Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής & Επικοινωνιών
Τμήμα Πληροφορικής

Ακαδ. έτος 2023-24 (χειμ. εξάμηνο)

ΣΥΣΤΗΜΑΤΑ ΔΙΑΧΕΙΡΙΣΗΣ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

5ο εξ., υποχρ. κατευθ. ΠΣΥ / ΤΛΕΣ

Διδάσκων:
καθ. Γιάννης Θεοδωρίδης

Εργαστηριακός βοηθός:
Γιάννης Κοντούλης (ΚΕΚΤ εργ. 205)

ΕΡΓΑΣΙΑ ΜΑΘΗΜΑΤΟΣ

(σε ομάδες των 2-3 ατόμων)

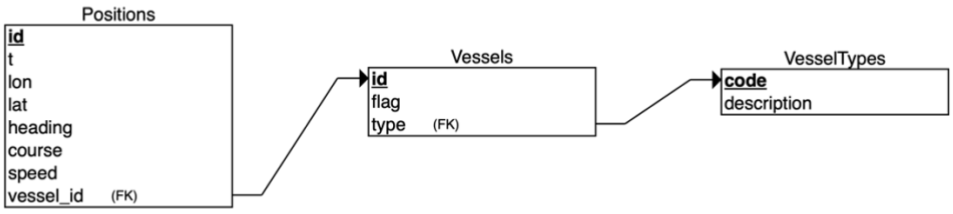
Εισαγωγή

Στο πλαίσιο της εργασίας του μαθήματος, θα εργαστείτε με ένα σύνολο πραγματικών δεδομένων σχετικό με τις θέσεις των πλοίων στον Πειραιά και τον Αργοσαρωνικό, όπως αυτές εκπέμπονται από τα πλοία και συλλέγονται από κατάλληλες κεραίες μέσω του Αυτόματου Συστήματος Αναγνώρισης (Automatic Identification System - AIS)[1]. Συγκεκριμένα, το σύνολο δεδομένων που θα χρησιμοποιήσετε αποτελείται από τρία (3) αρχεία CSV, ένα για κάθε πίνακα της ΒΔ. Ειδικότερα, το πρώτο αρχείο (Positions.csv) περιέχει στίγματα πλοίων για το χρονικό διάστημα 01/08/2019 - 30/08/2019, το δεύτερο (Vessels.csv) περιέχει στατικές πληροφορίες για τα πλοία (τύπο και σημαία) και, τέλος, το τρίτο (VesselTypes.csv) περιέχει σύντομη περιγραφή των διαφόρων τύπων πλοίων. Μπορείτε να κατεβάσετε το σύνολο δεδομένων από εδώ: https://datastories.cs.unipi.gr/index.php/s/ZEM86Fe6i4FeJCj.

- **Positions** (id, t, lon, lat, heading, course, speed, vessel_id) // τα στίγματα των πλοίων
- **Vessels** (id, flag, type) // πληροφορίες σχετικά με τα πλοία

[1] Για το σύστημα AIS, βλ. https://en.wikipedia.org/wiki/Automatic_identification_system. Για την κεραία AIS του Παν/μίου Πειραιώς βλ. https://www.datastories.org/univ-piraeus-ais-stream-visualization/. Ειδικότερα, τα δεδομένα που συλλέγει η κεραία οπτικοποιούνται σε πραγματικό χρόνο στην εφαρμογή https://www.datastories.org/unipi-ais/ και αποθηκεύονται σε τοπική ΒΔ.

- **VesselTypes** (<u>code</u>, description) // οι τύποι των πλοίων



Εικόνα 1: Το σχεσιακό σχήμα της ΒΔ

id	vessel_id	t	lon	lat	heading	course	speed
6894	c8fab22ae78c78461509e849f73e5e120426044a877f2e89228820669a5a91f5	2019-08-01 09:01:35	23.54545	37.88193	242	287.9	0.1
6895	106ec44c8e46b979ec3444d44764f292a02ce0bab1c75df97d013a8b7c24b7fa	2019-08-01 09:01:35	23.5982	37.95271	204	203.5	5.2
6896	b400a9d9f7c8f5984b521403ad0b4d4d0bcda8ac3ef6b1ffacbfa6829a374efc	2019-08-01 09:01:36	23.5681	37.9096	356	351.3	1.8

id	type	flag
fc9e55e7a03626160e0d708a81868e662ce1b18737407c0b2cf71ae1530cc182	37	United Kingdom
e19dc8a6f180ae92644a2dc46eb6d24a4a89f7b75209e49e5198ccb3f79fbe06	37	United Kingdom
7496449fd71ff3059e7857a5c8434d0efa95f96a5c43dfff76920d39385fea00	36	United Kingdom

code	description
49	High speed craft (HSC), No additional information
50	Pilot Vessel
51	Search and Rescue vessel

Εικόνα 2: Ενδεικτικές εγγραφές από τα αρχεία με τα οποία τροφοδοτούμε τη ΒΔ

Βάσει των παραπάνω, καλείστε να απαντήσετε στα ακόλουθα ερωτήματα 1-5. Προσοχή: 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 πρώτος χρόνος εκτέλεσης 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 1-5 πρέπει να γίνει με τη σειρά εμφάνισής τους, δηλαδή, θα απαντήσετε στο ερώτημα 2 αφού πρώτα έχετε απαντήσει στο 1 κοκ., έτσι ώστε οι αλλαγές που κάνατε (buffers, parallelism, κτλ.) στο ένα ερώτημα να συνεχίσουν να είναι ενεργές στα επόμενα.

Ερώτημα 1 (30 %)

Αφού φορτώσετε τα δεδομένα στην 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 αριθμό των στιγμάτων (lon, lat) ανά ημερολογιακή ημέρα και ταξινομήστε το αποτέλεσμα σε φθίνουσα σειρά (ως προς το πλήθος των στιγμάτων). Διευκρίνιση: ο συνδυασμός των χαρακτηριστικών lon, lat σε κάθε εγγραφή του πίνακα Positions συνιστά τη θέση του πλοίου τη συγκεκριμένη χρονική στιγμή t.

ii. Βρείτε πόσα πλοία με ελληνική σημαία ανά τύπο πλοίου είναι καταγεγραμμένα στη ΒΔ.

iii. Βρείτε ποια πλοία ανέπτυξαν κάποια στιγμή ταχύτητα άνω των 30 κόμβων[2], τι τύπου ήταν το κάθε πλοίο και πόσα ήταν αυτά τα πλοία ανά τύπο.

iv. Ειδικά για τα επιβατηγά πλοία (τύποι «passenger …»), πόσα στίγματα καταγράφηκαν ανά ημέρα την περίοδο 14/08/2019 - 18/08/2019; TIP: description LIKE…

v. Ποια πλοία τύπου cargo ήταν 'αγκυροβολημένα' (ταχύτητα μηδέν) κάποια στιγμή μέσα στην περίοδο 15/08/2019 - 18/08/2019; Ποια για ολόκληρη την περίοδο 12/08/2019 - 19/08/2019;

Ερώτημα 2 (15 %)

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).

Ερώτημα 3 (15 %)

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

Ερώτημα 4 (20 %)

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, καθώς και το πώς βοηθάει στη βελτίωση του χρόνου εκτέλεσης. Αν κάποιο ευρετήριο δεν βελτιώσει την απόδοση, εξηγείστε γιατί.

[2] Σημείωση: όπου αναφερόμαστε σε ταχύτητα, εννοούμε την πληροφορία που έχει καταγραφεί στη ΒΔ (πεδίο Positions.speed). Επίσης, θυμίζουμε ότι στη θάλασσα η συνήθης μονάδα μέτρησης της απόστασης είναι το ναυτικό μίλι (ν.μ.), άρα μονάδα μέτρησης της ταχύτητας θα είναι ο κόμβος (1 κόμβος = 1 ν.μ./ώρα).

Ερώτημα 5 (20 %)

Σπάστε το dataset σε shards/partitions χρησιμοποιώντας τη μέθοδο της επιλογής από αυτές που είδατε στις εργαστηριακές διαλέξεις (διαμέριση μέσω κληρονομικότητας μεταξύ πινάκων / δηλωτική διαμέριση). 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.

Παραδοτέο εργασίας:

Η εργασία θα παραδοθεί αποκλειστικά μέσω email στην ακόλουθη διεύθυνση ηλεκτρονικού ταχυδρομείου: ikontoulis@unipi.gr. Κάθε email θα έχει ως τίτλο "Εργασία DB2 2023-2024 - < ονοματεπώνυμα - ΑΜ μελών ομάδας>" και θα περιέχει τα ζητούμενα σε ένα zip αρχείο.

Το περιεχόμενο του παραδοτέου θα αποτελείται από τα εξής: (i) τεχνική αναφορά – κείμενο (σε PDF μορφή), στο οποίο θα περιγράφεται η υλοποίηση των ερωτημάτων, μαζί με τα κατάλληλα screenshots από τις εκτελέσεις των ερωτημάτων σε PostgreSQL. (ii) SQL script με τις εντολές SQL.

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία αφορά στην άσκηση μπορείτε να απευθυνθείτε στον εργαστηριακό βοηθό.

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.

Καλή Επιτυχία!