ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

ΣΥΝΤΟΜΗ ΑΝΑΦΟΡΑ

- ΑΘΑΝΑΣΙΑ ΜΑΡΙΑ ΠΑΠΑΘΑΝΑΣΙΟΥ (p3180147)
- ΑΙΚΑΤΕΡΙΝΗ ΟΡΟΥΤΖΟΓΛΟΥ (p3180135)

Για την υλοποίηση του μέρους Α της εργασίας ακολουθήσαμε τα παρακάτω βήματα:

- ➔ Δημιουργίας νέας βάσης για αποθήκευση των δεδομένων στο aws.

- ➔ Δημιουργήσαμε τους πίνακες δίνοντας κατάλληλα ονόματα στις στήλες τους, καθώς και τους κατάλληλους τύπους δεδομένων για την κάθε στήλη και για να τα εισάγουμε στη βάση χρησιμοποιήσαμε την εντολή: \i path_to_file/create_file.sql

- ➔ Για την αποφυγή προβλημάτων με την κωδικοποίηση των χαρακτήρων τρέξαμε την εντολή set client_encoding to ‘utf8’ για κάθε πίνακα για τον οποίο θέλαμε να εισάγουμε τα δεδομένα του. Για την εισαγωγή δεδομένων από το csv αρχείο στους πίνακες της βάσης μας χρησιμοποιήσαμε την εντολή \copy Table_name FROM 'path_to_file/file.csv' DELIMITER ',' CSV HEADER.

- ➔ Στην εργασία μας ζητήθηκε να διαγράψουμε τα διπλότυπα από όλους τους πίνακες(εκτός του πίνακα ratings_small). Για να γίνει αυτό, πρέπει αρχικά να βρούμε ποιοί πίνακες περιέχουν διπλότυπα. Για την εύρεση διπλότυπων μετρήσαμε τις γραμμές που εμφανίζονται πάνω από μία φορά, χρησιμοποιώντας το group by με having count(*)>1.

- ➔ Οι πίνακες στους οποίους εντοπίσαμε διπλότυπα είναι ο credits, keywords, movies_metadata. Για τη διαγραφή των διπλοτύπων χρησιμοποιήσαμε έναν ενδιάμεσο πίνακα στον οποίο εισάγαμε τα στοιχεία των αρχικών πινάκων που είχανε μοναδικό id. Διαγράψαμε τον αρχικό πίνακα και μετονομάσαμε τον ενδιάμεσο πίνακα όπως και ο αρχικός.

- ➔ Έπειτα, μας ζητήθηκε να διαγράψουμε τα δεδομένα ταινιών που δεν υπήρχαν στον πίνακα movies_metadata. Για να γίνει αυτό δημιουργήσαμε έναν ενδιάμεσο πίνακα ως σύζευξη του κάθε αρχικού πίνακα με το movies_metadata όπου το id του αρχικού πίνακα ήταν ίδιο με το id του movies_metadata. Διαγράψαμε τον αρχικό πίνακα και μετονομάσαμε τον ενδιάμεσο πίνακα όπως και ο αρχικός.

- ➔ Στη συνέχεια προσθέσαμε στους πίνακες τα κατάλληλα primary keys, δηλαδή τα id και movieId που ήταν μοναδικά. Ο πίνακας ratings_small επειδή περιέχει διπλότυπα δεν μπορεί να έχει primary key. Ακόμη προσθέσαμε foreign keys σε πίνακα για να συνδέσουμε τους πίνακες μεταξύ τους.