Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου

Εργαστηριακή Άσκηση 2022/23

|  |  |  |
| --- | --- | --- |
| **Ονομα** | **Επώνυμο** | **ΑΜ** |
| Βασιλική Ευαγγελία | Δούρου | 1072633 |
| Δημήτριος | Μπαλάφας | 1072499 |

Βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για το συγκεκριμένο μάθημα/σεμινάριο/πρόγραμμα σπουδών.

Έχω ενημερωθεί ότι σύμφωνα με τον εσωτερικό κανονισμό λειτουργίας του Πανεπιστημίου Πατρών άρθρο 50§6, τυχόν προσπάθεια αντιγραφής ή εν γένει φαλκίδευσης της εξεταστικής και εκπαιδευτικής διαδικασίας από οιονδήποτε εξεταζόμενο, πέραν του μηδενισμού, συνιστά βαρύ πειθαρχικό παράπτωμα.

|  |  |
| --- | --- |
| Υπογραφή    21 / 09 / 2023 | Υπογραφή    21 / 09 / 2023 |

**Συνημμένα αρχεία κώδικα**

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα:

|  |  |  |
| --- | --- | --- |
| Αρχείο | Αφορά το ερώτημα | Περιγραφή/Σχόλιο |
| csv\_to\_json.py | 2 | Περιέχει κώδικα για μετατροπή των .csv αρχείων από το Kaggle σε .json αρχεία. |
| json\_merge.py | 2 | Περιέχει κώδικα για δημιουργία αρχείου .json που περιέχει όλες τις πληροφορίες για τις συνταγές. |
| popular\_recipes\_inserts.py | 2 | Περιέχει κώδικα για τα insert στον πίνακα popular\_recipes. |
| recipe\_details\_inserts.py | 2 | Περιέχει κώδικα για τα insert στον πίνακα recipe\_details. |
| recipes\_by\_difficulty\_inserts.py | 2 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_difficulty. |
| recipes\_by\_name\_inserts.py | 2 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_name. |
| recipes\_by\_tag\_inserts.py | 2 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_tag. |
| select\_1.py | 3 | Περιέχει κώδικα για το 1ο select statement του 3ου ερωτήματος. |
| select\_2.py | 3 | Περιέχει κώδικα για το 2ο select statement του 3ου ερωτήματος. |
| select\_3.py | 3 | Περιέχει κώδικα για το 3ο select statement του 3ου ερωτήματος. |
| select\_4.py | 3 | Περιέχει κώδικα για το 4ο select statement του 3ου ερωτήματος. |
| select\_5.py | 3 | Περιέχει κώδικα για το 5ο select statement του 3ου ερωτήματος. |
| create\_for\_lab.txt | 4 | Περιέχει τα DDL statements για τη δημιουργία της βάσης στο εργαστήριο. |
| popular\_recipes\_inserts\_lab.py | 4 | Περιέχει κώδικα για τα insert στον πίνακα popular\_recipes για το εργαστήριο. |
| recipe\_details\_lab.py | 4 | Περιέχει κώδικα για τα insert στον πίνακα recipe\_details για το εργαστήριο. |
| recipes\_by\_difficulty\_lab.py | 4 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_difficulty για το εργαστήριο. |
| recipes\_by\_name\_lab.py | 4 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_name για το εργαστήριο. |
| recipes\_by\_tag\_inserts\_for\_lab.py | 4 | Περιέχει κώδικα για τα insert στον πίνακα recipes\_by\_tag για το εργαστήριο. |
| select\_1\_lab.py | 4 | Περιέχει κώδικα για την εκτέλεση του 1ου select statement στο εργαστήριο. |
| select\_2\_lab.py | 4 | Περιέχει κώδικα για την εκτέλεση του 2ου select statement στο εργαστήριο. |
| select\_3\_lab.py | 4 | Περιέχει κώδικα για την εκτέλεση του 3ου select statement στο εργαστήριο. |
| select\_4\_lab.py | 4 | Περιέχει κώδικα για την εκτέλεση του 4ου select statement στο εργαστήριο. |
| select\_5\_lab.py | 4 | Περιέχει κώδικα για την εκτέλεση του 5ου select statement στο εργαστήριο. |

Επίσης, επισυνάπτεται και το αρχείο secure-connect-recipes.zip για τη σύνδεση στη βάση.

# Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

Για όλα τα ερωτήματα της εργασίας δημιουργήθηκε ένα στιγμιότυπο της AstraDB στο cloud και δεν εγκαταστάθηκε τοπικά η Cassandra.

# Ερώτημα 1: Σχεδιασμός ΒΔ

A diagram of a diagram

Description automatically generatedΤο εννοιολογικό μοντέλο που προκύπτει από τα queries της εκφώνησης είναι το ακόλουθο:

Στο εννοιολογικό μοντέλο, υπάρχουν δύο οντότητες: ο χρήστης (user) και η συνταγή (recipe). Οι σχέσεις που συνδέουν αυτές τις δύο οντότητες είναι δύο:

1. Ένας χρήστης μπορεί να δημιουργήσει όσες συνταγές θέλει, αλλά κάθε συνταγή ανήκει σε έναν μόνο χρήστη (σχέση 1:Ν).
2. Ένας χρήστης μπορεί να αξιολογήσει Ν συνταγές και μία συνταγή μπορεί να αξιολογηθεί από Μ χρήστες (σχέση Ν:Μ). Αυτή η σχέση έχει επιπλέον δύο attributes: τη βαθμολογία που έβαλε ο χρήστης και την ημερομηνία αξιολόγησης.

Τα queries που προκύπτουν για τις ανάγκες των χρηστών που περιγράφονται στην εκφώνηση είναι τα εξής:

* Q1. View popular recipes.
* Q2. Search recipes by name.
* Q3. Search recipes by difficulty.
* Q4. View recipe details.
* Q5. Search recipes by tags.

Οπότε, με βάση τα παραπάνω, το application workflow που προκύπτει είναι το ακόλουθο:

A diagram of a recipe

Description automatically generated

Στο παραπάνω διάγραμμα, κάθε κουτί αντιπροσωπεύει ένα βήμα στο application workflow. Η αναζήτηση μίας συνταγής με οποιοδήποτε τρόπο, καθώς και η εμφάνιση των πιο δημοφιλών συνταγών, πρέπει να μπορεί να γίνει από την αρχική σελίδα της εφαρμογής, καθώς θα βοηθάει την εφαρμογή να επιστρέψει δεδομένα για πολλές συνταγές που πληρούν κάποιες προϋποθέσεις. Αντιθέτως, η εμφάνιση περισσότερων λεπτομερειών για μία συνταγή, μπορεί να γίνει με βάση το id της, το οποίο μπορεί να αντληθεί από τα παραπάνω queries.

Το Chebotko diagram που προκύπτει είναι τα ακόλουθο:

A screenshot of a computer

Description automatically generated

Το query Q1 είναι η αναζήτηση δημοφιλών συνταγών, όμως δεν γνωρίζουμε επακριβώς για ποιο χρονικό διάστημα. Άρα, αποθηκεύουμε τις συνταγές με βάση τα id τους, οπότε το recipe\_id θα είναι κομμάτι του primary key. Όμως, υπάρχουν πολλές κριτικές για μία συνταγή από πολλούς διαφορετικούς χρήστες. Για να έχουμε επομένως, μοναδικό partition για κάθε αξιολόγηση μίας συνταγής, θα προσθέσουμε το user\_id και το date ως clustering columns.

Το query Q2 είναι να βρούμε συνταγές με βάση κάποια λέξη κλειδί στο όνομα τους. Επομένως, η αναζήτηση γίνεται με βάση το όνομα, οπότε το name θα είναι μέρος του primary key. Όμως, υπάρχουν πολλές συνταγές με κοινές λέξεις στο όνομα τους, οπότε για να έχουμε ένα μοναδικό partition για κάθε συνταγή, θα είναι το recipe\_id clustering column.

Το query Q3 είναι η αναζήτηση συνταγών με βάση τη δυσκολία τους. Επομένως, το partition key θα είναι το category. Όμως, επειδή πολλές συνταγές ανήκουν σε μία κατηγορία για να διαχωριστούν, θα είναι και το recipe\_id clustering column. Επιπλέον, επειδή στα παρακάτω ερωτήματα θέλουμε η κατάταξη να γίνεται με βάση τη βαθμολογία, θα είναι και το rating κομμάτι του clustering column.

Το query Q4 είναι η εμφάνιση λεπτομέρειών μίας συνταγής. Αυτό θα γίνεται με βάση το id της συνταγής, το οποίο είναι μοναδικό. Οπότε μόνο το recipe\_id θα είναι το primary key.

Το query Q5 είναι η αναζήτηση συνταγών με βάση την ετικέτα τους. Επομένως, το partition key θα είναι το tags. Όμως, επειδή πολλές συνταγές έχουν ένα tag για να διαχωριστούν, θα είναι και το recipe\_id clustering column. Επιπλέον, επειδή στα παρακάτω ερωτήματα θέλουμε η κατάταξη να γίνεται με βάση τη βαθμολογία και την ημερομηνία υποβολής, θα είναι και το rating και το submitted κομμάτι του clustering column.

# Ερώτημα 2: Ερωτήματα DDL

Τα tables που δημιουργήθηκαν στη βάση μαζί με τα αντίστοιχα DDL statements είναι τα ακόλουθα:

|  |  |
| --- | --- |
| **Table** | popular\_recipes |
| **DDL statement** | CREATE TABLE popular\_recipes (  recipe\_id INT,  user\_id INT,  rating DOUBLE,  date DATE,  name TEXT,  PRIMARY KEY((recipe\_id),date,user\_id)  ) WITH CLUSTERING ORDER BY (date ASC,user\_id ASC);  CREATE INDEX ON popular\_recipes (date); |
| **Screenshot** |  |

|  |  |
| --- | --- |
| **Table** | recipes\_by\_name |
| **DDL statement** | CREATE TABLE recipes\_by\_name (  recipe\_id INT,  name TEXT,  PRIMARY KEY(name,recipe\_id)  ) WITH CLUSTERING ORDER BY (recipe\_id ASC); |
| **Screenshot** |  |

|  |  |
| --- | --- |
| **Table** | recipes\_by\_difficulty |
| **DDL statement** | CREATE TABLE recipes\_by\_difficulty (  recipe\_id INT,  rating DOUBLE,  n\_steps INT,  minutes INT,  category TEXT,  name TEXT,  PRIMARY KEY((category),rating,recipe\_id)  ) WITH CLUSTERING ORDER BY (rating DESC,recipe\_id ASC); |
| **Screenshot** |  |

|  |  |
| --- | --- |
| **Table** | recipes\_by\_tag |
| **DDL statement** | CREATE TABLE recipes\_by\_tag (  recipe\_id INT,  rating DOUBLE,  tags TEXT,  name TEXT,  submitted DATE,  PRIMARY KEY((tags),rating,submitted,recipe\_id)  ) WITH CLUSTERING ORDER BY (rating DESC, submitted DESC, recipe\_id ASC); |
| **Screenshot** |  |

|  |  |
| --- | --- |
| **Table** | recipe\_details |
| **DDL statement** | CREATE TABLE recipe\_details (  recipe\_id INT,  name TEXT,  contributor\_id INT,  submitted DATE,  rating DOUBLE,  tags TEXT,  n\_steps INT,  minutes INT,  steps TEXT,  n\_ingredients INT,  ingredients TEXT,  nutrition TEXT,  category TEXT,  description TEXT,  PRIMARY KEY(recipe\_id)  ); |
| **Screenshot** |  |

# Ερώτημα 3: Απαντήσεις ερωτημάτων

Τα αποτελέσματα της εκτέλεσης των python scripts είναι τα ακόλουθα:

|  |  |
| --- | --- |
| **Ερώτημα** | **Απάντηση** |
| Εμφάνιση των 30 συνταγών με την υψηλότερη μέση βαθμολογία μεταξύ 01/01/2012 και 31/05/2012 | recipe\_id: 480155 , name: dinosaur bbq style baked beans , average rating: 5.0  recipe\_id: 480075 , name: hot dog onions with only 4 easy ingredients , average rating: 5.0  recipe\_id: 479906 , name: lemon basil mashed potatoes , average rating: 5.0  recipe\_id: 479900 , name: fresh greek pasta salad , average rating: 5.0  recipe\_id: 479864 , name: warm date syrup tahini platter , average rating: 5.0 |
| Εμφάνιση όλων των λεπτομερειών για την συνταγή «chic greek salad» (κατηγορία δυσκολίας, διατροφικές αξίες, βήματα, περιγραφή, μέση βαθμολογία) | chic greek salad 205331 easy 398275 2007-01-14 4.333333333333333 ['15-minutes-or-less', 'time-to-make', 'course', 'main-ingredient', 'preparation', 'salads', 'side-dishes', 'seafood', 'vegetables', 'easy', 'fish', 'dietary', 'tuna', 'saltwater-fish'] ['arrange the cucumbers , tomatoes , onion , red pepper , artichoke hearts and parsley on a platter',  'squeeze the lemon juice over the platter', 'season with salt and pepper', 'place shallots , vinegar and oregano in a bowl and let stand for 5 minutes', 'whisk in olive oil in a slow stream', 'pour dressing over salad', 'garnish salad with chunks of tuna , pepperoncini peppers and olives'] [338.9, 30.0, 35.0, 90.0, 43.0, 14.0, 7.0] ['seedless cucumber', 'red ripe tomatoes', 'red onion', 'red bell pepper', 'water-packed artichoke hearts', 'italian parsley', 'kosher salt & freshly ground black pepper', 'lemon, juice of', 'shallot', 'red wine vinegar', 'oregano', 'extra virgin olive oil', 'tuna in vegetable oil', 'pepperoncini peppers', 'kalamata olive'] 7 12 15 a super, veggie-packed salad. courtesy  of rachael ray. |
| Εμφάνιση των συνταγών της κατηγορίας «εύκολη» ταξινομημένες ως προς τη μέση βαθμολογία τους | 46 a jad cucumber pickle 5.0  52 cafe cappuccino 5.0  59 lou s fabulous bruschetta 5.0  63 cabbage and sausage soup 5.0  67 bourbon pecan pound cake 5.0 |
| Εμφάνιση των συνταγών που περιέχουν την ετικέτα “crock-pot-slow-cooker” με ταξινόμηση ανά ημερομηνία προσθήκης (πιο πρόσφατες πρώτα) | 537429 red wine braised short ribs with fingerlings potatoes 2018-11-05  536990 slow cooker soy braised chicken 2018-09-18  536501 smoky kentucky chili 2018-07-27  536212 slow cooker cheesy bacon ranch potatoes 2018-07-08  533052 killer voodoo gumbo 2017-07-31 |
| Εμφάνιση των 20 συνταγών με την υψηλότερη μέση βαθμολογία για την ετικέτα “cocktails”. | 537459 bailey s chocotini 5.0  534270 island holiday 5.0  533130 masala martini 5.0  533118 cariad rhudd crimson love 5.0  532734 kaiserspritzer 5.0 |

Οι χρόνοι εκτέλεσης του κάθε ερωτήματος, όπως προκύπτουν με τη χρήση του trace, είναι οι ακόλουθοι:

* 1ο ερώτημα: 0:00:00.074807
* 2ο ερώτημα: 0:00:00.002364
* 3ο ερώτημα: 0:00:00.018269
* 4ο ερώτημα: 0:00:00.019513
* 5ο ερώτημα: 0:00:00.006670

# Ερώτημα 4Α: Χρόνοι εισαγωγής δεδομένων

Λόγω περιορισμένου χρόνου στο εργαστήριο, για την εισαγωγή δεδομένων χρησιμοποιήθηκε ένα αρχείο που περιέχει μικρότερο όγκο δεδομένων.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Επίπεδο write consistency** | | |
| **ALL** | **QUORUM** | **ONE** |
| **popular\_recipes** | 31s 67ms | 30s 14ms | 27s 81ms |
| **recipes\_by\_name** | 22s 19ms | 23s 55ms | 20s 44ms |
| **recipes\_by\_difficulty** | 8s 27ms | 8s 20ms | 7s 79ms |
| **recipes\_by\_tag** | 1m 25s 8ms | 1m 23s 12ms | 1m 21s 84ms |
| **recipe\_details** | 9s 40ms | 8s 97ms | 9s 32ms |
| **Μέσος όρος** | 31s 32ms | 30s 439ms | 28s 864ms |

# Ερώτημα 4Β: Χρόνοι ανάκτησης δεδομένων

Λόγω περιορισμένου χρόνου στο εργαστήριο, η εκτέλεση των ερωτημάτων έγινε στα δεδομένα ενός αρχείου που περιέχει μικρότερο όγκο δεδομένων.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Επίπεδο read consistency** | | |
| **ALL** | **QUORUM** | **ONE** |
| **Ερώτημα 1** | 0:00:00.0273568 | 0:00:00.0410637 | 0:00:00.0121926 |
| **Ερώτημα 2** | 0:00:00.0217547 | 0:00:00.0162184 | 0:00:00.0056232 |
| **Ερώτημα 3** | 0:00:00.0753127 | 0:00:00.058883 | 0:00:00.0347597 |
| **Ερώτημα 4** | 0:00:00.0267922 | 0:00:00.0249702 | 0:00:00.0130223 |
| **Ερώτημα 5** | 0:00:00.0145957 | 0:00:00.0092791 | 0:00:00.0088182 |
| **Μέσος όρος** | 0:00:00.03316242 | 0:00:00.03008288 | 0:00:00.0148832 |

# Ερώτημα 4Γ: Σχολιασμός αποτελεσμάτων

Η γραφική παράσταση του μέσου όρου του χρόνου εκτέλεσης των ερωτημάτων εισαγωγής των δεδομένων ανά write consistency level είναι ο ακόλουθος:

Ενώ η γραφική παράσταση του μέσου όρου του χρόνου εκτέλεσης των ερωτημάτων ανάκτησης των δεδομένων ανά read consistency level είναι ο ακόλουθος:

Παρατηρείται και στις δύο περιπτώσεις, πως το επίπεδο συνέπειας ONE εμφανίζει τους πιο γρήγορους χρόνους εκτέλεσης, ενώ το επίπεδο συνέπειας ALL εμφανίζει τους πιο αργούς.

Σύμφωνα με το θεώρημα CAP σε ένα κατανεμημένο σύστημα, μπορούν να συντρέχουν μόνο δύο από τις ιδιότητές του. Οι ιδιότητες είναι οι ακόλουθες :

* Consistency (Συνέπεια)
* Availability (Διαθεσιμότητα)
* Partition Tolerance (Ανοχή Διαμερισμού)

Επειδή χρειαζόμαστε οπωσδήποτε Partition Tolerance, διαπραγματευόμαστε ανάμεσα στα Consistency και Availability. Επομένως, η επιλογή ισχυρότερου επιπέδου συνέπειας, θα συνεπάγεται χαμηλότερο επίπεδο διαθεσιμότητας, δηλαδή υψηλότερους χρόνους εκτέλεσης των reads και writes, καθώς οι περισσότεροι (QUORUM) ή και όλοι (ALL) οι κόμβοι πρέπει να συμμετέχουν. Αντίθετα, η επιλογή χαμηλότερου επιπέδου συνέπειας (ONE) θα έχει ως αποτέλεσμα υψηλότερο επίπεδο διαθεσιμότητας, δηλαδή χαμηλότερους χρόνους εκτέλεσης των reads και writes, καθώς μόνο ένας κόμβος χρειάζεται να αποκρίνεται.

# Βιβλιογραφία

Οι πηγές που αξιοποιήθηκαν για την εκπόνηση της εργασίας είναι οι ακόλουθες:

* <https://www.oreilly.com/content/cassandra-data-modeling/>
* <https://docs.datastax.com/en/cql-oss/3.3/cql/ddl/dataModelingApproach.html>
* <https://docs.datastax.com/en/developer/python-driver/3.21/getting_started/>
* <https://docs.datastax.com/en/cql-oss/3.x/cql/cql_reference/cqlshConsistency.html>
* <https://docs.datastax.com/en/cassandra-oss/3.0/cassandra/dml/dmlConfigConsistency.html>
* <https://docs.datastax.com/en/cql-oss/3.3/cql/cql_using/useTracingTrace.html>
* <https://cassandra.apache.org/doc/latest/cassandra/data_modeling/intro.html>
* <https://docs.datastax.com/en/developer/python-driver/3.25/getting_started/#setting-a-consistency-level>
* <https://docs.datastax.com/en/developer/python-driver/3.25/faq/#how-do-i-trace-a-request>