
καθ. Γιάννης Θεοδωρίδης, επίκ. καθ. Νίκος Πελέκης

Εργαστηριακοί βοηθοί:
Γιάννης Κοντούλης, Πέτρος Πέτρου (ΚΕΚΤ εργ. 205)

ΕΡΓΑΣΙΑ ΜΑΘΗΜΑΤΟΣ

(σε ομάδες των 2-3 ατόμων)

Εισαγωγή

Έστω η Βάση Δεδομένων μιας ασφαλιστικής εταιρείας. Κάθε εγγραφή της ΒΔ, της μορφής (ID, Timestamp, Distance, RegionID) αποθηκεύει το αναγνωριστικό κάθε πελάτη, ημερομηνία και ώρα της μέτρησης, την απόσταση σε μέτρα που έχει διανύσει κάθε όχημα σε συγκεκριμένη χρονική διάρκεια, καθώς και το αναγνωριστικό της κάθε περιοχής που έγινε η μέτρηση, π.χ. (1, now, 500m, 1). Σας δίνεται ένα μεγάλο σύνολο από πραγματικά δεδομένα. Πάνω σε αυτό, απαντήστε τα παρακάτω ερωτήματα.

Ερώτημα 1 (60%). Βελτιστοποίηση Επερωτήσεων σε Κεντρικοποιημένη ΒΔ

Κάθε φορά που εκτελείτε μία από τις παρακάτω επερωτήσεις θα δείχνετε τον χρόνο εκτέλεσης (πάντα θα εκτελείτε την επερώτηση τουλάχιστον δύο φορές και θα κρατάτε τον τελευταίο χρόνο – ο λόγος που δεν κρατάμε απλά τον χρόνο εκτέλεσης της πρώτης φοράς είναι ότι δεν είναι αντιπροσωπευτικός γιατί τα buffers δεν έχουν προλάβει να αρχικοποιηθούν) καθώς και το πλάνο εκτέλεσης (χρησιμοποιώντας την εντολή EXPLAIN, screenshot). Σκοπός είναι κάθε φορά που αλλάζετε κάτι στη ΒΔ, με απώτερο στόχο να

βελτιώσετε τους χρόνους εκτέλεσης, να παρατηρείτε αν υπάρχει βελτίωση και πόση είναι αυτή αλλά και να εξηγείτε τη βελτίωση αυτή με βάση τη θεωρία και το πλάνο εκτέλεσης.

Η απάντηση στα ερωτήματα a, b, c, d, e πρέπει να γίνει με τη σειρά εμφάνισής τους, δηλαδή, θα απαντήσετε στο ερώτημα b αφού πρώτα έχετε απαντήσει στο a (και ούτω καθεξής), έτσι ώστε οι αλλαγές που κάνατε στο a (π.χ. buffers, parallelism, κτλ.) να συνεχίσουν να είναι ενεργές στο b και έπειτα.

(a) Αφού φορτώσετε τα δεδομένα στο Σύστημα Διαχείρισης ΒΔ της PostgreSQL (εντολή "COPY ... WITH CSV HEADER") και ανανεώσετε τα στατιστικά χρησιμοποιώντας την εντολή "VACUUM FULL ...", εκτελέστε τις παρακάτω επερωτήσεις (queries) χρησιμοποιώντας τις προεπιλεγμένες ρυθμίσεις της PostgreSQL και χωρίς να έχετε δημιουργήσει βοηθητικές δομές (π.χ. ευρετήρια).

   i. Βρείτε τον πελάτη ο οποίος είχε τη μεγαλύτερη χιλιομετρική απόσταση μέχρι μία συγκεκριμένη ημέρα-ώρα (timestamp) που θα ορίσετε εσείς (στο αποτέλεσμα θα εμφανίζεται ο κωδικός του πελάτη).

   ii. Βρείτε τη μέση χιλιομετρική απόσταση για τον τελευταίο μήνα από το σύνολο των δεδομένων (στο αποτέλεσμα θα εμφανίζεται ένας αριθμός, δηλαδή η μέση διανυθείσα απόσταση).

   iii. Βρείτε το μηνιαίο άθροισμα της χιλιομετρικής απόστασης για κάθε πελάτη (στο αποτέλεσμα θα εμφανίζονται ο κωδικός του πελάτη, ο μήνας και η μηνιαία χιλιομετρική απόσταση).

   iv. Βρείτε τη μέση διανυθείσα απόσταση για κάθε πελάτη (στο αποτέλεσμα θα εμφανίζονται ο κωδικός του πελάτη και η μέση κατανάλωση του).

   v. Βρείτε τη μέση χιλιομετρική απόσταση ανά περιοχή.

(b) Ρυθμίστε την PostgreSQL έτσι ώστε να χρησιμοποιεί ως buffer περισσότερη μνήμη από τη μνήμη RAM του υπολογιστή σας (ικανή ώστε να χωράει όσο γίνεται περισσότερο από το dataset, όλο αν είναι δυνατόν). Έπειτα, εκτελέστε πάλι τις παραπάνω επερωτήσεις και εξηγήστε τι παρατηρείτε. TIP: shared_buffers (π.χ. ALTER SYSTEM SET shared_buffers TO '256MB'; -- απαιτείται επανεκκίνηση του postgreSQL server).

(c) Ρυθμίστε την PostgreSQL έτσι ώστε να χρησιμοποιεί όλη την επεξεργαστική ισχύ του υπολογιστή σας. Έπειτα, εκτελέστε πάλι τις παραπάνω επερωτήσεις και εξηγήστε τι παρατηρείτε. TIP: max_parallel_workers_per_gather

(d) Δημιουργήστε τα κατάλληλα ευρετήρια στη ΒΔ για να τρέξουν οι παραπάνω επερωτήσεις πιο γρήγορα. Για κάθε ευρετήριο που θα δημιουργήσετε θα εξηγήσετε τους λόγους για τους οποίους επιλέξατε τον συγκεκριμένο τύπο ευρετηρίου και το πώς βοηθάει στην βελτίωση του χρόνου εκτέλεσης.

(e) Σπάστε το dataset σε shards/partitions χρησιμοποιώντας την κληρονομικότητα μεταξύ πινάκων. Υπάρχουν πολλοί τρόποι με τους οποίους μπορείτε να κάνετε το partitioning (π.χ. random, hash, range, κτλ.), κάθε ομάδα θα επιλέξει μόνο έναν τρόπο και θα επιχειρηματολογήσετε για την επιλογή σας. Έπειτα, εκτελέστε πάλι τις παραπάνω επερωτήσεις. TIP: Σε κάθε πίνακα παιδί μπορείτε να δημιουργήσετε τα

κατάλληλα ευρετήρια για την περαιτέρω βελτίωση του χρόνου εκτέλεσης των ερωτημάτων.

Ερώτημα 2 (40%). Υλοποίηση Ερωτημάτων σε Spark (RDD)

Αφού φορτώσετε το σύνολο δεδομένων που σας δίνεται στο Spark:

(a) Υπολογίστε κάποια στατιστικά για τα δεδομένα (ενδεικτικά, μπορείτε να υπολογίσετε πόσα είναι τα αυτοκίνητα, πόσα αυτοκίνητα πέρασαν από ποιες περιοχές, κλπ.).

(b) Επιλέξτε ένα υποερώτημα από 1.a και προσαρμόστε το σε Spark με τη χρήση των κατάλληλων εντολών που παρέχονται από την πλατφόρμα (π.χ. map, reduce, reducebykey, avg, min, max, κλπ.).

(c) Χρησιμοποιώντας ένα partitioning από την πλατφόρμα (range hash), ξανατρέξτε το προηγούμενο ερώτημα και παρατηρείστε τη διαφορά στο χρόνο εκτέλεσης. Βοήθησε το partitioning στην εκτέλεση του ερωτήματος;

Παραδοτέο εργασίας:

Το περιεχόμενο του παραδοτέου θα είναι η πλήρης εργασία εκτυπωμένη υπό μορφή τεχνικής αναφοράς (κείμενο, screenshots, queries, κτλ.). Συγκεκριμένα, για το ερώτημα 1 θα συμπεριλάβετε τα SLQ queries, καθώς και τα αποτελέσματά τους (screenshots και κώδικας). Αντίστοιχα, για το ερώτημα 2 θα συμπεριλάβετε τον Java κώδικα και τα παραγόμενα αποτελέσματα.

Τρόπος, τόπος και χρόνος παράδοσης:

Η εργασία θα παραδοθεί στη θυρίδα του κ. Θεοδωρίδη (γραφείο 501) μέχρι την ημερομηνία εξέτασης του μαθήματος κατά την εξεταστική Ιανουαρίου. Απαραίτητη διευκρίνιση: εργασίες δεν γίνονται δεκτές κατά την εξεταστική Σεπτεμβρίου.

Στο εξώφυλλο θα υπάρχουν τα στοιχεία:

Μάθημα: «Συστήματα Διαχείρισης Βάσεων Δεδομένων (5ο εξ.)»
Ομάδα εργασίας: (ΑΜ, ονοματεπώνυμο)

Επιπλέον της εκτυπωμένης εργασίας, θα αποστείλετε την εργασία, δηλ. την τεχνική αναφορά καθώς και τα συνοδευτικά αρχεία (SQL queries, Spark script), στους εργαστηριακούς βοηθούς του μαθήματος στις ακόλουθες διευθύνσεις ηλεκτρονικού ταχυδρομείου: ppetrou@unipi.gr, ikontoulis@unipi.gr. Κάθε email θα έχει ως τίτλο "Εργασία DB2 2017-2018 - <ΑΜ μελών ομάδας>" και θα περιέχει τα ζητούμενα σε ένα zip αρχείο.

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία αφορά στην άσκηση μπορείτε να απευθυνθείτε στους εργαστηριακούς βοηθούς.

Ζητήματα δεοντολογίας

Είναι προφανές ότι η βαθμολογία πρέπει να αντικατοπτρίζει το επίπεδο της γνώσης που αποκόμισε ο φοιτητής μέσα από το μάθημα και κατάφερε να μεταφέρει αυτή τη γνώση στην εργασία. Για να εξασφαλιστεί όσο είναι δυνατό η παραπάνω αρχή, (α) σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται, (β) σε περίπτωση αμφιβολίας για το κατά πόσο η ομάδα που αναγράφεται ήταν εκείνη που ανέπτυξε την εργασία, ενδέχεται να της ζητηθεί να την παρουσιάσει για τυχόν διευκρινίσεις.