ΑΝΑΦΟΡΑ ΔΙΑΔΙΚΑΣΙΑΣ ΑΠΟΘΗΚΕΥΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΒΑΣΗ

Αρχικά για την προπεξεργασία των δεδομένων χρησιμοποιήσαμε γλώσσα python και την βιβλιοθήκη pandas για να κάνουμε χρήση dataframes.Εκεί, για κάθε αρχείο που διαβάσαμε(εκτός του ratings) με την εντολή drop_duplicates() διαγράψαμε τα διπλότυπα και έπειτα με την εντολή isin ελέγξαμε αν όλες οι ταινίες των πινάκων βρίσκονταν και στον Movies_metadata αλλιώς τις διαγράφαμε.Τέλος,επιστρέφαμε το new αρχείο το οποίο θα έπρεπε να μπει και στην βάση μας.

Στα καινούργια αρχεία μας δημιουργήσαμε τα αρχεία sql για τα create tables μέσω του προγράμματος python gen_ddl_python3.Έπειτα,ελέγξαμε αυτά τα αρχεία διότι κάποιοι τύποι είχαν αλλαχθεί λόγω του dataframe και τους διορθώσαμε.(π.χ το Id στο credits ήταν Varchar αντί για integer).Στην διαδικασία εισαγωγής τους,χρειάστηκε να σβήσουμε απο το credits κάποιες κενές στήλες που προυπήρχαν.Επίσης,χρειάστηκε να σβήσουμε την πρώτη στήλη απο κάθε πίνακα καθώς ήταν και αυτές δημιουργήματα του dataframe που δεν μας χρειαζόντουσαν και δουλεύουμε με την ratings_small λόγω προβλημάτων μεγέθους.Τέλος,σε μεμονωμένα 2-3 περιστατικά σβήσαμε 2-3 σειρές απο πίνακες που εμφάνιζαν ακατανοήτα errors κατά την εισαγωγή τους στην βάση.