

# Ιόνιο Πανεπιστήμιο

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή Εργασία

**Μελέτη ερευνητικών θεμάτων σχετικά με τις μηχανές  
βάσεων δεδομένων (database engines)**

**Μιλτιάδης Τομαράς**

Π2016180

**Επιβλέπουσα :** Άννα Σωτηροπούλου

Σεπτέμβριος 2023

### **Επιβλέπουσα**

**Άννα Σωτηροπούλου**, Εργαστηριακό Διδακτικό Προσωπικό, Ιόνιο Πανεπιστήμιο

### **Τριμελής Επιτροπή**

**Άννα Σωτηροπούλου**, Εργαστηριακό Διδακτικό Προσωπικό, Ιόνιο Πανεπιστήμιο

**Ανδρέας Καναβός**, Αναπληρωτής Καθηγητής, Ιόνιο Πανεπιστήμιο

**Θεμιστοκλής Έξαρχος**, Επίκουρος Καθηγητής, Ιόνιο Πανεπιστήμιο

### **Αφιερώσεις**

Θα ήθελα να αφιερώσω αυτή την εργασία στην οικογένεια μου που πάντα βρίσκεται στο πλευρό μου και με στηρίζει στις εύκολες αλλά και στις δύσκολες στιγμές.

## **Περίληψη**

Σκοπός της παρούσας πτυχιακής εργασίας ήταν η βαθιά ανάλυση των μηχανών βάσεων δεδομένων. Δεδομένου ότι αποτελούν απαραίτητο εργαλείο για την κατασκευή και την λειτουργία των βάσεων δεδομένων, παρουσιάζει μεγάλο ενδιαφέρον στην ερευνητική κοινότητα. Έτσι, έχουν πραγματοποιηθεί πολλές έρευνες που αφορούν την δομή και τις λειτουργίες που εκτελούν οι μηχανές βάσεων δεδομένων διαχρονικά. Μέσω αυτών των ερευνών, υλοποιήθηκε η παρακάτω εργασία, δίνοντας έμφαση σε ζητήματα που αφορούν τις ιεραρχίες αποθήκευσης που χρησιμοποιούνται για την αποδοτική χρήση των διαθέσιμων πόρων, τις δομές αποθήκευσης τους για την αποτελεσματική αποθήκευση δεδομένων, τις λειτουργίες ομαδοποίησης των δεδομένων, ώστε τα παρόμοια δεδομένα να αποθηκεύονται σε κοντινά σημεία καθιστώντας έτσι την ανάκτηση τους γρήγορη και χαμηλή σε κόστος. Τέλος, αναλύονται συστήματα που αφορούν την εύρεση των δεδομένων σε μια βάση δεδομένων και σκοπό έχουν την ταχύτερη εκτέλεσης της λειτουργίας και παράλληλα την εξοικονόμηση χώρου. Πολλές από τις έρευνες που χρησιμοποιήθηκαν αφορούν νέες προσεγγίσεις και καινοτόμες τεχνικές που χρησιμοποιούνται για αποδοτικότερα αποτελέσματα.

## **Abstract**

The purpose of this thesis was to provide an in-depth analysis of database engines. Since they are an essential tool for the construction and operation of databases, they are of a great interest to the research community. Thus, many researches have been conducted on the structure and functions performed by database engines over time. Through these researches, the following work has been implemented, emphasizing on issues related to storage hierarchies used for efficient use of available resources, their storage structures for efficient data storage, data clustering functions so that similar data are stored in nearby locations thus making their retrieval fast and low cost. Finally, systems that involve the retrieval of data in a database and are designed to make the operation faster while saving space are analyzed. Many of the researches used involve new approaches and innovative techniques used for more efficient results.

## **Πρόλογος και Ευχαριστίες**

Η εργασία αυτή μας αναλύει τον ρόλο και την χρησιμότητα των μηχανών βάσεων δεδομένων σε συστήματα βάσεων δεδομένων που χρησιμοποιούνται σε πολλούς επιστημονικούς τομείς και όχι μόνο, καθημερινά.

Θα ήθελα να ευχαριστήσω την κυρία Σωτηροπούλου που ήταν ιδιαίτερα υποστηρικτική και με καθοδήγησε με υπομονή καθόλη την διάρκεια της εκτέλεσης της εργασίας, καθώς και όλους τους καθηγητές του τμήματος, που μου έδωσαν τα κατάλληλα εφόδια για την ολοκλήρωση των σπουδών μου αλλά και για το μέλλον.

## Περιεχόμενα

|                                                                                          |           |
|------------------------------------------------------------------------------------------|-----------|
| <b>Περιεχόμενα.....</b>                                                                  | <b>6</b>  |
| <b>1)Εισαγωγή.....</b>                                                                   | <b>10</b> |
| 1.1) Ορισμός Μηχανών Βάσεων Δεδομένων.....                                               | 10        |
| <b>2) Ιεραρχία αποθήκευσης μηχανών βάσης δεδομένων.....</b>                              | <b>11</b> |
| 2.1) Ιεραρχίες αποθήκευσης.....                                                          | 11        |
| 2.1.1) Αφθονος Ετερογενής Παραλληλισμός (Abundant Heterogeneous Parallelism).....        | 12        |
| 2.1.2) Ιεραρχίες βαθιάς αποθήκευσης (Deep Storage Hierarchies).....                      | 13        |
| 2.2) Αρχαιοθέτηση δεδομένων.....                                                         | 14        |
| 2.2.1) Oligoarchive architecture.....                                                    | 15        |
| 2.2.2) DNA σύστημα αποθήκευσης.....                                                      | 15        |
| 2.3) Ενοποιημένη ιεραρχία μνήμης-αποθήκευσης.....                                        | 16        |
| 2.3.1) Προκλήσεις.....                                                                   | 17        |
| 2.3.2) Λύσεις.....                                                                       | 17        |
| 2.3.3) Σχεδιασμός FlatFlash.....                                                         | 17        |
| <b>3) Δομές αποθήκευσης.....</b>                                                         | <b>19</b> |
| 3.1) Εποχή της παρατηρησιμότητας.....                                                    | 19        |
| 3.1.1) Προκλήσεις και λύσεις.....                                                        | 19        |
| 3.1.2) Εναλλακτικές λύσεις.....                                                          | 19        |
| 3.1.3) Επισκόπηση Mach.....                                                              | 20        |
| 3.2) Αποθήκευση δεδομένων σε μορφή γράφου.....                                           | 21        |
| 3.3) Αποδοτική αποθήκευση σε LSM βάσεις δεδομένων.....                                   | 23        |
| 3.3.1) LSM δέντρο.....                                                                   | 23        |
| 3.3.2) Δέντρο LSM με βάση τη στοίβα (Stack-based LSM tree).....                          | 25        |
| 3.3.3) Ισοπεδωμένο δέντρο LSM (Leveled LSM tree).....                                    | 25        |
| 3.3.4) Διαφορές μεταξύ Stack-based και Leveled LSM δέντρα.....                           | 26        |
| 3.4) Συμπαγείς δομές δεδομένων (Compact data structures).....                            | 27        |
| 3.4.1) K2-raster.....                                                                    | 27        |
| 3.4.2) Κατασκευή και δομές δεδομένων.....                                                | 28        |
| <b>4) Ομαδοποίηση δεδομένων (Data clustering).....</b>                                   | <b>31</b> |
| 4.1) Βαθιά ομαδοποίηση (Deep clustering).....                                            | 31        |
| 4.1.1) Ενότητα ομαδοποίησης (Clustering module).....                                     | 31        |
| 4.1.2) Βαθιά ομαδοποίηση με αντιστοίχιση σχέσεων (Relation matching deep clustering).... | 32        |
| 4.1.3) Βαθιά ομαδοποίηση μέσω ψευδο Ετικετοποίησης (Pseudo Labeling Deep Clustering)...  | 33        |
| 4.1.4) Βαθιά ομαδοποίηση με αυτοεκπαίδευση.....                                          | 34        |
| 4.1.5) Μεγιστοποίηση της αμοιβαίας πληροφορίας με βάση την ομαδοποίηση.....              | 35        |
| 4.1.6) Αντιθετική βαθιά ομαδοποίηση.....                                                 | 35        |
| 4.2) Ομαδοποίηση διαχρονικών δεδομένων με αργά μεταβαλλόμενες τάσεις.....                | 36        |
| 4.2.1) Μέθοδοι.....                                                                      | 36        |
| 4.2.2) Διαμήκης k-means (Longitudinal k-means).....                                      | 37        |
| 4.2.3) Ομαδοποίηση δύο βημάτων (Two-step clustering).....                                | 37        |
| 4.2.4) Μοντελοποίηση τροχιάς με βάση την ομάδα (Group-based trajectory modeling).....    | 38        |
| 4.2.5) Μοντελοποίηση μείγματος ανάπτυξης (Growth mixture modeling).....                  | 39        |

|                                                                                                         |           |
|---------------------------------------------------------------------------------------------------------|-----------|
| 4.2.6) Μοντελοποίηση μικτών χρονικά μεταβαλλόμενων επιδράσεων (Mixed time-varying effect modeling)..... | 39        |
| 4.3) Ομαδοποιήσεις στην ανάλυση και εξόρυξη δεδομένων (Clustering in data analysis and mining).....     | 40        |
| 4.3.1 Άλλες μέθοδοι.....                                                                                | 40        |
| 4.3.2) Αλγόριθμος MHTSASM.....                                                                          | 40        |
| 4.3.2.1) Εντατικοποίηση και διαφοροποίηση.....                                                          | 41        |
| 4.4) Δίκαιη ομαδοποίηση κατηγορικών δεδομένων.....                                                      | 45        |
| 4.4.1) Μέθοδος.....                                                                                     | 45        |
| <b>5) Ευρετηρίαση βάσεων δεδομένων.....</b>                                                             | <b>51</b> |
| 5.1) Ευρετηρίαση μεγάλων χωρικών βάσεων δεδομένων.....                                                  | 51        |
| 5.1.1) Ευρετηρίαση χωρικών δεδομένων.....                                                               | 51        |
| 5.1.2) Δομή δεδομένων X-tree.....                                                                       | 52        |
| 5.1.3 Προκλήσεις για την δημιουργία του aX-tree.....                                                    | 53        |
| 5.1.4) Κατασκευή aX-tree.....                                                                           | 53        |
| 5.1.5) Περιγραφή αλγορίθμου.....                                                                        | 54        |
| 5.2) Ευρετηρίαση χρονικών βάσεων δεδομένων.....                                                         | 54        |
| 5.2.1) Δομή δεδομένων TIndex.....                                                                       | 55        |
| 5.2.2) Δομή δεδομένων TIndex2.....                                                                      | 57        |
| 5.2.3) Πειράματα.....                                                                                   | 57        |
| 5.2.4) Σύνολα δεδομένων.....                                                                            | 58        |
| 5.2.5) Χρήση μνήμης.....                                                                                | 58        |
| 5.2.6) Χρόνοι εκτέλεσης.....                                                                            | 59        |
| 5.3) Ευρετήρια σε βάσεις δεδομένων γραφημάτων.....                                                      | 61        |
| 5.3.1) Προκλήσεις και υπάρχουσες τεχνικές.....                                                          | 62        |
| 5.3.2) GraphGrepSX.....                                                                                 | 63        |
| 5.3.2.1) Η φάση της προεπεξεργασίας.....                                                                | 63        |
| 5.3.2.2) Φιλτράρισμα και αντιστοίχιση.....                                                              | 64        |
| 5.3.2.3) Πειράματα και εφαρμογές.....                                                                   | 65        |
| 5.4) Ανανεώσιμος και προσαρμοστικός δείκτης μάθησης.....                                                | 67        |
| 5.4.1) Μαθησιακοί δείκτες B+tree.....                                                                   | 67        |
| 5.4.2) Εισαγωγή του ALEX.....                                                                           | 68        |
| 5.4.2.1) Σχεδιασμός του ALEX.....                                                                       | 69        |
| 5.4.2.2) Διάταξη.....                                                                                   | 69        |
| 5.4.3) Αλγόριθμοι ALEX.....                                                                             | 70        |
| 5.4.3.1) Αναζητήσεις και ερωτήματα.....                                                                 | 70        |
| 5.4.3.2) Λειτουργία εισαγωγής σε μη πλήρη κόμβο δεδομένων.....                                          | 70        |
| 5.4.3.3) Λειτουργία εισαγωγής σε πλήρη κόμβο δεδομένων.....                                             | 71        |
| 5.4.3.4) Διαγραφή και ενημερώσεις.....                                                                  | 72        |
| <b>6) Βιβλιογραφία.....</b>                                                                             | <b>73</b> |

## Κατάλογος Εικόνων

|                                                                                                                                                                                                                                                                                                                            |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Εικόνα 1:</b> Πλεονέκτημα χρήσης RDMA.....                                                                                                                                                                                                                                                                              | 12 |
| <b>Εικόνα 2:</b> Παράδειγμα προσαρμοστικότητας των rack-scale.....                                                                                                                                                                                                                                                         | 13 |
| <b>Εικόνα 3:</b> Η ικανότητα της GPU να χειριστεί πολύπλοκους υπολογισμούς.....                                                                                                                                                                                                                                            | 13 |
| <b>Εικόνα 4:</b> Διαφορετικά επίπεδα αποθήκευσης για διαφορετικές ανάγκες.....                                                                                                                                                                                                                                             | 14 |
| <b>Εικόνα 5:</b> Κατά την εκτέλεσης μιας λειτουργίας put, το σύστημα συνθέτει αλυσίδες DNA και τις αποθηκεύει σε μια βιβλιοθήκη. Κατά την εκτέλεση μιας λειτουργίας get, οι αλυσίδες DNA αλληλουξίζονται και οι αναγνώσεις επιστρέφονται. Η ανάκτηση θα πρέπει επιτυγχάνεται παρά τα σφάλματα που μπορεί να προκύψουν..... | 16 |
| <b>Εικόνα 6:</b> Αρχιτεκτονική FlatFlash.....                                                                                                                                                                                                                                                                              | 18 |
| <b>Εικόνα 7:</b> Η αρχιτεκτονική του Mach .....                                                                                                                                                                                                                                                                            | 20 |
| <b>Εικόνα 8:</b> Εκκαθάριση των MemTables και του πίνακα Delete.....                                                                                                                                                                                                                                                       | 24 |
| <b>Εικόνα 9:</b> Διαδικασία συγχώνευσης για δύο ευρετήρια.....                                                                                                                                                                                                                                                             | 24 |
| <b>Εικόνα 10:</b> Λειτουργία συγχώνευσης σε δέντρο LSM με βάση τη στοίβα.....                                                                                                                                                                                                                                              | 26 |
| <b>Εικόνα 11:</b> Λειτουργία συγχώνευσης σε ισοπεδωμένο δέντρο LSM.....                                                                                                                                                                                                                                                    | 26 |
| <b>Εικόνα 12:</b> Περίπτωση πίνακα ράστερ και εννοιολογική δενδροειδής αναπαράσταση.....                                                                                                                                                                                                                                   | 28 |
| <b>Εικόνα 13:</b> Βελτιωμένη αναπαράσταση με χρήση διαφορών για μέγιστες και ελάχιστες τιμές.....                                                                                                                                                                                                                          | 30 |
| <b>Εικόνα 14:</b> Μέθοδοι βαθιάς ομαδοποίησης.....                                                                                                                                                                                                                                                                         | 32 |
| <b>Εικόνα 15:</b> Το πλαίσιο του αλγορίθμου MHTSASM.....                                                                                                                                                                                                                                                                   | 41 |
| <b>Εικόνα 16:</b> Υπολογισμός των αποστάσεων μεταξύ των πολλαπλών συστάδων (πομπού και δέκτη).....                                                                                                                                                                                                                         | 49 |
| <b>Εικόνα 17:</b> (α) δομή ενός X-tree, (β) σχήματα ενός X-tree σε διαφορετικές διαστάσεις δεδομένων.....                                                                                                                                                                                                                  | 53 |
| <b>Εικόνα 18:</b> Η δομή ενός aX-tree σε δύο διαστάσεις.....                                                                                                                                                                                                                                                               | 54 |
| <b>Εικόνα 19:</b> TIndex για το μέρος P1, του πίνακα 14.....                                                                                                                                                                                                                                                               | 56 |
| <b>Εικόνα 20:</b> Παράδειγμα ενός πλήρες TIndex.....                                                                                                                                                                                                                                                                       | 56 |
| <b>Εικόνα 21:</b> Παράδειγμα ενός πλήρες TIndex2.....                                                                                                                                                                                                                                                                      | 57 |
| <b>Εικόνα 22:</b> Σύγκριση απαιτούμενης μνήμης μεταξύ TIndex και TIndex2.....                                                                                                                                                                                                                                              | 59 |
| <b>Εικόνα 23:</b> Σύγκριση αλγορίθμων στο σύνολο δεδομένων Chess.....                                                                                                                                                                                                                                                      | 59 |
| <b>Εικόνα 24:</b> Σύγκριση αλγορίθμων στο σύνολο δεδομένων Mushroom.....                                                                                                                                                                                                                                                   | 60 |
| <b>Εικόνα 25:</b> Σύγκριση αλγορίθμων στο σύνολο δεδομένων Retail.....                                                                                                                                                                                                                                                     | 60 |
| <b>Εικόνα 26:</b> Σύγκριση αλγορίθμων στο σύνολο δεδομένων Accidents.....                                                                                                                                                                                                                                                  | 61 |
| <b>Εικόνα 27:</b> Βάση δεδομένων γραφημάτων.....                                                                                                                                                                                                                                                                           | 61 |
| <b>Εικόνα 28:</b> Φιλτράρισμα και επαλήθευση στο GraphGrepSX.....                                                                                                                                                                                                                                                          | 64 |
| <b>Εικόνα 29:</b> Παράδειγμα δομής ενός Learned Index.....                                                                                                                                                                                                                                                                 | 67 |
| <b>Εικόνα 30:</b> Επέκταση ενός κόμβου.....                                                                                                                                                                                                                                                                                | 71 |

## Κατάλογος Πινάκων

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| Πίνακας 1: Η δομή των κορυφών.....                                                                     | 22 |
| Πίνακας 2: Η δομή των ακμών.....                                                                       | 22 |
| Πίνακας 3: Στα αριστερά είναι η αποθήκευση του γράφου και στα δεξιά είναι οι ιδιότητες του γράφου..... | 22 |
| Πίνακας 4: Τα αρχεία δεικτών.....                                                                      | 23 |
| Πίνακας 5: Σύγκριση μεταξύ διαφορετικών μεταερευνητικών αλγορίθμων.....                                | 45 |
| Πίνακας 6: Δείγμα παρατηρήσεων από ένα σύνολο δεδομένων.....                                           | 46 |
| Πίνακας 7: Κατανομή των συστάδων των χαρακτηριστικών.....                                              | 47 |
| Πίνακας 8: Μη κενές πολλαπλές συστάδες.....                                                            | 47 |
| Πίνακας 9: Πίνακας σύμπτωσης.....                                                                      | 48 |
| Πίνακας 10: Παράδειγμα συσχέτισης με μία μόνο σύμπτωση μεταξύ πολλαπλών συστάδων.....                  | 48 |
| Πίνακας 11: Παράδειγμα συσχέτισης με περισσότερες από μία συμπτώσεις μεταξύ πολλαπλών συστάδων.....    | 49 |
| Πίνακας 12: Πίνακας βέλτιστων πολυσυστοιχιών.....                                                      | 49 |
| Πίνακας 13: Παράδειγμα χρονική βάσης δεδομένων.....                                                    | 55 |
| Πίνακας 14: Πληροφορίες των βάσεων δεδομένων.....                                                      | 58 |
| Πίνακας 15: Χρήση μνήμης για το TIndex.....                                                            | 58 |
| Πίνακας 16: Χρήση μνήμης για το TIndex2.....                                                           | 58 |
| Πίνακας 17: Χρόνος που απαιτείται για την κατασκευή δεικτών (sec).....                                 | 65 |
| Πίνακας 18: Μέγεθος των δεικτών(Kb).....                                                               | 65 |
| Πίνακας 19: Χρόνος φιλτραρίσματος (sec).....                                                           | 66 |
| Πίνακας 20: Χρόνος αναζήτησης (sec).....                                                               | 66 |
| Πίνακας 21: Συνολικός αριθμός υποψηφίων.....                                                           | 66 |
| Πίνακας 22: Συνολικός χρόνος (sec).....                                                                | 66 |

## 1)Εισαγωγή

Στις μέρες μας, οι πληροφορίες αποτελούν ένα πολύ ουσιώδες εργαλείο για την επιτυχία όλων των επιχειρήσεων και όχι μόνο. Όσο γρηγορότερος είναι ο ρυθμός με τον οποίο αυξάνεται ο όγκος των δεδομένων που παράγονται, τόσο μεγαλύτερη είναι η ανάγκη για την αποτελεσματική αποθήκευση, ανάκτηση και διαχείριση των δεδομένων. Για την εξυπηρέτηση των αναγκών αυτών, είναι απαραίτητη η χρήση των Συστημάτων Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ). Τα ΣΔΒΔ υλοποιούν τις παραπάνω λειτουργίες μέσω των μηχανών βάσεων δεδομένων. Οι μηχανές βάσεων δεδομένων λοιπόν, αποτελούν σημαντικό μέρος των σύγχρονων εργαλείων διαχείρισης δεδομένων. Κάθε μηχανή βάσεων δεδομένων χαρακτηρίζεται από τα πλεονεκτήματα και τα μειονεκτήματά της, έτσι είναι σημαντική η επιλογή του ιδανικού τύπου, ανάλογα τον φόρτο εργασίας και το είδος των δεδομένων. Κάποιοι από τους βασικότερους παράγοντες που επηρεάζουν την επιλογή μιας τέτοιας μηχανής, είναι η επεκτασιμότητα, η ασφάλεια, η ευκολία χρήσης, η επίδοση και το κόστος.

Στην εργασία αυτή, θα παρουσιάσουμε κάποια από τα βασικά χαρακτηριστικά και τις λειτουργίες που εκτελούν οι μηχανές βάσεων δεδομένων. Αρχικά, θα δούμε κάποιες ιεραρχίες αποθήκευσης που σχηματίζονται, δηλαδή τον τρόπο με τον οποίο χρησιμοποιείται η μνήμη για την αποδοτικότερη χρήση των πόρων. Στην συνέχεια, θα παρουσιαστούν διαφορετικές δομές αποθήκευσης, οι οποίες κατασκευάζονται ώστε να υλοποιείται η αποθήκευση πολλών τύπων δεδομένων με αποτελεσματικούς τρόπους. Έπειτα, αναφέρονται προσεγγίσεις που στοχεύουν στην ομαδοποίηση των δεδομένων. Με την ομαδοποίηση, τα δεδομένα που παρουσιάζουν κοινά στοιχεία αποθηκεύονται σε κοντινά σημεία και καθιστά την ανάκτηση τους εύκολη και με μικρό κόστος. Στις τελευταίες ενότητες, αναλύονται τεχνικές εύρεσης δεδομένων που χρησιμοποιούνται για την βελτίωση της απόδοσης μιας βάσης δεδομένων, μειώνοντας την εξέταση των εγγραφών κατά την εκτέλεση ενός ερωτήματος και το συνολικό κόστος. Τα ευρήματα ενώ επιταχύνουν την πρόσβαση στα δεδομένα, απαιτείται να ενημερώνονται όταν αλλάζουν τα δεδομένα και επίσης δεσμεύουν χωρητικότητα στη βάση δεδομένων.

### 1.1) Ορισμός Μηχανών Βάσεων Δεδομένων

Μια μηχανή βάσεων δεδομένων είναι το λογισμικό που χρησιμοποιεί ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS) για τη δημιουργία, ανάγνωση, ενημέρωση και διαγραφή δεδομένων (CRUD) από μια βάση δεδομένων. Τα περισσότερα συστήματα διαχείρισης βάσεων δεδομένων, περιλαμβάνουν δική τους διεπαφή προγραμματισμού εφαρμογών (API) που επιτρέπει στον χρήστη να αλληλεπιδρά με την μηχανή, προσπερνώντας την διεπαφή χρηστών του DBMS.

Ο όρος “μηχανή βάσης δεδομένων” χρησιμοποιείται συχνά εναλλακτικά με τον όρο “διακομιστής βάσης δεδομένων” ή “σύστημα διαχείρισης βάσεων δεδομένων”.

## 2) Ιεραρχία αποθήκευσης μηχανών βάσης δεδομένων

Από τα πρώτα βήματα της πληροφορικής, η διαχείριση δεδομένων αποτέλεσε σημαντικό πεδίο σε έρευνες και στην βιομηχανία. Τα συστήματα διαχείρισης δεδομένων έχουν συμβάλει στην δημιουργία πολλών διακεκριμένων εφαρμογών, διαδικτυακών υπηρεσιών, όπως τα κοινωνικά δίκτυα (Twitter, Facebook) και οικονομικές εφαρμογές επεξεργασίας σύνθετων γεγονότων σε πραγματικό χρόνο, μέσα από την επιτυχή επεξεργασία συναλλαγών και την ανάλυση μεγάλου όγκου δεδομένων

Αυτές οι εφαρμογές που συνεχώς εξελίσσονταν ήταν η κινητήρια δύναμη για πολλές καινοτομίες στις βάσεις δεδομένων και την αρχιτεκτονική τους για πολλά χρόνια. Ήταν αναγκαίο, τα συστήματα να εξελιχθούν παράλληλα με τις εφαρμογές για την αποτελεσματική υποστήριξη τους. Οι συνεχώς αυξανόμενες απαιτήσεις των εφαρμογών, χρειάζονται εξειδικευμένα σχέδια συστημάτων για την ικανοποίησή τους. Αυτό οδήγησε σε ένα πλήθος σχεδίων συστημάτων, όπου το καθένα να είναι προσανατολισμένο προς μια διαφορετική εφαρμογή.

Οι επαγγελματίες, χρησιμοποιούν μηχανές ροής για δεδομένα που φτάνουν σε ροές υψηλής ταχύτητας, μηχανές προσανατολισμένες στη στήλη (**column-oriented engines**) για ανάλυση δεδομένων, συστήματα βασισμένα στην κύρια μνήμη για συναλλαγές κ.ά. Ωστόσο, ο χώρος του σχεδιασμού περιέχει συμβιβασμούς σε διάφορες διαστάσεις όπως:

1. **Την διάταξη των δεδομένων** (προσανατολισμένα σε σειρά, στη στήλη, ιεραρχική ή υβριδική).
2. **Την συσκευή αποθήκευσης** ( Δίσκος, SSD, κύρια μνήμη).

Η τάση εξειδίκευσης και η αυξανόμενη πολυπλοκότητα των εφαρμογών διαχείρισης βάσεων δεδομένων σε συνδυασμό με την εισαγωγή όλο και πιο εξελιγμένου hardware, εμπνέουν την χρήση προσαρμοσμένων αλγορίθμων, οι οποίοι ωστόσο συχνά κάνουν υποθέσεις σχετικά με τις μορφές δεδομένων, τις διανομές, τις ιδιότητες του hardware κ.ά. Έτσι, η ενσωμάτωση τους σε οποιοδήποτε άλλο σύστημα καθίσταται αδύνατη. Η επόμενη γενιά εφαρμογών απαιτεί προσαρμοσμένα ή ακόμα και προσαρμόσιμα συστήματα τα οποία θα επιλέγουν δυναμικά, τα βέλτιστα:

1. **Hardware** αποθήκευσης δεδομένων
2. **Αλγορίθμους** επεξεργασίας δεδομένων

### 2.1) Ιεραρχίες αποθήκευσης

Οι πρόσφατες τάσεις του hardware στους τομείς της επεξεργασίας και της αποθήκευσης επηρεάζουν έντονα τις εφαρμογές διαχείρισης δεδομένων. Παρακάτω θα προσδιορίσουμε τα χαρακτηριστικά τους με τη μεγαλύτερη επιρροή.

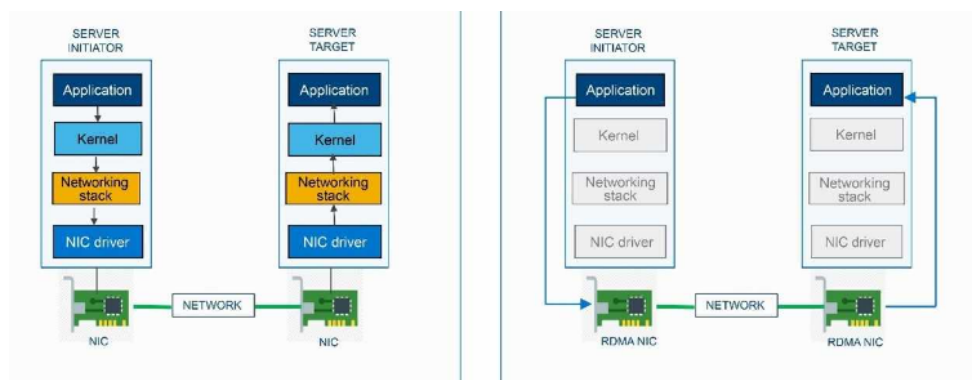
### 2.1.1) Άφθονος Ετερογενής Παραλληλισμός (Abundant Heterogeneous Parallelism)

Τα σχέδια των επεξεργαστών, γίνονται όλο και περισσότερο ετερογενή, αυξάνοντας την πολυπλοκότητα τους και ακολουθώντας μη ομοιόμορφες αρχιτεκτονικές. Σύμφωνα με τον νόμο του Moore, είναι σημαντικότερο να αυξάνουμε τον αριθμό των πυρήνων σε έναν επεξεργαστή, παρά την συχνότητα του ίδιου του επεξεργαστή. Στο σήμερα, ακόμα και χαμηλής ισχύος συστήματα διαθέτουν δεκάδες πυρήνες στο ίδιο τσιπ.

Η εκμετάλλευση του παραλληλισμού είναι ζωτικής σημασίας για την αξιοποίηση των διαθέσιμων αρχιτεκτονικών πόρων και την επιτάχυνση του διαθέσιμου λογισμικού. Παρόλα αυτά, ο σχεδιασμός επεκτάσιμων συστημάτων που μπορούν να επωφεληθούν από τον υποκείμενο παραλληλισμό του πυρήνα, παραμένει μια πρόκληση. Στην παραδοσιακή επεξεργασία συναλλαγών υψηλής απόδοσης, η εγγενής επικοινωνία οδηγεί σε δυσκολίες στην επεκτασιμότητα στο πολυπύρνηνο και πολύπριζο υλικό. Ακόμη και συστήματα τα οποία καταφέρνουν να κλιμακώνονται σε μια γενιά πολλών πυρήνων, πολλές φορές αποτυγχάνουν να συνεχίσουν την κλιμάκωση αυτή και σε επόμενες γενιές. Τέλος, οι βάσεις δεδομένων που σχεδιάστηκαν για μονοπύρνηνους επεξεργαστές, αποτυγχάνουν να εκμεταλλευτούν τον άφθονο παραλληλισμό που μπορεί να προσφέρει το σύγχρονο υλικό.

Καθώς ο αριθμός των πυρήνων επεξεργασίας αυξάνεται όλο και περισσότερο, οι διασυνδέσεις χαμηλής καθυστέρησης ελαχιστοποιούν τις αποστάσεις μεταξύ των διακομιστών.

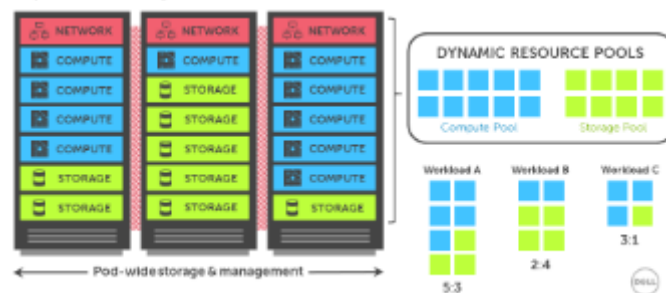
Με την τεχνολογία του **RDMA** (Remote Direct Memory Access), επιτρέπεται στις εφαρμογές να έχουν άμεση πρόσβαση στη μνήμη σε ένα απομακρυσμένο μηχάνημα, χωρίς να εμπλέκεται κάποιο λειτουργικό σύστημα ή επεξεργαστής. Με την χρήση του σε γρήγορες δομές δικτύου, η καθυστέρηση της πρόσβασης σε ένα απομακρυσμένο μηχάνημα είναι συγκρίσιμη με την καθυστέρηση πρόσβασης στην απομακρυσμένη μνήμη σε ένα παραδοσιακό μηχάνημα μεγάλης μνήμης.



Εικόνα 1: Πλεονέκτημα χρήσης RDMA.

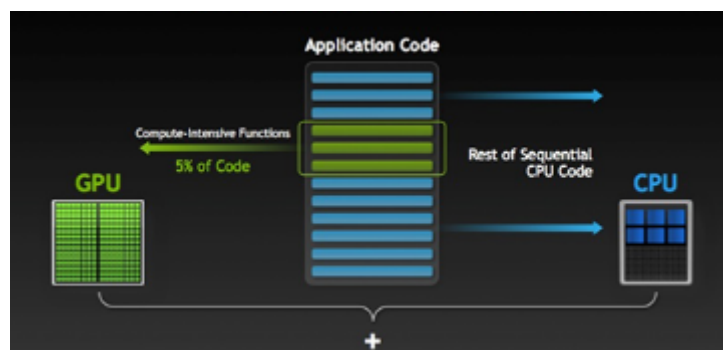
Μελλοντικά, συστήματα **rack-scale** όπως το “cluster-in-a-box” θα χρησιμοποιούν εξαιρετικά προσαρμοσμένους system-on-a-chip κόμβους, που θα περιέχουν μόνο πυρήνες επεξεργασίας, μνήμη, διεπαφές I/O και θα επικοινωνούν μέσω διασυνδέσεων χαμηλής καθυστέρησης. Τότε, αναπόφευκτα οι σχεδιαστές λογισμικών θα ενσωματώσουν τα συστήματα rack-scale στις εφαρμογές του, όπως στο

παρελθόν συνέβη όταν άρχισαν να χρησιμοποιούνται οι μηχανές κοινής μνήμης και οι υπερυπολογιστές για τομείς όπως η επιστήμη και η επιχειρηματικότητα.



Εικόνα 2: Παράδειγμα προσαρμοστικότητας των rack-scale.

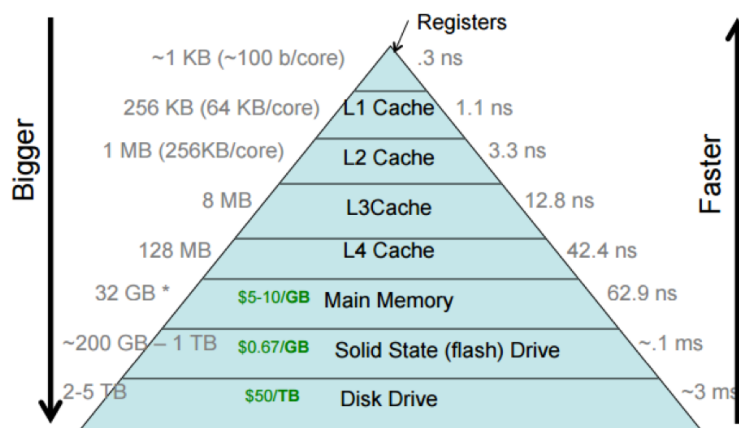
Έτσι, τα ετερογενή τσιπ που περιέχουν εξειδικευμένους επιταχυντές διαχείρισης δεδομένων γίνονται πιο ελκυστικά. Οι GPUs αποτελούν ένα πολύ γνωστό στόχο διαχείρισης δεδομένων για πολλά έτη. Ωστόσο, η ζήτηση για αποτελεσματικότερη χρήση της ενέργειας φέρνει προτάσεις που αφορούν προσαρμοσμένα τσιπ που επιταχύνουν μεμονωμένες λειτουργίες όπως η κατάτμηση και ο κατακερματισμός αλλά και σχέδια που εξειδικεύουν τα τσιπ ώστε να ολοκληρώσουν αναλυτικά ερωτήματα.



Εικόνα 3: Η ικανότητα της GPU να χειριστεί πολύπλοκους υπολογισμούς.

### 2.1.2) Ιεραρχίες βαθιάς αποθήκευσης (Deep Storage Hierarchies)

Γενικότερα, τα συστήματα διαχείρισης δεδομένων έχουν σχεδιαστεί ώστε να διαθέτουν τους δίσκους ως κύρια μνήμη αποθήκευσης, ενώ η κύρια μνήμη χρησιμοποιείται για την προσωρινή αποθήκευση και οι κασέτες (tape) για την δημιουργία αρχειοθέτησης. Με το κόστος της **DRAM** να μειώνεται όμως, η ιεραρχία αποθήκευσης έχει διαμορφωθεί έτσι ώστε η κύρια μνήμη να χρησιμεύει ως η κύρια αποθήκευση δεδομένων και οι κρυφές μνήμες του επεξεργαστή να παρακολουθούν την κατάσταση της επεξεργασίας δεδομένων. Παρόλα αυτά, σε περιπτώσεις όπου πραγματοποιείται μεγάλος αριθμός ταυτόχρονων και περίπλοκων αιτημάτων από χρήστες, η αποθήκευση γίνεται αναγκαστικά στο δίσκο. Αυτό σημαίνει πως τα νέα συστήματα διαχείρισης, θα πρέπει να μπορούν υπό περιπτώσεις να χρησιμοποιούν διαφορετικά επίπεδα της ιεραρχίας αποθήκευσης εναλλακτικά. (Ailamaki 2015)



Εικόνα 4: Διαφορετικά επίπεδα αποθήκευσης για διαφορετικές ανάγκες.

Οι **SSDs** αποτελούνται από έναν υπολογιστικό πυρήνα που συνδέεται απευθείας με τσιπ NVM. Το NVMe έχει αναδειχθεί ως το πρωτόκολλο αποθήκευσης για τους SSD και συνδέεται απευθείας μέσω PCIe ή απομακρυσμένα μέσω **Fabric Interconnect** (FI). Η έλευση των flash chips ωφέλησε τους SSD σε απόδοση και μείωση καθυστέρησης. Οι ανάγκες για προβλέψιμη απόδοση, υψηλή χρήση πόρων και αναδιαμορφώσιμες υποδομές, οδήγησαν σε SSD ανοιχτού καναλιού, εξοπλισμένο με ένα ελάχιστο ελεγκτή αποθήκευσης ο οποίος εκθέτει του πόρους του SSD αμέσως σε έναν κεντρικό υπολογιστή. (Bonnet, 2017)

Η σταθερή μνήμη **RAM** συνιστά μια τεχνολογία μνήμης που σκοπό έχει να προσφέρει τις αποτελεσματικότερες επιδόσεις σε ταχύτητα και πυκνότητα. Αν και οι παραπάνω τεχνολογίες ανταγωνίζονται μεταξύ τους και προσφέρουν διαφορετικά επίπεδα απόδοσης και αντοχής, έχουν κοινό χαρακτηριστικό την εμμονή και τις καθυστερήσεις στην πρόσβαση. Η RAM παρουσιάζει επίσης ασυμμετρία στις καθυστερήσεις ανάγνωσης και εγγραφής που είναι κοινές στους δίσκους, ενώ παρέχουν την δυνατότητα, να αποθηκευτούν δεδομένα σε συγκεκριμένες θέσεις της κύριας μνήμης (**byte addressability**). Οι ιδιότητες αυτές, τα καθιστούν ένα αναπόσπαστο μέρος της ιεραρχίας αποθήκευσης και προσφέρει ερευνητικές ευκαιρίες. (Ailamaki, 2015)

## 2.2) Αρχαιοθήκηση δεδομένων.

Τα τελευταία χρόνια, η τεράστια παραγωγή δεδομένων σε συνδυασμό με την ανάγκη διατήρησης αυτών των δεδομένων για μελλοντικές εργασίες, έχουν ως αποτέλεσμα να συναντάμε μια ραγδαία αύξηση αρχειοθετημένων δεδομένων από επιχειρήσεις και όχι μόνο. Η ταχύτητα με την οποία εξελίσσονται τα αποθηκευτικά μέσα ωστόσο είναι πολύ μικρότερη σε σύγκριση με τον ρυθμό αύξησης των παραγόμενων δεδομένων. Αυτό το συμβάν οδήγησε τους ερευνητές στην εύρεση νέων οικονομικών αρχιτεκτονικών και μέσων για την αποθήκευση δεδομένων που δεν αναζητούνται συχνά. Στην συνέχεια θα δούμε μια αρχιτεκτονική που ονομάζεται OligoArchive και χρησιμοποιεί το DNA,

το οποίο αποτελεί ένα μέσο αποθήκευσης και χαρακτηρίζεται από την ανθεκτικότητα και την υψηλή πυκνότητα του.

### 2.2.1) Oligoarchive architecture.

Οι μηχανές βάσεων δεδομένων χρησιμοποιούν μια ιεραρχία αποθήκευσης που αποτελείται από τρία επίπεδα και περιέχει συσκευές με μεγάλο εύρος τιμής και απόδοσης.

**Το επίπεδο της απόδοσης** αποθηκεύει τα δεδομένα στα οποία, πρόσβαση έχουν εφαρμογές ανάλυσης σε πραγματικό χρόνο και **OLTP**. Λαμβάνοντας υπόψη τις αυστηρές απαιτήσεις καθυστέρησης αυτών των εφαρμογών, οι DRAM και οι SSD αποτελούν τις καλύτερες επιλογές σε αυτή τη βαθμίδα.

**Το επίπεδο της χωρητικότητας** αποθηκεύει δεδομένα στα οποία, πρόσβαση διαθέτουν εφαρμογές μαζικής ανάλυσης. Το υλικό που χρησιμοποιείται συνήθως για την αποθήκευση σε αυτή τη βαθμίδα είναι οι **HDD** (μονάδες σκληρού δίσκου), λόγω του χαμηλού κόστους σε σχέση με άλλες επιλογές.

Τέλος, **το επίπεδο της αρχειοθέτησης** χρησιμοποιείται για την αποθήκευση δεδομένων που δέχονται πρόσβαση σε πολύ σπάνιες περιπτώσεις, όπως σε ελέγχους συμμόρφωσης με την ασφάλεια ή νομικούς ελέγχους. Το βασικό υλικό που χρησιμοποιείται εδώ είναι οι **VTL** (Virtual Tape Libraries), λόγω του χαμηλότερου κόστους και της υψηλότερης ανθεκτικότητας σε σχέση με τους HDD. (Appuswamy, 2019)

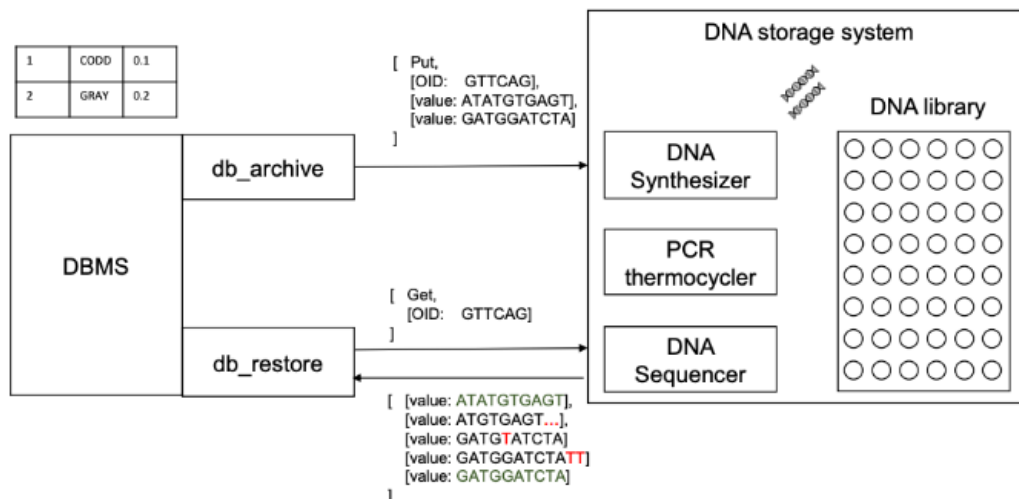
### 2.2.2) DNA σύστημα αποθήκευσης

Σύμφωνα με πρόσφατες μελέτες, μεγάλο ποσοστό δεδομένων το οποίο αυξάνεται συνεχώς θεωρείται “κρύο”, δηλαδή σπανίως δέχεται πρόσβαση. Αυτό συμβαίνει λόγω της γρήγορης σε ρυθμό ανάπτυξης των δεδομένων, γεγονός που επισκιάζει τις βελτιώσεις που παρουσιάζει το παραδοσιακό υλικό. Για την αντιμετώπιση του συγκεκριμένου προβλήματος και την δημιουργία αρχειοθέτησης, οι ερευνητές αποφάσισαν να διερευνήσουν τη χρήση τεχνικών σχεδιασμού μεταξύ υλικού και λογισμικού ή να χρησιμοποιούν υλικό με πολύ μεγαλύτερη ανθεκτικότητα και πυκνότητα όπως το **DNA**.

Το DNA χαρακτηρίζεται από τέσσερις ιδιότητες οι οποίες το κάνουν αποτελεσματικό στην αρχειακή αποθήκευση.

1. Πολύ πυκνό αποθηκευτικό μέσο.
2. Εξαιρετικά ανθεκτικό όταν αποθηκεύεται σε κατάλληλες συνθήκες.
3. Εύκολο, γρήγορο και οικονομικό στη χρήση.
4. Μηδενική τιμή **Kryder**.

Το DNA σύστημα αποθήκευσης είναι υπεύθυνο για την αποθήκευση των ολιγονουκλεοτιδίων (**oligos**). Μοντελοποιούμε μια συσκευή αποθήκευσης DNA ως μια αποθήκη αντικειμένων και υποθέτουμε ότι εσωτερικά αποτελείται από ένα συνθεσάιζερ, ένα θερμοκυκλοποιητή PCR και ένα προσδιοριστή αλληλουχίας. Στην αρχιτεκτονική OligoArchive, γίνεται ένας σαφής καταμερισμός εργασίας, μεταξύ των συσκευών αποθήκευσης DNA και της μηχανής βάσης δεδομένων.



**Εικόνα 5:** Κατά την εκτέλεσης μιας λειτουργίας **put**, το σύστημα συνθέτει αλυσίδες DNA και τις αποθηκεύει σε μια βιβλιοθήκη. Κατά την εκτέλεση μιας λειτουργίας **get**, οι αλυσίδες DNA αλληλουχίζονται και οι αναγνώσεις επιστρέφονται. Η ανάκτηση θα πρέπει επιτυγχάνεται παρά τα σφάλματα που μπορεί να προκύψουν.

Το DNA παρέχει την λειτουργικότητα της αλληλουχίας και της σύνθεσης, ενώ η κωδικοποίηση και η αποκωδικοποίηση γίνεται από τη βάση δεδομένων. Έτσι, η DNA συσκευή αποθήκευσης, παίρνει ένα αναγνωριστικό αντικείμενο που προσδιορίζει μοναδικά μια ομάδα ολιγονουκλεοτιδίων. Το αντικείμενο που ανακτάται, αποτελεί ένα αρχείο που αναπαριστά μέσω κειμένου τα ολιγονουκλεοτίδια τα οποία κωδικοποιούν τα δεδομένα που αντιστοιχούν σε ένα **primer** (εκκινητή).

Τέλος, το σύστημα αποθήκευσης χρησιμοποιεί το συνθεσάιζερ για τον χειρισμό των λειτουργιών **put**, για τη δημιουργία του συνθετικού DNA. Μία λειτουργία **get** χρησιμοποιείται για το φιλτράρισμα και την εξαγωγή όλων των κλώνων που αντιστοιχούν σε αυτόν τον εκκινητή.

### 2.3) Ενοποιημένη ιεραρχία μνήμης-αποθήκευσης.

Το μειωμένο κόστος των SSDs σε συνδυασμό με την τεράστια αποδοτικότητα τους, έχει φέρει στο προσκήνιο μια προσέγγιση όπου χρησιμοποιούνται ως κύρια μνήμη. Η κλιμάκωση τους μπορεί να φτάσει σε τεράστια επίπεδα ανά υποδοχή PCIe σε σχέση με την DRAM όπου παρουσιάζει χαμηλότερες δυνατότητας κλιμάκωσης ανά υποδοχή DIMM. Αυτή η ανησυχία για τις DRAM έχει

οδηγήσει σε προσεγγίσεις οι οποίες αξιοποιούν μηχανισμούς σελιδοποίησης και την χαρτογράφηση μνήμης, ενώ οι SSDs χρησιμοποιούνται ως υποστηρικτικές αποθηκεύσεις για την DRAM.

### 2.3.1) Προκλήσεις.

Παρά τα θετικά των παραπάνω προσεγγίσεων, προκύπτουν τρία βασικά μειονεκτήματα τα οποία θα δούμε.

1. Η επίδοση επιδέχεται επιβαρύνσεις μέσω του μηχανισμού σελιδοποίησης. Πιο συγκεκριμένα, όταν εκτελούνται προσπελάσεις μνήμης σε δεδομένα που δεν βρίσκονται στην DRAM, τότε αφού πρώτα ξεκινήσει ένα σφάλμα σελίδας, έπειτα το λογισμικό χειρισμού σελίδας μεταφέρει την προσπελάσιμη σελίδα του SSD στην κύρια μνήμη, επιβαρύνοντας χρονικά την εκτέλεση.
2. Επιπλέον, ο φόρτος εργασίας που αφορά εφαρμογές σε εντατικά δεδομένα που διαθέτουν μέγεθος πολύ μεγαλύτερο σε σχέση με την ελεύθερη χωρητικότητα της DRAM οδηγεί σε ένα πρόβλημα το οποίο ονομάζεται **thrashing**.
3. Με την διαδικασία της εκτέλεσης σελιδοποίησης στις προσπελάσεις των δεδομένων, δημιουργείται μια σημαντική αύξηση των κινήσεων I/O. Αυτό έχει ως αποτέλεσμα την μείωση της διάρκειας ζωής ενός SSD.

### 2.3.2) Λύσεις.

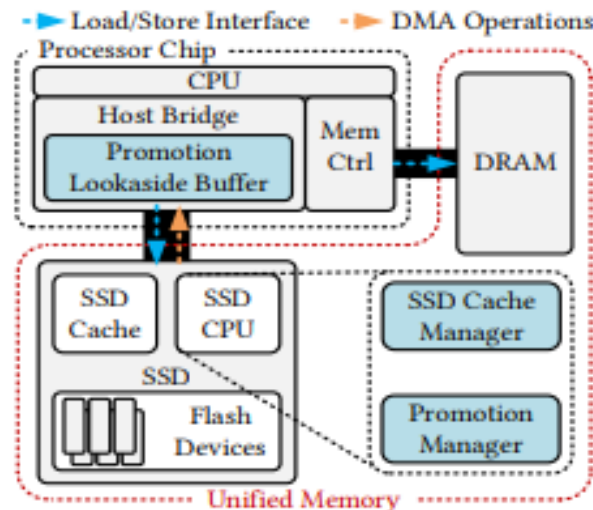
Σήμερα, καθώς είναι εφικτή η διεύθυνση byte μέσω του PCIe και σε συνδυασμό με την εσωτερική μνήμη των ελεγκτών των SSD, διαθέτουμε SSDs που παρέχουν διευθυνσιμότητα των bytes μέσω των καταχωρητών βασικής διεύθυνσης BAR. Έτσι, μπορούμε να διατηρούμε πρόσβαση σε έναν SSD σε byte αλλά και σε επίπεδα μπλοκ. Αυτό βέβαια έχει ως αποτέλεσμα την πολυπλοκότερη διαχείριση των SSDs. Στις παρακάτω ενότητες, θα αναλύσουμε μια ενοποιημένη ιεραρχία μνήμης-αποθήκευσης η οποία ονομάζεται **FlatFlash** και βασίζεται στην δυνατότητα πρόσβασης των SSD σε byte με σκοπό την επίλυση των παραπάνω προκλήσεων.

### 2.3.3) Σχεδιασμός FlatFlash.

Την λύση στις παραπάνω προκλήσεις που είδαμε, καλείται να δώσει το σύστημα FlatFlash, του οποίου η αρχιτεκτονική φαίνεται στην εικόνα 6, μέσα από την αξιοποίηση της πρόσβασης byte σε έναν SSD. Η διαδικασία αυτή χωρίζεται σε πέντε τμήματα.

1. Τις τεχνικές οι οποίες είναι απαραίτητες για την προσβασιμότητα byte των SSD στο σύστημα FlatFlash.

2. Ο συνδυασμός μεταξύ SSD και DRAM σε ένα ενοποιημένο χώρο μνήμης. Οπότε, μια σελίδα μνήμης μπορεί να αντιστοιχιστεί στην DRAM ή στον SSD. Έτσι, ο προγραμματισμός των SSD στους οποίους είναι εφικτή η προσπέλαση byte, γίνεται απλούστερος.
3. Μέσω ενός μηχανισμού προώθησης σελίδων από τον SSD στην DRAM, ο οποίος διασφαλίζει την συνοχή των δεδομένων, αυξάνεται η ταχύτητα των προσπελάσεων στην DRAM.
4. Το FlatFlash διαθέτει μηχανισμό προώθησης ο οποίος βοηθά στην αύξηση της απόδοσης για εφαρμογές. Ο μηχανισμός αυτός είναι απαραίτητος για την διαχείριση της προώθησης και μέσω αλληλεπίδρασης με την SSD-Cache, αποφασίζει ποιες σελίδες πρέπει να προωθηθούν.
5. Το FlatFlash μέσα από την δημιουργία δεδομένων σε αναλυτική-byte μορφή, μπορεί με ευκολότερο τρόπο να διατηρήσει σημαντικά δεδομένα που σχετίζονται με το λογισμικό συστημάτων και τις εφαρμογές. (Abulila, 2019)



Εικόνα 6: Αρχιτεκτονική FlatFlash

### 3) Δομές αποθήκευσης.

Στην παρακάτω ενότητα θα δούμε και θα αναλύσουμε την δομή, τα χαρακτηριστικά και την λειτουργία διαφορετικών δομών αποθήκευσης που απαιτούνται για την αποτελεσματική αποθήκευση δεδομένων. Ανάλογα με τον τομέα στον οποίο χρησιμοποιούνται, παρατηρούνται διαφοροποιήσεις.

#### 3.1) Εποχή της παρατηρησιμότητας.

Η παρατηρησιμότητα, αποτελεί ένα ιδιαίτερο χαρακτηριστικό το οποίο βοηθά στην αποτελεσματικότερη κατανόηση σύνθετων και πολύπλοκων συστημάτων σε πραγματικό χρόνο μέσα από την συλλογή δεδομένων, ιδιαίτερα μετρήσεων. Οι μετρήσεις αποτελούν ποσοτικές μετρήσεις και είναι απαραίτητες για πολλές εργασίες όπως η δημιουργία ειδοποιήσεων και η συμπλήρωση των πινάκων ελέγχου.

##### 3.1.1) Προκλήσεις και λύσεις.

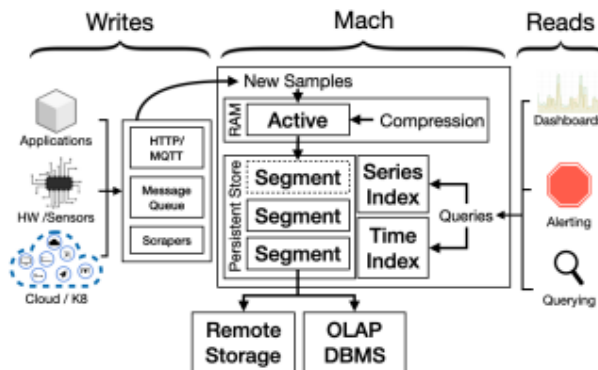
Στην πράξη, ο όγκος των μετρήσεων δεδομένων που παράγεται σε κάποιες περιπτώσεις είναι τεράστιος. Αυτό έχει ως αποτέλεσμα, τα ήδη υπάρχοντα συστήματα να μην μπορούν να ανταποκριθούν, καθώς δεν υπάρχει η δυνατότητα κλιμάκωσης σε τέτοιο βαθμό. Για την αντιμετώπιση αυτού του προβλήματος, κατασκευάστηκε μια φορητή μηχανή αποθήκευσης που ονομάζεται **Mach** και αποτελείται από μια αρχιτεκτονική εξειδικευμένη για τα χαρακτηριστικά των μετρικών δεδομένων, ενώ διαθέτει ένα εύκολης χρήσης API το οποίο βοηθά τους χρήστες να αποθηκεύουν και να λαμβάνουν δεδομένα. Τα παραπάνω καθιστούν το Mach πολύ αποδοτικότερο σε σύγκριση με τις ήδη υπάρχουσες εναλλακτικές. Στις επόμενες ενότητες θα αναλύσουμε την λειτουργία του Mach.

##### 3.1.2) Εναλλακτικές λύσεις.

Κάποιες από τις ήδη υπάρχουσες επιλογές που χρησιμοποιούνται για την διατήρηση των δεδομένων, είναι συστήματα που ονομάζονται ενσωματωμένες αποθήκες κλειδιών-τιμών όπως το WiredTiger, το RocksDB κ.ά. Το μειονέκτημα αυτών των βιβλιοθηκών, είναι ότι από την κατασκευή τους προορίζονταν για γενική χρήση, οπότε περιβάλλονται από αρκετούς περιορισμούς που μετέπειτα μειώνουν την απόδοση τους ενώ ταυτόχρονα δεν μπορούν να ανταποκριθούν σε εισαγωγές μεγάλου όγκου δεδομένων. Για τις μετρήσεις δεδομένων, το πιο ευρέως χρησιμοποιούμενο σύστημα ονομάζεται Prometheus. Το Prometheus αν και είναι πολύ αποδοτικό στην λήψη των μετρήσεων, υστερεί στην υποστήριξη πολυμεταβλητών δεδομένων τα οποία εμφανίζονται σε αρκετές κοινές εργασίες.

### 3.1.3) Επισκόπηση Mach.

Για την αλληλεπίδραση μεταξύ των χρηστών και του Mach, χρησιμοποιείται ένα API απλής μορφής. Έτσι, είναι εφικτή η αποθήκευση και η ανάκτηση των δειγμάτων. Το Mach αποτελεί εξειδικευμένο εργαλείο στο τομέα της παρατηρησιμότητας, απαραίτητο για τον αποτελεσματική υποστήριξη μεγάλου όγκου εγγραφών και αναγνώσεις με μικρές χρονικές καθυστερήσεις.



Εικόνα 7: Η αρχιτεκτονική του Mach.

Όσον αφορά τις λειτουργίες εγγραφής, οι πηγές μέσω ενός ενδιάμεσου φορέα εκπέμπουν τα δείγματα ή τα εισάγουν αμέσως χρησιμοποιώντας κάποια πρωτόκολλα. Τα πρόσφατα δείγματα εισάγονται σε ένα buffer το οποίο όταν φτάσει στην πλήρη χωρητικότητά του, ακολουθεί μια διαδικασία συμπίεσης και έπειτα γράφεται σε έναν μόνιμο αποθηκευτικό χώρο. Από την άλλη, τα τμήματα τα οποία συνήθως δεν χρησιμοποιούνται, μεταφέρονται σε ένα σύστημα OLAP ώστε να αναλυθούν ή σε μια απομακρυσμένη αποθήκευση με σκοπό την αρχειοθέτηση.

Για τις λειτουργίες ανάγνωσης από την άλλη, μέσα από την λήψη ταυτόχρονων ερωτημάτων μπορούν να αξιοποιηθούν η ετικέτα και ο χρόνος που κατασκευάστηκαν οι δείκτες κατά την εκτέλεση μιας εισαγωγής, με σκοπό την εύρεση των πηγών και των χρονικών στιγμών που ενδιαφέρουν.

Η αρχιτεκτονική του Mach όπως αναφέρθηκε και προηγουμένως, χαρακτηρίζεται από χαλαρότητα. Ο χαλαρός αυτός συντονισμός, αξιοποιείται με τους εξής τρεις τρόπους.

1. **Πολλαπλά ανεξάρτητα ενεργά τμήματα:** Στις υπόλοιπες μηχανές αποθήκευσης είναι εφικτό να υπάρχουν πολλοί συγγραφείς για μια πηγή δεδομένων, ενώ συντονίζονται μέσα από την χρήση ενός mutex για τον έλεγχο της ταξινόμησης των δειγμάτων. Η χρήση του mutex ωστόσο επιβαρύνει πολύ την συνολική εκτέλεση μιας εγγραφής. Από την άλλη, το Mach επιτρέπει πολλά ανεξάρτητα νήματα συγγραφέα σε ένα όχι πολύπλοκο σχεδιασμό. Τα νήματα έχουν την δυνατότητα να γράψουν δείγματα από πολλές διαφορετικές πηγές, ενώ οι πηγές καθιερώνονται σε ένα μόνο νήμα συγγραφέα. Η κατοχή των μεταδεδομένων εσωτερικά με βάση την ελάχιστη κοινή κατάσταση, βοηθάει στην αποτροπή του ανταγωνισμού. Έτσι, το Mach καταφέρνει να λειτουργεί σαν σύνδεσμος πολλών ανεξάρτητων μηχανών αποθήκευσης.

2. **Προσθήκη γρήγορης διαδρομής:** Για την εξασφάλιση ευελιξίας, οι περισσότερες μηχανές αποθήκευσης γενικών χρήσεων επιτρέπουν ενημερώσεις και εισαγωγές εκτός σειράς. Για εργασίες που σχετίζονται με μετρήσεις ωστόσο, η λήψη και η εκπομπή των δειγμάτων γίνεται με την σειρά. Η διατήρηση των δομών δεδομένων δεν αξιοποιείται σωστά και επιβαρύνεται το σύστημα. Από την άλλη πλευρά, το Mach παρέχει ένα γρήγορο μονοπάτι με σκοπό την προσθήκη ως επί το πλείστον γραφής (append-mostly writes). Λόγω του σχεδιασμού του Mach, το γρήγορο μονοπάτι μπορεί να συγκριθεί με την εκτέλεση μιας απλής προσθήκης σε μια λίστα ενώ δεν υπάρχουν επιβαρύνσεις που αφορούν τον συγχρονισμό. Το Mach δηλαδή καθώς υιοθετεί συμπεριφορές λίστας, έχει την δυνατότητα να εκτελεί αναγνώσεις χωρίς να χρειάζεται να σταματήσουν οι εκτελέσεις εγγραφών.
3. **Ελαστικότητα σε επίπεδο νήματος:** Ο φόρτος εργασίας μπορεί να αυξηθεί σε μεγάλα επίπεδα. Για την αντιμετώπιση αυτού του προβλήματος, απαιτείται η αντίστοιχη αύξηση των εκτελούμενων αποθηκευτικών παραδειγμάτων. Αυτή η διαδικασία μπορεί να αυξήσει ωστόσο την πολυπλοκότητα και να οδηγήσει σε μη αποδοτική χρήση των πόρων. Αν και υπάρχουν συστήματα τα οποία δημιουργήθηκαν για αυτόν τον σκοπό, σε περιπτώσεις τεράστιου φόρτου εργασίας, η κλιμάκωση τους δεν είναι η απαραίτητη. Τα νήματα στο Mach όμως, λειτουργώντας ως ανεξάρτητες μηχανές αποθήκευσης επιτρέπουν στο σύστημα να κλιμακώνεται στο επίπεδο που είναι απαραίτητο ανάλογα με τις περιπτώσεις του φόρτου εργασίας, αυξάνοντας ή μειώνοντας τον αριθμό των νημάτων εγγραφής. (Soleza, 2022)

### 3.2) Αποθήκευση δεδομένων σε μορφή γράφου

Πολλές υπάρχουσες βάσεις δεδομένων γράφων συνήθως χρησιμοποιούν λίστες γειννίας ή συνδυαστικές λίστες για να αποθηκεύσουν τα δεδομένα με δομή γράφου. Ωστόσο, τώρα θα δούμε πως χρησιμοποιούμε πίνακες για να αποθηκεύσουμε τα δεδομένα με δομή γράφου. Κάθε κορυφή και ακμή των γράφων ιδιοκτησιών έχει ένα byte με σταθερό μήκος στους πίνακες. Κάθε κορυφή έχει πρόσβαση στους γείτονές της μέσω της διεύθυνσης του πίνακα. Η πολυπλοκότητα του χρόνου είναι  $O(1)$  όταν αυτό το σύστημα διατρέχει γράφους ιδιοκτησιών. Η μορφή αποθήκευσης είναι ένα πολύ θετικό χαρακτηριστικό για queries στους γράφους. Διότι όλες οι λειτουργίες διαγραφής, ενημέρωσης και εισαγωγής στη βάση δεδομένων γράφων μας γίνονται με τρόπο προσθήκης, αυτό το σύστημα δεν έχει περιορισμούς στην αποθήκευση δεδομένων με δομή γράφου στους πίνακες. Ο γράφος αποτελείται από τρία μέρη: κορυφές, ακμές και τοπολογίες. Τα δεδομένα με δομή γράφου θα γίνουν σειριοποίηση από τον δίσκο στη μνήμη όταν τα ερωτήματα στους γράφους χρειάζεται να έχουν πρόσβαση στον καθορισμένο γράφο. Ωστόσο, απλά ερωτήματα στους γράφους μπορούν να λαμβάνουν τα δεδομένα απευθείας μέσω του δείκτη του γράφου χωρίς να απαιτείται ανακατασκευή του γράφου.

|               |              |            |               |        |
|---------------|--------------|------------|---------------|--------|
| type:         | uint32       | uint32     | uint32        | uint32 |
| vertex array: | Pid+VertexId | TopoOffset | Flag+OEOffset | Time   |

**Πίνακας 1: Η δομή των κορυφών**

Η δομή των κορυφών παρουσιάζεται στον παραπάνω Πίνακα. Όλες οι κορυφές ενός γράφου αποθηκεύονται σε έναν πίνακα κορυφών. Κάθε κορυφή έχει ένα byte με σταθερό μήκος στον πίνακα κορυφών. Το Pid είναι συντομογραφία για το αναγνωριστικό διαίρεσης (partition id). Διότι ένας μεγάλος γράφος θα χωριστεί σε πολλούς υπογράφους, το partition id είναι το αναγνωριστικό ενός υπογράφου. Το VertexId είναι ένας μοναδικός αριθμός κορυφής σε έναν υπογράφο. Το TopoOffset (Topology Offset) είναι ο δείκτης της τοπολογίας στον πίνακα τοπολογιών. Το πεδίο Flag είναι κρατημένο για συγκεκριμένους σκοπούς. Το OEOffset (OutEdge Offset) είναι μια μετατόπιση ως προς το TopoOffset. Αυτό το σύστημα χρειάζεται να διακρίνει τις εισερχόμενες ακμές και τις εξερχόμενες ακμές μέσω του OEOffset. Το Time είναι μια ιδιότητα της κορυφής. Το σύστημα μπορεί να παρατηρεί τις λεπτομερείς αλλαγές σε ένα γράφο μέσω του Time.

|             |           |             |        |         |        |
|-------------|-----------|-------------|--------|---------|--------|
| type:       | uint32    | uint32      | uint32 | uint32  | uint32 |
| edge array: | Pid+SrcId | Flag+DistId | EdgeId | RelType | Time   |

**Πίνακας 2: Η δομή των ακμών**

| Property Block Head | Graph Block Head |          |
|---------------------|------------------|----------|
| Prop_value          | Vertex Struct    | Prop_key |
| Prop_value          | Edge Struct      | Prop_key |
| Prop_value          | Topology Struct  |          |
| ...                 | ...              |          |

**Πίνακας 3: Στα αριστερά είναι η αποθήκευση του γράφου και στα δεξιά είναι οι ιδιότητες του γράφου.**

Η δομή της τοπολογίας παρουσιάζεται στον Πίνακα παραπάνω. Ο πίνακας τοπολογίας (topo array) είναι ένας πίνακας που περιέχει όλες τις τοπολογίες ενός γράφου. Το Label είναι ο αριθμός στο λεξικό των ετικετών (dictionary number of labels). Το OutEdgeId είναι το αναγνωριστικό των εξερχόμενων ακμών. Το InEdgeId είναι το αναγνωριστικό των εισερχόμενων ακμών. Η τοπολογία είναι συνδεδεμένη με την κορυφή. Η κατάσταση των τοπολογιών θα αλλάζει ανάλογα με την κατάσταση των κορυφών και των ακμών. Ταυτόχρονα, αυτό το σύστημα μπορεί να ολοκληρώσει ερωτήματα στον γράφο μέσα σε έναν περιορισμένο χρόνο μέσω της δομής της τοπολογίας.

Τα δεδομένα με δομή γράφου αποθηκεύονται σε αρχεία μεγέθους 4G (βλέπε πίνακα 3). Το μέγεθος των αρχείων είναι 4G λόγω του τύπου uint32. Ταυτόχρονα, το σύστημά μας θα συγχωνεύει

τα δεδομένα στο δίσκο περιοδικά για να επιταχύνει τα ερωτήματα στον γράφο. Είναι επίσης χρήσιμο να υλοποιηθούν πολλαπλά αντίγραφα με αρχεία μεγέθους 4G. Το τελευταίο αρχείο κρατά τα πιο πρόσφατα δεδομένα, διότι τα νέα δεδομένα προστίθενται στα υπάρχοντα. Οι πληροφορίες περιόδου των αρχείων μπορούν να βοηθήσουν αυτό το σύστημα να επιταχύνει την αναζήτηση στους γράφους προέλευσης. Αυτό το σύστημα μπορεί να έχει απευθείας πρόσβαση στις ιδιότητες των κορυφών και των ακμών μέσω του πεδίου "prop key".

| Index Head    |              |
|---------------|--------------|
| VertexId+Time | Block+offset |
| EdgeId+Time   | Block+offset |
| Topo_flag     | Block+offset |
| ...           |              |

Πίνακας 4: Τα αρχεία δεικτών.

### 3.3) Αποδοτική αποθήκευση σε LSM βάσεις δεδομένων.

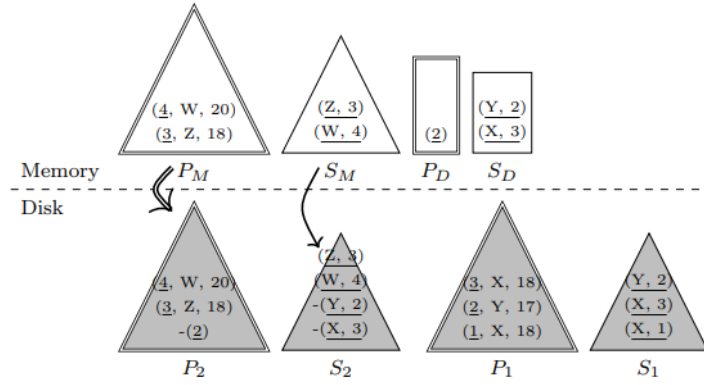
Βρισκόμαστε σε μια εποχή, όπου το δίκτυο χρησιμοποιείται από εκατομμύρια ανθρώπους ταυτόχρονα, ο αριθμός των παραγόμενων δεδομένων αυξάνεται ραγδαία και οι ηλεκτρονικές συσκευές και εφαρμογές αναπτύσσονται συνεχώς. Πολλές βάσεις δεδομένων χρησιμοποιούν αρχιτεκτονικές δομημένης συγχώνευσης αρχείων καταγραφής (LSM) για την εγγραφή με υψηλή απόδοση. Παρακάτω θα αναλύσουμε τη δομή ενός LSM δέντρου, τις υποκατηγορίες του και τις πολιτικές συγχώνευσης που υλοποιεί.

#### 3.3.1) LSM δέντρο

Η δομή ενός LSM δέντρου (**LSM tree**) αποτελείται από δύο στρώματα. Το πρώτο βρίσκεται στη μνήμη και διαθέτει ένα ενεργό MemTable (στοιχείο μνήμης), ενώ το δεύτερο βρίσκεται στο δίσκο όπου και τα δεδομένα είναι οργανωμένα σε μία ή περισσότερες ταξινομημένες εκτελέσεις.

Στις ταξινομημένες εκτελέσεις, βρίσκονται εγγραφές οι οποίες είναι διαδοχικά ταξινομημένες με βάση το δεικτοδοτούμενο κλειδί. Επιπροσθέτως, μια ταξινομημένη εκτέλεση μπορεί να κατέχει ένα ή πολλά αμετάβλητα SSTables (στοιχεία δίσκου), τα οποία τις περισσότερες φορές εφαρμόζονται με την χρήση δενδρικών δομών.

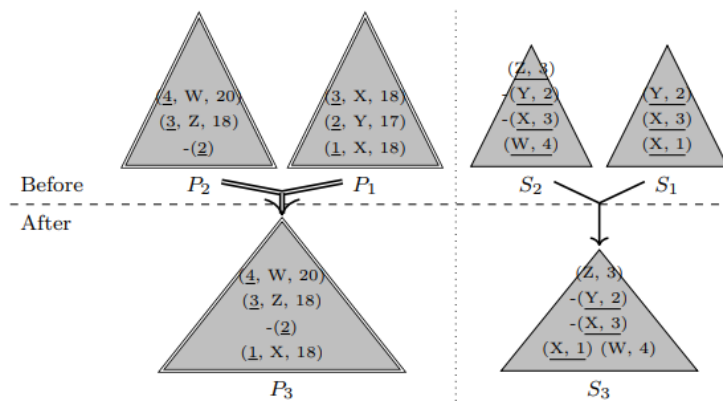
Οι δενδρικές δομές γενικότερα χωρίζουν τις εγγραφές τους σε μπλοκ ή σελίδες ως κόμβους. Το συγκεκριμένο είδος κατάτμησης ονομάζεται τοπική κατάτμηση σε ένα φυσικό αρχείο. Αντιθέτως, παγκόσμια κατάτμηση ονομάζεται η διαδικασία κατά την οποία μια ταξινομημένη διαδρομή που μπορεί να είναι ένα αρχείο, είναι καταναμημένο σε πολλά φυσικά αρχεία. (Mao, 2022)



Εικόνα 8: Εκκαθάριση των MemTables και του πίνακα Delete.

Κάθε εγγραφή που εισάγεται ή ενημερώνεται συλλέγεται στον πίνακα MemTable όπου και ομαδοποιείται με τα υπόλοιπα. Μόλις χρησιμοποιηθεί όλη η χωρητικότητα του MemTable, τότε ξεκινάει η λειτουργία εκκαθάρισης του προς το δίσκο, κατασκευάζοντας έναν πίνακα SSTable. Η λειτουργία εκκαθάρισης, ταξινομεί αρχικά τις εγγραφές που υπάρχουν στον MemTable PM και στην συνέχεια γράφει μαζικά τις ήδη ταξινομημένες εγγραφές σε έναν πίνακα SSTable P2.

Για την υλοποίηση της ταξινόμησης των εγγραφών στον πίνακα MemTable, χρησιμοποιείται ένας συγκριτής (comparator), ο οποίος συγκρίνει δύο κλειδιά. Ένα κλειδί τοποθετείται σε έναν διαφορετικό πίνακα διαγραφής PD στον χώρο της μνήμης, όταν μια εγγραφή διαγράφεται (π.χ 2). Για την εκκαθάριση του PD παράλληλα με τον PM, χρησιμοποιούμε εγγραφές αντι-ύλης στον πίνακα P2 (π.χ -2). (Mao, 2022)



Εικόνα 9: Διαδικασία συγχώνευσης για δύο ευρετήρια.

Όσο περισσότερο αυξάνεται ο αριθμός των SSTables, οι αναγνώσεις γίνονται όλο και πιο αργές. Για την αποφυγή αυτής της μείωσης της αποδοτικότητας στην ανάγνωση, θα πρέπει να υλοποιηθεί μια συγχώνευση στους SSTables μέσω μιας πολιτικής συγχώνευσης (**merge policy**).

Η διαδικασία που πραγματοποιείται μέσα από μια λειτουργία συγχώνευσης, συνιστά την σάρωση όλων των εγγραφών από όλους τους πίνακες συγχώνευσης και την δημιουργία ταξινομημένης ροής μέσω μιας ουράς προτεραιότητας και του ίδιου συγκριτή, ενώ τέλος την μαζική

γραφή των μοναδικών εγγραφών σε νέα SSTables. Οι εγγραφές παλαιάς έκδοσης διαγράφονται κατά τη διάρκεια της συγχώνευσης, αφήνοντας στη θέση τους τις νεότερες εκδόσεις. Μετά την συμμετοχή μιας ταξινομημένης εκτέλεσης σε μια συγχώνευση, υπάρχει η δυνατότητα διαγραφής μιας εγγραφής αντι-ύλης.

Τα ερωτήματα ανάγνωσης αρχικά ελέγχουν τα μεταδεδομένα των πινάκων και στην συνέχεια προσθέτουν όλους τους πίνακες των οποίων το εύρος κλειδιών περιέχει το κλειδί που επιθυμούμε σε μια διατεταγμένη λίστα λειτουργικών πινάκων. Οι πίνακες αυτοί διαθέτουν τις εγγραφές που χρειάζονται για την απάντηση των ερωτημάτων και ο συνολικός αριθμός τους μπορεί να χρησιμοποιηθεί για τον υπολογισμό της ενίσχυσης ανάγνωσης (read amplification) και εν τέλη την μέτρηση του κόστους ανάγνωσης στην χειρότερη περίπτωση. (Mao, 2022)

### 3.3.2) Δέντρο LSM με βάση τη στοίβα (Stack-based LSM tree)

Στη δομή ενός δέντρου LSM με βάση τη στοίβα, τα SSTables αποτελούν ταξινομημένες διαδρομές οι οποίες διατάσσονται με βάση το χρόνο που δημιουργήθηκε από εκκαθαρίσεις ή συγχωνεύσεις. Στην συγκεκριμένη δομή ωστόσο, οι πολιτικές συγχώνευσης λαμβάνουν υπόψη συνθήκες μεγέθους και με την υλοποίηση μιας συγχώνευσης, μπορούν να συγχωνευτούν μόνο διαδοχικά SSTables και ως αποτέλεσμα παράγεται μόνο ένα ενιαίο SSTable. Αυτού του είδους οι πολιτικές συγχώνευσης ονομάζονται στυλ κλιμάκωσης (**tiering style**).

Οι πολιτικές συγχώνευσης tiering style αποτελούν ένα μέρος μόνο των συνολικών πολιτικών συγχώνευσης με βάση τη στοίβα. Σε αντίθεση με τις γενικές υποκατηγορίες των πολιτικών συγχώνευσης με βάση τη στοίβα όπου δεν υπάρχει σύνδεση μεταξύ του μεγέθους των SSTables και της φρεσκάδας, στις tiering style πολιτικές, τα μεγέθη των SSTables είναι μη φθίνοντα σε σχέση με τον χρόνο δημιουργίας τους. Με αυτόν τον περιορισμό, σκοπός είναι οι αρχικοί SSTables να μην έχουν μικρότερο μέγεθος από τους νεότερους.

Τέλος, μια υποκατηγορία των πολιτικών που είναι βασισμένες στη στοίβα ονομάζονται πολιτικές περιορισμένου βάθους (**bounded-depth policies**), με τις οποίες περιορίζεται ο συνολικός αριθμός των SSTables σε έναν σταθερό αριθμό, με αποτέλεσμα την συγκράτηση της χειρότερης περίπτωσης ενίσχυσης ανάγνωσης.

### 3.3.3) Ισοπεδωμένο δέντρο LSM (Leveled LSM tree)

Στη δομή ενός δέντρου LSM με επίπεδα, τα επίπεδα αποτελούν ταξινομημένες διαδρομές που τις περισσότερες φορές χωρίζονται σε πολλαπλά διαχωρισμένα SSTables που διαθέτουν την ίδια χωρητικότητα.

*<< Ο αριθμός των SSTables στο επίπεδο  $i$  >= 2 είναι  $B$  φορές μεγαλύτερος από τον αριθμό στο επίπεδο  $i-1$ . Μπορεί επίσης να υπάρχει ένα ειδικό επίπεδο 0 το οποίο περιέχει  $B0$  SSTables ως*

πολλαπλά ταξινομημένα τρεξίματα. Όταν ένα επίπεδο φτάσει στη χωρητικότητα του ( $B0$  ή  $Bi$ ), επιλέγεται ένας *SSTable* και συγχωνεύεται με όλους του επικαλυπτόμενους *SSTables* στο επόμενο επίπεδο, δημιουργώντας έναν ή περισσότερους νέους *SSTables* στο επόμενο επίπεδο. >> (Mao, 2022)

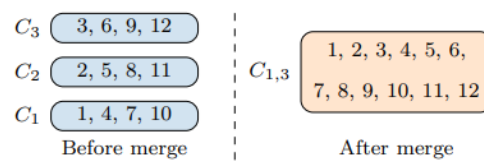
Ενώ όμως γενικότερα κάθε *SSTable* μπορεί να επιλεγεί, στο επίπεδο 0 πρέπει να επιλέγεται πάντα ο παλαιότερος.

Ένα σημειακό ερώτημα θα πρέπει με την σειρά του να ελέγξει όλους τους *SSTables* στο επίπεδο 0, και το πολύ 1 *SSTable* σε όλα τα επίπεδα  $i \geq 1$ .

Ένα ερώτημα εύρους από την άλλη, ίσως χρειαστεί να ελέγξει μερικά *SSTable* σε κάθε επίπεδο, ώστε να μειωθεί η συνολική χωρητικότητα που θα ελεγχθεί

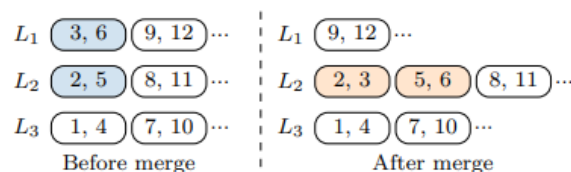
### 3.3.4) Διαφορές μεταξύ *Stack-based* και *Leveled LSM* δέντρα.

Στις παρακάτω εικόνες βλέπουμε κάποιες διαφορές μεταξύ των δύο δομών που αναλύσαμε νωρίτερα σχετικά με τις λειτουργίες συγχώνευσης.



Εικόνα 10: Λειτουργία συγχώνευσης σε δέντρο LSM με βάση τη στοιβία.

Τα μπλε ορθογώνια συμβολίζουν τα *SSTables* εισόδου, ενώ τα πορτοκαλί τα *SSTables* εξόδου. Σε ένα δέντρο LSM με βάση τη στοιβία, οι πίνακες *SSTables*  $C1-C2-C3$  μετά την συγχώνευση τους, δημιουργούν έναν ενιαίο πίνακα *SSTable*  $C1,3$ .



Εικόνα 11: Λειτουργία συγχώνευσης σε ισοπεδωμένο δέντρο LSM.

Στην αντίπερα όχθη, ο *SSTable* με περιεχόμενο  $([3,6])$  από το επίπεδο  $L1$  συγχωνεύεται με τον μοναδικό επικαλυπτόμενο *SSTable* με περιεχόμενο  $([2,5])$  που βρίσκεται στο επίπεδο  $L2$ . Επίσης, είναι ικανή η λειτουργία συγχώνευσης στους συνδυασμούς  $[8,11]$  και  $[7,10]$  ή  $[9,12]$  και  $[8,11]$  αφού δεν επικαλύπτονται από τα ήδη συγχωνευμένα *SSTables*.

### 3.4) Συμπαγείς δομές δεδομένων (Compact data structures).

Οι συμπαγείς δομές δεδομένων, αποτελούν δομές αποθήκευσης που επιτυγχάνουν την ανάκτηση δεδομένων, μέσω του συνδυασμού των μηχανισμών πρόσβασης και μιας συμπιεσμένης αναπαράστασης των δεδομένων. Έτσι, διατηρώντας τα δεδομένα σε συμπιεσμένη μορφή, είναι ευκολότερη η επεξεργασία τους.

Τα θετικά που μπορούμε να προσκομίσουμε μέσα από αυτήν την προσέγγιση είναι τα εξής:

1. Φόρτωση μεγάλου όγκου σύνολα δεδομένων στην κύρια μνήμη.
2. Αποδοτικότερη χρήση της ιεραρχίας της μνήμης.
3. Εξοικονόμηση χρόνου.

Οι συμπαγείς δομές δεδομένων, περιέχουν αρκετές διαφορές σε σχέση με δομές που αναλύσαμε προηγουμένως. Αυτό συμβαίνει καθώς μέσω των συγκεκριμένων δομών δεδομένων, δημιουργούνται δομές αποθήκευσης για δεδομένα **raster** τα οποία χρησιμοποιούνται κυρίως για την επεξεργασία χαρακτηριστικών του χώρου σε γεωγραφικά συστήματα πληροφοριών. Το συγκεκριμένο είδος δομής που θα αναλύσουμε, έχει τη δυνατότητα να ευρετηριάζει το περιεχόμενο της, εκτελώντας τα ερωτήματα.

Γενικότερα όμως, τα συστήματα αυτά αποδίδουν λιγότερο αποτελεσματικά σε μεγάλους αριθμούς διαφορετικών τιμών. (Ladra, 2017)

#### 3.4.1) K2-raster

Το K2-raster αποτελεί μια νέα και αποδοτική δομή αποθήκευσης η οποία αναπαριστά τα δεδομένα raster σε συμπιεσμένο χώρο όπως άλλες δομές αποθήκευσης, με την διαφορά ότι έχει την δυνατότητα παράλληλα να ευρετηριάζει το χώρο και τις τιμές που είναι αποθηκευμένες.

Αν θεωρήσουμε ότι διαθέτουμε ένα πίνακα raster  $P$  μεγέθους  $n \times n$ , με το  $n$  να είναι δύναμη του  $k$ , όπου κάθε κελί αποθήκευσης  $P_{ij}$  αποθηκεύει μια τιμή  $v \geq 0.3$ .

Σε αυτή την περίπτωση, το K2-raster θα διαιρέσει αναδρομικά τον πίνακα  $P$  σε  $k^2$  υποπίνακες, και έπειτα θα δημιουργήσει ένα δέντρο το οποίο θα αντιπροσωπεύει την αναδρομική υποδιαίρεση. Αυτή στρατηγική κατάτμησης είναι ίδια με αυτή που χρησιμοποιεί το αρχικό μοντέλο  $k^2$ -δέντρο ( $k^2$ -tree).

Την στιγμή που όλα τα κελιά ενός υποπίνακα έχουν την ίδια τιμή, η αναδρομική διαίρεση θα σταματήσει. Στην συνέχεια, η ελάχιστη και η μέγιστη τιμή του συγκεκριμένου υποπίνακα θα αποθηκευτούν στους κόμβους του δέντρου ώστε να υπάρχει η δυνατότητα εύρεσης τιμών από τα κελιά.

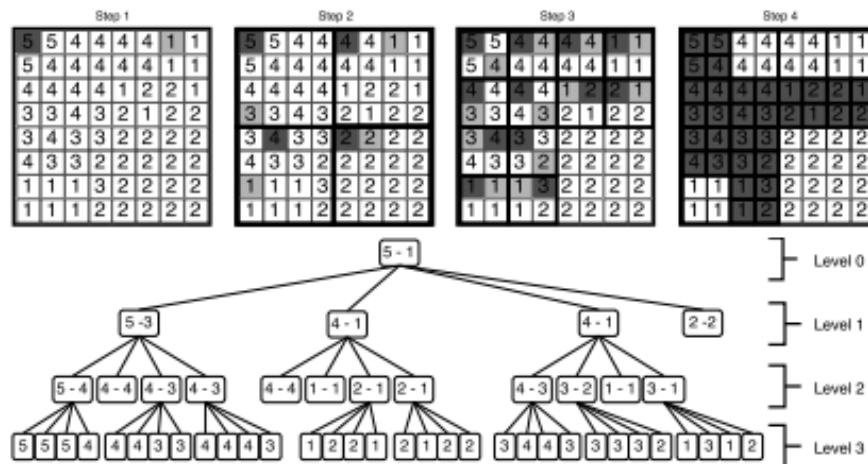
Με αυτό το τρόπο, ομαδοποιεί το χωρικό ευρετήριο του τετραδένδρου, την ελάχιστη και την μέγιστη ευρετηρίαση των raster και μια συμπιεσμένη αναπαράσταση των δεδομένων. Έτσι,

επιτυγχάνεται η αποτελεσματική αναπαράσταση των τιμών στα κελιά και στους κόμβους, αλλά και η ευρετηρίαση τιμών.

### 3.4.2) Κατασκευή και δομές δεδομένων.

Για την κατασκευή της δομής αυτής, η διαδικασία είναι η εξής. Αρχικά θα πρέπει να δημιουργηθεί ένας κόμβος-ρίζα όπου και θα αποθηκευτούν η ελάχιστη και η μέγιστη τιμή του πίνακα. Στην περίπτωση όπου οι τιμές αυτές είναι ίσες, τότε θα αποθηκευτεί μια τιμή ως μέγιστη και η διαδικασία θα ολοκληρωθεί. Αλλιώς, θα αποθηκευτούν και οι δύο τιμές και στην συνέχεια ο πίνακας θα διαιρεθεί σε k2 υποπίνακες.

Κάθε K2 υποπίνακας θα προσθέσει ένα κόμβο στη ρίζα. Η διαδικασία αυτή, επαναλαμβάνεται για κάθε υποπίνακα αναδρομικά, μέχρι η διάσπαση ενός υποπίνακα λάβει υποπίνακες ενός μοναδικού κελιού ή μέχρι η μέγιστη και η ελάχιστη τιμή γίνουν ίσες.



Εικόνα 12: Περίπτωση πίνακα ράστερ και εννοιολογική δενδροειδής αναπαράσταση.

Στην εικόνα 12, βλέπουμε ένα παράδειγμα που περιγράφει την διαδικασία της κατασκευής. Αρχικά στο 1ο βήμα, έχουμε έναν πίνακα 8x8 και από κάτω το k2-raster δέντρο. Ο κόμβος-ρίζα αποθηκεύει τις μέγιστες και τις ελάχιστες τιμές, εφόσον οι τιμές αυτές είναι διαφορετικές.

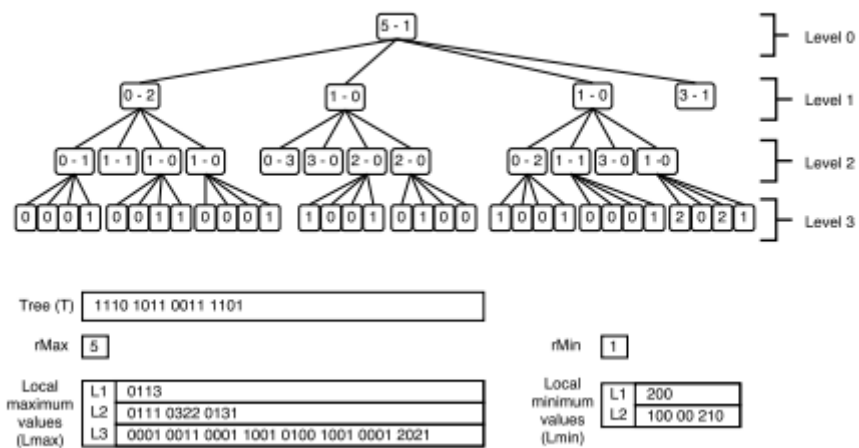
Στην συνέχεια, ο πίνακας θα χωριστεί σε 4 υποπίνακες μεγέθους 4x4. Στο 2ο βήμα παρουσιάζονται επίσης οι νέες μέγιστες και ελάχιστες τιμές. Κάθε φορά που υλοποιείται υποδιαίρεση στους πίνακες, ένας κόμβος προστίθεται στον κόμβο-ρίζα του δέντρου, παρέχοντας τις αντίστοιχες τιμές. Ο κάτω δεξιά πίνακας που φαίνεται να έχει ίδια ελάχιστη και μέγιστη τιμή γίνεται κόμβος φύλλο και δεν θα υποδιαιρεθεί ξανά.

Στο 3ο βήμα, βλέπουμε ότι υπόλοιποι πίνακες υποδιαιρούνται ξανά, ενώ ταυτόχρονα στο επίπεδο 2, υπάρχουν κόμβοι που παρουσιάζουν ίδιες ελάχιστες και μέγιστες τιμές, οι οποίοι επίσης δημιουργούν έναν κόμβο φύλλο.

Τέλος, στο 4ο βήμα, φτάνουμε στο σημείο όπου για κάθε υποδιαίρεση ο κόμβος διαθέτει ένα παιδί για το αντίστοιχο κελί του και αποθηκεύει την ίδια την τιμή του συγκεκριμένου κελιού.

Το παραπάνω παράδειγμα που αναλύσαμε ωστόσο αποτελεί μια περιγραφή υψηλού επιπέδου ενώ μια πραγματική αναπαράσταση περιέχει διάφορες στρατηγικές για την υλοποίηση της συμπίεσης. Δηλαδή:

1. Η αποθήκευση της τοπολογίας του δέντρου πραγματοποιείται ξεχωριστά από τα υπόλοιπα δεδομένα, έτσι ο k2-raster χρησιμοποιεί μια δομή δεδομένων που παρουσιάζει κοινά στοιχεία με τη δομή δεδομένων ενός k2-tree. Ωστόσο, υπάρχουν κάποιες βασικές διαφορές μεταξύ τους, όπως για παράδειγμα η ύπαρξη του 0 σε έναν κόμβο k2-raster υποδηλώνει ότι όλες οι τιμές του υποπίνακα είναι ίσες, ενώ το 1 σημαίνει ότι υπάρχουν διαφορετικές τιμές.
2. Ταυτόχρονα, όπως και η τοπολογία, έτσι και οι μέγιστες τιμές που είναι αποθηκευμένες στους κόμβους του δέντρου, αντιμετωπίζονται ξεχωριστά από τα υπόλοιπα δεδομένα. Η εξοικονόμηση χωρητικότητας είναι σημαντική, έτσι όλες οι τιμές, εκτός από την τιμή της ρίζας, κωδικοποιούνται έτσι ώστε να αποτελούν την διαφορά σε σχέση με τη μέγιστη τιμή των κόμβων-γονέων τους. Το αποτέλεσμα της διαφοράς αυτής, δεν είναι ποτέ αρνητικό καθώς η μέγιστη τιμή ενός κόμβου-γονέα είναι πάντα ίση ή μεγαλύτερη από ότι των παιδιών του. Το δέντρο το οποίο θα λάβουμε, περιέχει διαφορές οι οποίες αποθηκεύονται σε έναν ξεχωριστό πίνακα  $L_{max}$  όπου οι θέσεις των τιμών προσδιορίζονται από την πρώτη διάσχιση κατά πλάτος αυτής της δομής. Η περίπτωση αυτή, είναι ιδανική για τους **DACs** (Directly Addressable Codes) καθώς μπορούν να προσφέρουν αποδοτική συμπίεση και άμεση πρόσβαση σε οποιαδήποτε θέση. Τέλος, η μέγιστη τιμή της ρίζας αποθηκεύεται ως ένας ακέραιος αριθμός.
3. Οι ελάχιστες τιμές, κωδικοποιούνται παρομοίως με τις μέγιστες τιμές, μέσω της διαφοράς με την ελάχιστη τιμή του κόμβου-γονέα. Οι τιμές είναι εξίσου πάντα θετικές, για τον ίδιο λόγο με τις μέγιστες τιμές. Ωστόσο, καμία ελάχιστη τιμή δεν χρειάζεται να αποθηκευτεί στους κόμβους των φύλλων, καθώς η μέγιστη τιμή που είναι αποθηκευμένη είναι αρκετή για να την αναπαραστήσει. Παράλληλα, ονομάζουμε τον πίνακα που θα λάβουμε  $L_{min}$  και επίσης κωδικοποιείται μέσω DACs. Τέλος, η ελάχιστη τιμή της ρίζας αποθηκεύεται ως ένας ακέραιος αριθμός.



**Εικόνα 13: Βελτιωμένη αναπαράσταση με χρήση διαφορών για μέγιστες και ελάχιστες τιμές.**

Στην εικόνα 13, βλέπουμε την ίδια δομή που αναπαραστάθηκε προηγουμένως, με την διαφορά ότι στο συγκεκριμένο σχήμα υπάρχουν βελτιώσεις που αφορούν την κωδικοποίηση και την ξεχωριστή αποθήκευση των μέγιστων και των ελάχιστων τιμών ενός κόμβου μέσω των διαφορών σε σχέση με τις τιμές του κόμβου-γονέα του.

Σκοπός είναι η αποθήκευση των δομών δεδομένων (κάτω μέρος), ενώ το εννοιολογικό δέντρο χρησιμοποιείται κυρίως για την ευκολότερη κατανόηση της δομής.

#### 4) Ομαδοποίηση δεδομένων (Data clustering).

Η ομαδοποίηση αποτελεί ένα τομέα της μηχανικής μάθησης που έχει απασχολήσει τον επιστημονικό κλάδο. Με την κλασική έννοια της ομαδοποίησης, σκοπός είναι οργανωθούν σε ομάδες τα δεδομένα, έτσι ώστε να υπάρχουν συστάδες με δείγματα που περιέχουν ομοιότητες και συστάδες με ανόμοια δείγματα. Με το πέρασμα των χρόνων, τα δεδομένα γίνονται πιο πολύπλοκα, γεγονός που κάνει την διαδικασία χειρισμού των δεδομένων, πολύ δύσκολη για τις ήδη υπάρχουσες μεθόδους ομαδοποίησης.

##### 4.1) Βαθιά ομαδοποίηση (Deep clustering)

Μέσω της μεγάλης προσφοράς της βαθιάς μάθησης (**deep learning**) και ειδικότερα της βαθιάς μάθησης χωρίς επίβλεψη (**deep unsupervised learning**), προτάθηκαν πολλές τεχνικές κυρίως για τα αδόμητα και πολυδιάστατα δεδομένα με σκοπό την αντιμετώπιση του προβλήματος. Ωστόσο, η ύπαρξη πολύπλοκων σχέσεων μεταξύ δειγμάτων, πολλές φορές δεν αποτυπώνεται σωστά με αποτέλεσμα την μη αποδοτική ομαδοποίηση.

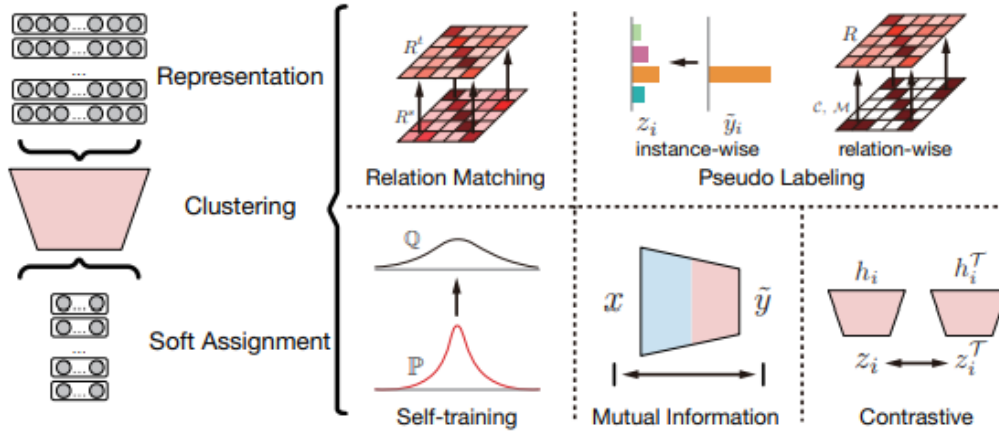
Η αντιμετώπιση των προβλημάτων απαιτεί την βελτιστοποίηση της βαθιάς αναπαράστασης μάθησης και ομαδοποίησης. Αυτό καλείται να πετύχει η βαθιά ομαδοποίηση (**Deep Clustering**) και εστιάζει κυρίως στα εξής: (Zhou, 2022)

1. Σωστή επιλογή διακριτικών αναπαραστάσεων που οδηγούν σε αποδοτικότερες ομαδοποιήσεις.
2. Πως να επιτυγχάνεται η βέλτιστη ομαδοποίηση και εκμάθηση αναπαραστάσεων.
3. Αποφυγή εμποδίων και αλληλοενίσχυση μεταξύ ομαδοποίησης και μάθησης αναπαράστασης.

##### 4.1.1) Ενότητα ομαδοποίησης (Clustering module).

Στην συνέχεια θα παρουσιάσουμε τις διαφορετικές ενότητες ομαδοποίησης που αφορούν την βαθιά ομαδοποίηση. Η διαδικασία ξεκινάει λαμβάνοντας ως είσοδο μια αναπαράσταση χαμηλής διάστασης και παράγοντας ως έξοδο, ετικέτες συστάδων που χρειάζονται σκληρή ομαδοποίηση.

*“Σε μια βαθιά ομαδοποίηση, τα διανυσματικά χαρακτηριστικά τροφοδοτούνται απευθείας μέσω βαθιών νευρωνικών δικτύων και μειώνουν την διάσταση στον αριθμό συστάδων, στη συνέχεια εφαρμόζεται η συνάρτηση softmax στο τελευταίο στρώμα, έτσι ώστε η κατανομή ανάθεσης να μπορεί να καθιερωθεί.” (Zhou, 2022)*



Εικόνα 14: Μέθοδοι βαθιάς ομαδοποίησης.

#### 4.1.2) Βαθιά ομαδοποίηση με αντιστοίχιση σχέσεων (Relation matching deep clustering).

Γενικότερα, κατά την διαδικασία της βαθιάς συσταδοποίησης, οι περιπτώσεις δειγμάτων μπορούν να αναπαρασταθούν σε δύο χώρους, στον d-διάστατο χώρο της ενσωμάτωσης και στον k-διάστατο χώρο των ετικετών.

Οι σχέσεις που συνδέουν τα δείγματα, είναι συνεπείς όσο μειώνονται οι διαστάσεις και αποτελούν σημαντικό εργαλείο για την σημασιολογική σύνδεση μεταξύ της μάθησης αναπαράστασης και της ομαδοποίησης. Υπάρχουν διαφόρων ειδών τρόποι, για την αποτελεσματική αντιστοίχιση των σχέσεων.

$$\mathcal{L}_{RM} = \sum_i \sum_j \ell(R_{ij}^S, R_{ij}^T) \quad (1)$$

“Όπου  $\ell$  είναι ένα μέτρο ταύτισης σχέσεων, π.χ η ομοιότητα συνημιτόνου ή η ευκλείδεια απόσταση,  $R^S$  και  $R^T$  είναι οι σχέσεις στο χώρο πηγής και στόχου, ο οποίος μπορεί να είναι είτε χώρος ενσωμάτωσης είτε χώρος ετικετών.” (Zhou, 2022)

Ωστόσο, για τον υπολογισμό των περιπτώσεων  $N^2$  ζευγών, επειδή αποτελεί μια πολύπλοκη υπολογιστικά διαδικασία, χρησιμοποιούνται μέθοδοι οι οποίες έχουν ως κανόνα την διατήρηση των σχέσεων μόνο των k-κοντινότερων γειτόνων για κάθε περίπτωση ή για τις σχέσεις με υψηλή εμπιστοσύνη.

Η επιλογή αυτή, είναι ιδανική όσον αφορά την αύξηση της απόδοσης στο κομμάτι της εκπαίδευσης του συστήματος, ωστόσο αποτελεί δύσκολη διαδικασία ο ορισμός της άλλης υπερ-παραμέτρου χωρίς επίβλεψη. Τέλος, στα αρχικά στάδια της εκπαίδευσης όπου η απόδοση είναι περιορισμένη, είναι συχνό φαινόμενο να υπάρχουν θορυβώδη ζεύγη σχέσεων.

Ένα από τα θέματα που απασχολεί ακόμα τους ερευνητές, αφορά την αύξηση της απόδοσης του συστήματος, μέσα από το φιλτράρισμα των επιθυμητών σχέσεων ενώ ταυτόχρονα απορρίπτονται οι θορυβώδεις σχέσεις των δειγμάτων.

#### 4.1.3) Βαθιά ομαδοποίηση μέσω ψευδο Ετικετοποίησης (Pseudo Labeling Deep Clustering).

Η συγκεκριμένη διαδικασία, αποτελεί άλλο ένα είδος αντιστοίχισης σχέσεων, με την διαφορά ότι σε αυτή την περίπτωση, οι σχέσεις είναι διακριτές με βάση τη συνέπεια των ετικετών. Οι υπάρχουσες μέθοδοι, διακρίνονται σε δύο κατηγορίες εξαιτίας του τρόπου με τον οποίο χρησιμοποιούνται οι ψευδο ετικέτες.

1. **Ψευδο Ετικετοποίηση κατά περίπτωση (Instance-wise pseudo labeling):** Φιλτράρει μόνο ένα υποσύνολο όλων των δειγμάτων που να διαθέτει μεγάλη εμπιστοσύνη και στην συνέχεια με επίβλεψη, εκπαιδεύει το δίκτυο με απώλεια διασταυρούμενης εντροπίας.

$$\mathcal{L}_{IPL} = - \frac{1}{|X^c|} \sum_i \sum_{k=1}^K y_{ik} \log(z_{ik}) \quad (2)$$

“Όπου  $\mathcal{L}_{IPL}$  δηλώνει την απώλεια της ψευδο ετικετοποίησης ανά περίπτωση,  $X^c$  δηλώνει το υποσύνολο των περιπτώσεων με υψηλή εμπιστοσύνη,  $y_{ik}$  και  $z_{ik}$  είναι η προβλεπόμενη σκληρή ετικέτα και η μαλακή ανάθεση συστάδας.” (Zhou, 2022)

2. **Ψευδο Ετικετοποίηση με βάση τη σχέση (Relation-wise pseudo labeling)**

Η βασική ιδέα της ψευδο ετικετοποίησης με βάση τη σχέση, αναφέρεται στον τρόπο με τον οποίο ομαδοποιούνται τα δείγματα. Δείγματα που διαθέτουν ίδιες ψευδο ετικέτες οργανώνονται σχετικά κοντά, ενώ σε περιπτώσεις με διαφορετικές ψευδο ετικέτες οργανώνονται μακριά το ένα από το άλλο. Επίσης, δημιουργούνται διακριτές σχέσεις μεταξύ των δειγμάτων, για την βελτίωσης της εκμάθησης της αναπαράστασης. Δηλαδή, μια υποχρεωτική σύνδεση μεταξύ των δειγμάτων που διαθέτουν ίδιες ψευδο ετικέτες και καμία σύνδεση για την αντίθετη περίπτωση δειγμάτων.

$$\mathcal{L}_{RPL} = \frac{1}{|C|} \sum_{\{i,k\} \in C} R_{ik} - \frac{1}{|M|} \sum_{\{i,j\} \in M} R_{ij} \quad (3)$$

“Όπου  $M$  είναι το σύνολο των σχέσεων που πρέπει να συνδέονται και  $C$  είναι το σύνολο των σχέσεων που δεν μπορούν να συνδέονται,  $R_{ij}$  είναι η ομοιότητα των περιπτώσεων  $x_i$  και  $x_j$  στον χαμηλής διάστασης χώρο ενσωμάτωσης” (Zhou, 2022)

Από ότι παρατηρούμε, στον χώρο της ομαδοποίησης, η ψευδο ετικετοποίηση έχει προσφέρει την μάθηση με ημι επίβλεψη. Παρόλα αυτά, η ποιότητα των φιλτραρισμένων ετικετών μπορεί να επηρεάσει αρνητικά την απόδοση του συστήματος.

#### 4.1.4) Βαθιά ομαδοποίηση με αυτοεκπαίδευση.

Η μέθοδος της αυτοεκπαίδευσης αναφέρεται στην βελτιστοποίηση της λειτουργίας που αφορά την ανάθεση των συστάδων μέσω της μείωσης της απόκλισης KL.

$$\mathcal{L} = D_{KL}(\mathbb{P}||\mathbb{Q}) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \quad (4)$$

“Όπου το  $\mathbb{Q}$  είναι η κατανομή ανάθεσης συστάδων και  $\mathbb{P}$  είναι η βοηθητική κατανομή.  $q_{ik}$  και  $p_{ik}$  δηλώνει την πιθανότητα η περίπτωση  $x_i$  να ανήκει στη συστάδα  $K$ .” (Zhou, 2022)

$$q_{ik} = \frac{(1+||h_i - c_k||_2^2/a)^{-\frac{a+1}{2}}}{\sum_j (1+||h_i - c_j||_2^2/a)^{-\frac{a+1}{2}}} \quad (5)$$

“Όπου  $h_i$  είναι η αναπαράσταση της περίπτωσης δεδομένων  $x_i$ , και  $c_k$  είναι η αναπαράσταση της συστάδας  $k$ ,  $a$  είναι ο βαθμός ελευθερίας της κατανομής  $t$  του μαθητή.” (Zhou, 2022)

$$p_{ik} = \frac{q_{ik}^2/f_k}{\sum_j q_{ij}^2/f_j} \quad (6)$$

“Όπου  $f_k = \sum_i q_{ik}$  είναι οι συχνότητες μαλακών συστάδων” (Zhou, 2022)

Η πολυτιμότητα της έχει να κάνει με δύο ιδιότητες που την χαρακτηρίζουν:

1. Το μοντέλο παρουσιάζει μεγαλύτερη προσοχή στα δείγματα με υψηλή εμπιστοσύνη, τα οποία στην συνέχεια, μειώνουν τον αντίκτυπο των δειγμάτων με χαμηλή εμπιστοσύνη.
2. Η πιθανότητα ένα δείγμα να βρίσκεται στην  $k$ -ωστή συστάδα αποτελεί το soft cluster frequencies  $f_k$ . Με αυτόν τον τρόπο, είναι εφικτή η αποφυγή του προβλήματος που θεωρεί ότι όλα τα δείγματα βρίσκονται στην ίδια συστάδα.

#### 4.1.5) Μεγιστοποίηση της αμοιβαίας πληροφορίας με βάση την ομαδοποίηση.

Η μέθοδος αυτή δεν περιέχει περιορισμούς σχετικά με την διάσταση των χαρακτηριστικών και μετράει τα επίπεδα εξάρτησης μεταξύ δειγμάτων και ανάθεσης συστάδας. Καταφέρνει να ενσωματώσει αποτελεσματικά τις τεχνικές εκμάθησης αναπαράστασης στην διαδικασία της ομαδοποίησης. Η μέγιστη απόδοση της μεθόδου με βάση την ομαδοποίηση περιγράφεται ως εξής:

$$I(X, Y^{\sim}) = D_{KL}(p(x, y^{\sim}) || p(x)p(y^{\sim})) = \sum_{x \in X} \sum_{y^{\sim} \in Y} p(x, y^{\sim}) \log \frac{p(x, y^{\sim})}{p(x)p(y^{\sim})} \quad (7)$$

Όπου το  $X$  είναι τα αρχικά δεδομένα,  $Y^{\sim}$  οι προβλεπόμενες ετικέτες,  $p(x, y^{\sim})$  η κοινή κατανομή και  $p(x), p(y^{\sim})$  η οριακή κατανομή. Η εκτίμηση της αμοιβαίας πληροφορίας είναι ίδια με εκείνη της μάθησης αναπαράστασης.

#### 4.1.6) Αντιθετική βαθιά ομαδοποίηση.

Βασικός στόχος αυτής της μεθόδου είναι, η έλξη των θετικών ζευγών κοντά ενώ ταυτόχρονα απωθεί τα αρνητικά ζεύγη. Τα ζευγάρια αυτά ωστόσο, χωρίζονται σε τρεις κατηγορίες:

1. Περίπτωση-Αντίθεση περίπτωσης (**Instance-Instance contrast**): Η ανάθεση συστάδας για τα δείγματα, χρησιμοποιείται ως αναπαράσταση και στην συνέχεια εκτελείται κατευθείαν η απώλεια της αντιθετικής εκμάθησης αναπαράστασης.

$$L_{IIC} = - \log \sum_{i=1}^N \frac{\exp(f(z_i, z_i^T)/t)}{\sum_{j=1}^N \exp(f(z_i, z_j^T)/t)} \quad (8)$$

Όπου,  $z_i$  είναι η ανάθεση συστάδας της περίπτωσης  $x_i$ , που προβλέπεται από τη μονάδα συσταδοποίησης.

2. Συστάδα-Αντίθεση συστάδας (**Cluster-Cluster contrast**): Η συγκεκριμένη κατηγορία έχει την τάση να χειρίζεται τις συστάδες ως δείγματα. Για αυτό τον λόγο, ο σκοπός της είναι να τραβήξει την συστάδα και την επανυξημένη εκδοχή της κοντά, ενώ ταυτόχρονα να διώξει άλλες συστάδες.

$$L_{CCC} = - \log \sum_{k=1}^K \frac{\exp(f(c_k, c_k^T)/t)}{\sum_{j=1}^K \exp(f(c_k, c_j^T)/t)} \quad (9)$$

Όπου  $c_k$  είναι η αναπαράσταση της συστάδας  $k$ . Αξίζει να σημειωθεί ότι η αντίθεση μεταξύ των συστάδων, ικανοποιεί τη βασική απαίτηση της συσταδοποίησης ότι κάθε συστάδα πρέπει να είναι ανόμοια, γεγονός που συμφωνεί με την εκμάθηση αναπαράστασης φιλικής προς τη συσταδοποίηση.

3. Περίπτωση-Αντίθεση συστάδας (**Instance-Cluster contrast**): Παρουσιάζει κοινά στοιχεία με την k-means, καθοδηγείται δηλαδή εξίσου μέσω των κεντροειδών της συστάδας. Καθώς η διαδικασία αναπαράστασης των δειγμάτων και του κεντροειδούς της συστάδας γίνεται στον ίδιο χώρο, αυτό σημαίνει ότι οποιοδήποτε δείγμα θα βρίσκεται κοντά στο κεντροειδές της αντίστοιχης συστάδας, ενώ όμως ταυτόχρονα βρίσκεται απομακρυσμένο από τα υπόλοιπα κεντροειδή της συστάδας.

$$L_{ICC} = - \log \sum_{i=1}^N \frac{\exp(f(h_i, c'_i)/t)}{\sum_{j=1}^K \exp(f(h_i, c_j)/t)} \quad (10)$$

Σε αντίθεση με τις περισσότερες μεθόδους που χρησιμοποιούνται για την ομαδοποίηση δειγμάτων, η αντιθετική βαθιά ομαδοποίηση βελτιώνεται όσο περισσότερο αυξάνεται ο αριθμός των δεδομένων. Όπου  $c'$  είναι το αντίστοιχο κεντροειδές της συστάδας  $x_i$ , το οποίο συνήθως εκτιμάται με μια εναλλακτική λύση μέθοδο συσταδοποίησης. Αυτό μπορεί επίσης να γίνει κατανοητό ως μεγιστοποίηση της αμοιβαίας πληροφορίας μεταξύ της αναπαράστασης και της ανάθεσης συστάδας με την αύξηση των δεδομένων. (Zhou, 2022)

#### 4.2) Ομαδοποίηση διαχρονικών δεδομένων με αργά μεταβαλλόμενες τάσεις.

Οι διαχρονικές ομαδοποιήσεις αποτελούν χρήσιμο εργαλείο για πολλούς τομείς όπως η ιατρική, η κοινωνιολογία κ.ά. Μέσω αυτής της λειτουργίας, είναι εφικτή η συλλογή μεγάλου όγκου δεδομένων από υποκείμενα για οποιαδήποτε χρονική στιγμή. Αυτό έχει ως αποτέλεσμα, την δημιουργία προφίλ για τα υποκείμενα, υπολογίζοντας επίσης την ετερογένεια και τις διαφορετικές τάσεις του κάθε υποκειμένου. Ο μεγάλος όγκος υποκειμένων και παρατηρήσεων βοηθάει στην λήψη καλύτερων συμπερασμάτων. Τα εντατικά διαχρονικά δεδομένα (ILD) ορίζονται ως τα δεδομένα υψηλής συχνότητας. Στην συνέχεια θα δούμε κάποιες μεθόδους που χρησιμοποιούνται για την διαχρονική ομαδοποίηση. (Teuling, 2021)

##### 4.2.1) Μέθοδοι

“Η τροχιά ενός υποκειμένου περιγράφεται  $i \in I$ , όπου  $I$  συμβολίζει το σύνολο των διαθέσιμων υποκειμένων. Οι μετρήσεις  $T$  από  $y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,T})$  όπου οι μετρήσεις  $y_{i,j}$  καταγράφονται τη στιγμή  $t_{i,j}$ . Μέσω του  $Ey_{i,j} = \psi_i(t_{i,j})$  περιγράφουμε το ILD  $\psi_i(t)$  για το υποκείμενο  $i \in I$  με την

μέτρηση  $y_{i,j}$  την χρονική στιγμή  $t_{i,j}$  με το  $j=1,2,\dots, T$ , και το  $\psi_i$  μια συνεχή συνάρτηση  $\psi_i: R_+ \rightarrow R$ . Θα

επικεντρωθούμε σε πολωνυμικά χρονικά προφίλ:  $\psi_i(t) = \sum_{r=0}^p \beta_{i,r} t^r$ . Εναλλακτικά, μπορούμε επίσης

να μελετήσουμε γραμμικά προφίλ κατά τεμάχια όταν εφαρμόζουμε αφελείς μεθόδους ομαδοποίησης, αλλά αυτά τα προφίλ θα αναπτυχθούν σε ομάδες συμμετεχόντων. Σε επίπεδο υποκειμένου, το γραμμικό μοντέλο ορίζεται ως εξής:

$$\psi_i(t) = \sum_{r=1}^T (a_{i,r} + \beta_{i,r} t) 1_{(t_{i,r-1}, t_{i,r})}(t)$$

με τους περιορισμούς  $a_{i,r} = a_{i,r-1} + \beta_{i,r-1} t_{i,r}$ . ” (Teuling, 2021)

#### 4.2.2) Διαμήκης k-means (Longitudinal k-means).

Η διαμήκης k-means αποτελεί μια αφελή προσέγγιση που παρόλα αυτά χρησιμοποιείται αρκετά. Σε αυτή την περίπτωση θεωρούμε ότι τα διανύσματα των παρατηρήσεων διαθέτουν ίσο μήκος και ευθυγραμμισμένα. Δηλαδή  $t_{m,j} = t_{n,j}$  για  $m, n \in I, j = 1, \dots, T$ .

“Ο αλγόριθμος *k-means* στοχεύει στην εύρεση της διαμέρισης  $I_1, I_2, \dots, I_G$  με  $\bigcup_{g=1}^G I_g = I$  και  $I_g \cap I_h = \emptyset$  όταν  $g \neq h$  που ελαχιστοποιεί τη διακύμανση εντός της συστάδας, η οποία μεγιστοποιεί τη διακύμανση μεταξύ των συστάδων. Η αντικειμενική συνάρτηση δίνεται από τη σχέση:

$$\arg \min_{I_1, I_2, \dots, I_G} \sum_{g=1}^G \sum_{i \in I_g} \|y_i - \mu_g\|^2$$

με το  $\mu$  να αποτελεί το μέσο διάνυσμα των στοιχείων της ομάδας, δηλαδή  $\mu_g = |I_g|^{-1} \sum_{i \in I_g} y_i$  ”

(Teuling, 2021)

Μέσω αυτού του αλγορίθμου, η αρχική και τυχαία κατάτμηση που προέρχεται από τις επαναληπτικές εκτελέσεις που γίνονται, βελτιώνεται μέχρι να ληφθεί η αποτελεσματικότερη επιλογή. Οι ομάδες που δημιουργούνται διακρίνονται από ομοιογένεια, τα υποκείμενα δηλαδή που βρίσκονται σε μια ομάδα, ακολουθούν την ίδια τροχιά.

#### 4.2.3) Ομαδοποίηση δύο βημάτων (Two-step clustering).

Το συγκεκριμένο μοντέλο ομαδοποίησης χρησιμοποιεί ένα μοντέλο καμπύλης ανάπτυξης με στόχο την μοντελοποίηση των τροχιών και την μέθοδο k-means για την ομαδοποίηση των παραμέτρων των υποκειμένων. Αποτελεί μια μικτή μέθοδο και περιγράφει τις ενιαίες πορείες των ομάδων (σταθερές επιδράσεις) αλλά και την απόκλιση των υποκειμένων από αυτή την ομαδική πορεία (τυχαίες επιδράσεις).

Οι τροχιές περιγράφονται συνήθως από ένα πολυώνυμο της τάξης 1 ή 2. Μια τροχιά που περιγράφεται με βάση ένα πολυώνυμο τάξης K και τυχαία αποτελέσματα σε όλους τους όρους δίνεται από την σχέση:

$$y_{i,j} = \sum_{k=0}^K \beta_{k,i} t_{i,j}^k + \varepsilon_{i,j},$$

$$\beta_{k,i} = \alpha_k + \zeta_{k,i}.$$

εδώ, το  $\alpha_k$  παρουσιάζει το k-οστό συντελεστή τάξης της πολυωνυμικής τροχιάς. Το  $\zeta_{k,i}$  παρουσιάζει την τυχαία επίδραση του υποκειμένου i για τον k-οστό συντελεστή και το  $\varepsilon_{i,j}$  παρουσιάζει το σφάλμα μέτρησης. (Teuling, 2021)

Οι τυχαίες επιδράσεις γενικότερα, διατηρούν μέση τιμή ίση με το 0, μη δομημένη διακύμανση, δεν συνδέονται με κάποιο τρόπο με το σφάλμα μέτρησης και αποτελούν φυσιολογικά κατανομημένες πολυμεταβλητές. Το σφάλμα μέτρησης από την άλλη, έχει επίσης μέση τιμή ίση με το 0 και κοινή διακύμανση ενώ ταυτόχρονα είναι κατανομημένο ανεξάρτητα.

Η χρήση γραμμικών αμερόληπτων προγνωστικών δεικτών (BLUPS) βοηθάει στον υπολογισμό των τυχαίων επιδράσεων, και στην συνέχεια μεταφέρει τα αποτελέσματα στον αλγόριθμο k-means ως διανύσματα εισόδου.

#### 4.2.4) Μοντελοποίηση τροχιάς με βάση την ομάδα (Group-based trajectory modeling).

Αυτό το μοντέλο (GBTM) αποτελεί μια όχι πολύπλοκη μέθοδο και εμφανίζει σχετικά κοινά στοιχεία με το μοντέλο k-means καθώς θεωρούμε ότι τα υποκείμενα που βρίσκονται μέσα σε μια ομάδα, ακολουθούν πλήρως το προφίλ της ομάδας.

<<Για μια δεδομένη τροχιά  $y_i$ , οι παρατηρήσεις της περιγράφονται από την ομαδική τροχιά της ομάδας g ως εξής:

$$y_{i,j}^{(g)} = \sum_{k=0}^K a_k^{(g)} t_{i,j}^k + \varepsilon_{i,j},$$

όπου το  $a_k^{(g)}$  δηλώνει το k-οστό συντελεστή για το πολυώνυμο της ομάδας g και το  $\varepsilon_{i,j}$  περιγράφει το υπόλοιπο την στιγμή  $t_{i,j}$ . Σε αυτό το πλαίσιο, οι τροχιές των υποκειμένων  $\psi_i(t_{i,j})$  είναι όλες ίδιες με

$\sum_{k=0}^K a_k^{(g)} t_{i,j}^k + \varepsilon_{i,j}$ , όταν το υποκείμενο i ανήκει στην ομάδα  $I_g$ . Το υπόλοιπο θεωρείται ότι είναι

ανεξάρτητα κανονικά κατανομημένο με μηδενική μέση τιμή. Ο οριακός μέσος όρος ενός GBTM δίνεται από τη σχέση:

$$E(y_{i,j}) = \sum_{g=1}^G \pi^{(g)} \sum_{k=0}^K a_k^{(g)} t_{i,j}^k.$$

Εδώ το  $\pi^{(g)}$  δηλώνει το ποσοστό της ομάδας  $g$  με  $0 \leq \pi^{(g)} \leq 1$  και  $\sum_{g=1}^G \pi^{(g)} = 1$  (Teuling, 2021)

#### 4.2.5) Μοντελοποίηση μείγματος ανάπτυξης (Growth mixture modeling).

Η παρούσα μέθοδος (**GMM**) θα μπορούσε να θεωρηθεί ως μια γενίκευση του μοντέλου GBTM και χρησιμοποιεί ένα μείγμα ανάπτυξης μοντέλων καμπύλης με το οποίο μπορεί να βρίσκει τις ομάδες με ετερογένεια.

<<Μια δεδομένη τροχιά  $y_i$  περιγράφεται από την ομάδα  $g$  με τον ακόλουθο τρόπο:

$$y_{i,j}^{(g)} = \sum_{k=0}^K \beta_{k,i}^{(g)} t_{i,j}^k + \varepsilon_{i,j}^{(g)},$$

$$\beta_{k,i}^{(g)} = a_k^{(g)} + \zeta_{k,i}^{(g)}.>> \text{ (Teuling, 2021)}$$

Το GMM αν και αποτελεί σημαντικά πολυπλοκότερη μέθοδο σε σχέση με το GBTM καθώς υπάρχει η επιπλέον εκτίμηση των τυχαίων επιδράσεων στο αντικείμενο των ομάδων, επιλέγεται τις περισσότερες φορές χάρις την ευελιξία του σχετικά με τον καθορισμό των τυχαίων επιδράσεων και των σχέσεων μεταξύ τους.

#### 4.2.6) Μοντελοποίηση μικτών χρονικά μεταβαλλόμενων επιδράσεων (Mixed time-varying effect modeling).

Η διαφορά που συναντάμε σε αυτήν την μέθοδο (**TVEM**) σε σχέση με τις άλλες είναι ότι υπάρχει η δυνατότητα, οι συντελεστες παλινδρόμησης που απεικονίζουν τη σχέση μεταξύ των συνδιακυμάνσεων και του αποτελέσματος να διαφοροποιούνται καθώς περνάει ο χρόνος.

“Η σχέση μεταξύ του χρόνου και της έκβασης περιγράφεται από μία ομαλή συνάρτηση  $\psi(t)$ , επομένως οι διαμήκεις τροχιές μπορούν να περιγραφούν από την TVEM που δίνεται από τη σχέση:

$$y_{i,j} = \psi(t_{i,j}) + \varepsilon_{i,j}.$$

Ένα P-spline σε συνδυασμό με πολυώνυμα χαμηλής τάξης μπορεί να παρουσιάσει χρονικά διαστήματα με τη δυνατότητα μετάβασης σε αυτά και χρησιμοποιείται για την μοντελοποίηση της

συνάρτησης  $\psi$ . Η εκτίμηση της αποτελεσματικότητας των P-splines μπορεί να γίνει μέσω της διαδικασίας των ελαχίστων τετραγώνων (**least-squares**) καθώς αναφερόμαστε σε γραμμικό μοντέλο.

#### **4.3) Ομαδοποιήσεις στην ανάλυση και εξόρυξη δεδομένων (Clustering in data analysis and mining).**

Αν και υπάρχουν πολλές τεχνικές για την επιτυχή ομαδοποίηση δεδομένων, η συμμετρία είναι μια ιδιότητα που βοηθάει στην ανακατασκευή σχημάτων και την αναγνώριση δεδομένων με σκοπό την προσφορά στον τομέα της ανάλυσης και εξόρυξης δεδομένων, ενώ ωθεί τις μεθόδους προς αναζήτηση συστάδων που είναι συμμετρικές ως προς το κέντρο τους. (Alotaibi, 2022)

##### **4.3.1 Άλλες μέθοδοι.**

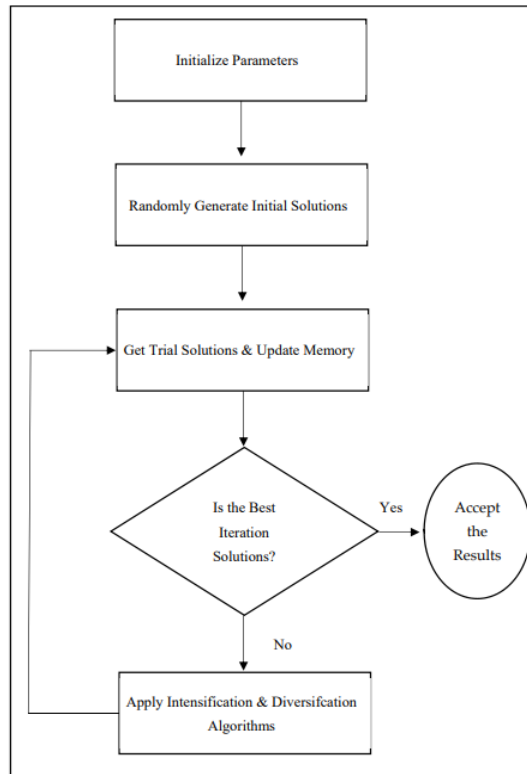
Ένας από τους πιο συνηθισμένους αλγορίθμους για την ομαδοποίηση δεδομένων είναι ο K-means. Παρόλα αυτά, κάποιες αδυναμίες του όπως η ευαισθησία στην αρχικοποίηση και η παγίδευση σε τοπικά βέλτιστα, τον καθιστούν όχι ιδιαίτερα αποτελεσματικό. Για την αποφυγή αυτών των περιορισμών και γενικότερα σύνθετων προβλημάτων, χρησιμοποιούνται μεταερευτικοί αλγόριθμοι. Οι μεταερευτικοί αλγόριθμοι μπορούν να εξερευνήσουν και να εκμεταλλευτούν τον χώρο αναζήτησης.

Ο αλγόριθμος Tabu Search (TS) αποτελεί έναν μεταερευτικό αλγόριθμο ο οποίος βασιζόμενος σε αντικείμενα μνήμης, εκμεταλλεύεται και εξερευνά τον χώρο με χαμηλό κόστος, αποφεύγοντας τις επισκέψεις σε περιοχές που έχουν ήδη επισκεφθεί.

Στις επόμενες ενότητες θα παρουσιαστεί ένας νέος μεταερευτικός αλγόριθμος που ονομάζεται MHTSASM και αποτελεί μια ανεπτυγμένη μορφή του TS.

##### **4.3.2) Αλγόριθμος MHTSASM.**

Ο αλγόριθμος MHTSASM που θα αναλύσουμε στην συνέχεια, αποτελεί μια μέθοδο για την ομαδοποίηση δεδομένων και περιέχει στοιχεία από τους αλγορίθμους TS και K-means. Η λειτουργία του αλγορίθμου ξεκινάει με μια αρχικά τυχαία λύση. Στην συνέχεια, αφού λάβει όλους τους γείτονες (δοκιμαστικές λύσεις) της αρχικής αυτής λύσης, πραγματοποιεί ενημερώσεις στα στοιχεία της μνήμης για κάθε επανάληψη που υλοποιείται. Επίσης, αυξάνει την απόδοση της αναζήτησης μέσω τεχνικών εντατικοποίησης και διαφοροποίησης.



Εικόνα 15: Το πλαίσιο του αλγορίθμου MHTSASM.

#### 4.3.2.1) Εντατικοποίηση και διαφοροποίηση.

Οι δύο αυτές τεχνικές αποτελούν σημαντικές συνιστώσες στον αλγόριθμο MHTSASM. Η τεχνική της εντατικοποίησης έχει την τάση να μεταναστεύει στις αποτελεσματικότερες λύσεις. Μπορεί να επιστρέφει σε αξιοσημείωτες περιοχές με σκοπό την εκτενέστερη εξερεύνηση τους και καθιστούν απαραίτητη την αποθήκευση των καλύτερων λύσεων (**elite solutions**) στο EL, με σκοπό τη δυνατότητα έρευνας των γειτόνων τους. Η ανταποκρινόμενη μνήμη ASMu δίνει την ευχέρεια στο MHTSASM να επιτηρεί τις καλύτερες καταταμήσεις στον χώρο αναζήτησης.

Στην αντίπερα όχθη, μέσω της στρατηγικής της διαφοροποίησης, ο αλγόριθμος μπορεί να επεκτείνει την διαδικασία εξερεύνησης όσον αφορά περιοχές που δεν έχουν επισκεφθεί προηγουμένως και να παραχθεί μια νέα λύση. Έτσι, διαλέγει καταταμήσεις με χαμηλό αριθμό επισκέψεων.

Ο υπολογισμός των νέων κέντρων  $C^j$ ,  $j=1, \dots, K$  όπου αναθέτονται τα  $x$  αντικείμενα δεδομένων πραγματοποιείται μέσω της μεθόδου προσαρμογής του κέντρου.

$$j^* = \arg \min_{j = 1, \dots, K} ||x^i - C^{(j)}||$$

$$C^{(j)} = \frac{1}{|S_j|} \sum_{x \in S_j} x$$

### Αλγόριθμος 1:

Προσαρμογή κέντρου ( $K$ , διάνυσμα  $C$ ,  $d$ ,  $D$ )

1. Ορίστε  $S_j$ ,  $j = 1, \dots, K$  να είναι κενό
2. Επαναλάβετε τα βήματα (2.1-2.2) για  $i = 1, \dots, n$ .
  - 2.1 Υπολογίστε την ευκλείδεια απόσταση μεταξύ του αντικειμένου δεδομένων  $x_i$  και κάθε κέντρου διάνυσμα  $C^{(j)}$ ,  $j = 1, \dots, K$ .
  - 2.2 Προσθήκη  $x_i$  στο  $S_j^*$ , όπου υπολογισμός της εξίσωσης ( $j^*$ ).

Οπότε,  $S_j^* = S_j \cup x^i$
3. Υπολογίστε ένα νέο σύνολο κέντρων  $\tilde{C} = \{C^{(1)}, \dots, C^{(K)}\}$ ; κάθε κέντρο  $C^{(j)} \in C$  υπολογίζεται σύμφωνα με την ακόλουθη εξίσωση: Υπολογισμός της εξίσωσης ( $\tilde{C}^{(j)}$ ).
4. Επιστροφή  $\tilde{C}^{(j)}$ ,  $j = 1, \dots, K$ . (Alotaibi, 2022)

### Αλγόριθμος 2:

Ο αλγόριθμος της εντατικοποίησης ( $\delta$ ,  $C$ ,  $ASM_u$ ,  $ASM_v$ ,  $d$ ) έχει ως εξής:

1. Για κάθε κέντρο στο  $\tilde{C}$ , επανέλαβε τα ακόλουθα βήματα (1.1-1.4)
  - 1.1. Επιλέξτε  $\delta$  χαρακτηριστικά με την υψηλότερη τιμή βελτίωσης ενημέρωσης στο  $ASM_u$  από τα  $d$  χαρακτηριστικά δεδομένων.
  - 1.2. Επιλέξτε μια κατάτμηση από τις  $p$  κατατμήσεις για κάθε επιλεγμένο χαρακτηριστικό με την υψηλότερη τιμή βελτίωσης ενημέρωσης για το τρέχον κέντρο στο  $ASM_u$ .
  - 1.3. Ενημερώστε την τιμή του επιλεγμένου χαρακτηριστικού υπολογίζοντας μια τυχαία τιμή στο εύρος μεταξύ της τρέχουσας τιμής και της τιμής της επιλεγμένης κατάτμησης.
  - 1.4 Ενημέρωση των στοιχείων μνήμης  $ASM_v$ .
2. Επιστροφή των ενημερωμένων κέντρων  $C, ASM_v$ .

### Αλγόριθμος 3:

<<H μέθοδος της διαφοροποίησης ( $\delta$ ,  $C$ ,  $ASM_v$ ,  $d$ )

1. Για κάθε κέντρο στο  $\tilde{C}$ , επανέλαβε τα ακόλουθα βήματα (1.1-1.4)

- 1.1 Επιλέξτε  $\delta$  χαρακτηριστικά από τα χαρακτηριστικά με τον μικρότερο αριθμό επισκέψεων στο ASMν απλο τα  $d$  χαρακτηριστικά των δεδομένων.
- 1.2 Επιλέξτε μια κατάτμηση από τις  $p$  κατατμήσεις για κάθε επιλεγμένο χαρακτηριστικό με τον μικρότερο αριθμό επισκέψεων για το τρέχον κέντρο στο ASMν.>> (Alotaibi, 2022)

Η μέθοδος της παραγωγής δοκιμαστικών λύσεων:

#### **Αλγόριθμος 4:**

<<Μέθοδος δοκιμαστικών λύσεων ( $C, \mu, ASMν, d$ )

1. Επαναλάβετε τα ακόλουθα βήματα (1.1-1.3) για  $j = 1, \dots, \mu$ .
  - 1.1 Ορίστε το  $C^{(j)}$  ίσο με τα κέντρα εισόδου  $C$ .
  - 1.2 Επιλέξτε  $\eta$  χαρακτηριστικά τυχαία από τα  $d$  χαρακτηριστικά των δεδομένων
  - 1.3 Επαναλάβετε τα βήματα (1.3.1-1.3.2) για  $m = 1, \dots, \eta$ .
    - 1.3.1 Ενημερώστε τις τιμές των επιλεγμένων χαρακτηριστικών στο  $C^{(j)}$  δημιουργώντας νέες τιμές τυχαία.
    - 1.3.2 Ενημέρωση των στοιχείων μνήμης ASMν.
2. Επιστροφή του ενημερωμένου συνόλου κέντρων του  $C^{(j)}$ ,  $j = 1, \dots, \mu$ , ASMν.>> (Alotaibi, 2022)

#### **Αλγόριθμος 5:**

<<Μέθοδος βελτίωσης ( $C, \mu, ASMν, ASMu, EL, d$ )

1. Εφαρμόστε τον αλγόριθμο της παραγωγής δοκιμαστικών λύσεων για την δημιουργία  $\mu$  κέντρων στο  $C^{\sim}$  κοντά στο σύνολο των κέντρων  $C$ .
2. Ενημερώστε τα στοιχεία της μνήμης,
3. Εάν επιτευχθεί η βελτίωση, τότε μεταβείτε στο βήμα 1. Διαφορετικά, μεταβείτε στο βήμα 4.
4. Επιστροφή των  $C, ASMν, ASMu, EL$ .>> (Alotaibi, 2022)

Και τέλος, ο αλγόριθμος MHTSASM:

#### **Αλγόριθμος 6:**

#### <<Αλγόριθμος MHTSASM

1. Αρχικοποίηση. Ορίστε τιμές για τον αριθμό των συστάδων  $\mathbf{K}$ , τον αριθμό των δοκιμών γειτονιάς, και ορίστε τα στοιχεία μνήμης TL, ASMv, ASMu και EL να είναι κενά.
2. Δημιουργήστε  $\mathbf{K}$  τυχαία αρχικά κέντρα  $\mathbf{C}$ .
3. Αξιολογήστε τα αρχικά κέντρα  $\mathbf{C}$  και ενημερώστε τα στοιχεία μνήμης TL και EL.
4. Κύριος βρόγχος.
  - 4.1. Δημιουργία  $\mu$  δοκιμαστικών λύσεων  $\tilde{\mathbf{c}}^{(j)}$ ,  $j = 1, \dots, \mu$  εφαρμόζοντας τον αλγόριθμο παραγωγής δοκιμαστικών λύσεων.
  - 4.2 Επαναλάβετε τα ακόλουθα βήματα (4.2.1-4.2.2) για  $j = 1, \dots, \mu$ .
    - 4.2.1. Ρύθμιση του συνόλου των κέντρων διάνυσμα  $\mathbf{C}^{(j)}$  εφαρμόζοντας τον αλγόριθμο προσαρμογής κέντρου: υπολογισμός της αντικειμενικής συνάρτησης  $f(\mathbf{C}^{(j)})$  εφαρμόζοντας την εξίσωση του  $\mathbf{C}^{(j)}$ .
    - 4.2.2. Εάν η αντικειμενική συνάρτηση  $f(\mathbf{C}^{(j)})$  δεν είναι καλύτερη από την  $f(\mathbf{C})$ , τότε πηγαίνετε στο βήμα 4.2.1, διαφορετικά, θέστε το  $\mathbf{C}$  ίσο με  $\tilde{\mathbf{c}}$  και ενημερώστε το ASMu.
  - 4.3. Ενημερώστε τα στοιχεία μνήμης TL και EL.
5. Τερματισμός. Εάν ικανοποιούνται οι συνθήκες τερματισμού, τότε μεταβείτε στο βήμα 8. Διαφορετικά, μεταβείτε στο βήμα 6.
6. Εντατικοποίηση. Εάν απαιτείται εντατική λύση, τότε εφαρμόστε τον αλγόριθμο της διαφοροποίησης για να δημιουργήσετε νέα κέντρα  $\mathbf{C}$ . Ενημερώστε τα στοιχεία μνήμης και μεταβείτε στο βήμα 4.
7. Διαφοροποίηση. Εάν απαιτείται διαφοροποιημένη λύση, τότε εφαρμόστε τον αλγόριθμο της εντατικοποίησης για να δημιουργήσετε νέα κέντρα  $\mathbf{C}$ . Ενημερώστε τα στοιχεία μνήμης και μεταβείτε στο βήμα 4.
8. Τελική βελτίωση. Εφαρμόστε τον αλγόριθμο της μεθόδου βελτίωσης για να βελτιώσετε την καλύτερη λύση που έχει βρεθεί μέχρι στιγμής.>> (Alotaibi, 2022)

| Algorithm Name | Implementation Complexity Level |        |     | Computational Cost Level |        |     |
|----------------|---------------------------------|--------|-----|--------------------------|--------|-----|
|                | High                            | Medium | Low | High                     | Medium | Low |
| K-M            |                                 |        | ✓   |                          |        | ✓   |
| TS             |                                 |        | ✓   |                          |        | ✓   |
| GA             |                                 | ✓      |     | ✓                        |        |     |
| SA             |                                 | ✓      |     |                          | ✓      |     |
| Algorithm 1    |                                 | ✓      |     |                          | ✓      |     |
| J-M+           |                                 |        | ✓   |                          |        | ✓   |
| VNS            |                                 | ✓      |     |                          | ✓      |     |
| VNS-1          |                                 | ✓      |     |                          | ✓      |     |
| G1             |                                 | ✓      |     |                          |        | ✓   |
| G2             |                                 | ✓      |     |                          |        | ✓   |
| KHM            |                                 | ✓      |     |                          | ✓      |     |
| PSO            |                                 | ✓      |     |                          | ✓      |     |
| PSOKHM         | ✓                               |        |     |                          | ✓      |     |
| ACO            | ✓                               |        |     |                          | ✓      |     |
| ACOKHM         | ✓                               |        |     |                          | ✓      |     |
| Proposed       |                                 |        |     |                          |        |     |
| MHTSASM        |                                 |        | ✓   |                          |        | ✓   |
| Approach       |                                 |        |     |                          |        |     |

Πίνακας 5 : Σύγκριση μεταξύ διαφορετικών μεταερευνητικών αλγορίθμων.

#### 4.4) Δίκαιη ομαδοποίηση κατηγορικών δεδομένων.

Μια νέα μεγάλη ανησυχία έχει προκύψει με την συνεχή χρήση αλγορίθμων ταξινόμησης καθημερινά για κοινωνικούς λόγους κ.ά Τα αυτοματοποιημένα αποτελέσματα που παράγονται, καθιστούν ουσιαστικά τον άνθρωπο αμέτοχο σε σημαντικά ζητήματα. Μια λανθασμένη έξοδος ενδέχεται να προκληθεί εξαιτίας της πολυπλοκότητας της διαδικασίας ή του τεράστιου όγκου δεδομένων και μπορεί να προκαλέσει προβλήματα σε ορισμένες κοινωνικές ομάδες που διακρίνονται από κάποια συγκεκριμένα χαρακτηριστικά. Η ανάγκη για την αποφυγή αυτών των σφαλμάτων, οδήγησε στην δημιουργία δίκαιων αλγορίθμων ομαδοποίησης. (C. Santos-Mangudo, 2022)

##### 4.4.1) Μέθοδος.

Αρχικά, θεωρούμε πως το σύνολο των δεδομένων διακρίνεται από ένα προστατευόμενο χαρακτηριστικό και μια αναλογία μεταξύ των τιμών του. Η λειτουργία του αλγορίθμου στοχεύει στην διάσπαση της βάσης δεδομένων σε ομάδες που διακρίνονται από ομοιογένεια καθώς αποτελούνται από συγγενικές παρατηρήσεις.

Η λειτουργία της “δίκαιης πολυ-συστοιχίας” χωρίζεται σε 3 βήματα:

##### Βήμα 1ο:

1. Αρχικά, θα πρέπει να καθορίσουμε τις ανάλογες συστάδες για κάθε χαρακτηριστικό και τις τιμές του ξεχωριστά.
2. Στην συνέχεια, μέσω μιας διαδικασίας συγχώνευσης, όπου ενοποιούνται όλες οι συστάδες που πιθανόν να περιέχει ένα συγκεκριμένο χαρακτηριστικό, στόχος είναι δημιουργία ενός αρχικού συμπλέγματος συστάδων. Αν δηλαδή, έχουμε δύο χαρακτηριστικά με τις τιμές Q, W, E και R, T, Y αντίστοιχα, τότε θα δημιουργηθούν 9 πολυσυστάδες: “QR, QT, QY, WR, WT, WY, ER, ET, EY. Αυτό αποτελεί βέβαια τον μέγιστο αριθμό συστάδων που μπορεί να

χρησιμοποιηθούν. Τις περισσότερες φορές ωστόσο, εφόσον δεν είναι απαραίτητο, πολλά από αυτά δεν χρησιμοποιούνται και άρα μένουν κενά.

## Βήμα 2ο:

1. Έπειτα, πρέπει να εκτιμήσουμε τις συνολικές συμπτώσεις που υπάρχουν μεταξύ των συστάδων. Μεταξύ δύο συστάδων που περιέχουν τα χαρακτηριστικά QR και QT αντίστοιχα, ο αριθμός των συμπτώσεων είναι 1 (Q), ενώ αν οι συστάδες περιείχαν τα χαρακτηριστικά WR και ET αντίστοιχα, ο αριθμός των συμπτώσεων ισούται με 0. Τα αποτελέσματα βρίσκονται στον πίνακα συμπτώσεων (**Coincidence Matrix**).
2. Μέσω του πίνακα συμπτώσεων, θα πρέπει να συγχωνευθούν οι πολυσυστοιχίες με τις περισσότερες συμπτώσεις καθώς διαθέτουν τα περισσότερα χαρακτηριστικά. Εάν υπάρχουν δύο ή περισσότερες περιπτώσεις με τον ίδιο μέγιστο αριθμό συμπτώσεων, τότε χρησιμοποιούμε τον συντελεστή του **Fleiss Kappa** που μπορεί να υπολογίσει τους βαθμούς ομοιότητας. Συστάδες που έχουν συγκριθεί, δεν συγκρίνονται ξανά.

## Βήμα 3ο:

1. Στην συνέχεια, για την επίτευξη της αποδοτικότερης κατανομής του χαρακτηριστικού, θα δημιουργήσουμε ένα πίνακα που θα περιέχει τα βέλτιστα αποτελέσματα της προηγούμενης διαδικασίας και θα συνδέσουμε τις συστάδες μεταξύ τους ανά δύο. Έτσι, θα λάβουμε τις πιο επιθυμητές αναλογίες τιμών του προστατευόμενου χαρακτηριστικού.
2. Τελικά, θα πρέπει να επαναλάβουμε το προηγούμενο βήμα μέχρι να χρησιμοποιηθούν όλες οι συστάδες που θέλουμε. Το αποτέλεσμα που θα παραχθεί, αποτελείται από ένα σύνολο συστάδων το οποίο διέπεται από ομοιογένεια και δικαιοσύνη.

Με την βοήθεια του πίνακα 6, θα δούμε τον τρόπο με τον οποίο λειτουργεί ο αλγόριθμος.

| Observation | Status account | Credit history | Purpose | Savings | Employment | Installment | Gender | Debtors | Residence | Property | Others | Housing | Credits | Job  | Number people | Telephone | Foreign |
|-------------|----------------|----------------|---------|---------|------------|-------------|--------|---------|-----------|----------|--------|---------|---------|------|---------------|-----------|---------|
| 775         | A13            | A34            | A40     | A63     | A71        | 2           | M      | A101    | 4         | A124     | A141   | A153    | 3       | A171 | 1             | A191      | A201    |
| 204         | A11            | A32            | A48     | A61     | A74        | 4           | M      | A101    | 4         | A122     | A143   | A151    | 1       | A173 | 1             | A191      | A201    |
| 699         | A14            | A34            | A43     | A61     | A73        | 4           | M      | A101    | 2         | A123     | A143   | A152    | 2       | A173 | 1             | A191      | A201    |
| 250         | A14            | A32            | A43     | A61     | A71        | 3           | F      | A102    | 4         | A121     | A143   | A151    | 1       | A173 | 1             | A191      | A201    |
| 593         | A14            | A32            | A49     | A64     | A75        | 4           | F      | A101    | 4         | A121     | A141   | A152    | 1       | A172 | 1             | A191      | A201    |
| 661         | A13            | A32            | A43     | A61     | A73        | 3           | M      | A101    | 4         | A121     | A143   | A151    | 1       | A173 | 1             | A191      | A201    |
| 463         | A12            | A32            | A42     | A61     | A72        | 3           | F      | A101    | 1         | A121     | A143   | A151    | 1       | A174 | 1             | A191      | A201    |
| 828         | A14            | A30            | A49     | A61     | A73        | 2           | M      | A101    | 2         | A123     | A142   | A152    | 2       | A173 | 2             | A191      | A201    |
| 845         | A14            | A32            | A42     | A61     | A75        | 4           | M      | A101    | 4         | A122     | A141   | A152    | 3       | A173 | 2             | A192      | A201    |
| 11          | A12            | A32            | A40     | A61     | A72        | 3           | F      | A101    | 1         | A123     | A143   | A151    | 1       | A173 | 1             | A191      | A201    |
| 896         | A14            | A33            | A41     | A65     | A74        | 3           | M      | A101    | 2         | A123     | A142   | A152    | 1       | A174 | 2             | A192      | A201    |
| 379         | A12            | A32            | A40     | A61     | A75        | 4           | M      | A101    | 2         | A124     | A143   | A153    | 1       | A174 | 1             | A192      | A201    |
| 989         | A11            | A32            | A41     | A61     | A71        | 4           | M      | A101    | 2         | A124     | A143   | A153    | 1       | A174 | 1             | A192      | A201    |
| 323         | A11            | A32            | A41     | A61     | A74        | 2           | M      | A101    | 1         | A124     | A143   | A153    | 1       | A174 | 1             | A192      | A201    |
| 716         | A14            | A34            | A41     | A65     | A75        | 1           | M      | A101    | 4         | A123     | A143   | A152    | 2       | A173 | 1             | A191      | A201    |
| 619         | A12            | A32            | A42     | A62     | A73        | 2           | F      | A102    | 4         | A123     | A143   | A151    | 1       | A173 | 1             | A191      | A201    |
| 908         | A12            | A32            | A46     | A65     | A73        | 2           | M      | A101    | 2         | A123     | A143   | A152    | 1       | A173 | 1             | A191      | A201    |
| 955         | A11            | A32            | A40     | A61     | A73        | 4           | F      | A103    | 4         | A122     | A143   | A152    | 1       | A173 | 1             | A192      | A201    |
| 209         | A11            | A32            | A49     | A61     | A73        | 2           | M      | A101    | 2         | A123     | A142   | A152    | 1       | A172 | 1             | A191      | A201    |
| 304         | A11            | A34            | A40     | A61     | A74        | 4           | M      | A102    | 3         | A122     | A143   | A152    | 2       | A173 | 1             | A192      | A201    |

Πίνακας 6: Δείγμα παρατηρήσεων από ένα σύνολο δεδομένων.

Το πρώτο κομμάτι του βήματος 1 του αλγορίθμου, παρουσιάζεται στον πίνακα 76, βάζοντας τις τιμές των χαρακτηριστικών σε αύξουσα σειρά, ανάλογα με το μέγεθος τους και χαρακτηρίζοντας το φύλο (gender) ως προστατευόμενο χαρακτηριστικό με τις τιμές Male και Female.

| <table> <tr><th>Status account</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A13</td><td>2</td><td>1</td></tr> <tr><td>A12</td><td>5</td><td>2</td></tr> <tr><td>A11</td><td>6</td><td>3</td></tr> <tr><td>A14</td><td>7</td><td>4</td></tr> </table>                                     | Status account | Frequency | Cluster | A13  | 2 | 1 | A12  | 5  | 2 | A11                                                                                                                                                                                                            | 6       | 3         | A14                                                                                                                                                                                                                                                   | 7    | 4         | <table> <tr><th>Credit history</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A30</td><td>1</td><td>1</td></tr> <tr><td>A33</td><td>1</td><td>2</td></tr> <tr><td>A34</td><td>4</td><td>3</td></tr> <tr><td>A32</td><td>14</td><td>4</td></tr> </table> | Credit history | Frequency | Cluster                                                                                                                                                                                                                                                                                           | A30        | 1         | 1       | A33                                                                                                                                                                                                                                            | 1         | 2         | A34     | 4  | 3 | A32                                                                                                                                                                                                            | 14            | 4         | <table> <tr><th>Purpose</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A46</td><td>1</td><td>1</td></tr> <tr><td>A48</td><td>1</td><td>2</td></tr> <tr><td>A42</td><td>3</td><td>3</td></tr> <tr><td>A43</td><td>3</td><td>4</td></tr> <tr><td>A49</td><td>3</td><td>5</td></tr> <tr><td>A41</td><td>4</td><td>6</td></tr> <tr><td>A40</td><td>5</td><td>7</td></tr> </table> | Purpose | Frequency | Cluster | A46  | 1  | 1                                                                                                                                                                                                                                                | A48         | 1         | 2       | A42 | 3 | 3 | A43 | 3 | 4 | A49 | 3 | 5 | A41 | 4 | 6 | A40 | 5 | 7 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|-----------|---------|------|---|---|------|----|---|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|-----------|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-----------|---------|----|---|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|-----------|---------|------|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|-----------|---------|-----|---|---|-----|---|---|-----|---|---|-----|---|---|-----|---|---|
| Status account                                                                                                                                                                                                                                                                                  | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A13                                                                                                                                                                                                                                                                                             | 2              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A12                                                                                                                                                                                                                                                                                             | 5              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A11                                                                                                                                                                                                                                                                                             | 6              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A14                                                                                                                                                                                                                                                                                             | 7              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Credit history                                                                                                                                                                                                                                                                                  | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A30                                                                                                                                                                                                                                                                                             | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A33                                                                                                                                                                                                                                                                                             | 1              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A34                                                                                                                                                                                                                                                                                             | 4              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A32                                                                                                                                                                                                                                                                                             | 14             | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Purpose                                                                                                                                                                                                                                                                                         | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A46                                                                                                                                                                                                                                                                                             | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A48                                                                                                                                                                                                                                                                                             | 1              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A42                                                                                                                                                                                                                                                                                             | 3              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A43                                                                                                                                                                                                                                                                                             | 3              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A49                                                                                                                                                                                                                                                                                             | 3              | 5         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A41                                                                                                                                                                                                                                                                                             | 4              | 6         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A40                                                                                                                                                                                                                                                                                             | 5              | 7         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| <table> <tr><th>Savings</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A62</td><td>1</td><td>1</td></tr> <tr><td>A63</td><td>1</td><td>2</td></tr> <tr><td>A64</td><td>1</td><td>3</td></tr> <tr><td>A65</td><td>3</td><td>4</td></tr> <tr><td>A61</td><td>14</td><td>5</td></tr> </table> | Savings        | Frequency | Cluster | A62  | 1 | 1 | A63  | 1  | 2 | A64                                                                                                                                                                                                            | 1       | 3         | A65                                                                                                                                                                                                                                                   | 3    | 4         | A61                                                                                                                                                                                                                                                          | 14             | 5         | <table> <tr><th>Employment</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A72</td><td>2</td><td>1</td></tr> <tr><td>A71</td><td>3</td><td>2</td></tr> <tr><td>A74</td><td>4</td><td>3</td></tr> <tr><td>A75</td><td>4</td><td>4</td></tr> <tr><td>A73</td><td>7</td><td>5</td></tr> </table> | Employment | Frequency | Cluster | A72                                                                                                                                                                                                                                            | 2         | 1         | A71     | 3  | 2 | A74                                                                                                                                                                                                            | 4             | 3         | A75                                                                                                                                                                                                                                                                                                                                                                                | 4       | 4         | A73     | 7    | 5  | <table> <tr><th>Installment</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>1</td><td>1</td><td>1</td></tr> <tr><td>3</td><td>5</td><td>2</td></tr> <tr><td>2</td><td>6</td><td>3</td></tr> <tr><td>4</td><td>8</td><td>4</td></tr> </table> | Installment | Frequency | Cluster | 1   | 1 | 1 | 3   | 5 | 2 | 2   | 6 | 3 | 4   | 8 | 4 |     |   |   |
| Savings                                                                                                                                                                                                                                                                                         | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A62                                                                                                                                                                                                                                                                                             | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A63                                                                                                                                                                                                                                                                                             | 1              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A64                                                                                                                                                                                                                                                                                             | 1              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A65                                                                                                                                                                                                                                                                                             | 3              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A61                                                                                                                                                                                                                                                                                             | 14             | 5         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Employment                                                                                                                                                                                                                                                                                      | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A72                                                                                                                                                                                                                                                                                             | 2              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A71                                                                                                                                                                                                                                                                                             | 3              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A74                                                                                                                                                                                                                                                                                             | 4              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A75                                                                                                                                                                                                                                                                                             | 4              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A73                                                                                                                                                                                                                                                                                             | 7              | 5         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Installment                                                                                                                                                                                                                                                                                     | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 1                                                                                                                                                                                                                                                                                               | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 3                                                                                                                                                                                                                                                                                               | 5              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 2                                                                                                                                                                                                                                                                                               | 6              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 4                                                                                                                                                                                                                                                                                               | 8              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| <table> <tr><th>Gender</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>F</td><td>6</td><td>1</td></tr> <tr><td>M</td><td>14</td><td>2</td></tr> </table>                                                                                                                                    | Gender         | Frequency | Cluster | F    | 6 | 1 | M    | 14 | 2 | <table> <tr><th>Debtors</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A103</td><td>1</td><td>1</td></tr> <tr><td>A102</td><td>3</td><td>2</td></tr> <tr><td>A101</td><td>16</td><td>3</td></tr> </table> | Debtors | Frequency | Cluster                                                                                                                                                                                                                                               | A103 | 1         | 1                                                                                                                                                                                                                                                            | A102           | 3         | 2                                                                                                                                                                                                                                                                                                 | A101       | 16        | 3       | <table> <tr><th>Residence</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>3</td><td>1</td><td>1</td></tr> <tr><td>1</td><td>3</td><td>2</td></tr> <tr><td>2</td><td>7</td><td>3</td></tr> <tr><td>4</td><td>9</td><td>4</td></tr> </table> | Residence | Frequency | Cluster | 3  | 1 | 1                                                                                                                                                                                                              | 1             | 3         | 2                                                                                                                                                                                                                                                                                                                                                                                  | 2       | 7         | 3       | 4    | 9  | 4                                                                                                                                                                                                                                                |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Gender                                                                                                                                                                                                                                                                                          | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| F                                                                                                                                                                                                                                                                                               | 6              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| M                                                                                                                                                                                                                                                                                               | 14             | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Debtors                                                                                                                                                                                                                                                                                         | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A103                                                                                                                                                                                                                                                                                            | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A102                                                                                                                                                                                                                                                                                            | 3              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A101                                                                                                                                                                                                                                                                                            | 16             | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Residence                                                                                                                                                                                                                                                                                       | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 3                                                                                                                                                                                                                                                                                               | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 1                                                                                                                                                                                                                                                                                               | 3              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 2                                                                                                                                                                                                                                                                                               | 7              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 4                                                                                                                                                                                                                                                                                               | 9              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| <table> <tr><th>Property</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A121</td><td>4</td><td>1</td></tr> <tr><td>A122</td><td>4</td><td>2</td></tr> <tr><td>A124</td><td>4</td><td>3</td></tr> <tr><td>A123</td><td>8</td><td>4</td></tr> </table>                                       | Property       | Frequency | Cluster | A121 | 4 | 1 | A122 | 4  | 2 | A124                                                                                                                                                                                                           | 4       | 3         | A123                                                                                                                                                                                                                                                  | 8    | 4         | <table> <tr><th>Others</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A141</td><td>3</td><td>1</td></tr> <tr><td>A142</td><td>3</td><td>2</td></tr> <tr><td>A143</td><td>14</td><td>3</td></tr> </table>                                                | Others         | Frequency | Cluster                                                                                                                                                                                                                                                                                           | A141       | 3         | 1       | A142                                                                                                                                                                                                                                           | 3         | 2         | A143    | 14 | 3 | <table> <tr><th>Housing</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A153</td><td>4</td><td>1</td></tr> <tr><td>A151</td><td>6</td><td>2</td></tr> <tr><td>A152</td><td>10</td><td>3</td></tr> </table> | Housing       | Frequency | Cluster                                                                                                                                                                                                                                                                                                                                                                            | A153    | 4         | 1       | A151 | 6  | 2                                                                                                                                                                                                                                                | A152        | 10        | 3       |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Property                                                                                                                                                                                                                                                                                        | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A121                                                                                                                                                                                                                                                                                            | 4              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A122                                                                                                                                                                                                                                                                                            | 4              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A124                                                                                                                                                                                                                                                                                            | 4              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A123                                                                                                                                                                                                                                                                                            | 8              | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Others                                                                                                                                                                                                                                                                                          | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A141                                                                                                                                                                                                                                                                                            | 3              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A142                                                                                                                                                                                                                                                                                            | 3              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A143                                                                                                                                                                                                                                                                                            | 14             | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Housing                                                                                                                                                                                                                                                                                         | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A153                                                                                                                                                                                                                                                                                            | 4              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A151                                                                                                                                                                                                                                                                                            | 6              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A152                                                                                                                                                                                                                                                                                            | 10             | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| <table> <tr><th>Credits</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>3</td><td>2</td><td>1</td></tr> <tr><td>2</td><td>4</td><td>2</td></tr> <tr><td>1</td><td>14</td><td>3</td></tr> </table>                                                                                           | Credits        | Frequency | Cluster | 3    | 2 | 1 | 2    | 4  | 2 | 1                                                                                                                                                                                                              | 14      | 3         | <table> <tr><th>Job</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A171</td><td>1</td><td>1</td></tr> <tr><td>A172</td><td>2</td><td>2</td></tr> <tr><td>A174</td><td>5</td><td>3</td></tr> <tr><td>A173</td><td>12</td><td>4</td></tr> </table> | Job  | Frequency | Cluster                                                                                                                                                                                                                                                      | A171           | 1         | 1                                                                                                                                                                                                                                                                                                 | A172       | 2         | 2       | A174                                                                                                                                                                                                                                           | 5         | 3         | A173    | 12 | 4 | <table> <tr><th>Number people</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>2</td><td>3</td><td>1</td></tr> <tr><td>1</td><td>17</td><td>2</td></tr> </table>                                            | Number people | Frequency | Cluster                                                                                                                                                                                                                                                                                                                                                                            | 2       | 3         | 1       | 1    | 17 | 2                                                                                                                                                                                                                                                |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Credits                                                                                                                                                                                                                                                                                         | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 3                                                                                                                                                                                                                                                                                               | 2              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 2                                                                                                                                                                                                                                                                                               | 4              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 1                                                                                                                                                                                                                                                                                               | 14             | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Job                                                                                                                                                                                                                                                                                             | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A171                                                                                                                                                                                                                                                                                            | 1              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A172                                                                                                                                                                                                                                                                                            | 2              | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A174                                                                                                                                                                                                                                                                                            | 5              | 3         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A173                                                                                                                                                                                                                                                                                            | 12             | 4         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Number people                                                                                                                                                                                                                                                                                   | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 2                                                                                                                                                                                                                                                                                               | 3              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| 1                                                                                                                                                                                                                                                                                               | 17             | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| <table> <tr><th>Telephone</th><th>Frequency</th><th>Cluster</th></tr> <tr><td>A192</td><td>7</td><td>1</td></tr> <tr><td>A191</td><td>13</td><td>2</td></tr> </table>                                                                                                                           | Telephone      | Frequency | Cluster | A192 | 7 | 1 | A191 | 13 | 2 |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| Telephone                                                                                                                                                                                                                                                                                       | Frequency      | Cluster   |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A192                                                                                                                                                                                                                                                                                            | 7              | 1         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |
| A191                                                                                                                                                                                                                                                                                            | 13             | 2         |         |      |   |   |      |    |   |                                                                                                                                                                                                                |         |           |                                                                                                                                                                                                                                                       |      |           |                                                                                                                                                                                                                                                              |                |           |                                                                                                                                                                                                                                                                                                   |            |           |         |                                                                                                                                                                                                                                                |           |           |         |    |   |                                                                                                                                                                                                                |               |           |                                                                                                                                                                                                                                                                                                                                                                                    |         |           |         |      |    |                                                                                                                                                                                                                                                  |             |           |         |     |   |   |     |   |   |     |   |   |     |   |   |     |   |   |

**Πίνακας 7: Κατανομή των συστάδων των χαρακτηριστικών.**

Το δεύτερο κομμάτι του βήματος 1, το οποίο αφορά τον υπολογισμό του συνολικού αριθμού πολυσυστοιχιών μέσα από την συγχώνευση συστάδων, εμφανίζεται στον πίνακα 8. Όπως βλέπουμε, ενώ ο μέγιστος αριθμός συστάδων θα ήταν τεράστιος, μόνο τα 20 από αυτά δεν είναι κενά.

| Observation | Status account | Credit history | Purpose | Savings | Employment | Installment | Gender | Debtors | Residence | Property | Others | Housing | Credits | Job | Number people | Telephone | Multicluseter    |
|-------------|----------------|----------------|---------|---------|------------|-------------|--------|---------|-----------|----------|--------|---------|---------|-----|---------------|-----------|------------------|
| 775         | 1              | 3              | 7       | 2       | 2          | 3           | 2      | 3       | 4         | 3        | 1      | 1       | 1       | 1   | 2             | 2         | 1372232343111122 |
| 204         | 3              | 4              | 2       | 5       | 3          | 4           | 2      | 3       | 4         | 2        | 3      | 2       | 3       | 4   | 2             | 2         | 3425342342323422 |
| 699         | 4              | 3              | 4       | 5       | 5          | 4           | 2      | 3       | 3         | 4        | 3      | 3       | 2       | 4   | 2             | 2         | 4345542334332422 |
| 250         | 4              | 4              | 4       | 5       | 2          | 2           | 1      | 2       | 4         | 1        | 3      | 2       | 3       | 4   | 2             | 2         | 4445221241323422 |
| 593         | 4              | 4              | 5       | 3       | 4          | 4           | 1      | 3       | 4         | 1        | 1      | 3       | 3       | 2   | 2             | 2         | 4453441341133222 |
| 661         | 1              | 4              | 4       | 5       | 5          | 2           | 2      | 3       | 4         | 1        | 3      | 2       | 3       | 4   | 2             | 2         | 1445522341323422 |
| 463         | 2              | 4              | 3       | 5       | 1          | 2           | 1      | 3       | 2         | 1        | 3      | 2       | 3       | 3   | 2             | 2         | 2435121321323322 |
| 828         | 4              | 1              | 5       | 5       | 5          | 3           | 2      | 3       | 3         | 4        | 2      | 3       | 2       | 4   | 1             | 2         | 4155532334232412 |
| 845         | 4              | 4              | 3       | 5       | 4          | 4           | 2      | 3       | 4         | 2        | 1      | 3       | 1       | 4   | 1             | 1         | 4435442342131411 |
| 11          | 2              | 4              | 7       | 5       | 1          | 2           | 1      | 3       | 2         | 4        | 3      | 2       | 3       | 4   | 2             | 2         | 2475121324323422 |
| 896         | 4              | 2              | 6       | 4       | 3          | 2           | 2      | 3       | 3         | 4        | 2      | 3       | 3       | 3   | 1             | 1         | 4264322334233311 |
| 379         | 2              | 4              | 7       | 5       | 4          | 4           | 2      | 3       | 3         | 3        | 3      | 1       | 3       | 3   | 2             | 1         | 2475442333313321 |
| 989         | 3              | 4              | 6       | 5       | 2          | 4           | 2      | 3       | 3         | 3        | 3      | 1       | 3       | 3   | 2             | 1         | 3465242333313321 |
| 323         | 3              | 4              | 6       | 5       | 3          | 3           | 2      | 3       | 2         | 3        | 3      | 1       | 3       | 3   | 2             | 1         | 3465332323313321 |
| 716         | 4              | 3              | 6       | 4       | 4          | 1           | 2      | 3       | 4         | 4        | 3      | 3       | 2       | 4   | 2             | 2         | 4364412344332422 |
| 619         | 2              | 4              | 3       | 1       | 5          | 3           | 1      | 2       | 4         | 4        | 3      | 2       | 3       | 4   | 2             | 2         | 2431531244323422 |
| 908         | 2              | 4              | 1       | 4       | 5          | 3           | 2      | 3       | 3         | 4        | 3      | 3       | 3       | 4   | 2             | 2         | 2414532334333422 |
| 955         | 3              | 4              | 7       | 5       | 5          | 4           | 1      | 1       | 4         | 2        | 3      | 3       | 3       | 4   | 2             | 1         | 3475541142333421 |
| 209         | 3              | 4              | 5       | 5       | 5          | 3           | 2      | 3       | 3         | 4        | 2      | 3       | 3       | 2   | 2             | 2         | 3455532334233222 |
| 304         | 3              | 3              | 7       | 5       | 3          | 4           | 2      | 2       | 1         | 2        | 3      | 3       | 2       | 4   | 2             | 1         | 3375342212332421 |

**Πίνακας 8: Μη κενές πολλαπλές συστάδες.**

Μέσα από τα αποτελέσματά του πίνακα 8, δημιουργείται ο πίνακας συμπτώσεων που παρουσιάζεται στον πίνακα 9.

| Multiclust       | Freq | 1372232343111122 | 1445522341323422 | 2414532334333422 | 2431531244323422 | 2435121321323322 | 2475121324323422 | 2475442333313321 | 3375342212332421 | 3425342342323422 | 3455532334233222 | 3465242333313321 | 3465332323313321 | 3475541142333421 | 4155532334232412 | 4264322334233311 | 4345542334332422 | 4364412344332422 | 4435442342131411 | 4445221241323422 | 4453441341133222 |
|------------------|------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 1372232343111122 | 1    |                  | 6                | 5                | 4                | 3                | 4                | 6                | 4                | 5                | 5                | 6                | 6                | 3                | 4                | 2                | 5                | 6                | 5                | 4                | 5                |
| 1445522341323422 | 1    |                  |                  | 9                | 9                | 10               | 10               | 7                | 5                | 11               | 8                | 7                | 7                | 8                | 6                | 4                | 9                | 7                | 6                | 12               | 7                |
| 2414532334333422 | 1    |                  |                  |                  | 10               | 7                | 9                | 8                | 5                | 8                | 11               | 7                | 7                | 7                | 9                | 7                | 10               | 9                | 5                | 6                | 6                |
| 2431531244323422 | 1    |                  |                  |                  |                  | 9                | 10               | 5                | 4                | 8                | 7                | 4                | 5                | 8                | 5                | 2                | 6                | 6                | 4                | 10               | 6                |
| 2435121321323322 | 1    |                  |                  |                  |                  |                  | 13               | 8                | 3                | 8                | 6                | 7                | 8                | 6                | 3                | 4                | 5                | 4                | 4                | 10               | 7                |
| 2475121324323422 | 1    |                  |                  |                  |                  |                  |                  | 8                | 5                | 9                | 7                | 6                | 7                | 8                | 5                | 4                | 7                | 6                | 4                | 10               | 6                |
| 2475442333313321 | 1    |                  |                  |                  |                  |                  |                  |                  | 7                | 8                | 7                | 13               | 11               | 8                | 4                | 6                | 7                | 5                | 7                | 5                | 6                |
| 3375342212332421 | 1    |                  |                  |                  |                  |                  |                  |                  |                  | 9                | 5                | 7                | 7                | 10               | 5                | 4                | 9                | 7                | 7                | 5                | 3                |
| 3425342342323422 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  | 8                | 9                | 9                | 10               | 5                | 4                | 8                | 7                | 8                | 9                | 7                |
| 3455532334233222 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 8                | 8                | 7                | 11               | 7                | 9                | 6                | 5                | 5                | 8                |
| 3465242333313321 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 13               | 8                | 4                | 7                | 7                | 5                | 6                | 6                | 5                |
| 3465332323313321 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 7                | 4                | 7                | 5                | 5                | 5                | 4                |
| 3475541142333421 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 4                | 3                | 7                | 5                | 8                | 7                |
| 4155532334232412 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 8                | 11               | 8                | 7                | 4                | 5                |
| 4264322334233311 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 6                | 7                | 6                | 3                | 4                |
| 4345542334332422 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 11               | 7                | 7                | 6                |
| 4364412344332422 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 7                | 6                | 7                |
| 4435442342131411 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 5                | 8                |
| 4445221241323422 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 8                |
| 4453441341133222 | 1    |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  | 17               |

**Πίνακας 9: Πίνακας σύμπτωσης.**

Στην συνέχεια, μέσα από τον πίνακα συμπτώσεων, θα συγχωνευθούν οι συστάδες με τις περισσότερες συμπτώσεις. Το αποτέλεσμα της διαδικασίας αυτής το βλέπουμε στον πίνακα 10, θεωρώντας ότι το ζεύγος συστάδων με τις περισσότερες συμπτώσεις είναι μόνο 1.

| Multiclust | Freq | 1372232343111122 | 1445522341323422 | 2414532334333422 | 2431531244323422 | 2435121321323322 |
|------------|------|------------------|------------------|------------------|------------------|------------------|
|            |      |                  |                  | 9                | 9                | 10               |
| Multiclust | Freq | 2475121324323422 | 2475442333313321 | 3375342212332421 | 3425342342323422 | 3455532334233222 |
|            |      | 10               | 7                | 5                | 11               | 8                |
| Multiclust | Freq | 3465242333313321 | 3465332323313321 | 3475541142333421 | 4155532334232412 | 4264322334233311 |
|            |      | 7                | 7                | 8                | 6                | 4                |
| Multiclust | Freq | 4345542334332422 | 4364412344332422 | 4435442342131411 | 4445221241323422 | 4453441341133222 |
|            |      | 9                | 7                | 6                | 12               | 7                |

**Πίνακας 10: Παράδειγμα συσχέτισης με μία μόνο σύμπτωση μεταξύ πολλαπλών συστάδων.**

Ενώ στην περίπτωση όπου υπάρχουν περισσότερα ζεύγη συστάδων, το αποτέλεσμα της συγχώνευσης που θα εξαρτηθεί από τον συντελεστή Fleiss Kappa, απεικονίζεται στον πίνακα 11.

|                  |      |                  |                  |                  |                  |                  |
|------------------|------|------------------|------------------|------------------|------------------|------------------|
| Multicluster     | Freq | 1372232343111122 | 1445522341323422 | 2414532334333422 | 2431531244323422 | 2435121321323322 |
| 1372232343111120 | 1    |                  | 6                | 5                | 4                | 3                |
| Multicluster     | Freq | 2475121324323422 | 247544233313321  | 3375342212332421 | 3425342342323422 | 3455532334233222 |
| 1372232343111120 | 1    | 4                | 6                | 4                | 5                | 5                |
| Multicluster     | Freq | 346524233313321  | 346533232313321  | 3475541142333421 | 4155532334232412 | 4264322334233311 |
| 1372232343111120 | 1    | 6                | 6                | 3                | 4                | 2                |
| Multicluster     | Freq | 4345542334332422 | 4364412344332422 | 4435442342131411 | 4445221241323422 | 4453441341133222 |
| 1372232343111120 | 1    | 5                | 6                | 5                | 4                | 5                |

**Πίνακας 11: Παράδειγμα συσχέτισης με περισσότερες από μία συμπτώσεις μεταξύ πολλαπλών συστάδων.**

Αφού λοιπόν ολοκληρωθούν οι παραπάνω διαδικασίες, τότε θα έχουμε τον πίνακα των βέλτιστων πολυ-συστοιχιών. (πίνακας 12).

| Optimal multiclusters | Frequency |
|-----------------------|-----------|
| 3455532334233220      | 1         |
| 346524233313320       | 1         |
| 4155532334232410      | 1         |
| 4345542334332420      | 1         |
| 2475121324323420      | 2         |
| 346533232313320       | 2         |
| 4364412344332420      | 2         |
| 4435442342131410      | 2         |
| 4445221241323420      | 2         |
| 3475541142333420      | 3         |
| 4453441341133220      | 3         |

**Πίνακας 12: Πίνακας βέλτιστων πολυσυστοιχιών.**

Με το τελευταίο στάδιο του αλγορίθμου, θα πρέπει να εντάσσεται κάθε συστάδα (**Transmitter**) στην επόμενη (**Receiver**) και να υπολογίζεται η ευκλείδεια απόσταση των αναλογιών του προστατευόμενου χαρακτηριστικού μεταξύ τους, με την σειρά (από πάνω προς τα κάτω). Κάθε ένταξη που πραγματοποιείται στην συνέχεια, τα αποτελέσματα της θα συγκρίνονται με την επόμενη. Τέλος, η διαδικασία αυτή αφορά όλες τις συστάδες, μέχρι να συνδεθούν οι κατάλληλες συστάδες και οι αναλογίες των τιμών να είναι οι πλησιέστερες των επιθυμητών (Εικόνα 16 ).

| Optimal Multicluster (Transmitter) | Optimal Multiclusters (Receiver) | Distance      |
|------------------------------------|----------------------------------|---------------|
| 346524233313320                    | 346524233313320                  | na            |
|                                    | 4155532334232410                 | 0,4243        |
|                                    | 4345542334332420                 | 0,4243        |
|                                    | 2475121324323420                 | 0,5185        |
|                                    | 346533232313320                  | 0,4243        |
|                                    | 4364412344332420                 | 0,4243        |
|                                    | 4435442342131410                 | <b>0,0471</b> |
|                                    | 4445221241323420                 | 0,4243        |
|                                    | 3475541142333420                 | 0,0707        |
|                                    | 4453441341133220                 | 0,2828        |

**Εικόνα 16: Υπολογισμός των αποστάσεων μεταξύ των πολλαπλών συστάδων (πομπού και δέκτη).**

## **5) Ευρετηρίαση βάσεων δεδομένων.**

Η ευρετηρίαση δεδομένων, μέσα σε μια βάση δεδομένων αποτελεί μια απαραίτητη λειτουργία ενός συστήματος η οποία μέσω μιας μεγάλης ποικιλίας προσεγγίσεων, αντιμετωπίζει ζητήματα μεγάλου εύρους. Στην συνέχεια της εργασίας, θα δούμε που και με ποιες τεχνικές χρησιμοποιείται η ευρετηρίαση μέσα σε μια βάση δεδομένων. (Diakoff, 2004)

### **5.1) Ευρετηρίαση μεγάλων χωρικών βάσεων δεδομένων.**

Για διαδικασίες αποθήκευσης και αναζήτησης δεδομένων σε σχέση με τον γεωμετρικό τους χώρο, χρησιμοποιούνται οι χωρικές βάσεις δεδομένων. Τα δεδομένα αυτά, έχουν σκοπό την παρατήρηση αντικειμένων σχετικά με την θέση τους σε χρονικές στιγμές. Ο όγκος αυτών των δεδομένων είναι τεράστιος και σε συνδυασμό με την πολυπλοκότητα λόγω της σχέσης μεταξύ χώρου και χρόνου, έχει ως αποτέλεσμα να είναι απαραίτητη η ανάπτυξη εξειδικευμένων συστημάτων για την αποτελεσματική διαχείριση τους.

Τα περισσότερα συστήματα που υπάρχουν για την διαχείριση δεδομένων, δεν πληρούν τις προϋποθέσεις για την αποδοτική υποστήριξη των χωρικών και των χωροχρονικών δεδομένων. Έτσι, για την αποφυγή προσωρινών λύσεων που παρουσιάζουν αδυναμίες, όπως οι ήδη υπάρχουσες τεχνικές **X-trees**, οι οποίες διαθέτουν την δυνατότητα εφαρμογής λειτουργιών όπως η εισαγωγή και η διαγραφή αλλά παρουσιάζουν αδυναμίες στον τομέα της ανάκτησης δεδομένων, προτείνεται μια νέα παραλλαγμένη τεχνική. Στην συνέχεια θα αναλύσουμε τη δομή ενός δέντρου X-tree και τέλος μια βελτιστοποιημένη παραλλαγή του που ονομάζεται aX-tree και μέσω μιας μεθόδου συσκευασίας, ξεπερνάει τους περιορισμούς και προσφέρει αποδοτικότερη διαχείριση στις χωρικές βάσεις δεδομένων. (Samson, 2018)

#### **5.1.1) Ευρετηρίαση χωρικών δεδομένων.**

Σημαντικός παράγοντας στην ευρετηρίαση χωροχρονικών δεδομένων σε μεγάλες βάσεις δεδομένων είναι η ταξινόμηση η οποία έχει εφαρμοστεί στα δεδομένα και μέσω του ευρετηρίου, γίνονται πιο προσβάσιμα. Τις περισσότερες φορές, κάθε αντικείμενο της βάσης δεδομένων συνδέεται με ένα διάνυσμα χαρακτηριστικών το οποίο στην συνέχεια χρησιμοποιείται για την αναπαράσταση του αντικειμένου. Παρότι, η βασική λειτουργία των δεικτών αποτελεί την ταξινόμηση των δεδομένων, στην συγκεκριμένη περίπτωση, όπου υπάρχουν περισσότερες διαστάσεις, ο ρόλος των δεικτών προσφέρει την ταξινόμηση των χωρικών αντικειμένων σε σχέση με τον χώρο που καταλαμβάνουν.

Μια σιωπηρή ταξινόμηση, προσφέρει σταθερότητα, καθώς με την αλλαγή των ερωτημάτων, δεν θα είναι απαραίτητη η δημιουργία νέου ευρετηρίου. Αποτελεσματικό εργαλείο για την βελτιστοποίηση μεγάλων δυναμικών βάσεων δεδομένων, αποτελούν οι τεχνικές χωρικής

ευρετηρίασης μέσα από την χρήση τεχνικών ταξινόμησης. Αυτό συμβαίνει, καθώς οι λειτουργίες ταξινόμησης έχουν την δυνατότητα να αντιστοιχίσουν πολυδιάστατα δεδομένα σε μια διάσταση χωρίς να επηρεαστεί η τοπικότητα των δεδομένων σημείων. Στην συνέχεια, αφού εκτελεστούν οι κατάλληλες ενέργειες ταξινόμησης στα δεδομένα, δημιουργείται μια χωρική δομή δεδομένων και πραγματοποιούνται βελτιώσεις στα αποτελέσματα των ερωτημάτων εάν χρειάζεται.

#### **5.1.2) Δομή δεδομένων X-tree**

Η δομή του X-tree αποτελεί παραλλαγή του R-tree και η δημιουργία του είχε ως σκοπό την ευρετηρίαση μεγάλου όγκου σημειακών και χωρικών δεδομένων σε υψηλές διαστάσεις. Το X-tree αποτελεί έναν συνδυασμό γραμμικού πίνακα και ιεραρχικού καταλόγου (όπως στο R-tree) και αντιμετωπίζει μια από τις μεγαλύτερες ανησυχίες στις μεγάλες διαστάσεις, την επικάλυψη των πλαισίων οριοθέτησης. Αν η αποφυγή αυτού του προβλήματος ωστόσο δεν είναι εφικτή, τότε κατά την εισαγωγή νέων εγγραφών στη δομή, θα δημιουργηθούν υπερκόμβοι. Η ύπαρξη των υπερκόμβων καθιστά την δομή του X-tree αποδοτικότερη σε σχέση με τις υπόλοιπες μεθόδους πρόσβασης.

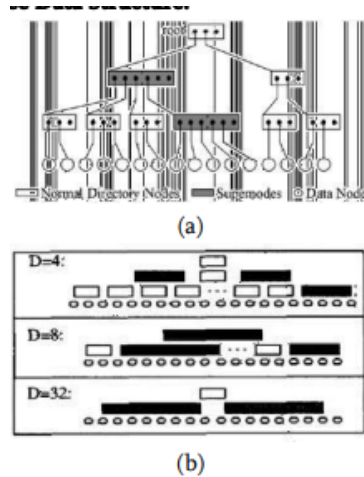
Οι υπερκόμβοι χαρακτηρίζονται από τις εξής ιδιότητες-πλεονεκτήματα:

1. Μικρότερος αριθμός διαχωρισμών, άρα και μεγαλύτερος μέσος όρος αποθηκευτικής χρήσης.
2. Μειωμένο ύψος δομής.

Το X-tree, μέσω ενός αλγορίθμου χωρίς επικαλύψεις, επιτυγχάνει την γρήγορη πρόσβαση στα διανύσματα χαρακτηριστικών των αντικειμένων. Αυτό συμβαίνει καθώς διαθέτει ταυτόχρονα μια ιεραρχική δομή καταλόγου και μια γραμμική δομή για διανύσματα χαμηλής και υψηλής διάστασης αντίστοιχα.

Παρόλα αυτά, η απόδοση ενός X-tree γίνεται ιδιαίτερα χαμηλή σε λειτουργίες που αφορούν την ανάκτηση δεδομένων, λόγω της πολυδιαστατικότητας. Επίσης, οι δενδρικές δομές εμφανίζουν δεσμεύσεις όπως:

1. Υπερχείλιση στους υπερκόμβους.
2. Δυσχέραση εξαιτίας των πολλαπλών προσβάσεων στο δίσκο.
3. Πλήρης γραμμικοποίηση της δομής.



Εικόνα 17 : (α) δομή ενός X-tree, (β) σχήματα ενός X-tree σε διαφορετικές διαστάσεις δεδομένων.

### 5.1.3 Προκλήσεις για την δημιουργία του aX-tree.

Ο αλγόριθμος aX-tree αποτελεί μια νέα τεχνική συσκευασίας του X-tree για την διαχείριση στατικών χωρικών βάσεων δεδομένων. Σε αυτή την περίπτωση, ο αλγόριθμος προτείνεται για δεδομένα σε επίπεδο πολλών διαστάσεων, γεγονός που το ξεχωρίζει από τις προηγούμενες μεθόδους. Επιπλέον, χρησιμοποιείται διαφορετική μέθοδος εισαγωγής, αφού πριν από το πακετάρισμα του X-tree, υλοποιείται μια δευτερεύουσα ταξινόμηση.

Στόχος του aX-tree, είναι η μείωση του κόστους που προκαλείται από την εξέταση των γεωμετρικών αντικειμένων. Αυτό συμβαίνει λόγω της μεγάλης επικάλυψης μεταξύ των MBR, έτσι η επαναληπτική ταξινόμηση των υπερτριγώνων που διασταυρώνονται, πιθανότατα θα βελτιώσει την αποδοτικότητα των ερωτημάτων, αφού ο αριθμός των MBR που θα τέμνονται, θα έχει μειωθεί.

### 5.1.4) Κατασκευή aX-tree.

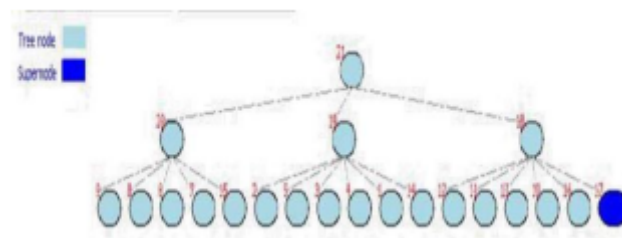
Η διαδικασία της συσκευασίας έχει ως στόχο, την εύρεση και την χρήση του βέλτιστου τρόπου συσκευασίας, άρα με την μεγαλύτερη πυκνότητα. Δηλαδή, σε όσο το δυνατόν μικρότερο δοχείο, να χωρέσουν όσο το δυνατόν περισσότερα αντικείμενα. Ένας αλγόριθμος με αναδρομική ταξινόμηση κεραμιδιών (**sort-tille-recursive**) εφαρμόζεται σε ένα X-tree σε επίπεδο 2D. Η τεχνική αυτή, εκτελείται με δύο διαφορετικούς τρόπους σε σημεία και αντικείμενα.

1. Υπολογίζεται η έκταση των αντικειμένων, μέσω των ελάχιστων και μέγιστων τιμών τους. Αυτό επιτυγχάνεται με την χρήση των ελάχιστων ορθογωνίων οριοθέτησης, τα οποία περιλαμβάνονται από τα σημεία των ελάχιστων και μέγιστων τιμών.
2. Μέσω μιας πιο κοστοβόρας τεχνικής η οποία θέτει το αντικείμενο προς ανάλυση σε ιδανικότερες συνιστώσες.

Όταν χρησιμοποιείται η πρώτη προσέγγιση, τότε η κατασκευή του δέντρου εξελίσσεται με τον τρόπο που θα δούμε παρακάτω.

#### 5.1.5) Περιγραφή αλγορίθμου.

Το πρώτο στάδιο υλοποίησης του αλγορίθμου περιλαμβάνει την προεπεξεργασία, τον καθαρισμό δηλαδή του πίνακα και την αλλαγή των LAT/LONG ή των συντεταγμένων X/Y σε ένα τυποποιημένο γεωμετρικό σχήμα της βάσης δεδομένων με σκοπό την κατασκευή της έκτασης. Έπειτα, μέσω του αλγορίθμου, δημιουργείται ένα οριοθετημένο πλαίσιο (envelop) που περιβάλλει την έκταση. Στην συνέχεια, θα πρέπει να υπολογιστεί το μέσο σημείο αυτού του πλαισίου, τα οποία θα ταξινομηθούν με βάση την συντεταγμένη του κεντρικού τους σημείου και τον συνολικό αριθμό κόμβων που είναι απαραίτητοι ( $\beta$ ). Ως **R** ορίζεται το σύνολο των αντικειμένων, ενώ το **M** συμβολίζει τον μέγιστο αποθηκευτικό χώρο ενός κόμβου. Η διαδικασία της συσκευασίας ξεκινάει με τον υπολογισμό του  $\beta^{1/d}$  όπου με το **d** συμβολίζεται η διάσταση. Έτσι υπολογίζεται ο συνολικός αριθμός διαμερισμάτων που χρειάζεται. Ανάλογα με το αποτέλεσμα της κατάτμησης, τα ορθογώνια σε συνδυασμό με κάποια βασικά στοιχεία όπως το **ID** τους, το ID του αντικειμένου, και η μέγιστη τιμή ενός κόμβου φορτώνονται σε σελίδες των M ως είσοδος. Έτσι, επιστρέφεται το ID της σελίδας και το MBR των χωρικών αντικειμένων, το οποίο χρησιμοποιείται για την ταξινόμηση. Μετά την εκτέλεση της κατάτμησης, το σύνολο των δεδομένων υφίσταται σάρωση και τα αντικείμενα προσθέτονται στην κατάλληλη κατάτμηση ανάλογα με το εύρος τους. (Samson, 2018)



Εικόνα 18 : Η δομή ενός aX-tree σε δύο διαστάσεις.

#### 5.2) Ευρετηρίαση χρονικών βάσεων δεδομένων.

Η εξαγωγή χρήσιμων πληροφοριών από τεράστιο όγκο δεδομένων είναι ένα πολύ ουσιαστικό θέμα έρευνας στις μέρες μας. Όσο το μέγεθος και η πολυπλοκότητα των δεδομένων αυξάνεται, η ανάγκη για αποδοτικούς αλγόριθμους ανάλυσης είναι πολύ μεγάλη. Σε αυτήν την ενότητα, παρουσιάζεται η δομή δεδομένων TIndex για την ευρετηρίαση συναλλαγών, σε συστήματα χρονικών βάσεων δεδομένων (**indexing temporal databases**) και τα μειονεκτήματά της στην υλοποίηση. Στη συνέχεια παρουσιάζεται μια ανανεωμένη παραλλαγή του, το TIndex2, το οποίο δίνει λύσεις σε μερικές από τις αδυναμίες του TIndex. (Fouad, 2015)

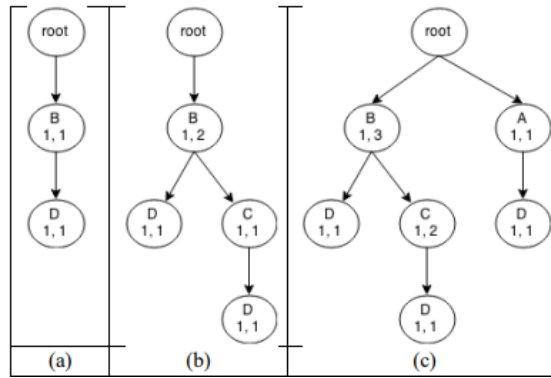
### 5.2.1) Δομή δεδομένων TIndex.

Μια νέα δενδρική δομή δεδομένων, ονομάζεται TIndex και χρησιμοποιείται για την ευρετηρίαση (**indexing**) συναλλαγών σε μια χρονική βάση δεδομένων. Το TIndex περιέχει δύο κύρια συστατικά: Το Δέντρο και τον Πίνακα Επικεφαλίδας (**Tree and Header Table.**). Η δομή δέντρου αποθηκεύει τις συναλλαγές στη βάση δεδομένων και ο πίνακας κεφαλίδων περιέχει τις πληροφορίες για όλα τα στοιχεία στην βάση δεδομένων και έναν σύνδεσμο προς τον πρώτο του κόμβο στο δέντρο. Κάθε κόμβος στο δέντρο είναι μια πλειάδα  $\langle \text{item, part, support} \rangle$ , όπου το item αναπαριστά το όνομα ενός στοιχείου, το part αποτελεί τον τρέχων αριθμό τμήματος και το support συμβολίζει τη συχνότητα του στοιχείου αυτού στο part. Στην τελική μορφή του TIndex, η συναλλαγή αντιστοιχίζεται σε μια διαδρομή από τη ρίζα στο φύλλο και κάθε κόμβος περιέχει ένα στοιχείο από αυτή τη συναλλαγή και το support της σε αυτή τη διαδρομή. Η απεικόνιση ενός παραδείγματος χρονικής βάσης δεδομένων, φαίνεται στον παρακάτω πίνακα.

|                | TID | Transaction |
|----------------|-----|-------------|
| P <sub>1</sub> | 1   | B D         |
|                | 2   | B C D       |
|                | 3   | B C         |
|                | 4   | A D         |
| P <sub>2</sub> | 5   | B C E       |
|                | 6   | D E         |
|                | 7   | A B C       |
|                | 8   | C D E       |
| P <sub>3</sub> | 9   | B C E F     |
|                | 10  | B F         |
|                | 11  | A D         |
|                | 12  | B D F       |

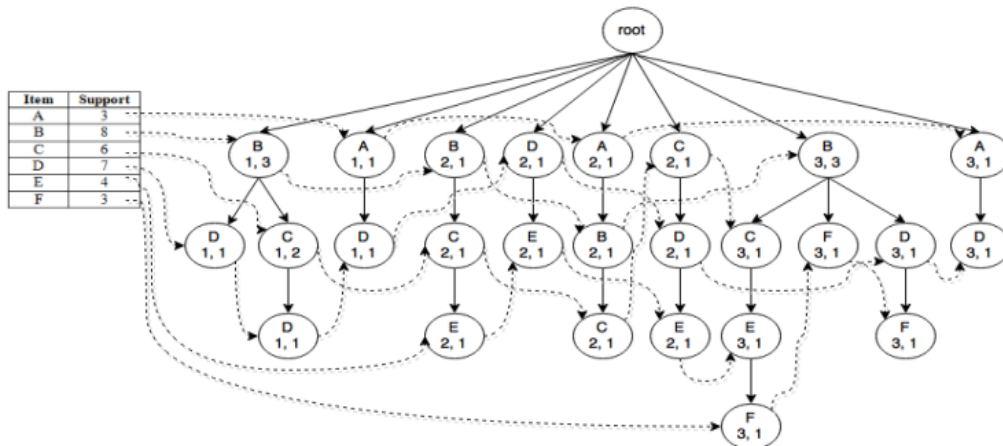
Πίνακας 13: Παράδειγμα χρονική βάσης δεδομένων.

Αρχικά, όπως βλέπουμε στη συναλλαγή TID-1, προστίθεται ένας νέος κόμβος  $\langle B, 1, 1 \rangle$  ως θυγατρικός κόμβος στη ρίζα και στην συνέχεια ένας δεύτερος κόμβος  $\langle D, 1, 1 \rangle$  εξίσου ως θυγατρικός στον κόμβο B αυτή τη φορά, όπως φαίνεται στην εικόνα 19 (α). Για την συναλλαγή TID-2, η ρίζα ήδη έχει θυγατρικό κόμβο  $\langle B, 1 \rangle$ , για αυτό το λόγο το support του αυξάνεται σε 2. Στη συνέχεια, προστίθεται νέο θυγατρικό κόμβο  $\langle C, 1, 1 \rangle$  καθώς ο κόμβος B δεν έχει αντίστοιχο παιδί. Τέλος, για το στοιχείο D, προστίθεται ένας νέος κόμβος  $\langle D, 1, 1 \rangle$  ως θυγατρικός κόμβος Γ όπως στην εικόνα 19 (b). Για την συναλλαγή TID-3, απλώς αυξάνεται το support του κόμβου  $\langle B, 1 \rangle$  και ο θυγατρικός του κόμβος  $\langle C, 1 \rangle$  καθώς βρίσκονται ήδη στο ίδιο μονοπάτι. Για την τελευταία συναλλαγή του P1, TID-4, ο κόμβος ρίζα δεν έχει παιδί με ετικέτα  $\langle A, 1 \rangle$ , για αυτό το λόγο, ένας νέος κόμβος  $\langle A, 1, 1 \rangle$  προστίθεται ως θυγατρικός στον κόμβο ρίζα. Στη συνέχεια, προστίθεται ένας νέος κόμβος  $\langle D, 1, 1 \rangle$  ως παιδί στο  $\langle A, 1 \rangle$ . Η εικόνα 19 (c) απεικονίζει την τελική μορφή του TIndex μετά την επιτυχή ευρετηρίαση όλων των συναλλαγών στο P1.



Εικόνα 19: TIndex για το μέρος P1, του πίνακα 14.

Η διαδικασία που παρακολουθήσαμε, εκτελείται με τον ίδιο τρόπο και για τα μέρη P2 και P3. Στη συναλλαγή TID-5 προστίθεται ένας νέος κόμβος  $\langle B, 2, 1 \rangle$  ως θυγατρικός στον κόμβο ρίζα. Έπειτα, ένας καινούριος κόμβος, εφόσον δεν ταιριάζει με τον κόμβο  $\langle B, 1 \rangle$ , θα προστεθεί στον κόμβο ρίζα. Με την ολοκλήρωση όλων των λειτουργιών, παράγεται ο τελικός δείκτης TIndex που απεικονίζεται στην εικόνα 20.



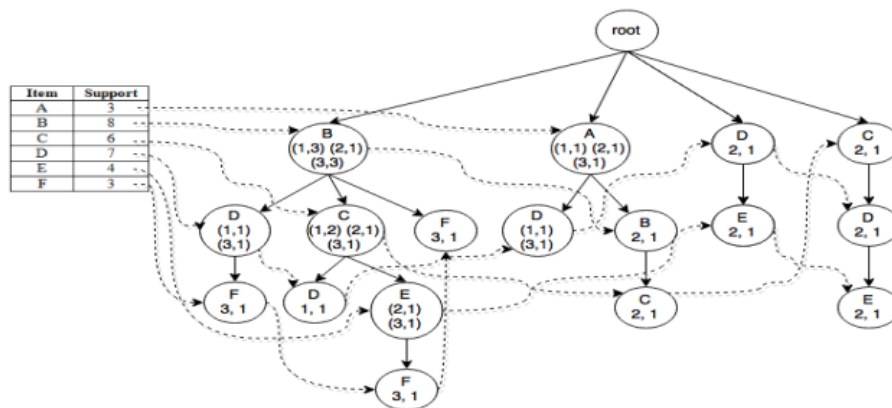
Εικόνα 20: Παράδειγμα ενός πλήρους TIndex.

Η υποστήριξη του συνόλου στοιχείων της εικόνας 20, υπολογίζεται μέσω της αναζήτησης του πρώτου στοιχείου στον πίνακα Header, για τον καθορισμό των πρώτων κόμβων. Έπειτα, θα πρέπει να αναζητηθούν τα υπόλοιπα στοιχεία του συνόλου στα υποδέντρα. Αν δηλαδή θεωρήσουμε ότι πρέπει να μετρηθεί η υποστήριξη του συνόλου  $\{B, D\}$ , τότε αρχικά θα πρέπει να εξετασθεί ο πίνακας Header για τον εντοπισμό του πρώτου κόμβου του στοιχείου  $\{B\}$  ο οποίος είναι  $\langle B, 1, 3 \rangle$  και αποτελεί τον κόμβο ρίζα του πρώτου υποδέντρου κατά την διαδικασία αναζήτησης του  $\{D\}$ . Ο επόμενος κόμβος που συνδέεται το  $\{B\}$  είναι ο  $\langle B, 2, 1 \rangle$ , χωρίς όμως να υπάρχει η δυνατότητα εύρεσης του  $\{D\}$  σε αυτό το υποδέντρο. Έτσι, με τον τελευταίο σύνδεσμο βρίσκει τον κόμβο  $\langle B, 3, 3 \rangle$  όπου υπάρχει το  $\{D\}$  στο (B-D-F) μονοπάτι. Το συνολικό μέγεθος της υποστήριξης υπολογίζεται με την πρόσθεση του βαθμού υποστήριξης σε όλα τα μονοπάτια που βρέθηκαν μεταξύ τους. Κύριο

μέλημα του TIndex, όπως είδαμε και στο παραπάνω παράδειγμα, είναι ο περιορισμός του χώρου αναζήτησης κατά την διάρκεια μέτρησης της υποστήριξης. Έτσι, αντί να χρησιμοποιηθούν όλες οι συναλλαγές, ο χώρος αναζήτησης περιορίστηκε σε τρεις μόνο συναλλαγές, εξοικονομώντας χρόνο. Αυτό είναι ιδιαίτερα σημαντικό, ειδικότερα σε περιπτώσεις βάσεων δεδομένων με τεράστιο όγκο δεδομένων. Βασικό μειονέκτημα της δομής αυτής ωστόσο, αποτελεί η χρήση χωρητικότητας μνήμης για την αποθήκευση του ίδιου του δέντρου ευρετηρίου και του πίνακα Header.

### 5.2.2) Δομή δεδομένων TIndex2.

Για την αντιμετώπιση του προβλήματος σχετικά με τον επιπλέον χώρο μνήμης, χρησιμοποιούμε το TIndex2 το οποίο διαθέτει κάποιες αλλαγές σε σύγκριση με το TIndex και συνίσταται για τον αποδοτικό χειρισμό των συναλλαγών με κοινό πρόθεμα που βρίσκονται σε διαφορετικά μέρη. Αρχικά, οι κόμβοι σε αυτή την περίπτωση, έχουν την δυνατότητα αποθήκευσης της υποστήριξης κάθε στοιχείου σε οποιοδήποτε μέρος, μέσα από την πρόσθεση καταλόγου μερών σε όλους τους κόμβους. Έτσι δηλαδή, ο κόμβος {A}, θα διαθέτει τον κατάλογο υποστήριξης του, σε όλα τα μέρη που βρίσκεται (1,1), (2,1), (3,1). Στην εικόνα 21, βλέπουμε το παράδειγμα ενός πλήρους TIndex2, το οποίο παρατηρούμε ότι αποτελεί μια πιο συμπυκνωμένη εκδοχή του TIndex καθώς ενώνει τα κοινά προθέματα από όλα τα μέρη. Αυτή η λειτουργία στο τέλος θα παίξει καθοριστικό ρόλο, καθώς θα μειώσει τον χώρο μνήμης που απαιτείται για το τελικό ευρετήριο και αυτός ο παράγοντας σε μεγάλες βάσεις δεδομένων είναι ζωτικής σημασίας. (Fouad, 2015)



Εικόνα 21: Παράδειγμα ενός πλήρους TIndex2.

### 5.2.3) Πειράματα.

Για την αξιολόγηση και την σύγκριση μεταξύ των TIndex και TIndex2 υλοποιήθηκαν κάποια πειράματα. Οι συμμετέχοντες των πειραμάτων ήταν τέσσερα σύνολα δεδομένων, με διαφορετικά μεγέθη και χαρακτηριστικά που κυμαίνονταν σε διάφορα μεγέθη βάσεων δεδομένων. Πρώτο στάδιο του πειράματος αποτελεί η μέτρηση της απαιτούμενης μνήμης. Δεύτερο στάδιο, η χρήση του

αλγορίθμου Apriori για την σύγκριση του απαιτούμενου χρόνου εκτέλεσης για την λειτουργία εξόρυξης ανάλογα με την χρήση ή όχι της τεχνικής ευρετηρίασης.

#### 5.2.4) Σύνολα δεδομένων.

Στον πίνακα 14, περιέχονται οι στατιστικές πληροφορίες των βάσεων δεδομένων που χρησιμοποιήθηκαν στα πειράματα. Απεικονίζεται ο συνολικός αριθμός των συναλλαγών(#Trans), των διακριτών στοιχείων(#Items), και ο μέσος αριθμός στοιχείων ανά συναλλαγή (TransSize).

| Dataset   | #Trans<br>(DB) | #Trans<br>(db) | #Items | TransSize |
|-----------|----------------|----------------|--------|-----------|
| Chess     | 2,696          | 500            | 75     | 37        |
| Mushroom  | 7,125          | 1000           | 119    | 23        |
| Retail    | 80,000         | 8,163          | 57     | 13        |
| Accidents | 306,183        | 34,000         | 468    | 33.8      |

Πίνακας 14: Πληροφορίες των βάσεων δεδομένων.

#### 5.2.5) Χρήση μνήμης.

Προηγουμένως αναφέρθηκε το πρόβλημα σχετικά με την απώλεια της χωρητικότητας εξαιτίας των συναλλαγών με κοινό πρόθεμα σε διαφορετικά μέρη. Στον πίνακα 15 και 16, υπολογίζεται ο χώρος μνήμης για όλα τα σύνολα δεδομένων για τα TIndex και TIndex 2 αντίστοιχα.

| Dataset   | Dataset<br>size<br>(MB) | TIndex<br>size<br>(MB) | Reduction<br>Ratio (%) |
|-----------|-------------------------|------------------------|------------------------|
| Chess     | 18.443                  | 10.465                 | 43.3                   |
| Mushroom  | 9.169                   | 6.887                  | 24.9                   |
| Retail    | 92.372                  | 64.284                 | 30.4                   |
| Accidents | 677.814                 | 511.631                | 24.5                   |

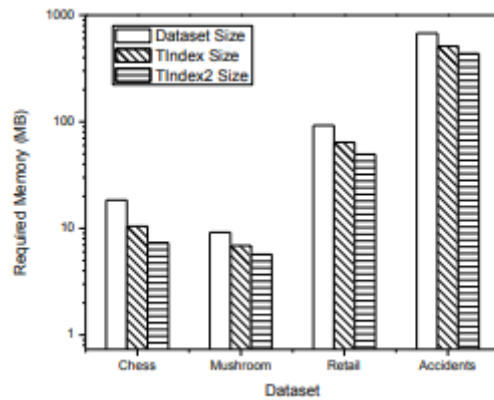
Πίνακας 15: Χρήση μνήμης για το TIndex.

Παρατηρούμε ότι το TIndex, επιτυγχάνει διαφορετικά ποσοστά μείωσης σε κάθε βάση δεδομένων, ανάλογα με τα χαρακτηριστικά που την διέπουν. Στην περίπτωση ενός αραιού συνόλου δεδομένων όπως το Mushroom, όπου δεν υπάρχει μεγάλος αριθμός συναλλαγών που να διατηρούν το ίδιο πρόθεμα, το ποσοστό μείωσης είναι χαμηλό. Αντίθετα, στην περίπτωση του Chess, επιτυγχάνεται μεγάλο ποσοστό μείωσης της χρήσης μνήμης, καθώς είναι περισσότερες οι συναλλαγές που διατηρούν το ίδιο πρόθεμα.

| Dataset   | Dataset<br>size<br>(MB) | TIndex2<br>size<br>(MB) | Reduction<br>Ratio (%) |
|-----------|-------------------------|-------------------------|------------------------|
| Chess     | 18.443                  | 7.331                   | 60.2                   |
| Mushroom  | 9.169                   | 5.696                   | 37.9                   |
| Retail    | 92.372                  | 49.627                  | 46.3                   |
| Accidents | 677.814                 | 435.398                 | 35.8                   |

Πίνακας 16: Χρήση μνήμης για το TIndex2.

Από την άλλη, στον πίνακα 16, απεικονίζεται το ποσοστό μείωσης του TIndex2 το οποίο είναι εμφανώς βελτιωμένο καθώς συνδυάζει τις συναλλαγές με κοινά προθέματα από κάθε μέρος. Στις πυκνότερες βάσεις δεδομένων, η μείωση που επιτυγχάνεται είναι μεγαλύτερη ενώ στις μικρότερες βάσεις δεδομένων του παραδείγματος, για το TIndex απαιτείται ένα μικρό μόνο μέρος του μέγεθος του συνόλου δεδομένων.

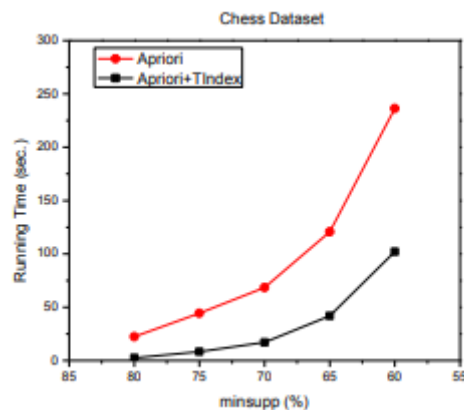


Εικόνα 22: Σύγκριση απαιτούμενης μνήμης μεταξύ TIndex και TIndex2

#### 5.2.6) Χρόνοι εκτέλεσης.

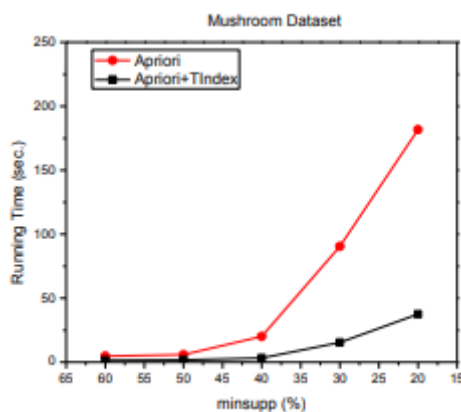
Στο πείραμα που εκτελέστηκε για τον υπολογισμό της απόδοσης της τεχνικής TIndex2, χρησιμοποιήθηκαν δύο αλγόριθμοι, ο Apriori και μια παραλλαγμένη έκδοση του, η Apriori-TIndex. Στις εικόνες που ακολουθούν, θα παρατηρήσουμε τους χρόνους εκτέλεσης με διαφορετικές **minsupp** τιμές κατωφλιού στα τέσσερα σύνολα δεδομένων μας.

Σε πυκνές περιπτώσεις συνόλου δεδομένων όπως το Chess, ο χρόνος εκτέλεσης που απαιτείται είναι μεγαλύτερος καθώς παράγεται μεγάλος αριθμός υποψηφίων. Στην εικόνα 23 βλέπουμε ότι η χρήση του TIndex σε συνδιασμό με τον αλγόριθμο Apriori, δε επιτυγχάνει μεγάλη αναλογία επιτάχυνσης. Αυτό συμβαίνει συγκεκριμένα επειδή, το TIndex σε περιπτώσεις βάσεων δεδομένων όπως αυτή, προσθέτει χρόνο εκτέλεσης πριν ξεκινήσει η λειτουργία του αλγορίθμου εξόρυξης.



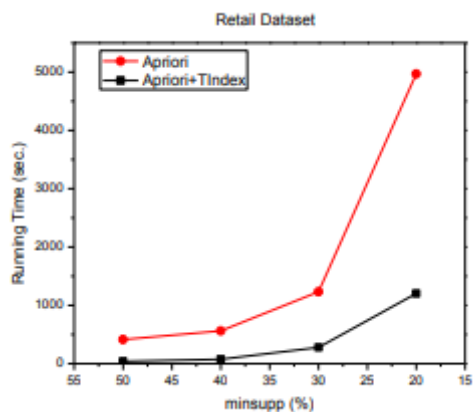
**Εικόνα 23: Σύγκριση αλγορίθμων στο σύνολο δεδομένων Chess.**

Όπως παρατηρούμε στην εικόνα 24, οι χρόνοι εκτέλεσης παρουσιάζουν μεγάλες διαφορές. Ο Apriori-TIndex είναι πολύ ταχύτερος και αυτό συμβαίνει επειδή ο παραδοσιακός Apriori δημιουργεί ένα τεράστιο όγκο υποψηφίων και έτσι απαιτείται περισσότερος χρόνος για τον υπολογισμό της υποστήριξης.



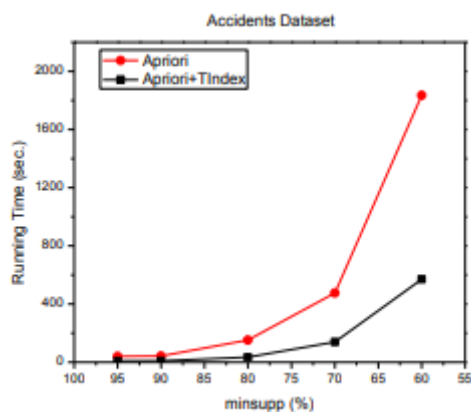
**Εικόνα 24: Σύγκριση αλγορίθμων στο σύνολο δεδομένων Mushroom.**

Στο σύνολο δεδομένων Retail στην εικόνα 25, η χρήση του αλγορίθμου, αυξάνει δραστικά την απόδοση του TIndex καθώς υπάρχουν επίσης μεγάλος αριθμός υποψηφίων, γεγονός που απαιτεί περισσότερο χρόνο για τον υπολογισμό της υποστήριξης. Έτσι, ο Apriori-TIndex εμφανίζει μεγάλη αναλογία επιτάχυνσης σε σχέση με την απλή έκδοση Apriori.



**Εικόνα 25: Σύγκριση αλγορίθμων στο σύνολο δεδομένων Retail.**

Τέλος, το Accidents παρά το μεγάλο μέγεθος του και τον επιπλέον χρόνο που απαιτείται για αυτό το λόγο, παράγει μικρό αριθμό υποψηφίων. Έτσι, στην εικόνα 26 φαίνεται πως το TIndex αυξάνει την ταχύτητα εκτέλεσης κατά μεγάλο βαθμό. (Fouad, 2015)



Εικόνα 26: Σύγκριση αλγορίθμων στο σύνολο δεδομένων Accidents.

### 5.3) Ευρετήρια σε βάσεις δεδομένων γραφημάτων.

Η αναζήτηση δεδομένων, αποτελεί ένα απαραίτητο εργαλείο για όλων των ειδών τις βάσεις δεδομένων. Αυτό ισχύει και για τις βάσεις δεδομένων γραφημάτων που χρησιμοποιούνται ιδιαίτερα σε επιστημονικούς τομείς όπως η βιοπληροφορική και η χημειοπληροφορική, όπου τα δεδομένα απεικονίζονται ως γραφήματα, οι κόμβοι αποτελούν βασικά στοιχεία και οι ακμές τους παρουσιάζουν τις σχέσεις που τα συνδέουν. Για την εξέλιξη λοιπόν αυτών των τμημάτων της επιστήμης απαιτούνται βάσεις δεδομένων που να περιέχουν μόρια που έχουν εξεταστεί στο παρελθόν, με σκοπό την δημιουργία μιας ταξινόμησης η οποία θα απεικονίζει αν ένα μόριο είναι αποδοτικό ή όχι. Τα πιο αποδοτικά στοιχεία, θα χρησιμοποιούνται για μεταγενέστερες μελέτες και δοκιμές. (Bonnici, 2010)

| Graph Database    | $G_1$                           | $G_2$ | $G_3$ |
|-------------------|---------------------------------|-------|-------|
|                   |                                 |       |       |
| Exact Query       | Number of Occurrences in Graphs |       |       |
| $Q_1$             |                                 | 2     |       |
| $Q_2$             |                                 | 2     | 3     |
| Approximate Query | Graphs Containing the Query     |       |       |
| $Q_3$             |                                 | 1     | 1     |

Εικόνα 27: Βάση δεδομένων γραφημάτων.

### 5.3.1) Προκλήσεις και υπάρχουσες τεχνικές.

Ας θεωρήσουμε ότι έχουμε μια βάση δεδομένων γραφημάτων  $X = \{G_1, G_2, \dots, G_n\}$  και ένα ερώτημα  $Q$  το οποίο μας καλεί να βρούμε όλα τα γραφήματα της βάσης δεδομένων  $X$  που να διαθέτουν το  $Q$  ως υπογράφημα. Αν και θα έπρεπε να εντοπίζονται όλες οι εμφανίσεις του  $Q$ , το πρόβλημα του ισομορφισμού γραφημάτων αποτρέπει την παραγωγή μιας αποδοτικής λύσης. Επιπροσθέτως, ο χρόνος που απαιτείται για την λειτουργία της αναζήτησης είναι τεράστιος, γεγονός που ώθησε τους ερευνητές σε διαδικασίες φιλτραρίσματος του χώρου αναζήτησης που δεν περιέχουν το ερώτημα. Όταν βρεθούν οι υποψήφιοι γράφοι, εκτελείται μια χρονοβόρα αναζήτηση σε αυτού με την χρήση κλασικών αντιστοίχισης υπογράφων.

Η διαδικασία του φιλτραρίσματος, οδηγεί την αναζήτηση προς τους υποψήφιους (γραφήματα) και μόνο. Στόχος είναι η δημιουργία ενός παγκόσμιου ευρετηρίου, το οποίο θα περιέχει όλα τα δομικά χαρακτηριστικά των γράφων τα οποία θα μπορούν να συγκριθούν με τα δομικά χαρακτηριστικά ενός ερωτήματος που θα προκύψει.

Τα παραδοσιακά συστήματα συνήθως χρησιμοποιούν υπογράφους (μονοπάτια) μικρού μεγέθους. Το **SING** αποτελεί ένα απαραίτητο εργαλείο για την χρήση παραδοσιακών συστημάτων σε μεγάλους γράφους, αφού χρησιμοποιείται για τον σχηματισμό της έννοιας των χαρακτηριστικών, καθώς αποθηκεύει τον αρχικό τους κόμβο. Επιπλέον μετατροπές στον αλγόριθμο αντιστοίχισης, τον ωθούν στην έναρξη της όχι τυχαίας αναζήτησης, αλλά από επιλεγμένους κόμβους που διαθέτουν την ετικέτα του ερωτήματος. Παρόλα αυτά, απαιτείται πολύς χρόνος για την δημιουργία του ευρετηρίου λόγω του τεράστιου μεγέθους του και ως επίπτωση η χρήση των κανόνων που περικλείουν την εξόρυξη υψηλής υποστήριξης, πραγματοποιείται μόνο σε συχνούς και μη περιττούς υπογράφους. Αναλυτικότερα, το **gIndex** αποθηκεύει τους συχνούς υποψήφιους σε μια δομή συμπαγές δέντρου, ενώ το **FGIndex**, μέσω δύο δεικτών που αποθηκεύονται στην κύρια μνήμη και στο δίσκο αντίστοιχα. Για την ανάθεση ενός χαρακτηριστικού στο ευρετήριο, η διαδικασία έχει ως εξής: Το ερώτημα γράφος αρχικά θα πρέπει να εκτελεστεί στο ευρετήριο που βρίσκεται στην κύρια μνήμη και σε περίπτωση που δεν υπάρξει κάποιο αποτέλεσμα, τότε το ίδιο ευρετήριο χρησιμοποιείται για την εύρεση των μπλοκ που βρίσκονται στον δίσκο. Τέλος, το **GraphFind** στοχεύει στον περιορισμό του μεγέθους του ευρετηρίου, μέσω μιας τεχνικής χαμηλής υποστήριξης εξόρυξης δεδομένων. Η χρήση των παραπάνω απαιτεί ένα σύστημα εξόρυξης δεδομένων με μεγάλος κόστος.

Το **CTree** με την χρήση ενός αποτελεσματικού, προσεγγιστικού αλγορίθμου που ονομάζεται **neighbor biased mapping**, υλοποιεί ένα κλείσιμο γράφου στους γράφους που βρίσκονται στη βάση δεδομένων, ευθυγραμμίζοντας κορυφές και ακμές και στην συνέχεια αποθηκεύει τον συνθετικό γράφο σε μια δενδρική δομή τύπου R-tree. Έπειτα, εκτελείται μια διαδικασία αντιστοίχισης στους γράφους κλεισίματος από πάνω προς τα κάτω, που απαιτεί τον περισσότερο χρόνο του Ctree.

Στην παρακάτω ενότητα, θα παρουσιαστεί μια νέα προσέγγιση που προτείνεται, η οποία περιέχει στοιχεία από την **GraphGrep** και ονομάζεται **GraphGrepSX**. Το GraphGrepSX αποθηκεύει σε μια δομή Suffix tree, χαρακτηριστικά τα οποία αποτελούν μονοπάτια περιορισμένου μήκους. Τα κοινά προθέματα των μονοπατιών, διευκολύνουν τον αλγόριθμο στο να αποφύγει τους πλεονασμούς και να αναπαριστά το ευρετήριο με έναν πιο σταθερό τρόπο. Αυτή η τεχνική, είναι αποδοτικότερη σε γράφους με μικρό χώρο ετικέτας, καθώς θα ήταν συχνός ο συνδυασμός ιδίων ετικετών. Η προσέγγιση του GraphGrepSX έχει την δυνατότητα να επιταχύνει λειτουργίες σχετικά με την κατασκευή του ευρετηρίου και το φιλτράρισμα της αναζήτησης.

### 5.3.2) GraphGrepSX.

Στην συνέχεια θα προβληθούν οι φάσεις που ακολουθούνται για την υλοποίηση της μεθόδου.

#### 5.3.2.1) Η φάση της προεπεξεργασίας.

Η αρχική αυτή φάση της προεπεξεργασίας, μεταφέρει τα χαρακτηριστικά της βάσης δεδομένων στο παγκόσμιο ευρετήριο. Στην συνέχεια, κάθε κόμβος  $v_j$  ενός γράφου  $G_i$  που βρίσκεται στη βάση δεδομένων, δέχεται μια αναζήτηση βάθους. Όσο εκτελείται αυτή η διαδικασία, όλα τα μονοπάτια που διαθέτουν μήκος ίσο με  $lp$ , εξάγονται. Οι ετικέτες κόμβων χρησιμοποιούνται για την αναπαράσταση των μονοπατιών. Έτσι, κάθε μονοπάτι θα αντιστοιχεί με τις αντίστοιχες ετικέτες. Τα υπομονοπάτια  $(v_i, v_k, \dots, v_j)$  για  $1 \leq i \leq j \leq lp$  ενός μονοπατιού  $(v_1, v_2, \dots, v_j)$  αποτελούν χαρακτηριστικά τα οποία θα βρίσκονται στο συνολικό ευρετήριο. Κάθε φορά που εξάγεται ένα μονοπάτι, παρατηρούμε τις φορές που θα εμφανιστεί στους γράφους της βάσης δεδομένων. Αυτά τα χαρακτηριστικά έπειτα αποθηκεύονται σε ένα Suffix δέντρο. Ένα Suffix δέντρο κατασκευάζεται και δέχεται ενημερώσεις κατά την διαδικασία της αναζήτησης σε βάθος. Στο Suffix δέντρο, τα μονοπάτια που λαμβάνονται κατά τη διάρκεια της αναζήτησης σε βάθος, συμβολίζονται με έναν κόμβο. Οι κόμβοι πέρα από αυτό, αποθηκεύουν τον κατάλογο των γραφημάτων και τον αριθμό εμφανίσεων του μονοπατιού σε όλους τους γράφους. Επίσης, κάθε μονοπάτι μπορεί να ανακατασκευαστεί μέσω των προγόνων του στο Suffix δέντρο. Το Suffix δέντρο χρησιμοποιείται ως ένα N-αριαίο δέντρο, όπου τα παιδιά των κόμβων απεικονίζονται από μια λίστα. Οι εμφανίσεις των χαρακτηριστικών των γράφων βρίσκονται στον κατάλογο των εμφανίσεων, ο οποίος αποθηκεύεται σε ένα δυαδικό δέντρο, ενώ παράλληλα αποθηκεύεται και σε ένα αρχείο αρχειοθέτησης μαζί με το Suffix δέντρο.

Για τον συμβολισμό του χειρότερου κόστους, δηλαδή την ύπαρξη του μεγαλύτερου αριθμού διαφορετικών ετικετών κατά την αναζήτηση του παιδιού ενός κόμβου στο Suffix δέντρο, χρησιμοποιούμε το  $|ls|$ .

Για την ενημέρωση του δυαδικού δέντρου που είναι αποθηκευμένος ο κατάλογος εμφανίσεων κατά μία τιμή, το κόστος ισούται με  $\log |D|$ . Ενώ, το κόστος που αναλογεί για κάθε επίσκεψη σε

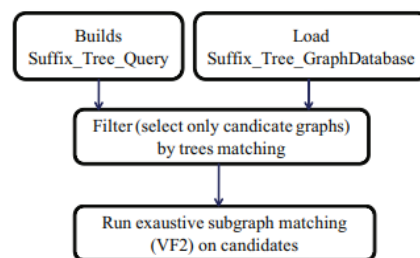
βάθος είναι  $n_i m_i^l$ , όπου τα  $n_i$  και  $m_i$  συμβολίζουν αντίστοιχα τον συνολικό αριθμό των κόμβων και την μέγιστη αξία τους στο γράφημα.

Τέλος, για την κατασκευή του ευρετηρίου, το συνολικό κόστος ισούται με  $O(\sum_I^{|D|} (n_i m_i^l |L| \log |D|))$ .

(Bonnici, 2010)

### 5.3.2.2) Φιλτράρισμα και αντιστοίχιση.

Η φάση του φιλτραρίσματος ασχολείται με τους γράφους που δεν θα μπορούσαν να αντιστοιχηθούν με το γράφημα ενός ερωτήματος Q. Η διαδικασία αυτή πραγματοποιείται σε δύο βήματα. Αρχικά, ο γράφος ερωτήματος q υπόκειται σε επεξεργασία και τα χαρακτηριστικά που συλλέγονται μεταφέρονται σε ένα Suffix δέντρο. Σε αυτήν την περίπτωση, έχουμε περάσει την φάση της προεπεξεργασίας και ερευνούμε μόνο τα μέγιστα μονοπάτια που επισκέφθηκαν κατά την λειτουργία της αναζήτησης βάθους στον γράφο του ερωτήματος. Ένα μονοπάτι, φτάνει το μέγιστο μέγεθος του όταν το μήκος του είναι ίσο με  $l_p$  ή όταν η περαιτέρω επέκταση του καθίσταται αδύνατη λόγω του τερματισμού της αναζήτησης σε βάθος. Οι εμφανίσεις των μέγιστων μονοπατιών αποθηκεύονται σε συγκεκριμένους κόμβους του δέντρου Suffix, οι οποίοι προηγουμένως έχουν σημειωθεί καθώς περιέχουν το τελικό σημείο ενός μέγιστου μονοπατιού. Έπειτα, γίνεται αντιστοίχιση μεταξύ του Suffix δέντρου του ερωτήματος και του Suffix δέντρου του παγκόσμιου ευρετηρίου. Η διαδικασία αυτή ονομάζεται κλάδεμα των υποψήφιων γραφημάτων. Οι κόμβοι που έχουν σημειωθεί, αντιπροσωπεύουν ένα σημειωμένο επίσης μονοπάτι και αναζητούνται στο Suffix δέντρο του παγκόσμιου ευρετηρίου. Σε περιπτώσεις γραφημάτων που δεν διαθέτουν επισημασμένο μονοπάτι ή διαθέτουν μόνο ένα με αριθμό εμφάνισης συγκριτικά μικρότερο σε σχέση με τον αριθμό εμφάνισης του ερωτήματος, δεν γίνονται αποδεκτά. Τα αποδεκτά γραφήματα αποτελούν τα υποψήφια γραφήματα που ταιριάζουν με το γράφημα του ερωτήματος. Η χρήση ενός αλγορίθμου συνδυαστικής αναζήτησης που ονομάζεται VF2 είναι απαραίτητη για τον ισομορφισμό των υπογράφων αλλά και για το κλάδεμα του χώρου αναζήτησης καθώς διαθέτει κανόνες τοπολογικής και σημασιολογικής εφικτότητας.



Εικόνα 28: Φιλτράρισμα και επαλήθευση στο GraphGrepSX

Το Suffix δέντρο του ερωτήματος εμφανίζεται ως  $T_q$ , ενώ ο αριθμός των κόμβων που περιέχει ως  $|T_q|$ . Το κόστος κατασκευής είναι  $O(n_q m_q^l |L_s|)$  όπου το  $n_q$  είναι ο συνολικός αριθμός των κόμβων στο δέντρο,  $m_q$  είναι ο μέγιστος βαθμός ενός κόμβου και  $|L_s|$  είναι ο μέγιστος αριθμός διαφορετικών ετικετών στο γράφημα ερωτήματος. Το κόστος του βήματος κλαδέματος δίνεται από το χρόνο αντιστοίχισης του Suffix δέντρου του ερωτήματος με το ευρετήριο της βάσης δεδομένων, δηλαδή  $O(|T_q| |L_s|)$ , συν το μέρο όρο του χρόνου της επαλήθευσης των εμφανίσεων, δηλαδή  $O(|D| \log |D|)$ . Επομένως ο συνολικός χρόνος είναι  $O(|T_q| |L_s| |D| \log |D|)$  και, έστω  $C$  το σύνολο των υποψήφιων γραφημάτων, το κόστος για κάθε επαλήθευση του υποψηφίου  $C_i$  είναι  $O(|V[C_i]| |V[C_i]|)$ . (Bonnici, 2010)

### 5.3.2.3) Πειράματα και εφαρμογές.

Παρακάτω, θα παρατηρήσουμε την απόδοση της προσέγγισης του GraphGrepSX σε σύγκριση με άλλα συστήματα αναζήτησης γράφων όπως το GraphGrep, το CTree, το GCoding και το SING. Για την υλοποίηση του πειράματος, χρησιμοποιήθηκε η βάση δεδομένων του AIDS, που περιέχει χαρακτηριστικά για 42.000 χημικές ενώσεις, δοκιμασμένες ενδείξεις και μικρό αριθμό κόμβων. Οι γράφοι ερωτήματα παράχθηκαν τυχαία από τη βάση δεδομένων και διαλέγοντας μια κορυφή  $v$ , πραγματοποιούσαν την επίσκεψη. Επίσης, το συνολικό μέγεθος των 42.000 χημικών ενώσεων μοιράστηκε σε τρία υποσύνολα (8.000, 24.000, 42.000).

| DB dim. | GraphGrepSX | GraphGrep | CTree | GCoding | SING |
|---------|-------------|-----------|-------|---------|------|
| 8000    | 16.51       | 550.4     | 8.21  | 632.21  | 22   |
| 24000   | 38          | 10399.39  | 25.34 | 1956.36 | 66   |
| 42000   | 66          | 45600.49  | 42.42 | 2944.8  | 108  |

Πίνακας 17: Χρόνος που απαιτείται για την κατασκευή δεικτών (sec).

Όπως παρατηρούμε στον πίνακα 17, μόνο το CTree αποδίδει σχεδόν στον ίδιο βαθμό με το GraphGrepSX στον χρόνο κατασκευής του ευρετηρίου, ενώ τα μεγέθη των ευρετηρίων που προέκυψαν απεικονίζεται στον πίνακα 18. Η συμπαγής δομή του Suffix δέντρου, αποτρέπει τους πλεονασμούς στο ευρετήριο, με αποτέλεσμα να απαιτεί λιγότερη χωρητικότητα σε σχέση με τις άλλες προσεγγίσεις.

| DB dim. | GraphGrepSX | GraphGrep | CTree | GCoding | SING  |
|---------|-------------|-----------|-------|---------|-------|
| 8000    | 3684        | 293992    | 13884 | 6687    | 8445  |
| 24000   | 11020       | 928912    | 41372 | 20088   | 22279 |
| 42000   | 18668       | 1577012   | 70208 | 30651   | 42830 |

Πίνακας 18: Μέγεθος των δεικτών(Kb).

Στον πίνακα 19, βλέπουμε τον χρόνο που χρειάζεται κάθε προσέγγιση για την εκτέλεση του φιλτραρίσματος, όπου όλες οι προσεγγίσεις είναι αποδοτικές.

| Query dim. | GraphGrepSX | GraphGrep | CTree | GCoding | SING  |
|------------|-------------|-----------|-------|---------|-------|
| 4          | 0.05        | 0.012     | 1.34  | 0.0042  | 0.51  |
| 8          | 0.05        | 0.006     | 1.57  | 0.01    | 0.76  |
| 16         | 0.041       | 0.005     | 1.51  | 0.026   | 0.17  |
| 32         | 0.04        | 0.014     | 1.01  | 0.059   | 0.071 |

**Πίνακας 19: Χρόνος φιλτραρίσματος (sec).**

Στον πίνακα 20, απεικονίζεται ο συνολικός χρόνος που απαιτείται για την αναζήτηση των υποψηφίων.

| Query dim. | GraphGrepSX | GraphGrep | CTree | GCoding | SING  |
|------------|-------------|-----------|-------|---------|-------|
| 4          | 14.9        | 15.41     | 13.27 | 23.61   | 12.4  |
| 8          | 17.5        | 7.1       | 44.24 | 15.79   | 15.24 |
| 16         | 2.08        | 12.78     | 51.07 | 5.39    | 0.798 |
| 32         | 1.07        | 3.56      | 50.91 | 1.25    | 0.136 |

**Πίνακας 20: Χρόνος αναζήτησης (sec).**

Μετά τις διαδικασίες του φιλτραρίσματος και της αναζήτησης, βρέθηκαν κάποιοι υποψήφιοι που εμφανίζονται στον πίνακα 21. Όπως βλέπουμε, τα CTree και GCoding παράγουν τον μικρότερο αριθμό υποψηφίων καθώς χρειάζονται πιο πολύ χρόνο εξαιτίας των λειτουργιών εξόρυξης κατά το φιλτράρισμα.

| Query dim. | GraphGrepSX | GraphGrep | CTree | GCoding | SING  |
|------------|-------------|-----------|-------|---------|-------|
| 4          | 26865       | 29196     | 16704 | 16188   | 23170 |
| 8          | 21337       | 13920     | 5840  | 8567    | 14012 |
| 16         | 1629        | 7053      | 289   | 1648    | 214   |
| 32         | 142         | 3193      | 3     | 142     | 4     |

**Πίνακας 21: Συνολικός αριθμός υποψηφίων.**

Στον τελευταίο πίνακα 22, συγκρίνουμε τους συνολικούς χρόνους που απαιτούνται από τα GraphGrepSX και SING για την εκτέλεση ενός ερωτήματος, με τον χρόνο που χρειάζεται το πρώτο να είναι εμφανώς μικρότερος σε σχέση με του δεύτερου αντίστοιχα. (Bonnici, 2010)

| Query dim. | GraphGrepSX | SING   |
|------------|-------------|--------|
| 4          | 15.89       | 23.39  |
| 8          | 18.56       | 26.42  |
| 16         | 3.07        | 1f1.45 |
| 32         | 2.04        | 10.68  |

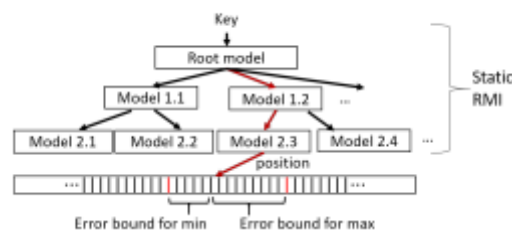
**Πίνακας 22: Συνολικός χρόνος (sec).**

#### 5.4) Ανανεώσιμος και προσαρμοστικός δείκτης μάθησης.

Έρευνες που έχουν δημοσιοποιηθεί, εμφανίζουν τα Learned Index τα οποία αποτελούν μια ιεραρχία μοντέλων μηχανικής μάθησης (ML) ως πιο αποδοτική λύση σε σχέση με τις τυπικές δομές ευρετηρίων στις βάσεις δεδομένων, καθώς εκπαιδεύονται από τα δεδομένα. Στην εικόνα 29, παρατηρούμε ότι αν θεωρήσουμε πως έχουμε ένα κλειδί, ο κόμβος-μοντέλο ενός ενδιάμεσου επιπέδου στην ιεραρχία γνωρίζει ποιος θυγατρικός κόμβος-μοντέλο θα χρησιμοποιηθεί, ενώ στο κατώτερο επίπεδο της ιεραρχίας, ένας κόμβος φύλλο αποτελεί μοντέλο υπεύθυνο για την εύρεση του κλειδιού σε έναν πυκνό πίνακα.

Ωστόσο, η χρήση τους μπορούν να χειριστούν μικρό εύρος λειτουργιών όπως οι αναζητήσεις σε δεδομένα που είναι μόνο προς ανάγνωση, χωρίς να υπάρχει η δυνατότητα ενημερώσεων. Το γεγονός αυτό αποτελεί το μεγαλύτερο πρόβλημα στα Learned Index καθώς τα καθιστά ανίκανα σε συνηθισμένες περιπτώσεις που σχετίζονται με την ανάγνωση και την εγγραφή σε δυναμικό φόρτο εργασίας.

Στόχος λοιπόν, είναι η δημιουργία ενός ευρετηρίου που να παράγει υψηλή απόδοση σε αυτά χρησιμοποιώντας τις γνώσεις από τον πυρήνα του Learned Index σε συνδυασμό με αποτελεσματικές τεχνικές αποθήκευσης και ευρετηρίασης. Ο συνδυασμός αυτός θα παρέχει αξιοσημείωτες επιδόσεις σε χρόνο και χωρητικότητα. Η επίτευξη αυτού το στόχου πραγματοποιείται με την δημιουργία του ALEX, μια δυναμική δομή δεδομένων, κατάλληλο για αποτελεσματικές σημειακές αναζητήσεις, εισαγωγές και διαγραφές δεδομένων, ενημερώσεις και μαζική φόρτωση που θα παρουσιάσουμε στις παρακάτω ενότητες.



Εικόνα 29: Παράδειγμα δομής ενός Learned Index.

##### 5.4.1) Μαθησιακοί δείκτες B+tree.

Ερευνητές αντιλήφθηκαν πως οι δομές δεικτών **B+tree** μπορούν να χαρακτηριστούν ως μοντέλα. Αν θεωρήσουμε ότι υπάρχει ένα κλειδί, μπορούν να βρουν τη θέση του στο τελευταίο επίπεδο της ιεραρχίας ενός ταξινομημένου πίνακα. Όπως όλα τα μοντέλα, έτσι και τα ευρετήρια που αποτελούν μοντέλα, μπορούν μέσω τεχνικών ML να διδαχθούν βελτιώνοντας την απόδοση του δείκτη σε αυτή την κατανομή δεδομένων, μαθαίνοντας το **CDF**. Το CDF αποτελεί μια πολύπλοκη αθροιστική συνάρτηση κατανομής, η οποία καθιστά ένα μοντέλο ML μη ακριβές. Για την επίλυση αυτού του προβλήματος, χρησιμοποιείται ένας δείκτης μοντέλο που ονομάζεται **RMI**. Το RMI

αποτελείται από επίπεδα, συνήθως δύο ή τρία, με το μοντέλο που βρίσκεται στο υψηλότερο επίπεδο να επιλέγει κάθε φορά το θυγατρικό του. Όταν επιλεγθεί το μοντέλο που βρίσκεται στο τελευταίο επίπεδο, τότε εκείνο προβλέπει τη θέση που βρίσκεται το κλειδί. Έτσι, φαίνεται πως το RMI αντικαθιστά τους κόμβους της δενδρικής δομής B+tree σε μοντέλα και άρα οι διακλαδώσεις προς το τελευταίο επίπεδο του αποτελούν πλέον συμπεράσματα μοντέλου σε ένα μαθησιακό δείκτη. Αν η πρόβλεψη του μοντέλου που βρίσκεται στο τελευταίο επίπεδο, είναι σωστή, τότε η χρήση μιας τοπικής αναζήτησης είναι αποτελεσματικότερη από την δυαδική αναζήτηση σε όλα τα στοιχεία του πίνακα. Επιπλέον, το RMI μπορεί να περιέχει διαφόρων τύπων μοντέλα. Η χωρητικότητα που απαιτείται για την αποθήκευση μοντέλων στο RMI είναι ιδιαίτερα μικρός σε σύγκριση με την χωρητικότητα που απαιτείται στο B+tree. Παρά την διαφορά μεγέθους χώρου αποθήκευσης στη μνήμη, ένα Learned Index αποδίδει πολύ καλύτερα σε λειτουργίες αναζήτησης σε σχέση με ένα B+tree. Όπως αναφέρθηκε και προηγουμένως βέβαια, τα Learned Index είναι ανίκανα να πραγματοποιήσουν τροποποιήσεις, όπως εισαγωγές, διαγραφές και ενημερώσεις.

#### 5.4.2) Εισαγωγή του ALEX.

Η προσέγγιση του ALEX, χρησιμοποιεί δύο παράγοντες. Αρχικά, έναν συμβιβασμό μεταξύ χώρου και χρόνου που έχει ως αποτέλεσμα οι αναζητήσεις να είναι αποδοτικότερες σε χρόνο και η δομή δεδομένων να δέχεται ενημερώσεις, άρα να είναι ανανεώσιμη. Για την έρευνα αυτού του συμβιβασμού, χρησιμοποιείται ένα Gapped Array (GA), που τακτοποιεί τους κόμβους των φύλλων. Έπειτα, το SRMI (στατικό RMI) υποστηρίζεται από ένα Learned Index, μόνο σε περιπτώσεις όπου ο αριθμός των επιπέδων και ο αριθμός των μοντέλων δεν τροποποιείται κατά την αρχικοποίηση. Σε καταστάσεις όπου η μοντελοποίηση της κατανομής δεδομένων είναι πολύπλοκη, η απόδοση του SRMI είναι μηδαμινή.

Το ALEX έχει την δυνατότητα να ενημερωθεί κατά την διάρκεια εκτέλεσης και μπορεί να γνωρίζει τις καθυστερήσεις από την υλοποίηση λειτουργιών αναζήτησης και εισαγωγής μέσα από την χρήση γραμμικών μοντέλων κόστους τα οποία λειτουργούν με βάση στατιστικά στοιχεία που υπολογίζονται από ένα RMI. Τα γραμμικά μοντέλα κόστους επίσης χρησιμοποιούνται για την αρχικοποίηση και την προσαρμογή της δομής RMI, ανάλογα με τον όγκο των εργασιών.

Το ALEX σε σύγκριση με τις υπόλοιπες δομές δεικτών, έχει ως στόχο να:

1. Ο χρόνος που απαιτείται για εισαγωγές να είναι συγκρίσιμος με του B+tree.
2. Ο χρόνος που απαιτείται για αναζητήσεις να είναι μικρότερος από ότι στα B+tree και στα Learned Index.
3. Ο χώρος αποθήκευσης που απαιτείται θα πρέπει να είναι μικρότερος και από τις δύο άλλες δομές αντίστοιχα.
4. Ο χώρος αποθήκευσης που περιέχει τα δεδομένα, δηλαδή στο τελευταίο επίπεδο της ιεραρχίας (φύλλα) θα πρέπει να είναι ανταγωνιστικός με το δυναμικό B+tree.

Η δυνατότητα αποθήκευσης ευρετηρίων σε μικρότερο χώρο, καθιστά δυνατή την αποθήκευση μεγαλύτερου αριθμού ευρετηρίων στον ίδιο χώρο. Παρακάτω θα δούμε με ποιον τρόπο το ALEX υλοποιεί αυτούς τους στόχους.

#### 5.4.2.1) Σχεδιασμός του ALEX.

Το ALEX παρουσιάζει κάποιες ανομοιότητες σε σχέση με τα άλλες δομές δεικτών Learned Index όσον αφορά την δομή δεδομένων που χρησιμοποιείται για την αποθήκευση δεδομένων στο επίπεδο των φύλλων. Το ALEX διαθέτει έναν κόμβο ανά φύλλο, με αποτέλεσμα οι μεμονωμένοι κόμβοι να είναι πιο ευέλικτοι και μειώνει τις μετατοπίσεις που πραγματοποιούνται κατά την διαδικασία μιας εισαγωγής. Για την αποδοτικότερη εισαγωγή και αναζήτηση, το ALEX θέτει κενά σε σημεία στον πίνακα στοιχείων για νέες εισαγωγές.

Μια ακόμη διαφορά είναι ότι το ALEX, επιτυγχάνει την εύρεση κλειδιών στα φύλλα μέσω εκθετικής αναζήτησης, για την αποφυγή μη έγκυρων προβλέψεων από το RMI. Μέσω πειραμάτων έχει αποδειχθεί ότι μια εκθετική αναζήτηση χωρίς όρια αποδίδει καλύτερα συγκριτικά με μια δυαδική αναζήτηση με όρια, καθώς η πρόβλεψη αποδοτικών μοντέλων φτάνει πολύ κοντά στον στόχο. Επίσης, το ALEX αντιθέτως με τα Learned Index, υλοποιεί εισαγωγή βάση μοντέλου, δηλαδή προσθέτει κλειδιά στους κόμβους στους οποίους υπολογίζει ένα μοντέλο ότι βρίσκεται το κλειδί. Επιπλέον, ανάλογα με τον όγκο εργασίας, το ALEX μπορεί να προσαρμόσει χαρακτηριστικά του RMI όπως το σχήμα και το ύψος.

Τέλος, σε αντίθεση με τα Learned Index δεν διαθέτει παραμέτρους που θα πρέπει να διαμορφώνονται ανάλογα με τις περιστάσεις.

#### 5.4.2.2) Διάταξη.

Οι κόμβοι δεδομένων, αποτελούν τους κόμβους των φύλλων όπου και αποθηκεύονται τα δεδομένα. Επίσης, εκεί αποθηκεύεται ένα μοντέλο γραμμικής παλινδρόμησης το οποίο χρησιμοποιείται για την αντιστοίχιση του κλειδιού με την θέση του και δύο Gapped Arrays για τα κλειδιά και τα ωφέλιμα φορτία αντίστοιχα. Τα κλειδιά και τα ωφέλιμα φορτία σχεδόν πάντα διαθέτουν σταθερό μέγεθος ενώ ταυτόχρονα χρησιμοποιούνται περιορισμοί για το μέγιστο επιτρεπτό μέγεθος ενός κόμβου. Όπως αναφέρθηκε και προηγουμένως, τα κενά ανάμεσα στα στοιχεία του πίνακα δεδομένων δημιουργούνται μέσω του Gapped Array το οποίο χρησιμοποιεί εισαγωγές βάση του μοντέλου σε αντίθεση με την δομή ενός B+tree το οποίο διαθέτει τα κενά στο τέλος του πίνακα του. Επίσης, το Gapped Array προσθέτει το πιο κοντινό κλειδί σε όλα τα κενά του πίνακα για την εξασφάλιση της εκθετικής αναζήτησης. Ωστόσο, για να μην χάνεται χρόνος την στιγμή της σάρωσης, τα κενά θα πρέπει να αγνοούνται. Για αυτό τον λόγο, οι κόμβοι δεδομένων περιέχουν ένα γρήγορο **bitmap**, ο οποίος απαιτεί λίγο χώρο και εξετάζει αν οι θέσεις του πίνακα αποτελούνται από κλειδί ή είναι κενές.

Οι εσωτερικοί κόμβοι από την άλλη αποθηκεύουν στο εσωτερικό τους, ένα μοντέλο γραμμικής παλινδρόμησης και έναν πίνακα με δείκτες σε θυγατρικούς κόμβους. Στο ALEX, οι εσωτερικοί κόμβοι εκτιμούν τη θέση του δείκτη του παιδιού. Επίσης βάζουμε έναν περιορισμό για το μέγιστο μέγεθος ενός εσωτερικού κόμβου. Στόχος των εσωτερικών κόμβων είναι να υλοποιήσει την κατάτμηση του χώρου όπου βρίσκονται τα κλειδιά με έναν πιο ευπροσάρμοστο τρόπο σε σύγκριση με τα Learned Index.

#### **5.4.3) Αλγόριθμοι ALEX.**

Παρακάτω θα αναλύσουμε τους αλγόριθμους που χρησιμοποιεί ο ALEX για κάποιες από τις λειτουργίες του όπως αναζητήσεις, εισαγωγές, και διαγραφές.

##### **5.4.3.1) Αναζητήσεις και ερωτήματα.**

Η διαδικασία για την αναζήτηση ενός κλειδιού, ξεκινάει με την κατ'εξακολούθηση χρήση του μοντέλου για την εύρεση μιας θέσης στον πίνακα δεικτών και να οδηγηθούμε μέσω του δείκτη στο θυγατρικό κόμβο, μέχρις ότου καταλήξουμε στον κόμβο δεδομένων. Μέσα από το μοντέλο στο κόμβο δεδομένων, προσπαθούμε να υπολογίσουμε την θέση του κλειδιού αναζήτησης που βρίσκεται στον πίνακα κλειδιών και για την εύρεση της πραγματικής θέσης του κλειδιού, εκτελούμε εκθετική αναζήτηση αν θεωρηθεί απαραίτητο. Αν η αναζήτηση κριθεί επιτυχής και βρούμε το κλειδί, τότε διαβάζουμε την τιμή του στην αντίστοιχη θέση ενός πίνακα που ονομάζεται payloads και επιστρέφουμε την εγγραφή. Αν όχι, τότε επιστρέφουμε μια εγγραφή που ισούται με το 0. Ένα ερώτημα εύρους που υποβάλλεται ξεκινάει μια εκτέλεση αναζήτησης με σκοπό να βρει τη θέση και τον κόμβο δεδομένων του αρχικού κλειδιού και έπειτα σαρώνει κατά μέτωπο μέχρις ότου καταλήξει στην τιμή τέλους τους εύρους. Απαραίτητη σε αυτό το σημείο είναι η χρήση του bitmap για την αγνόηση των κενών, ενώ ίσως χρειαστούν δείκτες που βρίσκονται στον κόμβο για την σύνδεση με τον επόμενο κόμβο δεδομένων.

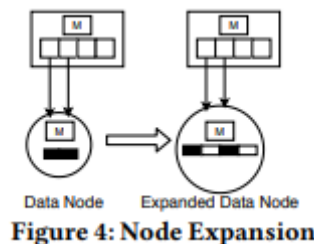
##### **5.4.3.2) Λειτουργία εισαγωγής σε μη πλήρη κόμβο δεδομένων.**

Η διαδικασία εισαγωγής εμφανίζει ομοιότητες με την διαδικασία αναζήτησης που αναλύθηκε προηγουμένως. Ο αλγόριθμος έχει ως σκοπό την επιλογή της σωστής θέσης για την εισαγωγή. Έτσι, το μοντέλο του κόμβου δεδομένων χρησιμοποιείται για τον υπολογισμό της θέσης εισαγωγής. Αν το αποτέλεσμα του υπολογισμού δεν είναι το επιθυμητό, τότε υλοποιούμε εκθετική αναζήτηση για την εύρεση της ακριβούς θέσης εισαγωγής. Όταν η θέση εισαγωγής που έχουμε είναι κενή, τότε μπορούμε να εισάγουμε το στοιχείο εκεί και η διαδικασία τελειώνει. Αν η θέση εισαγωγής δεν είναι κενή, τότε θα πρέπει να μετακινήσουμε τα στοιχεία κατά μια θέση, προς το κοντινότερο κενό του πίνακα και έπειτα να προσθέσουμε το στοιχείο στη θέση που έμεινε κενή. Ο χρόνος εισαγωγής που απαιτείται συμβολίζεται με  $O(\log n)$ .

#### 5.4.3.3) Λειτουργία εισαγωγής σε πλήρη κόμβο δεδομένων.

Για την αποφυγή προβληματισμών σε περίπτωση που ένας κόμβος δεδομένων γεμίσει, το ALEX μέσω των επεκτάσεων και των διαχωρισμών που αποτελούν δύο μηχανισμούς που εκτελεί, έχει την δυνατότητα να αυξήσει την χωρητικότητα. Μοντέλα κόστους είναι αυτά που υποδεικνύουν στο ALEX ποιους μηχανισμούς να χρησιμοποιήσει ανά περίπτωση. Στην συνέχεια θα αναλύσουμε τον όρο της πληρότητας, τους μηχανισμούς επέκτασης και διάσπασης και τα μοντέλα κόστους.

1. **Κριτήρια πληρότητας:** Όσο περισσότερο γεμίζει ένας κόμβος δεδομένων, τόσο η απόδοση μιας εισαγωγής σε ένα Gapped Array μειώνεται, αφού τα κενά λιγοστεύουν. Για αυτό το λόγο ο ALEX δεν αφήνει τους κόμβους δεδομένων να γεμίσει πλήρως. Προσδιορίζονται όρια ανώτατης και κατώτατης πυκνότητας στον Gapped Array:  $d_l, d_u \in (0, 1]$  θέτοντας ότι  $d_l \leq d_u$ . Ως πυκνότητα εννοούμε το σύνολο των θέσεων που γεμίζουν από εισαγωγές στοιχείων. Αντιλαμβανόμαστε ότι ένας κόμβος είναι πλήρης, όταν με την προσθήκη μιας νέας εισαγωγής το  $d_l$  υπερβαίνει το  $d_u$ .
2. **Μηχανισμός επέκτασης:** Έστω ότι έχουμε ένα κόμβο με  $n$  κλειδιά, για την επέκταση του κόμβου υποστηρίζεται ένα μεγαλύτερο Gapped Array με  $n/d_l$  θέσεις. Στην συνέχεια θα πρέπει να επανεκπαιδεύσουμε το μοντέλο γραμμικής παλινδρόμησης, ώστε μέσω αυτού να ξεκινήσουμε να υλοποιούμε τις εισαγωγές των στοιχείων σε αυτόν τον καινούριο και μεγαλύτερης χωρητικότητας κόμβο. Στην εικόνα 30, βλέπουμε έναν κόμβο ο οποίος έχει επεκταθεί και πλέον διαθέτει 4 υποδοχές.



Εικόνα 30: Επέκταση ενός κόμβου.

3. **Μηχανισμός διαχωρισμού:** Για να καταφέρουμε να διαχωρίσουμε έναν κόμβο δεδομένων σε δύο μέρη, θα πρέπει να ταξινομήσουμε τα κλειδιά στους δύο νέους κόμβους δεδομένων που θα χρησιμοποιηθούν. Για την λειτουργία του διαχωρισμού, το ALEX χρησιμοποιεί δύο μεθόδους. Η πρώτη μέθοδος ονομάζεται πλευρικός διαχωρισμός (**Splitting sideways**). Αν ο εσωτερικός κόμβος γονέας του κόμβου που διαιρέθηκε δεν είναι πλήρης, τότε μετατρέπουμε τους δείκτες του κόμβου γονέα προς τον διαιρεμένο κόμβο σε δείκτες προς

τους δύο νέους κόμβους δεδομένων. Αν τυχόν υπήρχαν πλεονάζοντες δείκτες, τότε επίσης μεταφέρονται στους νέους κόμβους που δημιουργήθηκαν.

Αν ο εσωτερικός κόμβος γονέας του κόμβου που διαιρέθηκε είναι πλήρης, τότε μπορεί να εφαρμοστεί διάσπαση του γονικού εσωτερικού κόμβου.

Η δεύτερη μέθοδος ονομάζεται διάσπαση προς τα κάτω (**Splitting down**) και υλοποιείται με την τροποποίηση ενός κόμβου δεδομένων. Ο κόμβος δεδομένων γίνεται εσωτερικός κόμβος που ακολουθείται από δύο θυγατρικούς κόμβους δεδομένων, οι οποίοι εκπαιδεύονται στα κλειδιά τους.

4. **Μοντέλα κόστους:** Τα γραμμικά μοντέλα κόστους υπολογίζουν τους χρόνους που απαιτούνται για λειτουργίες αναζήτησης και εισαγωγής μέσω δύο παραγόντων. Τον μέσο αριθμό επαναλήψεων εκθετικής αναζήτησης και τον μέσο αριθμό μετατοπίσεων που χρειάζονται για μια εισαγωγή. Τα μοντέλα αυτά είναι που “συμβουλεύεται” το ALEX για να επιλέξει τον κατάλληλο μηχανισμό ανά περίπτωση.

#### 5.4.3.4) Διαγραφή και ενημερώσεις.

Αν θέλουμε να διαγράψουμε ένα κλειδί, προηγουμένως θα πρέπει να εκτελέσουμε μια αναζήτηση ώστε να βρούμε την θέση του και έπειτα να το αφαιρέσουμε. Η εκτέλεση μιας διαγραφής είναι πολύ ευκολότερη σε σύγκριση με μια εισαγωγή καθώς σε αυτή την περίπτωση δεν μετακινούνται υπάρχοντα κλειδιά. Όταν ένας κόμβος δεδομένων, μετά από επανειλημμένες διαγραφές φτάσει στο χαμηλότερο όριο πυκνότητας του  $d_r$ , τότε ο κόμβος αυτός συρρικνώνεται.

Ενας συνδυασμός εισαγωγής και διαγραφής χρειάζεται για την εκτέλεση ενημέρωσης σε ένα κλειδί. Από την άλλη, αν η ενημέρωση στοχεύει ένα ωφέλιμο φορτίο, τότε θα πρέπει να αναζητηθεί το κλειδί και έπειτα να γραφτεί η καινούρια τιμή του ωφέλιμου φορτίου.

## 6) Βιβλιογραφία

- 1) A. Ailamaki, “Databases and hardware: the beginning and sequel of a beautiful friendship,” Proceedings of the VLDB Endowment, vol. 8, no. 12, pp. 2058–2061, 2015.
- 2) P. Bonnet. “What's Up with the Storage Hierarchy?,” Conference on Innovative Data Systems Research, 2017.
- 3) R. Appuswamy, “OligoArchive: Using DNA in the DBMS storage hierarchy,” Biennial Conference on Innovative Data Systems Research, 2019.
- 4) A. Abulila, “FlatFlash: Exploiting the Byte-Accessibility of SSDs within a Unified Memory-Storage Hierarchy,” Association for Computing Machinery, pp. 971-985, 2019.
- 5) F. Soleza, “Mach: A Pluggable Metrics Storage Engine for the Age of Observability,,” CIDR, 2022.
- 6) L. Changhong, “A Graph Database Storage Engine for Provenance Graphs,” DBKDA 2020, The Twelfth International Conference on Advances in Databases, Knowledge, and Data Applications, 2020.
- 7) Q. Mao, “Efficient Storage Design in Log-Structured Merge (LSM) Tree Databases,” UC Riverside, 2022.
- 8) S. Ladra, “Scalable and queryable compressed storage structure for raster data,” Information Systems, vol. 72, pp. 179-204, 2017.
- 9) S. Zhou, “A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions,” ArXiv, 2022.
- 10) D. Teuling, “A comparison of methods for clustering longitudinal data with slowly changing trends,” Commun Stat Simul Comput, 2021.
- 11) Y. Alotaibi, “A New Meta-Heuristics Data Clustering Algorithm Based on Tabu Search and Adaptive Search Memory,” Symmetry, vol. 14, no. 3, 2022.
- 12) C. Santos-Mangudo, “A fair-multicluster approach to clustering of categorical data,” Central European Journal of Operations Research, vol. 31, pp. 1-22, 2022.

- 13) H. Diakoff, "Database indexing: yesterday and today," *The Indexer: The International Journal of Indexing*, 2004.
- 14) G. Samson, "Large Spatial Database Indexing with aX-tree," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 3, pp. 759-773, 2018.
- 15) M. Fouad, "An Efficient Technique for Indexing Temporal Databases," *International Journal of Computer Applications*, vol. 127, no. 13, pp. 32-37, 2015.
- 16) V. Bonnici, "Enhancing Graph Database Indexing by Suffix Tree Structure," *Lecture Notes in Computer Science*, vol. 6282, pp. 195–203, 2010.
- 17) J. Ding, "ALEX: An Updatable Adaptive Learned Index," *Association for Computing Machinery*, pp. 969-984, 2020.