ΣΜΥΡΝΙΟΥΔΗΣ ΝΙΚΟΛΑΟΣ – 3170148
ΧΟΥΛΙΑΡΑΣ ΔΗΜΗΤΡΙΟΣ – 3170180

Αναφορά παράδοσης

Project μαθήματος βάσεων δεδομένων – μέρος 1

Τα βήματα επεξεργασίας δεδομένων Zillow μέσω Apache Beam.

Αρχικά δημιουργούνται 5 drain στο pipeline path, ένα για κάθε αρχείο Zillow. Ύστερα εφαρμόζεται σε καθένα απο αυτα το Ptransform ExtractRowsFn με διαφορετικά ορίσματα γιατί σε κάθε αρχείο η αρχική ημερομηνία και η τελική διαφέρουν και επίσης κάθε αρχείο αναπαριστά τιμές οικιών με διαφορετικό αριθμό δωματίων. Αυτό το Ptransform διασπάει μια γραμμή του αρχείου σε τόσες εγγραφές της μορφής (zipcode,...,sizerank,date,price,bedrooms) όσες και οι υπάρχουσες τιμές στην αρχική τιμή.Αυτά τα πέντε διαφορετικά pipeline path συνδέονται μαζί με ένα Flatten. Μετά εφαρμόζεται το Ptransform Distinct που φιλτράρει τυχόν διπλότυπες εγγραφές. Τέλος φιλτράρονται μέσω Filter.by οι εγγραφές οπου το 6ο πεδίο τους ,δηλαδή η ημερομηνία, είναι απο το 2016 και μετά και απο το 2018-01 και πρίν , μετατρέπονται σε String μέσω της ParDo.of(FormatAsTextFn) και τέλος γράφονται στο αρχείο εξόδου rental_price.csv .

Τα βήματα της επεξεργασίας δεδομένων airbnb.

Τα δεδομένα φορτώνονται μέσω του προγράμματος Main.java. Η Main ορίζει απο πού διαβάζονται τα αρχεία, δημιουργεί δυαδικα δέντρα(TreeSet) για τα foreignKeys και τα primaryKeys των πινάκων και ορίζει σε εναν ArrayListReceiver πως θα επεξεργαστεί η κάθε εγγραφή.

Ο ArrayListReceiver ανοίγει ένα αρχείο διαβάζει ένα αρχείο, δημιουργεί ένα Map<String,String> για κάθε λογική εγγραφή του αρχείο και εκτελεί τις ενέργειες που έχει ορίσει η Main.

Συνολικά, η Main διαβάζει για κάθε πίνακα της βάσεις όλα τα δεδομένα απο όλες τις πόλεις και εξάγει απο αυτά τα δεδομένα που δεν ανήκουν στα ήδη υπάρχοντα στην βάση (της πόλης austin)

Για κάθε αρχείο csv που φορτώνεται, το πρόγραμμα δημιουργεί ένα νέο δυαδικό δέντρο TreeSet<String> που περιέχει τα περιεχόμενα των primaryKey. Όταν φορτώνεται ένα αρχείο, για κάθε εγγραφή του προστίθεται στο δέντρο ένα ακόμα primaryKey. Έτσι όταν βρεθεί μια διπλότυπη εγγραφή σε κάποιο αρχείο,επειδή το primaryKey της θα υπάρχει ήδη στο δέντρο του αρχείο, το πρόγραμμα την εντοπίζει και δεν την τοποθετεί στα νεα δεδομένα.

Για να ελεγχθούν οι περιορισμού ξένου κλειδιού, σε κάθε εγγραφή ενός πίνακα ελεγχονται αν υπάρχουν τα πεδία που αποτελούν το ξένο κλειδι στο δέντρο με τα primaryKey ενος αλλου πίνακα ο οποίος πρέπει να έχει φορτωθεί πριν τον συγκεκριμένο. Αν δεν υπάρχει το foreignKey του ενός στα primaryKey του άλλου τότε αυτή η εγγραφή ξεσκαρτάρεται.

Ως προς την σειρά με την οποία φορτώνονται οι πίνακες για να ληφθούν υπόψη τα foreign keys αυτή είναι η εξής:

neighborhood → host → listing → calendar → review → calendar_summary → summary_listing → summary_review → amenity → has_amenity