ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ – ΑΝΑΦΟΡΑ PROJECT 2020

ΤΩΝ:

Δεληγιάννης Αντώνιος – 3180038

Παναγιώτου Παναγιώτης – 3180139

Τσιομπίκας Δημήτριος – 3180223

1)Γενικά

Το project είχε να κάνει με την προσομοίωση μιας βάσης δεδομένων της ιστοσελίδας IMBd (Internet Movie Database) στο δικό μας Postgres server με τη χρήση SQL. Οι πληροφορίες που χρειαζόμασταν μας δόθηκαν σε .csv αρχεία από τα οποία φτιάξαμε τους απαραίτητους Πίνακες και προσθέσαμε περιορισμούς κλειδιού, ωστέ να γίνεται δυνατή η σύνδεση τους και το τρέξιμο διαφόρων query για την λήψη στατιστικών στοιχείων. Πριν, από όλα αυτά όμως, έπρεπε να ακολουθήσουμε κάποια επιπλέον βήματα, ώστε να φροντίσουμε πως δεν ύπαρχουν διπλότυπα που θα μας περιορίζαν σημαντικά.

2)Διαδικάσια Στησίματος Βάσης

Ας σημειωθεί πως δεν εγίνε καμία επεξεργασία των .csv αρχείων με τη χρήση Excel ή οποιουδήποτε άλλου προγραμμάτος και πως όλες οι αλλαγές στα δεδομένα έγιναν αφού αυτά ανέβηκαν στη βάση (πολλές φορές το Excel μπορεί να χαλάσει τη μορφή των μεγαλών .csv αρχείων ή να μη τα φορτώσει καν!).

Πρώτο βήμα ήταν να φτιάξουμε 5 προσωρινούς πίνακες , με τις στήλες που θα έχουν οι κανονικοί μας αλλά με το ονομα τους να τελείωνει σε ‘Temp’ π.χ. “CreditsTemp”. Σε αυτούς τους πίνακες προσθέσαμε τα δεδομένα των .csv με το γνωστό μας τρόπο. Η μεταφόρτωση τους έγινε επιτυχώς, και αυτό δείχνει πως διαλέξαμε τους κατάλληλους τύπους δεδομένων για τις στήλες των πινάκων.

Στο επόμενο βήμα , αρχίσαμε να φτιάχνουμε τους κανονικούς πίνακες μας , αυτή τη φορά το όνομα τους ήταν το σωστό, π.χ. “Credits” , τους οποίους γεμίσαμε με τα δεδομένα των προσωρινών πινάκων. Η χρήση της εντολής DISTINCT είναι η λύση στο πρόβλημα του να μην περάσουν διπλότυπες πλειάδες.

Έπειτα έπρεπε να σιγουρέψουμε πως τα δεδομένα των πινάκων: “Credits”, “Keywords”, “Links”. “Ratings” δεν αντιστοιχούν σε ταίνιες που δεν υπάρχουν στον πίνακα “Movies_Metadata”. Οπότε χρησιμοποιήσαμε την εντολή:
Πχ:`DELETE FROM "Table_Name" WHERE id_column NOT IN (SELECT m.id FROM "Movies_Metadata" m)`

Όμως τα παραπάνω δεν αρκούσαν , καθώς υπήρχαν ακόμα κάποιες πλειάδες που είχαν ίδιες όλες τις στήλες εκτός από μια (πιθανώς λάθος του IMDb…) και αυτό εντοπίστηκε στην προσπάθεια μας να προσθέσουμε primary keys. Υπήρχαν κάμποσες πλειάδες στους πίνακες “Movies_Metadata” και “Credits” που είχαν αυτό το πρόβλημα, και εμείς το λύσαμε βάζοντας συνθήκη όταν σβήνουμε για να να παραμείνει μόνο μία πλειάδα για το κάθε id.
Πχ:`DELETE FROM "Credits" T1 USING "Credits" T2 WHERE T1.crew < T2.crew AND T1.id=T2.id;`
Εδώ κρατάμε την πλειάδα T2, που έχει μεγαλύτερο crew (δηλαδή περισσότερο μήκος συμβολοσειράς στο crew)από το T1 , αφού αναφέρονται στην ίδια ταίνια. Η επιλογή έγινε βάση ρεαλισμού.

Τέλος, διαγράψαμε τους προσωρινούς πίνακες μας και πραγματοποιήσαμε την προσθήκη των Primary και Foreign Key επιτυχώς. Πλέον οι πίνακες συνδέονται όπως φαίνεται στο ER διάγραμμα μας.

- Σημείωση 1: Σε κάθε εντολή που αφορά τους πίνακας μας βάζουμε εισαγωγικά (“”) στο όνομα τους , π.χ. “Movies_Metadata”, για να είναι case sensitive.
- Σημείωση 2: Το “insight” που παίρνουμε από τον πίνακα view_table για το δεύτερο μέρος της εργασίας είναι γραμμένο κάτω κάτω ως σχόλιο στο partB.sql.
- Σημειωση 3: Χρησιμοποιήσαμε το ratings_small.csv λόγω αργής σύνδεσης στο διαδίκτυο.