Συστήματα Διαχείρισης Δεδομένων Μεγάλου Όγκου

Εργαστηριακή Άσκηση 2021/22

|  |  |  |
| --- | --- | --- |
| **Όνομα** | **Επώνυμο** | **ΑΜ** |
| Αλέξανδρος-Παναγιώτης | Σοϊλεμεζίδης | 1067522 |
| Κυριακή | Λιούμη | 1067410 |

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

Αλέξανδρος-Παναγιώτης Σοϊλεμεζίδης

4 / 7 / 2022

Υπογραφή

Κυριακή Λιούμη

4 / 7 / 2022

**Συνημμένα αρχεία κώδικα**

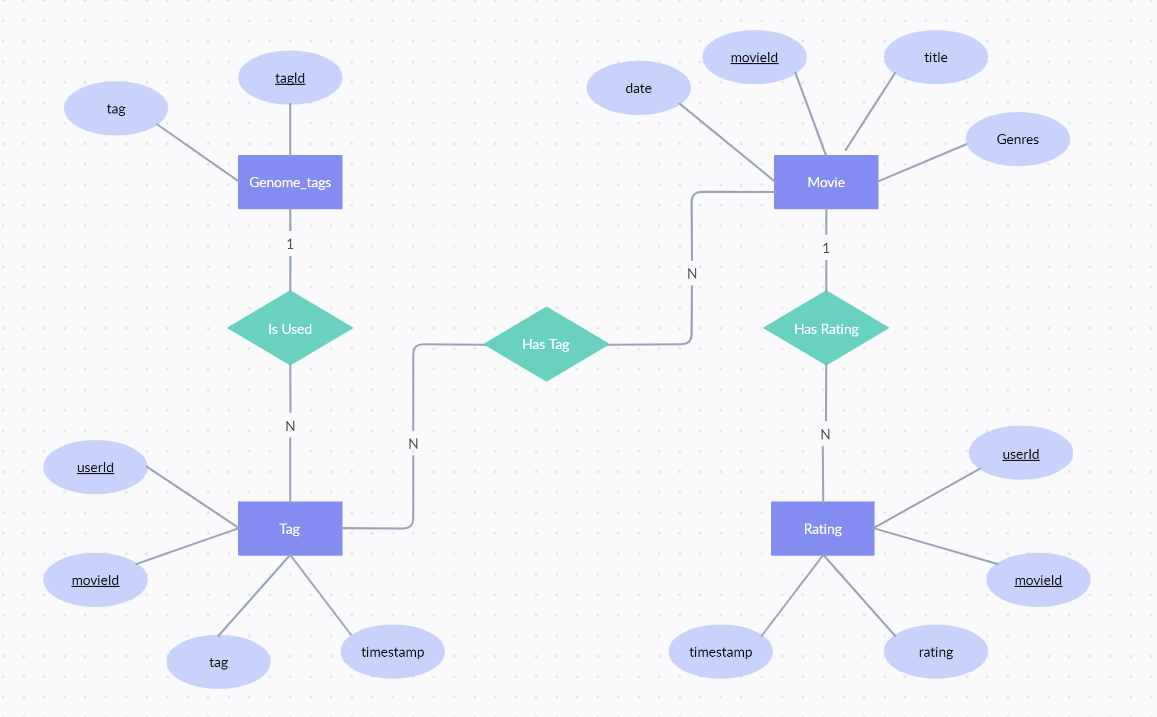
Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

|  |  |  |
| --- | --- | --- |
| **Αρχείο** | **Αφορά το ερώτημα** | **Περιγραφή/Σχόλιο** |
| connect\_db.py | 2,3 | Σύνδεση με cassandra |
| load.py | 2,3 | Φόρτωση των csv, επεξεργασία και δημιουργία νέων αρχείων, τα οποία περιλαμβάνουν ό,τι χρειάζεται για κάθε query |
| main.py | 2,3 | Δημιουργία table για κάθε query και insert |

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

Τα τρέξαμε σε Astra DB

Ερώτημα 1: Σχεδιασμός ΒΔ

Εννοιολογικό Μοντέλο

Εξηγήσεις: Φτιάξαμε 4 οντότητες, όσα και τα csv που μας δόθηκαν, με τα γνωρίσματα που περιέχονται σε κάθε αρχείο csv. Τα κλειδιά φαίνονται με υπογράμμιση. Για τις συσχετίσεις έχουμε ότι 1 ταινία θα έχει πολλά (Ν) ratings, 1 genome tag έχει πολλά (Ν) tags και πολλές (Ν) ταινίες έχουν πολλά (Ν) tags.

Application Workflow

Εικόνα που περιέχει κείμενο, ουρανός, στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

Εξηγήσεις: Χωρίσαμε το query 3 σε δύο μέρη. Στο Q3.1 έχουμε τις ταινίες ανά κατηγορία, ταξινομημένες με βάση την μέση βαθμολογία και στο Q3.2 έχουμε τις ταινίες ανά κατηγορία, ταξινομημένες με βάση το έτος που βγήκαν. Καθώς το Q4 μας φάνηκε το πιο ειδικό από όλα, κάναμε τις συνδέσεις όπως φαίνεται παραπάνω.

Chebotko Diagram

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, εσωτερικό

Περιγραφή που δημιουργήθηκε αυτόματα

Εξηγήσεις: Για κάθε query έχουμε φτιάξει έναν πίνακα. Όπου ‘Κ ’ είναι το κλειδί και όπου ‘C’ είναι η ταξινόμηση που χρησιμοποιούμε. DESC για φθίνουσα και ASC για αύξουσα. Οι συνδέσεις εξηγούνται παραπάνω.

Ερώτημα 2: Ερωτήματα DDL

|  |  |
| --- | --- |
| **Keyspace** | Movies |
| **DDL statement** |  |
| **Screenshot** |  |

Ερώτημα 3: Απαντήσεις ερωτημάτων

|  |  |
| --- | --- |
| **Ερώτημα** | **Απάντηση** |
| Εμφάνιση των 30 ταινιών με την υψηλότερη μέση βαθμολογία μεταξύ 01/01/2015 και 15/01/2015 |  |
| Εμφάνιση όλων των λεπτομερειών για την ταινία Jumanji (κατηγορία, μέση βαθμολογία, top-5 ετικέτες |  |
| Εμφάνιση των ταινιών της κατηγορίας “adventure” ταξινομημένες ως προς το έτος παραγωγής |  |
| Εμφάνιση των ταινιών που περιέχουν τη λέξη “star” |  |
| Εμφάνιση των 20 ταινιών με την υψηλότερη μέση βαθμολογία για την ετικέτα “comedy” |  |

Βιβλιογραφία

* <https://cassandra.apache.org/doc/latest/cassandra/data_modeling/data_modeling_queries.html>
* <https://cassandra.apache.org/doc/latest/cassandra/data_modeling/data_modeling_logical.html>
* <https://cassandra.apache.org/_/quickstart.html>