

# DataBase Project 1

Τζένη Μπολένα p3170117  
Κωνσταντίνος Νικολούτσος p3170122

## 1 Τι εργαλεία χρησιμοποιήσαμε

Να τονίσουμε ότι στην επεξεργασία δεδομένων χρησιμοποιήθηκαν τα εξής API:

- Apache Beam (Using Direct runner)
- Apache Commons csv parser
- PrintWriter
- BufferedReader

Επιλέξαμε ως γλώσσα προγραμματισμού την Java SE 12

## 2 Πως κάναμε την επεξεργασία δεδομένων (ETL)

Οπως και στην πραγματική ζωή τα δεδομένα της βάσης μας δεν ήταν στην επιθυμητή μορφή. Για να τα φέρουμε στην μορφή που θέλαμε ακολουθήσαμε τα εξής βήματα:

### 2.1 File merging

Σε αυτό το βήμα αρχικά κατηγοριοποιήσαμε στον υπολογιστή μας csv αρχεία ανάλογα με την ιδιότητα τους(Τα host μονα τους, τα listing μονα τους κλπ). Έπειτα φτιάξαμε ένα πρόγραμμα java το οποίο χρησιμοποιεί το Beam API με σκοπό τη συγχώνευση των αρχείων ανά κατηγορίες.

## 2.2 Duplicate removal

Εφόσον έχουμε κάνει Merge σε όλα τα csv ανα κατηγορία ειμαστε σε θέση να αφαιρέσουμε τα διπλότυπα που εμφανίζονται. Για να γίνει αυτο φτιάχτηκα πρόγραμμα σε java το οποίο χρησιμοποιεί διαβάζει διαδοχικά καθε tuple και στην συνεχεια βάζει το primary key σε ένα HashMap. Σε περίπτωση που το πρόγραμμα διαβάσει ένα tuple με primary key που υπάρχει ήδη στο HashMap τότε διαγράφει αυτο το tuple. Με αυτον τον naive τρόπο καταφαινουμε να αφαιρέσουμε όλα τα duplicates απο τα merged csvs μας!

## 2.3 Extra changes

Σε αυτό το βήμα κάνουμε κάποιες έξτρα αλλαγές στα δεδομένα μας ετσι ωστε να μην χαλάσει το σχήμα της βάσης μας (foreign keys κλπ). Πιο συγκεκριμένα φτιάχνουμε προγράμματα στην java τα οποία λύνουν τα ακατάλληλα δεδομένα για την βάση μας. Μερικα απο αυτά είναι:

- Πρόσθεση ενός amenity στο boston\_amenity.csv με id 161 – γιατι υπάρχουν listing που δείχνουν σε αυτο (Listing\_amenity\_connection.csv)
- Διαγραφή tuple στο summary\_listing, calendar, listing\_amenity\_connection, review, summary\_review, summary\_listing τα οποία έχουν listing\_id το οποίο δεν υπάρχει στο Listing πίνακα – Ετσι ώστε να διατηρήσουμε τα foreign key.
- Διαγραφή ενός column στο austin\_listing.csv που δεν υπάρχει στα boston, denver, portland
- Ανανέωση του πίνακα neighborhood με δεδομένα τα οποία λείπουν – Πρεπει για να διατηρήσουμε το foreign key με listing
- Μερικά zip\_code στο summary\_listing έχουν λάθος τιμή δηλαδή αντι να έχουν zipcode έχουν ονομα περιοχής. Για να το φτιάξουμε αυτο κάναμε cross validation με τα στοιχεία του Listing. – Πολύ σημαντικό για το inner join που θα κανουμε στην βάση μας!
- Καταλληλη μετατροπή των 5 csv zillow σε 1 – Με αυτο τον τρόπο κάνουμε πιο απλό το query για την ανακτηση της μετρικής
- Αλλαγή των zipcode στα listing του Boston ετσι ωστε να φύγει το πρωτο 0 – Το κανουμε αυτο διοτι φαίνεται αυτα τα δεδομένα να ήταν σφαλμένα.




















Τα ονόματα έχουν των αρχείων είναι γραμμένα αφαιρετικά, ελπίζουμε να μην είναι τόσο δύσκολο να διαβαστούν!

Αφού λοιπόν κάνουμε όλα τα παραπάνω, είμαστε έτοιμα να ανεβάσουμε τα δεδομένα στην βάση!

## 2.4 Εισαγωγή δεδομένων στην βάση

Τα δεδομένα σε csv αρχεία ανέβηκαν όπως και στις προηγούμενες εργασίες. Πιο αναλυτικά χρησιμοποιήσαμε το command line μέσω psql. Μπορείτε να βρείτε τι ακριβώς γράψαμε στον φάξελο με τον κώδικα μας!

## 2.5 Αποτελέσματα μετρικής

- ▼  Tables (11)
  - >  amenity
  - >  calendar
  - >  calendar\_summary
  - >  host
  - >  listing
  - >  listing\_amenity\_connection
  - >  neighborhood
  - >  review
  - >  summary\_listing
  - >  summary\_review
  - >  zillow\_median\_rental\_price
- >  Trigger Functions
- >  Types
- ▼  Views (4)
  - >  austin\_v\_revenue\_crossover
  - >  boston\_v\_revenue\_crossover
  - >  denver\_v\_revenue\_crossover
  - >  portland\_v\_revenue\_crossover

Airbnb on MasterUser@DbInstance

Query Editor

Query History

1

select \* from austin\_v\_revenue\_crossover

Data Output

Explain

Messages

Notifications

	zipcode character varying (50)	date timestamp with time zone	bedrooms integer	median_per_day double precision	zillow_monthly_price real	city character varying (100)	revenue_crossover_point double precision
1	78759	2017-03-01 00:00:00+00	2	199	1315	Austin	6.60804020100502
2	78753	2017-07-01 00:00:00+00	1	99	908	Austin	9.17171717171717
3	78751	2017-05-01 00:00:00+00	2	195	1595	Austin	8.17948717948718
4	78741	2017-03-01 00:00:00+00	2	225	1240	Austin	5.51111111111111
5	78745	2017-04-01 00:00:00+00	3	300	1850	Austin	6.16666666666667
6	78705	2017-05-01 00:00:00+00	2	240	1700	Austin	7.08333333333333
7	78702	2017-07-01 00:00:00+00	2	265	2318.5	Austin	8.74905660377359
8	78750	2017-09-01 00:00:00+00	1	140	934	Austin	6.67142857142857
9	78753	2017-12-01 00:00:00+00	2	110.5	1163	Austin	10.5248868778281
10	78751	2018-01-01 00:00:00+00	1	115	1292	Austin	11.2347826086957
11	78758	2017-04-01 00:00:00+00	3	179	1595	Austin	8.91061452513966
12	78735	2017-10-01 00:00:00+00	1	149.5	1279	Austin	8.55518394648829
13	78730	2017-05-01 00:00:00+00	1	174	1005	Austin	5.77586206896552