

## Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου

### Εργαστηριακή Άσκηση 2022/23

Όνομα	Επώνυμο	ΑΜ
Πάρης	Σεργιάννης	1067467
Στυλιανός	Στυλιανάκης	1059713

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

16 / 2 / 2024

Υπογραφή

16 / 2 / 2024

#### Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	Αφορά το ερώτημα	Περιγραφή/Σχόλιο
main.py	όλα	Σε αυτο το αρχείο καλούμε τις συναρτήσεις από το db_functions.py, για να εκτελέσουμε τα inserts και τα selects στην βάση.

		Επιλέγουμε mode αλλάζοντας την μεταβλητή operation_mode.
db_functions.py	όλα	Περιέχει όλες τις συναρτήσεις που αφορούν τις βασικές λειτουργίες της βασης, οι οποίες καλούνται από την main.py.
query_generation.py	2, 4	Περιέχει τα table schemas και χρησιμοποιείται για την κατασκευή ερωτημάτων.
data_shortening.py	όλα	Script που χρησιμοποιείται για την κατασκευή csv με ένα υποσύνολο δεδομένων, με σκοπό την ταχύτερη υλοποίηση των ερωτημάτων.

## Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

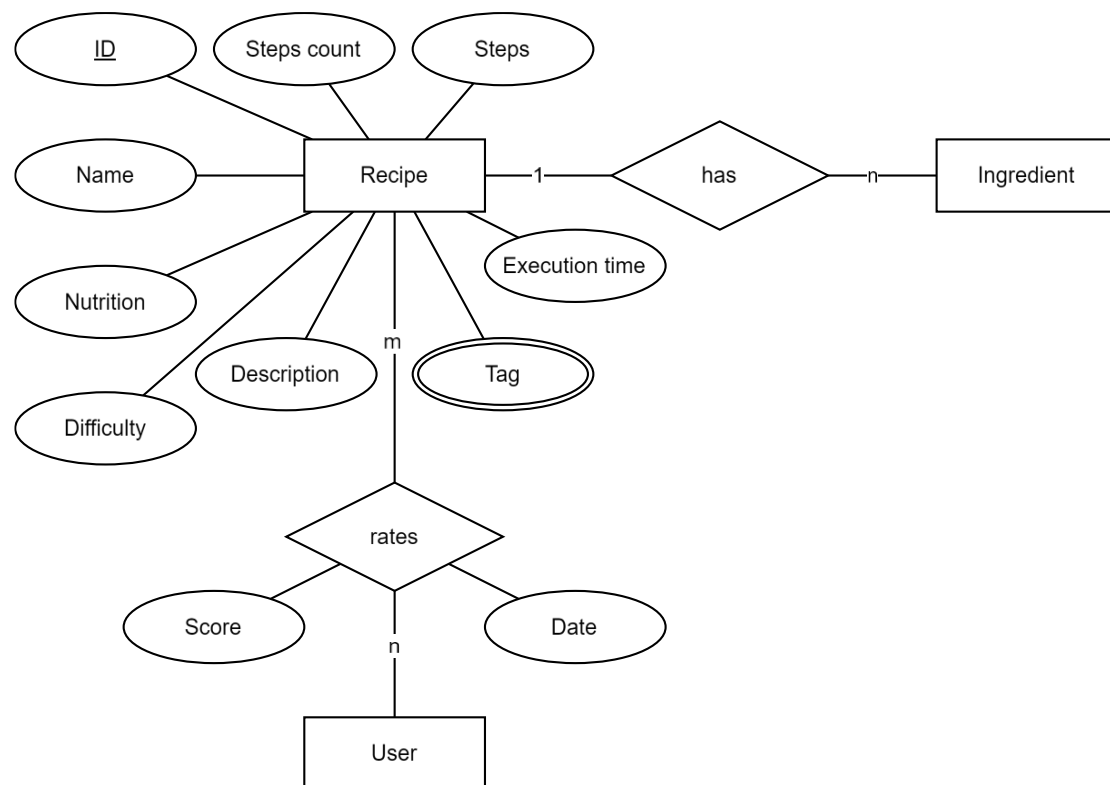
Για την υλοποίηση της εργασίας χρησιμοποιήθηκε αποκλειστικά το Astra DB.

## Ερώτημα 1: Σχεδιασμός ΒΔ

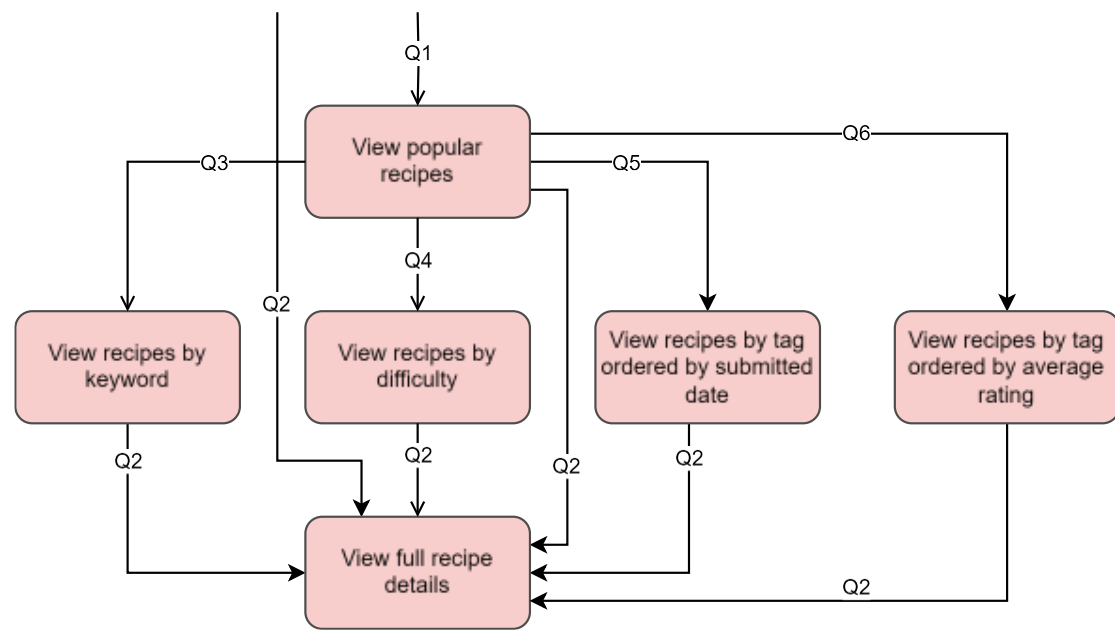
### Απαιτήσεις χρήστη:

- 1) Εμφάνιση των συνταγών που είναι δημοφιλείς (έχουν καλή βαθμολογία) εντός ενός χρονικού διαστήματος (π.χ. τους τελευταίους 3 μήνες) – αυτό μπορεί να αποτελεί κάλλιστα την αρχική οθόνη «προτάσεων» προς το χρήστη.
- 2) Να αναζητήσει την/τις συνταγές που περιέχουν κάποιες λέξεις – κλειδιά στον τίτλο.
- 3) Να αναζητήσει συνταγές με βάση την κατηγορία δυσκολίας (π.χ. χρόνος παρασκευής ή πλήθος βημάτων) και να τις λάβει με βάση κάποια ταξινόμηση (π.χ. μέση βαθμολογία).
- 4) Να δει τις λεπτομερείς πληροφορίες για κάποια συνταγή (κατηγορία δυσκολίας, μέση βαθμολογία, ετικέτες, υλικά, διαδικασία παρασκευής).
- 5) Να δει τις top-n συνταγές που σχετίζονται με κάποια ετικέτα.

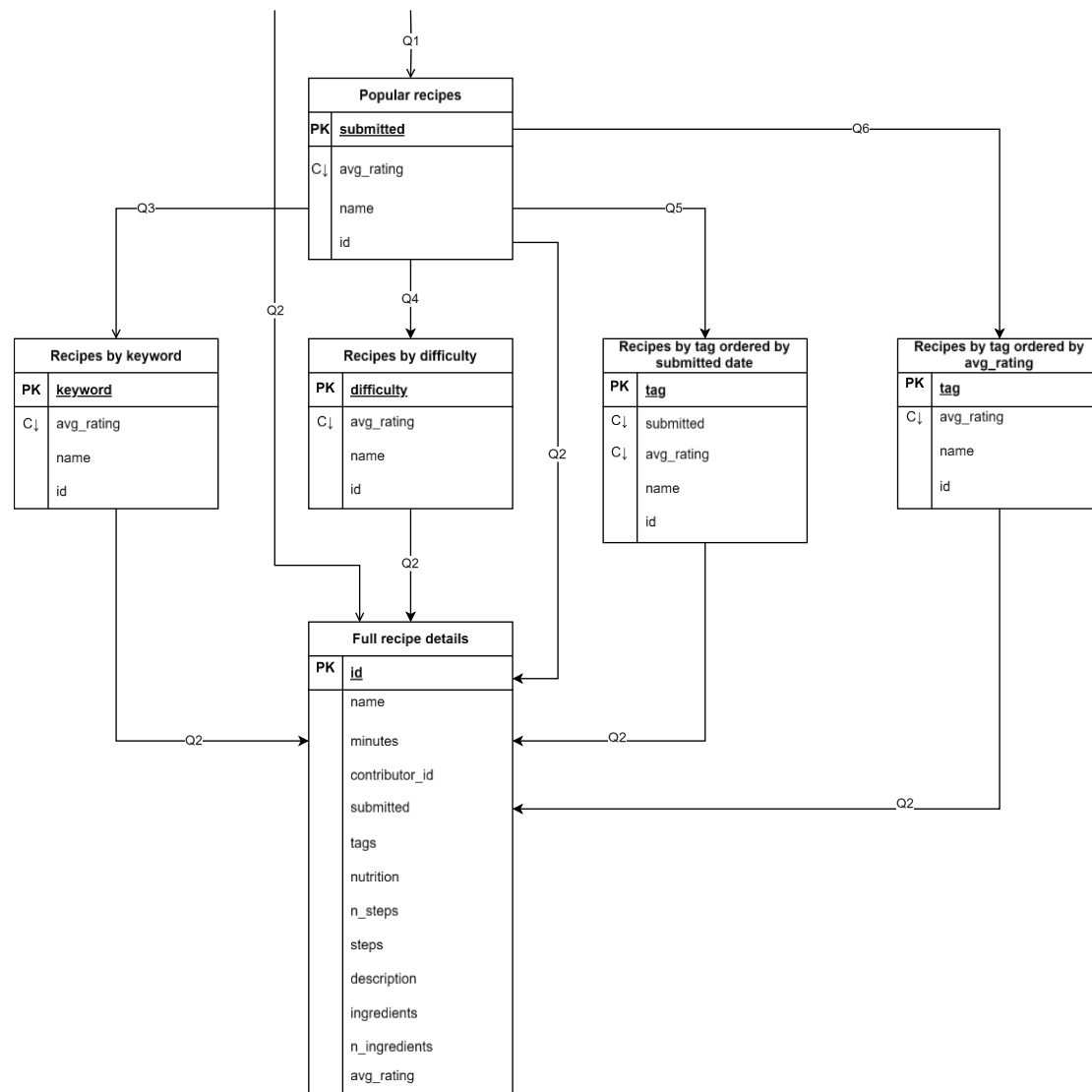
### Εννοιολογικό μοντέλο



## Application workflow



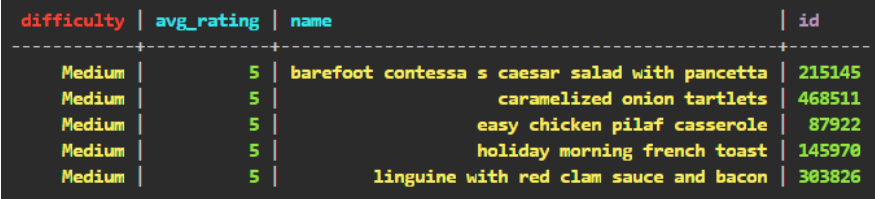
## Chebotko diagram

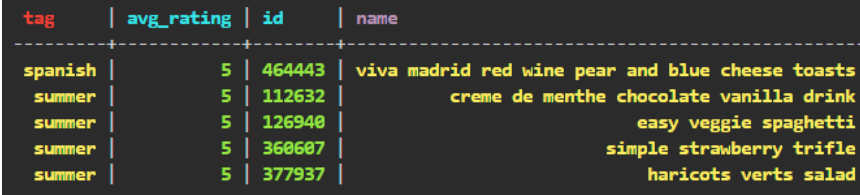


## Ερώτημα 2: Ερωτήματα DDL

[επαναλαμβάνετε τον παρακάτω πίνακα για κάθε keyspace στη ΒΔ σας]

<b>Keyspace</b>	recipes.popular_recipes
<b>DDL statement</b>	<pre>CREATE TABLE recipes.popular_recipes (   id int,   submitted date,   avg_rating float,   name text,   PRIMARY KEY (submitted, avg_rating, id) ) WITH CLUSTERING ORDER BY (avg_rating DESC, id ASC);</pre>
<b>Screenshot</b>	<pre>submitted   avg_rating   id   name -----+-----+-----+----- 2007-12-03   1   269651   sweet n sour ribs 2008-05-22   5   304570   sunshine punch 2006-06-21   4.85714   174358   chocolate snickerdoodles 2002-12-18   3.84   49189   christmas hard candy 2002-08-26   4.53333   38502   crock pot ginger chicken</pre>

<b>Keyspace</b>	recipes.recipes_difficulty
<b>DDL statement</b>	CREATE TABLE recipes.recipes_difficulty ( difficulty text, id int, avg_rating float, name text, PRIMARY KEY (difficulty, avg_rating, name) ) WITH CLUSTERING ORDER BY (avg_rating DESC, name ASC);
<b>Screenshot</b>	 <pre> difficulty   avg_rating   name   id ----- ----- ----- ----- Medium   5   barefoot contessa s caesar salad with pancetta   215145 Medium   5   caramelized onion tartlets   468511 Medium   5   easy chicken pilaf casserole   87922 Medium   5   holiday morning french toast   145970 Medium   5   linguine with red clam sauce and bacon   303826 </pre>

<b>Keyspace</b>	recipes.recipes_tag_rating
<b>DDL statement</b>	CREATE TABLE recipes.recipes_tag_rating ( tag text, id int, avg_rating float, name text, PRIMARY KEY (tag, avg_rating, id) ) WITH CLUSTERING ORDER BY (avg_rating DESC, id ASC);
<b>Screenshot</b>	 <pre> tag   avg_rating   id   name ----- ----- ----- ----- spanish   5   464443   viva madrid red wine pear and blue cheese toasts summer   5   112632   creme de menthe chocolate vanilla drink summer   5   126940   easy veggie spaghetti summer   5   360607   simple strawberry trifle summer   5   377937   haricots verts salad </pre>

<b>Keyspace</b>	recipes.recipes_tag_submitted
<b>DDL statement</b>	CREATE TABLE recipes.recipes_tag_submitted ( tag text, submitted date, avg_rating float, id int, name text, PRIMARY KEY (tag, submitted, avg_rating, id) ) WITH CLUSTERING ORDER BY (submitted DESC, avg_rating DESC, id ASC);

Screenshot	<pre> tag   submitted   avg_rating   id   name ----- spanish   2011-09-19   5   464443   viva madrid red wine pear and blue cheese toasts summer   2009-06-20   5   377937   haricots verts salad summer   2009-03-12   5   360607   simple strawberry trifle summer   2005-06-21   5   126940   easy veggie spaghetti summer   2005-03-04   5   112632   creme de menthe chocolate vanilla drink </pre>
------------	--

Keyspace	recipes.recipes_keywords
DDL statement	<pre> CREATE TABLE recipes.recipes_keywords (   id int,   keywords set&lt;text&gt;,   avg_rating float,   name text,   PRIMARY KEY (id, avg_rating, name) ) WITH CLUSTERING ORDER BY (avg_rating DESC, name ASC); </pre>
Screenshot	<pre> id   avg_rating   name   keywords ----- 86683   4   leah chase s green beans with ham and potatoes   {'and', 'beans', 'chase', 'green', 'ham', 'leah', 'potatoes', 's', 'with'} 145970   5   holiday morning french toast   {'french', 'holiday', 'morning', 'toast'} 118947   5   peaches and cream in phyllo   {'and', 'cream', 'in', 'peaches', 'phyllo'} 304570   5   sunshine punch   {'punch', 'sunshine'} 176990   4   oven roasted mussels   {'mussels', 'oven', 'roasted'} </pre>

Keyspace	recipes.recipes_details
DDL statement	<pre> CREATE TABLE recipes.recipes_details (   id int,   name text,   minutes int,   contributor_id int,   submitted date,   tags set&lt;text&gt;,   nutrition list&lt;float&gt;,   n_steps smallint,   steps list&lt;text&gt;,   description text,   ingredients set&lt;text&gt;,   n_ingredients smallint,   avg_rating float,   difficulty text,   keywords set&lt;text&gt;,   PRIMARY KEY (name) ); </pre>

Screenshot	<p>The screenshot shows a SQL query result for a recipe named 'sunshine punch'. The query is: <code>takeout@lghio select * from recipes_recipes_details limit 5;</code>. The result is a table with columns: <code>name</code>, <code>avg_rating</code>, <code>contributor_id</code>, <code>description</code>, <code>difficulty</code>, <code>id</code>, <code>ingredients</code>, <code>keywords</code>, <code>minutes</code>, <code>n_ingredients</code>, <code>n_steps</code>, <code>nutrition</code>, <code>steps</code>, <code>submitted</code>, and <code>tags</code>. The row for 'sunshine punch' has an average rating of 5.0, was submitted on 2012-05-12, and is categorized as 'Hard'. The description mentions it's a family favorite and includes ingredients like 'frozen limeade concentrate' and 'frozen orange juice concentrate'. The steps section describes how to make the punch by combining ingredients and chilling it.</p>
------------	---

### Ερώτημα 3: Απαντήσεις ερωτημάτων

Ερώτημα	Απάντηση
Εμφάνιση των 30 συνταγών με την υψηλότερη μέση βαθμολογία μεταξύ 01/01/2012 και 31/05/2012	<pre>SELECT * FROM popular_recipes WHERE submitted &gt;= '2010-01-01' AND submitted &lt;= '2012-05-31' ALLOW FILTERING submitted avg_rating id name 2011-11-19 5.0 468511 caramelized onion tartlets 2010-01-19 5.0 409298 creamy cheesy scrambled eggs with basil 2011-09-19 5.0 464443 viva madrid red wine pear and blue cheese toasts 2010-03-23 5.0 417697 rosemary basil chicken 2011-07-23 5.0 460738 s mores cookies</pre>
Εμφάνιση όλων των λεπτομερειών για την ταινία «curried bean salad» (κατηγορία δυσκολίας, διατροφικές αξίες, βήματα, περιγραφή, μέση βαθμολογία)	<pre>SELECT * FROM recipes_details WHERE name = 'curried bean salad'; {   'avg_rating': 5.0,   'contributor_id': 300249,   'description': 'serve this flavorful and refreshing salad as a main dish or '     'side dish. we're a family of spice wimps, so it works nicely '     'with just mildly spiced. (don't tell that it's healthy, '     'vegan, low fat, and probably gluten free.)',   'difficulty': 'Easy',   'id': 429010,   'ingredients': SortedSet(['black beans', 'cooked brown rice', 'creamed corn', 'diced tomatoes', 'dried cilantro', 'garbanzo beans', 'ginger paste', 'lemon juice', 'mild curry powder', 'onion', 'raisins', 'rice cakes']),   'keywords': SortedSet(['bean', 'curried', 'salad']),   'minutes': 20,   'n_ingredients': 12,   'n_steps': 4,   'name': 'curried bean salad',   'nutrition': [256.0, 2.0, 40.0, 18.0, 18.0, 1.0, 18.0],   'steps': ['drain &amp; rinse beans',     'stir all ingredients together',     'i microwave it until it simmers , about 9 minutes , but you can '     'probably skip this step',     'chill &amp; serve'],   'submitted': Date(14760),   'tags': SortedSet(['30-minutes-or-less', 'beans', 'black-beans', 'chick-peas-garbanzos', 'course', 'curried', 'dietary', 'free-of-something', 'gluten-free', 'healthy', 'healthy-2', 'low-calorie', 'low-cholesterol', 'low-fat', 'low-in-something', 'low-saturated-fat', 'main-dish', 'main-ingredient', 'preparation', 'salad', 'time-to-make', 'vegan', 'vegetarian'])}</pre>
Εμφάνιση των ταινιών της κατηγορίας «εύκολη» ταξινομημένες ως προς τη μέση βαθμολογία τους	<pre>SELECT * FROM recipes_difficulty WHERE difficulty = 'Easy'; difficulty avg_rating id name 0 Easy 5.000000 banana walnut oatmeal 119510 1 Easy 5.000000 barefoot contessa pork loin ina garten 170396 2 Easy 5.000000 creamy cheesy scrambled eggs with basil 409298 3 Easy 5.000000 creme de menthe chocolate vanilla drink 112632 4 Easy 5.000000 curried bean salad 429010</pre>
Εμφάνιση των συνταγών που περιέχουν την ετικέτα “course” με ταξινόμηση	<pre>SELECT * FROM recipes_tag_submitted WHERE tag = 'course'; tag submitted avg_rating id name 0 course 2016-08-07 5.000000 527825 frozen soufflé amaretto windows on the world 1 course 2014-02-10 5.000000 513170 mini chicken breast sliders 2 course 2013-11-03 5.000000 508830 turkish coffee mocha 3 course 2012-06-03 0.000000 480377 szechuan noodles raw vegan 4 course 2011-11-19 5.000000 468511 caramelized onion tartlets</pre>



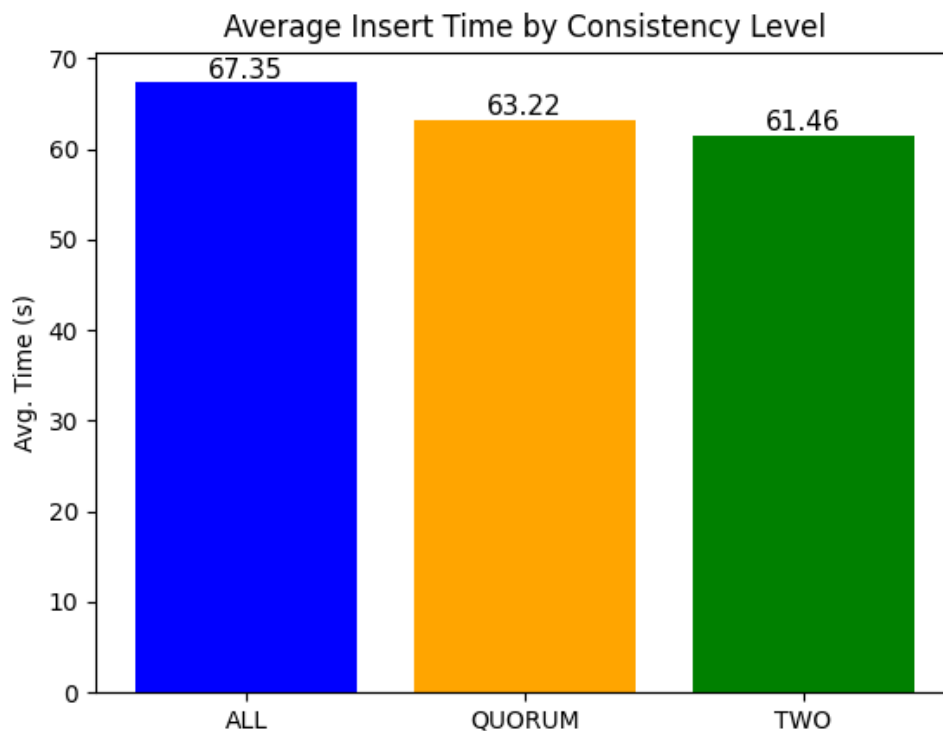
ανά ημερομηνία προσθήκης (πιο πρόσφατες πρώτα)																															
Εμφάνιση των 20 συνταγών με την υψηλότερη μέση βαθμολογία για την ετικέτα “30-minutes-or-less”.	<pre>SELECT * FROM recipes_tag_rating WHERE tag = '30-minutes-or-less' LIMIT 20;</pre> <table><thead><tr><th></th><th>tag</th><th>avg_rating</th><th>id</th><th>name</th></tr></thead><tbody><tr><td>0</td><td>30-minutes-or-less</td><td>5.000000</td><td>294579</td><td>hot mulled cider aprs ski slammer</td></tr><tr><td>1</td><td>30-minutes-or-less</td><td>5.000000</td><td>360607</td><td>simple strawberry trifle</td></tr><tr><td>2</td><td>30-minutes-or-less</td><td>5.000000</td><td>374447</td><td>upside down popover pizza</td></tr><tr><td>3</td><td>30-minutes-or-less</td><td>5.000000</td><td>377937</td><td>haricots verts salad</td></tr><tr><td>4</td><td>30-minutes-or-less</td><td>5.000000</td><td>417697</td><td>rosemary basil chicken</td></tr></tbody></table>		tag	avg_rating	id	name	0	30-minutes-or-less	5.000000	294579	hot mulled cider aprs ski slammer	1	30-minutes-or-less	5.000000	360607	simple strawberry trifle	2	30-minutes-or-less	5.000000	374447	upside down popover pizza	3	30-minutes-or-less	5.000000	377937	haricots verts salad	4	30-minutes-or-less	5.000000	417697	rosemary basil chicken
	tag	avg_rating	id	name																											
0	30-minutes-or-less	5.000000	294579	hot mulled cider aprs ski slammer																											
1	30-minutes-or-less	5.000000	360607	simple strawberry trifle																											
2	30-minutes-or-less	5.000000	374447	upside down popover pizza																											
3	30-minutes-or-less	5.000000	377937	haricots verts salad																											
4	30-minutes-or-less	5.000000	417697	rosemary basil chicken																											

## Ερώτημα 4Α: Χρόνοι εισαγωγής δεδομένων

**ΣΗΜΕΙΩΣΗ:** Η AstraDB απαγορεύει inserts με consistency level ONE επειδή τα θεωρεί κακή πρακτική. Γι αυτό τον λόγο εμείς τα εκτελέσαμε με consistency level TWO και παραθέτουμε τα αποτελέσματά μας.

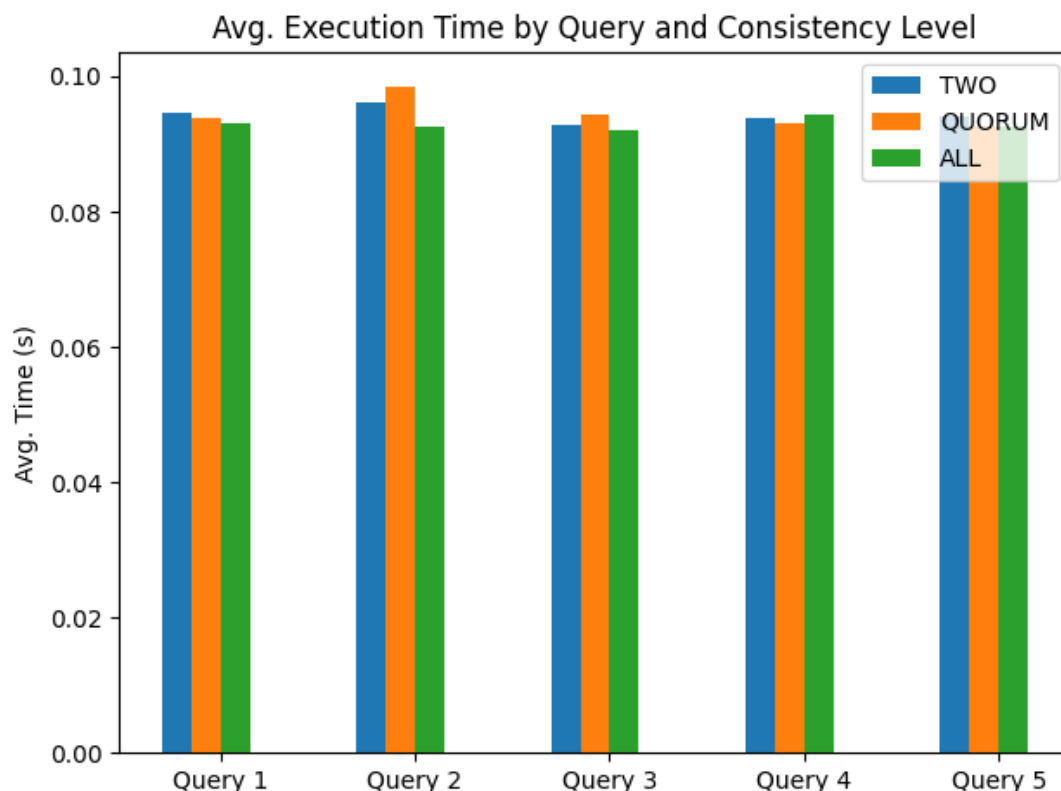
Εδώ κάναμε απευθείας τον μέσο όρο των insert operations στον κώδικα και τα παραθέτουμε απευθείας.

	Επίπεδο write consistency		
	ALL	QUORUM	TWO
<b>recipes</b>	61.46	63.22	67.35
<b>Μέσος όρος</b>	61.46	63.22	67.35



## Ερώτημα 4B: Χρόνοι ανάκτησης δεδομένων

	Επίπεδο write consistency		
	ALL	QUORUM	TWO
Ερώτημα 1	0.0948	0.0930	0.0929
Ερώτημα 2	0.0921	0.0928	0.0979
Ερώτημα 3	0.0927	0.0939	0.0945
Ερώτημα 4	0.0935	0.1002	0.0922
Ερώτημα 5	0.0939	0.0924	0.0919
Μέσος όρος	0.0934	0.0945	0.0939



## Ερώτημα 4Γ: Σχολιασμός αποτελεσμάτων

Στα ερωτήματα εγγραφής βλέπουμε ξεκάθαρη ανοδική τάση όσο το consistency level γίνεται αυστηρότερο. Αυτό είναι εντελώς αναμενόμενο, αφού αυξάνονται τα μηνύματα acknowledge που ανταλλάσσουν μεταξύ τους οι server της βάσης πρώτου βεβαιώσουν ένα write. Αυτό στην ουσία είναι το tradeoff του θεωρήματος CAP, σύμφωνα με το οποίο όσο αυξάνουμε το Consistency θα πρέπει να αναμένουμε μειωμένο Availability (κρατώντας σταθερό το Partition Tolerance). Στα ερωτήματα ανάγνωσης βλέπουμε παρόμοιο trend, με λίγη αστάθεια, ίσως λόγω αυξημένου traffic στους server της AstraDB.

## Βιβλιογραφία

- DataStax Cassandra Query Language Documentation:  
<https://docs.datastax.com/en/cql-oss/3.x/cql/cqlIntro.html>

- Pandas python library documentation: <https://pandas.pydata.org/docs/index.html>
- Drawio diagram tool: <https://app.diagrams.net/>