Project - Βάσεις Δεδομένων (Μέρος 1)

Προθεσμία: Τετάρτη 17/5/2019 23:59

Σκοπός

Στο project θα ασχοληθούμε με μία ανάλυση και οπτικοποίηση της δραστηριότητας Airbnb σε μερικές πόλεις της Αμερικής. Εκτός από τα Airbnb δεδομένα για την πόλη του Austin που ήδη έχουμε στη βάση από τις προηγούμενες εργασίες, θα χρησιμοποιήσουμε δεδομένα και από τις πόλεις Boston, Denver και Portland. Επίσης, θα έχουμε ένα ακόμα σετ δεδομένων από την εταιρεία Zillow που ασχολείται με μακροπρόθεσμες μισθώσεις ακινήτων στην Αμερική.

Συνδυάζοντας πληροφορία από τις δύο πηγές δεδομένων, μας ενδιαφέρει να απαντήσουμε σε ερωτήματα, όπως πόσα ακίνητα διατίθενται μέσω Airbnb σε κάθε πόλη, πόσα έσοδα παράγουν, αν συμφέρει για έναν ιδιοκτήτη να νοικιάζει το ακίνητό του μακροπρόθεσμα ή βραχυπρόθεσμα και πώς οι δείκτες αυτοί αλλάζουν μέσα στο χρόνο. Επίσης, θα οπτικοποιήσουμε με χρήση κατάλληλων εργαλείων πτυχές των δεδομένων που θεωρούμε ενδιαφέρουσες.

Το project αποτελείται από **δύο μέρη**:

- **Το πρώτο**, του οποίου το παραδοτέο θα είναι μέχρι τις 15/5/2019, ασχολείται με την προεπεξεργασία των δεδομένων ώστε να μπορούν να εισαχθούν στη βάση και τον υπολογισμό μια μετρικής που περιγράφεται παρακάτω.
- **Το δεύτερο** μέρος θα επικεντρώνεται στην οπτικοποίηση ενδιαφέρουσας πληροφορίας που προκύπτει από τα δεδομένα.

Δεδομένα

Τα δεδομένα του project βρίσκονται στο παρακάτω link.
https://drive.google.com/open?id=1xRK26IYn28_yjjQ7eW_6PY2d_bYu9ns_

Δεδομένα Airbnb:
Υπάρχουν 3 σετ δεδομένων Airbnb. Τα δεδομένα αφορούν τις πόλεις Boston, Denver και Portland. Στη βάση σας θα πρέπει να υπάρχουν τα δεδομένα για την πόλη Austin που είχαν δοθεί στην 2η εργασία.

Κάθε σετ δεδομένων περιλαμβάνει τα εξής αρχεία:

```
amenity.csv
calendar_summary.csv
```

```
calendar.csv
host.csv
listing.csv
listing2amenity.csv
neighborhood.csv
review.csv
summary_listing.csv
summary_review.csv
```

Δεδομένα Zillow:
Το σετ δεδομένων Zillow περιλαμβάνει τα εξής αρχεία:

```
Zip_MedianRentalPrice_1Bedroom.csv
Zip_MedianRentalPrice_2Bedroom.csv
Zip_MedianRentalPrice_3Bedroom.csv
Zip_MedianRentalPrice_4Bedroom.csv
Zip_MedianRentalPrice_5BedroomOrMore.csv
```

Τα αρχεία αυτά διατίθεται σε δύο μορφές: στη μία έχουν τις κεφαλίδες των πεδίων στην πρώτη γραμμή, ενώ στην άλλη όχι. Τα αρχεία με τις κεφαλίδες προορίζονται για να ανοιχτούν σε κάποιον text editor και να είναι πιο ευανάγνωστα για τον άνθρωπο. Τα αρχεία χωρίς κεφαλίδες προορίζονται για πιο εύκολη επεξεργασία από κάποιο πρόγραμμα/framework.

Προεπεξεργασία δεδομένων

Τα δεδομένα που μας δίνονται δεν είναι πάντα στη κατάλληλη μορφή για να εισαχθούν στη βάση. Πολλές φορές πρέπει να γράψουμε κώδικα για να τα επεξεργαστούμε και να τα φέρουμε στην κατάλληλη μορφή. Διαδικασίες όπως αυτή είναι γνωστές με τον όρο ETL (Extract, Transform, Load). Μετασχηματισμούς με σκοπό την εισαγωγή τους στη βάση θα εφαρμόσουμε και στα δεδομένα Zillow.

Μετασχηματισμοί και φίλτρα:

- Το πεδίο “RegionName” στα .csv αρχεία πρέπει να μετονομαστεί σε zipcode πριν περαστεί στους Rental_Price πίνακες.
- Τα πεδία year-month στα .csv αρχεία (π.χ. 2015-01, κλπ.) πρέπει να χρησιμοποιηθούν για να κατασκευαστεί ένα data type τη μορφής YYYY-MM-DD (π.χ. 2015-01-01 κλπ.). Χρησιμοποιήστε “01” για τη μέρα στις ημερομηνίες, καθώς η μέρα δεν διατίθενται στα αρχικά δεδομένα.
- Εξάγετε από τα .csv αρχεία μόνο τα δεδομένα που αφορούν τα χρόνια 2016 και μετά και τελειώνουν στο 2018-01. Παραλείψτε τα δεδομένα πριν το 2016.

Τα παραπάνω βήματα μπορείτε να τα γράψετε σε Java ή Python. Αν σας ενδιαφέρει να το ψάξετε παραπάνω, μπορείτε να χρησιμοποιήσετε το framework Apache Beam, το οποίο παρέχει ένα προγραμματιστικό μοντέλο για τη δημιουργία data processing pipelines. Έχει APIs σε Java και Python. Τρέχει σε διαφορετικές μηχανές εκτέλεσης, όπως Apache Flink, Apache Spark κ.α., αλλά και locally. Η εξοικείωση με το framework σίγουρα θα χρειαστεί 1-2 μέρες αλλά θα είναι μία καινούρια γνώση που δεδομένης της απήχησης των big data frameworks είναι πολύ πιθανό ότι θα σας είναι χρήσιμη και μελλοντικά. Επίσης, όσες ομάδες επιλέξουν να δουλέψουν με Beam θα βαθμολογηθούν με διευρυμένη κατά 0.5 μονάδες κλίμακα στο project. Δηλαδή, ενώ υπό κανονικές συνθήκες το project πιάνει 2 από τις 10 μονάδες του μαθήματος, για τις ομάδες με Beam το άριστα θα είναι το 2.5 με την μισή μονάδα bonus (10.5 μονάδες).

Ιδιότητα μοναδικότητας:
Οι εγγραφές των πινάκων πρέπει να είναι μοναδικές. Αυτό συνεπάγεται ότι ο κώδικάς σας θα αφαιρεί τις διπλότυπες εγγραφές πριν την εισαγωγή των δεδομένων στους πίνακες της βάσης.

Σχήμα:
Όσον αφορά το σχήμα, μπορείτε να σχεδιάσετε τους πίνακες για τα δεδομένα Zillow όπως θεωρείτε καλύτερα για τα ζητούμενα της εργασίας. Λάβετε υπόψη σας ότι ανάλογα με το σχήμα μπορεί να χρειαστεί να κάνετε επιπλέον μετασχηματισμούς στα δεδομένα Zillow εκτός από αυτούς που αναφέρονται παραπάνω.

Μετρικές

Κάντε join τα δεδομένα Airbnb και Zillow στα πεδία date, zipcode και bedroom για να υπολογίσετε τη μετρική Revenue Crossover Point, η οποία ορίζεται ως εξής:

Revenue Crossover Point = ceiling of (Zillow's median rental price per month / Airbnb's median rental price per day).

Η μετρική αυτή αναπαριστά τον αριθμό των ημερών ανά μήνα που ένας ιδιοκτήτης Airbnb πρέπει να νοικιάζει το ακίνητό του ώστε να κερδίσει το ίδιο εισόδημα με το αν το διέθετε για μακροχρόνια ενοικίαση. Το Revenue Crossover Point πρέπει να υπολογιστεί για κάθε συνδυασμό ημερομηνίας, Τ.Κ. και αριθμό κρεβατιών (date, zipcode, bedrooms).

Καθώς οι ημερομηνίες στα Zillow δεδομένα είναι σε επίπεδο μήνα, θα πρέπει να μετατρέψετε τις ημερομηνίες στα δεδομένα Airbnb (στον πίνακα Calendar) κατάλληλα ώστε να συμπίπτουν με αυτό το επίπεδο. Για παράδειγμα, η ημερομηνία ‘2017-05-16’ πρέπει να μετατραπεί σε ‘2017-05-01’. Μπορείτε να χρησιμοποιήσετε τη συνάρτηση της postgreSQL `date_trunc()` για να κάνετε αυτή τη μετατροπή.

Επιπλέον, η τιμή ενοικίασης ενός Airbnb πρέπει να προέλθει από το πεδίο Calendar.price αν αυτό έχει τιμή, αλλιώς από το Listing.price. Σημειώστε ότι η Calendar.price είναι πιο ακριβής από την Listing.price καθώς λαμβάνει υπόψη της διαφοροποιήσεις με βάση την εποχή, αλλά δεν περιλαμβάνεται πάντα στα δεδομένα. Μπορείτε να χρησιμοποιήσετε μία `SQL case expression` για να υλοποιήσετε τη λογική αυτής της συνθήκης.

Βγάλτε εκτός όλα τα Airbnbs που αφορούν την ενοικίαση ενός δωματίου σε κοινόχρηστο σπίτι/διαμέρισμα. Χρησιμοποιήστε τα Listing.room_type='Entire home/apt' and Listing.bedrooms > 0 ως κριτήρια φιλτραρίσματος. Επίσης, βγάλτε εκτός ημερομηνίες, Τ.Κ. και αριθμούς δωματίων (date, zipcode, bedrooms) που είναι NULL.

Υπολογίστε τη μέση τιμή ενοικίασης Airbnb ανά ημέρα (Airbnb's median rental price per day) χρησιμοποιώντας τη συνάρτηση της PostgreSQL `percentile_cont()`. Σώστε τα αποτελέσματα των queries για τον υπολογισμό του Revenue Crossover Point σε όψεις (views) στη βάση. Ο ορισμός της όψης πρέπει να είναι ο εξής:

```
v_Revenue_Crossover(date DATE, zipcode INT, bedrooms INT,
airbnb_price_day FLOAT, zillow_price_month FLOAT, crossover_pt FLOAT)
```

Δημιουργήστε ένα τέτοιο view για κάθε μία από τις 4 πόλεις και δώστε ένα χαρακτηριστικό όνομα π.χ. austin_v_Revenue_Crossover.

Χρήσιμα Links:

Μια συνοπτική περιγραφή των ETL διαδικασιών:
https://en.wikipedia.org/wiki/Extract,_transform,_load

Apache Beam:
https://beam.apache.org/

Συνάρτηση date_trunc():
https://www.postgresql.org/docs/9.6/functions-datetime.html

Συνάρτηση percentile_cont():
https://www.postgresql.org/docs/9.6/functions-aggregate.html

Median:
https://en.wikipedia.org/wiki/Median

Παραδοτέα:

- Δημιουργήστε ένα .txt αρχείο στο οποίο θα αναγράφονται το endpoint του AWS instance σας (μπορείτε να το δείτε στο AWS console, *RDS* > *Databases* > *db_identifier* > *Connectivity section*), το όνομα της βάσης σας και το username και το password ενός 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:

  Endpoint: <name_of_the_endpoint>
  Username: <username>
  Password: <password>
  Database: <name_of_the_database>

- Γράψτε ένα σύντομο report (.pdf αρχείο) για τα βήματα που ακολουθήσατε κατά την επεξεργασία των δεδομένων σας ώστε να αποθηκευτούν στη βάση.
- Βάλτε οποιοδήποτε SQL, Python ή Java αρχείο χρησιμοποιήσατε για τα ζητούμενα του project σε ένα φάκελο. Σε κάθε αρχείο γράψτε ένα σύντομο σχόλιο στην αρχή του αρχείου για το τι κάνει ο κώδικας. Ειδικά για τον ορισμό της view του Revenue Crossover Point γράψτε σε σχόλιο πόσες εγγραφές έχει. 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.
- Κάντε υποβολή το .zip αρχείο στο eclass στην ενότητα *Εργασίες / Project*.