2η Εργασία: Airbnb Staging Tables

Προθεσμία: Πέμπτη 14/03/2019 17:00

Σκοπός:

Σε αυτή την εργασία θα φτιάξουμε μια βάση δεδομένων για την αποθήκευση και εξερεύνηση δεδομένων που προέρχονται από την Airbnb. Η βάση δεδομένων θα αποτελείται από μία συλλογή πινάκων και τις μεταξύ τους σχέσεις. Για να ορίσουμε το σχήμα της θα βασιστούμε στη μορφή των δεδομένων εισόδου, το οποίο στη συνέχεια θα απεικονίσουμε σε ένα διάγραμμα ER. Θα δημιουργήσουμε τους πίνακες χρησιμοποιώντας SQL και θα εισάγουμε δεδομένα σε αυτούς με την εντολή `\copy`. Επίσης, θα δημιουργήσουμε περιορισμούς ξένου κλειδιού για αναφορική ακεραιότητα.

Δεδομένα:

Κατεβάστε το Airbnb dataset από παρακάτω link:
https://drive.google.com/file/d/1rT7fhzgTObuM_KRipCziW3CbI6qcsSzy/view?usp=sharing

Τι θα φτιάξουμε:

- Ένα διάγραμμα ER με τους Airbnb πίνακες, τις σχέσεις και τις ιδιότητές τους.
- Τη βάση δεδομένων Airbnb σε ένα Postgres Cloud instance.
- Η βάση αυτή θα πρέπει να περιέχει πίνακες για τους οποίους θα ισχύουν τα εξής:
    - κάθε πίνακας να αντιστοιχεί σε ένα αρχείο .csv του Airbnb dataset.
    - να έχουν εισαχθεί σε αυτόν τα αντίστοιχα δεδομένα.
    - να περιέχει περιορισμούς πρωτεύοντος και ξένου κλειδιού.

Απαραίτητα εργαλεία:

- Draw.io
- AWS RDS Postgres instance
- Postgres psql client / pgAdmin

Οδηγίες:

- Το πρώτο γράμμα του ονόματος κάθε πίνακα να ξεκινάει με κεφαλαίο(π.χ. Calendar κτλ.).

- Τα ονόματα των πεδίων των πινάκων (attributes) να ξεκινάνε με μικρό (π.χ. host_id κτλ.).
- Τοποθετήστε κάθε εντολή ***`create table`*** σε ξεχωριστό αρχείο. Για παράδειγμα create_calendar.sql.
- **Αν δουλεύετε στο terminal με psql**, μπορείτε να φτιάξετε ένα script create_tables.sql που να καλεί τα ξεχωριστά **`create table`** αρχεία χρησιμοποιώντας την εντολή **`\i`** που αναφέρεται παρακάτω.
- Τοποθετήστε όλα τις εντολές **`alter table`** σε ένα αρχείο, alter_tables.sql.
- Το διάγραμμα ER να γίνει στο draw.io και να σωθεί στον υπολογιστή σας σαν εικόνα. Το όνομα του αρχείου να είναι airbnb_ERD.png.

Συμβουλές για την υλοποίηση:

- Συνδεθείτε στη βάση σας στο AWS για να δημιουργήσετε τους πίνακες που ζητά η άσκηση.
- Σε κάθε πίνακα που φτιάχνετε να ορίζετε πρωτεύον κλειδί.
- Προσθέστε τον περιορισμό πρωτεύοντος κλειδιού στην εντολή **`create table`**.
- Επειδή μερικές εντολές **`create table`** έχουν πολλά πεδία, όπως αυτή για τον πίνακα Listing, μπορείτε να τρέξετε το python πρόγραμμα **`gen_ddl_python3.py`** (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 3 στον υπολογιστή σας ή το **`gen_ddl_python2.py`** (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 2, το οποίο παίρνει ως παράμετρο το .csv αρχείο των δεδομένων, π.χ. listings.csv για τον πίνακα Listing, και παράγει ένα αρχείο .sql με την εντολή **`create table`** για τον αντίστοιχο πίνακα. Ελέγξτε την παραγόμενη εντολή και προσθέστε τους περιορισμούς για πρωτεύοντα κλειδιά.
- Χρησιμοποιήστε την εντολή **`\i <filename>`** στην psql για να εκτελέσετε τον κώδικα SQL που έχετε αποθηκεύσει σε ένα αρχείο. Για παράδειγμα **`\i create_tables.sql.`** Εναλλακτικά, στο pgAdmin επιλέξτε τη βάση, πατήστε το query tool (κεραυνός πάνω αριστερά) και τρέξτε ένα sql script ως εξής: πατήστε το “Open file” (εικονίδιο φακέλου πάνω αριστερά στο query tool), επιλέξτε το sql script από τον υπολογιστή σας και πατήστε τον “κεραυνό” στη μπάρα του query tool.
- Χρησιμοποιήστε την εντολή **`\copy`** στην psql ή τη λειτουργία **`Import/Export`** στο pgAdmin για να εισάγετε τα δεδομένα. Λάβετε υπόψη σας ότι η πρώτη γραμμή στα .csv αρχεία είναι ο header. Δε θέλουμε να εισάγουμε τον header στον πίνακα. Θέστε την κατάλληλη παράμετρο είτε στο **`\copy`** είτε στην λειτουργία **`Import/Export`** ώστε να προσπεράσετε αυτή τη γραμμή. Παράδειγμα εντολής \copy:
  `\copy Listing FROM 'airbnb/listings.csv' DELIMITER ',' CSV HEADER;`
- Πριν εκτελέσετε την εντολή **`\copy`**, τρέξτε την εντολή **`set client_encoding to 'utf8';`** στην psql για να αποφύγετε προβλήματα με την κωδικοποίηση των χαρακτήρων.

- Προσθέστε τους περιορισμούς ξένου κλειδιού μετά την εισαγωγή των δεδομένων στους πίνακες.
- Για τον πίνακα Listings επιλέξτε ένα μικρό υποσύνολο των πεδίων του (περιέχει περίπου 90) για το ERD διάγραμμα σας.

Χρήσιμα links:

Εντολή create table:
https://www.postgresql.org/docs/9.6/sql-createtable.html

Εντολή copy:
https://www.postgresql.org/docs/9.6/sql-copy.html

Εντολή alter table:
https://www.postgresql.org/docs/9.6/sql-altertable.html

Postgres meta commands, όπως η \copy:
https://www.postgresql.org/docs/9.2/app-psql.html

pgAdmin query tool:
https://www.pgadmin.org/docs/pgadmin4/dev/query_tool.html

pgAdmin import:
https://www.pgadmin.org/docs/pgadmin4/dev/import_export_data.html

Παραδοτέα:

- Δημιουργήστε ένα .txt αρχείο στο οποίο θα αναγράφονται τα εξής στοιχεία: ονοματεπώνυμα και αριθμοί μητρώου των μελών της ομάδας, το endpoint του AWS instance σας (μπορείτε να το βρείτε στο AWS console, *RDS* > *Databases* > *db_identifier* > *Connectivity section*), το όνομα της βάσης σας και το username και το password του χρήστη examiner ή ενός άλλου χρήστη με read-only δικαιώματα, ώστε να μπορούμε να δούμε τους πίνακες της βάσης σας. Το .txt αρχείο θα πρέπει να έχει την παρακάτω μορφή:

<Ονοματεπώνυμο 1> - <A.M. 1>
<Ονοματεπώνυμο 2> - <A.M. 2>
Endpoint: <name_of_the_endpoint>
Username: <username>
Password: <password>

Database: <name_of_the_database>

- Βάλτε όλα τα .sql αρχεία, το αρχείο .txt και το αρχείο του διαγράμματος ER σε ένα φάκελο. Το όνομα του φακέλου πρέπει να αποτελείται από τους αριθμούς μητρώου σας χωρισμένους με παύλα, δηλαδή *αριθμός_μητρώου_1-αριθμός_μητρώου_2*. Δημιουργήστε ένα .zip αρχείο αυτού του φακέλου, το οποίο θα έχει το ίδιο όνομα με τον φάκελο.
- Ανεβάστε το .zip αρχείο στο eclass στην ενότητα *Εργασίες / 2η Εργασία*.