Βάσεις Δεδομένων Project 1

Κωνσταντίνος-Ηλίας Κωνσταντόπουλος 3170086

Λαπάκης Γεράσιμος 3170089

Η Προεπεξεργασία Δεδομένων AirBNB και Zillow έγινε με την χρήση του Apache Beam.

Για την μετονομασία του πεδίου RegionName σε Zipcode χρησιμοποιήσαμε την συνάρτηση mutateHeader, η οποία δέχεται ως όρισμα το header με την μορφή string και το χωρίζει σε ένα List. Έπειτα όταν βρει το πεδίο με όνομα RegionName το αλλάζει σε zipcode και το αποθηκεύει στο string result. Όμοια αυτή η συνάρτηση βάζει στο τέλος των ημερομηνιών το -01 και τα αποθηκεύει στο result. Τέλος, το φιλτράρισμα των ημερομηνιών υλοποιήθηκε στην findIndexes και στην runWordCount. Η 1η επιστρέφει το index της εμφάνισης του 2016-01-01, δηλαδή της πρώτης ημερομηνίας που μας ενδιαφέρει. Η 2η δημιουργεί ένα pipeline στο οποίο φορτώνουμε διαδοχικά καθένα από τα 5 zillow και εφαρμόζουμε το φίλτρο το οποίο κρατάει τα δεδομένα από το 2016-01-01 και έπειτα εξάγεται το διορθωμένο αρχείο Zillow.

Διαγραφή Διπλοτύπων με την χρήση Apache Commons

- Η διαγραφή των διπλοτύπων έγινε με την βοήθεια της συνάρτησης primaryKeys η οποία βρίσκει το/τα PK κάθε αρχείου και της checkDuplicates, η οποία με την βοήθεια της 1ης προσπερνά τις διπλές εγγραφές. Επιπλέον, η checkDuplicatesWithBase προσπερνά τις διπλές εγγραφές των πινάκων.

- Κατά την εισαγωγή των πινάκων Host παρατηρήθηκε η ύπαρξη διπλοτύπων επομένως κρίθηκε αναγκαίο, κατά την εισαγωγή νέων πινάκων host να ελέγχεται η ύπαρξη τους στη βάση. Κατά την εισαγωγή νέων host, η εγγραφή αποθηκεύεται στη στατική λίστα primary keys το id των host για την αποφυγή διπλοτύπων στο μέλλον. Χρησιμοποιήθηκαν οι συναρτήσεις: existsHost, validateHost και fillHost.

- Η getPairs και η updateSumList λύνουν το εξης πρόβλημα: Στο πεδίο zipcode των SummaryListing υπήρχε το neighbourhood και κάποια δεν είχαν αντιστοιχία Listing οπότε έπρεπε να προσπεραστούν.

- Παρατηρήσαμε ότι οι Listing είχαν μια παραπάνω στήλη από αυτόν που έχουμε στην βάση μας, την calculated_host_listings_count, οπότε με την βοήθεια της deleteColumn την διαγράψαμε.

- Διαπιστώθηκε ότι τα amenity και τα ids τους δεν ήταν σε αντιστοιχία με την βάση μας. Χρησιμοποιήσαμε 3 συναρτήσεις έτσι ώστε να ανέβουν στην βάση. Η scanAmenities αποθήκευε προσωρινά τα amenity_name και amenity_id της βάσης για την αναπροσαρμογή των νέων εγγραφών. Η AmenityExists καλείται επαναληπτικά από την checkAmenities η οποία είναι υπεύθυνη για την μετατροπή και την προσθήκη νέων amenity, που δεν υπήρχαν στην βάση.

- Αποφασίσαμε το Schema της βάσης μας να περιέχει επιπλέον έναν ακόμη πίνακα ο οποίος περιέχει τα δεδομένα Zillow, προσαρμοσμένο στις ανάγκες της εργασίας με Header :zipcode,City,State,Metro,CountyName,SizeRank,Date,Price,Bedrooms. Σε αυτό μας βοήθησε η συνάρτηση new Table στην οποία εισάγεται καθένα από τα Zillow και παρουσιάζονται με τον νέο τρόπο που έχουμε ορίσει.

Σημείωση: Λόγω του ότι ζητήθηκε σύντομη περιγραφή δεν περιγράψαμε αναλυτικά των τρόπο λειτουργίας των μεθόδων.

Στην main του αρχείου java έχει γίνει comment out εφαρμογή των μεθόδων διότι απαιτούν το filepath των αρχείων.

Δεν καταφέραμε να βρούμε τρόπο να φτιάξουμε το script ώστε να λειτουργεί ανεξάρτητα.

Για την περίπτωση που θελήσετε να τρέξετε το πρόγραμμά μας να ελέγξετε τη λειτουργικότητα του, θα πλοηγηθείτε στο directory ./word-beam και θα το τρέξετε εισάγοντας :

```
mvn compile exec:java -D exec.mainClass=org.apache.beam.examples.WordCount `
 -D exec.args="--inputFile=”αρχείο δοκιμής” --output=αρχείο_εξαγωγής.csv" -P direc
t-runner
```