Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής
ΠΜΣ «Κυβερνοασφάλεια και Επιστήμη Δεδομένων»,
2023-2024

Big Data Management

Διδάσκοντες: Νίκος Πελέκης, Γιώργος Παπαστεφανάτος
Εργαστηριακοί βοηθοί: Γ. Αλεξίου, Σ. Μαρούλης

1η Εργαστηριακή Άσκηση
(ατομική)

Αντικείμενο της άσκησης

Σκοπός της άσκησης είναι η εξοικείωση με NoSQL συστήματα διαχείρισης βάσεων δεδομένων και πιο συγκεκριμένα με τη MongoDB. Το σύνολο δεδομένων που θα χρησιμοποιηθεί προέρχεται από το TripAdvisor και περιλαμβάνει πληροφορίες για ευρωπαϊκά εστιατόρια. Συγκεκριμένα, περιλαμβάνει 1.083.397 ευρωπαϊκά εστιατόρια, με αναλυτικά στοιχεία όπως τοποθεσία, μέση βαθμολογία, αριθμός κριτικών, ώρες λειτουργίας, τύποι κουζίνας, και άλλα. Μπορείτε να κατεβάσετε τα δεδομένα σε μορφή JSON από εδώ.

Μεθοδολογία υλοποίησης της άσκησης

Ερώτημα 1. Εισαγωγή Δεδομένων

Εισάγετε το αρχείο με τα δεδομένα στη βάση, χρησιμοποιώντας είτε κατάλληλες εντολές (hint: mongoimport), είτε το εργαλείο MongoDB Compass. Το παραδοτέο του ερωτήματος θα είναι οι εντολές που χρησιμοποιήθηκαν ή screenshots από τα βήματα που ακολουθήθηκαν στο MongoDB Compass, καθώς και screenshots που θα παρουσιάζουν δείγμα των δεδομένων στη βάση.

Ερώτημα 2. Ανάκτηση δεδομένων από τη ΒΔ

Θα πρέπει να υλοποιήσετε τα παρακάτω queries:

1. Εντοπίστε τα 10 κορυφαία mid-range εστιατόρια (price_level = "€€-€€€") στην Αθήνα που προσφέρουν ιταλική ή ισπανική κουζίνα (πεδίο 'cuisines'). Η κατάταξη βασίζεται στον μέσο όρο της αξιολόγησής τους (avg_rating), με προϋπόθεση την ύπαρξη περισσότερων από 100 συνολικών αξιολογήσεων

(total_reviews_count). Εμφανίστε τα ονόματα, τις διευθύνσεις, τη μέση αξιολόγηση, το συνολικό αριθμό αξιολογήσεων, και τους τύπους κουζίνας που προσφέρει κάθε εστιατόριο, ταξινομημένα πρώτα με βάση τη μέση αξιολόγηση και στη συνέχεια με βάση το συνολικό αριθμό αξιολογήσεων.

2. Εντοπίστε τους 10 πιο δημοφιλείς τύπους κουζίνας (πεδίο 'cuisines') στην Ελλάδα βάσει του αριθμού των εστιατορίων που προσφέρουν κάθε τύπο. Παρουσιάστε το όνομα κάθε τύπου κουζίνας μαζί με τον αντίστοιχο αριθμό των εστιατορίων που τον προσφέρουν, καθώς και το μέσο όρο του αριθμού αξιολογήσεων αυτών.
3. Βρείτε τις 10 χώρες πλην της Ελλάδας με τον μεγαλύτερο αριθμό εστιατορίων που προσφέρουν ελληνική κουζίνα. Εμφανίστε τις χώρες αυτές, καθώς και το πλήθος αυτών των εστιατορίων.
4. Εντοπίστε το εστιατόριο στην Ελλάδα με το μεγαλύτερο αριθμό βραβείων. Χρησιμοποιείστε το πεδίο 'awards' για να υπολογίσετε το συνολικό αριθμό βραβείων για κάθε εστιατόριο. Στην περίπτωση που υπάρχουν περισσότερα από ένα εστιατόριο με τον ίδιο μέγιστο αριθμό βραβείων, εμφανίστε αυτό με το μεγαλύτερο αριθμό αξιολογήσεων. Τυπώστε το όνομα και τη διεύθυνση του εστιατορίου, καθώς και τη λίστα με τα βραβεία του.
5. Βρείτε τα 10 εστιατόρια στην Ελλάδα με την υψηλότερη αναλογία 'excellent' αξιολογήσεων σε σχέση με τον συνολικό αριθμό αξιολογήσεων (total_reviews_count), λαμβάνοντας υπόψιν εκείνα με πάνω από 100 αξιολογήσεις. Εμφανίστε τα ονόματα και τις διευθύνσεις των εστιατορίων, το ποσοστό των 'excellent' αξιολογήσεων και το συνολικό αριθμό αξιολογήσεων, ταξινομημένα με βάση το ποσοστό 'excellent' αξιολογήσεων.

Για τα παραπάνω queries, χρησιμοποιήστε τις μεθόδους **find()** και **aggregate()** της MongoDB. Το παραδοτέο του ερωτήματος θα είναι τα queries που απαντούν τις παραπάνω ερωτήσεις στη βάση καθώς και screenshots των αποτελεσμάτων.

Ερώτημα 3. Optimization

Δημιουργήστε κατάλληλα ευρετήρια για την επιτάχυνση των ερωτημάτων του Ερωτήματος 2. Αξιολογήστε την απόδοση των queries πριν και μετά την εφαρμογή των ευρετηρίων (χρησιμοποιώντας τη μέθοδο **explain()**). Καταγράψτε και συγκρίνετε τα αποτελέσματα μέσω screenshots των πλάνων εκτέλεσης και σχολιάστε τις διαφορές.

Στην αναφορά σας, συπεριλάβετε τις εντολές δημιουργίας των ευρετηρίων και αιτιολογήστε την επιλογή σας. Εξηγήστε γιατί επιλέξατε συγκεκριμένα πεδία για την ευρετηρίαση, λαμβάνοντας υπόψη στοιχεία όπως π.χ. τους τελεστές που χρησιμοποιήθηκαν στα ερωτήματα και την πληθικότητα των τιμών στα επιλεγμένα πεδία.

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία σχετικά με την άσκηση μπορείτε να απευθύνεστε στον κ. Σταύρο Μαρούλη (stavmars@athenarc.gr) με email.

Ημερομηνία και τρόπος παράδοσης

Η αναφορά σας για την εργασία να αποσταλεί σε μορφή **PDF** στο email stavmars@athenarc.gr, μέχρι και τις 22/12/2023. Παρακαλείσθε να έχετε την παρακάτω πρόταση ως θέμα στο email που θα στείλετε: **"CDS110 - Εργασία MongoDB".**