ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ
ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2021-2022

ΡΗΓΑΤΟΣ ΔΙΟΝΥΣΙΟΣ - P3200262

ΠΑΠΑΠΟΣΤΟΛΟΥ ΧΡΙΣΤΟΦΟΡΟΣ - P3150208

2η Εργασία

Δημιουργία Αρχικών Πινάκων

Αρχικά , δημιουργήσαμε μια νέα βάση “MovieLens” στο Postgres Cloud Instance μας στο Azure. Έπειτα, αφού κατεβάσαμε το dataset, χρησιμοποιήσαμε το python script “gen_ddl_python3.py", το οποίο μας έδωσε τα .sql αρχεία για τη δημιουργία των περισσότερων πινάκων. Κάναμε τις απαιραίτητες αλλαγές για την εισαγωγή Primary Keys και με την βοήθεια του pgAdmin τρέξαμε τα .sql αρχεία και δημιουργήσαμε τους πίνακες στην βάση MovieLens.

Εξαγωγή JSON και δημιουργία CSV

Το επόμενο βήμα ήταν να εξάγουμε τα εμφωλευμένα JSON και να δημιουργήσουμε τα αρχεία CSV που θα αντιστοιχούν στους πίνακες keyword και movie_keyword. Σε αυτό το σημείο αναπτύξαμε ένα δικό μας python script, στο οποίο διαβάζαμε τα εμφωλευμένα JSON στο CSV που μας δόθηκε στο dataset και με κατάλληλη επεξεργασία τα εξάγαμε και γράψαμε τις πληροφορίες στα αρχεία keyword.csv και movie_keyword.csv. Τέλος, με χρήση αλλη μια φορά του gen_ddl_python3.py , πήραμε τα .sql αρχεία για την δημιουργία των πινάκων keyword και Movie_keyword και μετά τις κατάλληλες αλλαγές για primary keys , δημιουργήσαμε τους πίνακες στην βάση.

Αφαίρεση Διπλοτύπων και δημιουργία Foreign Keys

Επειδή το αρχείο keywords περιείχε διπλότυπα, πριν περάσουμε τα δεδομένα στον πίνακα, κάναμε χρήση μεσάζοντα πίνακα (χωρίς primary keys) απο τον οποίο, με τη χρήση του query που φάινεται απο κάτω , αφαιρέσαμε τα διπλότυπα και φορτώσαμε τα δεδομένα στον πίνακα keyword. Τέλος, δημιουργήσαμε τα Foreign Keys στους πίνακες που έπρεπε και ελέγξαμε την ορθότητα των δεδομένων μας με μερικά COUNT queries.

```
delete from keywordf
where ctid in
(select m1.ctid
from keywordf m1
natural join keywordf m2
where m1.ctid > m2.ctid)
```

Κώδικας Python

Σε αυτό το σημείο θα εξηγήσουμε αναλύτικα την υλοποίηση του json_extract.py , το script που δημιουργήθηκε για την εξαγωγή εμφωλευμένον JSON. Αρχικά, δημιουργούμε ένα αντικείμενο csvreader για την ανάγνωση των εμφωλευμένων JSON από το αρχείο keywords.csv, δημιουργούμε επίσης δύο αντικείμενα csvwriter για την εγγραφή των keyword.csv και movie_keyword.csv.

Για κάθε γραμή που διαβάζουμε , παίρνουμε το movie_id απο το πρώτη στήλη και το γράφουμε στο νέο αρχείο με το csvwriter. Από τη δεύτερη στήλη, με τη βοήθεια της βιβλιοθήκης ast , παίρνουμε ενα list απο dictionaries. Κάθε dictionary περιέχει json πληροφορία, που πρέπει να εξαχθεί στην ίδια γραμμή. Άρα κάθε γραμμή στα νέα csv αντιστοιχεί σε ένα dictionary απο τη λίστα.

```
for row in csvreader:
    movie_id = row[0]
    data = ast.literal_eval(row[1])
      for dictionary in data:
          keyword_id = dictionary['id']
          name = dictionary['name']
          csvwriter.writerow([keyword_id, name])
```