ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προδραστική Ανάλυση Δεδομένων μέσω Ιστορικών Ερωτημάτων σε Περιβάλλον Edge Computing

ΤΣΑΛΙΑΚΟΣ ΒΑΣΙΛΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΚΟΛΟΜΒΑΤΣΟΣ
Αναπληρωτής Καθηγητής

Λαμία, Ιούνιος 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προδραστική Ανάλυση Δεδομένων μέσω Ιστορικών Ερωτημάτων σε Περιβάλλον Edge Computing

ΤΣΑΛΙΑΓΚΟΣ ΒΑΣΙΛΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΚΟΛΟΜΒΑΤΣΟΣ
Αναπληρωτής Καθηγητής

Λαμία, Ιούνιος 2026

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

Proactive Data Analysis through Historical Queries in Edge Computing Environments

TSALIAGKOS VASILIS

FINAL THESIS

ADVISOR

KONSTANTINOS KOLOMVATSOS
Associate Professor

Lamia, June 2026

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 19/06/2026

Ο – Η Δηλ.
ΤΣΑΛΙΑΓΚΟΣ ΒΑΣΙΛΗΣ

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία εξετάζει τη δυνατότητα αξιοποίησης ιστορικών ερωτημάτων για την υποστήριξη μηχανισμών Proactive Data Analysis σε περιβάλλοντα Edge Computing. Η συνεχής αύξηση του όγκου των δεδομένων που παράγονται από εφαρμογές Internet of Things (IoT), αισθητήρες και κατανεμημένα πληροφοριακά συστήματα δημιουργεί την ανάγκη για αποδοτικότερες τεχνικές επεξεργασίας, οι οποίες να μειώνουν την καθυστέρηση απόκρισης και το υπολογιστικό κόστος χωρίς σημαντική απώλεια ακρίβειας.

Για τον σκοπό αυτό αναπτύχθηκε ένας προσομοιωτής edge nodes, ο οποίος δημιουργεί συνθετικά δεδομένα βασισμένα στο σύνολο δεδομένων Iris, παράγει πολυδιάστατα range queries και προσομοιώνει διαφορετικά σενάρια φόρτου εργασίας. Η προτεινόμενη προσέγγιση συνδυάζει τεχνικές Approximate Query Processing και ανάλυσης ιστορικών ερωτημάτων. Συγκεκριμένα, χρησιμοποιούνται histograms ως συνοπτικές αναπαραστάσεις των δεδομένων για την εκτίμηση των αποτελεσμάτων των ερωτημάτων, ενώ εφαρμόζεται subtractive clustering για την ομαδοποίηση παρόμοιων queries και τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης.

Η αξιολόγηση του μοντέλου πραγματοποιήθηκε μέσω δεικτών απόδοσης που περιλαμβάνουν την κάλυψη δεδομένων, τον χρόνο εκτέλεσης και το σφάλμα προσέγγισης. Επιπλέον, εξετάστηκαν διαφορετικοί τύποι workload, ενώ πραγματοποιήθηκε ανάλυση ευαισθησίας ως προς τον αριθμό των edge nodes και τον αριθμό των histogram bins, με στόχο τη διερεύνηση της επίδρασής τους στην απόδοση και την ακρίβεια του συστήματος.

Τα αποτελέσματα έδειξαν ότι η χρήση συνοπτικών αναπαραστάσεων μπορεί να μειώσει σημαντικά τον χρόνο εκτέλεσης των ερωτημάτων σε σχέση με την ακριβή επεξεργασία, διατηρώντας παράλληλα αποδεκτό επίπεδο σφάλματος. Παράλληλα, η ανάλυση ιστορικών ερωτημάτων επέτρεψε τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης, τα οποία μπορούν να αξιοποιηθούν για την υποστήριξη μηχανισμών προδραστικής ανάλυσης δεδομένων. Τα ευρήματα της εργασίας υποδεικνύουν ότι ο συνδυασμός τεχνικών Edge Computing, Approximate Query Processing και Clustering αποτελεί μία υποσχόμενη προσέγγιση για την αποδοτικότερη διαχείριση και ανάλυση δεδομένων σε κατανεμημένα περιβάλλοντα.

ABSTRACT

This thesis investigates the use of historical queries to support Proactive Data Analysis mechanisms in Edge Computing environments. The continuous growth of data generated by Internet of Things (IoT) applications, sensors, and distributed information systems has created the need for more efficient data processing techniques that reduce response time and computational cost while maintaining an acceptable level of accuracy.

To address this challenge, an edge node simulator was developed, generating synthetic data based on the Iris dataset, producing multidimensional range queries, and simulating different workload scenarios. The proposed approach combines Approximate Query Processing techniques with historical query analysis. Specifically, histograms are used as data summaries for estimating query results, while subtractive clustering is applied to group similar queries and identify recurring access patterns.

The evaluation of the proposed model was conducted using performance metrics including data coverage, execution latency, and approximation error. In addition, different workload types were examined, and a sensitivity analysis was performed with respect to the number of edge nodes and the number of histogram bins in order to investigate their impact on system performance and estimation accuracy.

The experimental results demonstrated that the use of summary-based approximations can significantly reduce query execution time compared to exact processing while maintaining an acceptable level of error. Furthermore, the analysis of historical queries enabled the identification of recurring access patterns that can be exploited to support proactive data analysis mechanisms. The findings indicate that the combination of Edge Computing, Approximate Query Processing, and Clustering techniques represents a promising approach for efficient data management and analysis in distributed computing environments.

Table of Contents

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

1.1 Το Πρόβλημα και η Σημασία του

Η ραγδαία ανάπτυξη των τεχνολογιών Internet of Things (IoT), των έξυπνων συσκευών, των ασύρματων αισθητήρων και των κυβερνοφυσικών συστημάτων έχει οδηγήσει σε εκθετική αύξηση του όγκου των παραγόμενων δεδομένων. Σε καθημερινή βάση, δισεκατομμύρια συσκευές συλλέγουν και μεταδίδουν πληροφορίες που σχετίζονται με το περιβάλλον, την ανθρώπινη δραστηριότητα, τις βιομηχανικές διαδικασίες και τις μεταφορές. Η συνεχής αυτή παραγωγή δεδομένων δημιουργεί νέες απαιτήσεις ως προς την αποθήκευση, τη διαχείριση και κυρίως την επεξεργασία τους.

Παραδοσιακά, η ανάλυση των δεδομένων πραγματοποιούνταν σε κεντρικά υπολογιστικά νέφη (Cloud Computing), όπου τα δεδομένα μεταφέρονταν από τις συσκευές παραγωγής προς απομακρυσμένα κέντρα δεδομένων. Παρότι η προσέγγιση αυτή προσφέρει σημαντική υπολογιστική ισχύ και μεγάλη αποθηκευτική ικανότητα, παρουσιάζει περιορισμούς σε εφαρμογές που απαιτούν απόκριση σε πραγματικό χρόνο. Η συνεχής μεταφορά μεγάλου όγκου δεδομένων προς το cloud αυξάνει την καθυστέρηση επικοινωνίας, επιβαρύνει το δίκτυο και ενδέχεται να δημιουργήσει προβλήματα κλιμάκωσης.

Για την αντιμετώπιση των περιορισμών αυτών αναπτύχθηκε το Edge Computing, ένα υπολογιστικό μοντέλο στο οποίο μέρος της επεξεργασίας μεταφέρεται κοντά στην πηγή παραγωγής των δεδομένων. Οι edge κόμβοι λειτουργούν ως ενδιάμεσα επίπεδα μεταξύ των συσκευών και του cloud, επιτρέποντας την τοπική επεξεργασία, αποθήκευση και ανάλυση των πληροφοριών. Με τον τρόπο αυτό μειώνεται η καθυστέρηση, περιορίζεται η κατανάλωση δικτυακών πόρων και βελτιώνεται η συνολική απόδοση του συστήματος.

Ωστόσο, τα περιβάλλοντα edge computing χαρακτηρίζονται από περιορισμένους υπολογιστικούς και αποθηκευτικούς πόρους. Σε αντίθεση με τα κεντρικά datacenters, οι edge κόμβοι διαθέτουν περιορισμένη μνήμη, μικρότερη επεξεργαστική ισχύ και συχνά περιορισμένη ενεργειακή αυτονομία. Επομένως, η συνεχής εκτέλεση μεγάλου αριθμού ερωτημάτων πάνω στα δεδομένα μπορεί να οδηγήσει σε σημαντική αύξηση του υπολογιστικού κόστους.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα range queries, δηλαδή τα ερωτήματα που αναζητούν εγγραφές εντός συγκεκριμένων ορίων τιμών σε μία ή περισσότερες διαστάσεις. Τα ερωτήματα αυτά αποτελούν βασικό μηχανισμό αναζήτησης σε συστήματα βάσεων δεδομένων, εφαρμογές ανάλυσης δεδομένων, γεωγραφικά πληροφοριακά συστήματα και εφαρμογές Internet of Things. Καθώς ο αριθμός των ερωτημάτων αυξάνεται, η συνεχής ακριβής αξιολόγησή τους μπορεί να καταστεί ιδιαίτερα δαπανηρή.

Μία εναλλακτική προσέγγιση είναι η αξιοποίηση της γνώσης που παράγεται από τα ιστορικά ερωτήματα. Τα ερωτήματα που έχουν εκτελεστεί στο παρελθόν αποτελούν πολύτιμη πηγή πληροφορίας, καθώς αποτυπώνουν τα πρότυπα πρόσβασης των χρηστών και τις περιοχές των δεδομένων που παρουσιάζουν αυξημένο ενδιαφέρον. Μέσω της ανάλυσης αυτών των πληροφοριών καθίσταται δυνατός ο εντοπισμός επαναλαμβανόμενων μοτίβων, τα οποία μπορούν να αξιοποιηθούν για τη βελτίωση της απόδοσης του συστήματος.

Η λογική αυτή οδηγεί στην έννοια του Proactive Data Analysis, σύμφωνα με την οποία το σύστημα δεν λειτουργεί αποκλειστικά αντιδραστικά, αλλά προσπαθεί να προετοιμαστεί για

μελλοντικές ανάγκες επεξεργασίας. Αντί να περιμένει την υποβολή ενός νέου ερωτήματος, μπορεί να αξιοποιήσει ιστορικά δεδομένα και πρότυπα χρήσης ώστε να εκτιμήσει πιθανές μελλοντικές απαιτήσεις και να προετοιμάσει εκ των προτέρων χρήσιμες πληροφορίες.

Η ανάπτυξη μηχανισμών προδραστικής ανάλυσης παρουσιάζει ιδιαίτερο ενδιαφέρον στα σύγχρονα πληροφοριακά συστήματα, καθώς μετατοπίζει το επίκεντρο από την απλή επεξεργασία ερωτημάτων στην αξιοποίηση της γνώσης που παράγεται κατά τη λειτουργία του συστήματος. Τα ιστορικά ερωτήματα αποτελούν πολύτιμη πηγή πληροφορίας σχετικά με τη συμπεριφορά των χρηστών και τις περιοχές των δεδομένων που εμφανίζουν αυξημένο ενδιαφέρον. Η αξιοποίηση αυτής της πληροφορίας μπορεί να συμβάλει στη βελτίωση της αποδοτικότητας του συστήματος χωρίς την ανάγκη πρόσθετων υπολογιστικών πόρων.

Παράλληλα, η χρήση προσεγγιστικών τεχνικών επεξεργασίας δεδομένων αποκτά ολοένα και μεγαλύτερη σημασία σε περιβάλλοντα όπου η ταχύτητα απόκρισης είναι κρίσιμη. Σε πολλές περιπτώσεις, μία γρήγορη και αξιόπιστη εκτίμηση μπορεί να είναι περισσότερο χρήσιμη από μία απολύτως ακριβή απάντηση που απαιτεί σημαντικά μεγαλύτερο χρόνο επεξεργασίας. Η συνδυαστική αξιοποίηση ιστορικών ερωτημάτων και Approximate Query Processing δημιουργεί νέες δυνατότητες για την ανάπτυξη αποδοτικών μηχανισμών υποστήριξης αποφάσεων σε περιβάλλοντα Edge Computing.

Η παρούσα εργασία εντάσσεται ακριβώς σε αυτό το πλαίσιο και εξετάζει κατά πόσο η ανάλυση ιστορικών ερωτημάτων μπορεί να συμβάλει στη μείωση του υπολογιστικού κόστους και στη βελτίωση της αποδοτικότητας της επεξεργασίας δεδομένων σε περιβάλλοντα Edge Computing.

1.2 Σκοπός και Στόχοι της Εργασίας

Σκοπός της παρούσας εργασίας είναι η διερεύνηση της δυνατότητας αξιοποίησης ιστορικών range queries για την υποστήριξη μηχανισμών Proactive Data Analysis σε περιβάλλοντα Edge Computing. Η βασική ιδέα είναι ότι τα ιστορικά ερωτήματα περιέχουν πληροφορία η οποία μπορεί να αξιοποιηθεί για τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης και τη μελλοντική βελτιστοποίηση της επεξεργασίας δεδομένων.

Για την επίτευξη του σκοπού αυτού αναπτύχθηκε ένα περιβάλλον προσομοίωσης edge κόμβων, μέσα στο οποίο δημιουργούνται συνθετικά δεδομένα, παράγονται ερωτήματα εύρους τιμών και αξιολογούνται διαφορετικές τεχνικές ανάλυσης. Η χρήση ενός ελεγχόμενου περιβάλλοντος επιτρέπει τη συστηματική μελέτη της συμπεριφοράς του προτεινόμενου μοντέλου και τη μέτρηση της αποτελεσματικότητάς του.

Παράλληλα, η εργασία εξετάζει τη χρήση τεχνικών Approximate Query Processing μέσω συνοπτικών αναπαραστάσεων δεδομένων (summaries). Αντί να πραγματοποιείται πλήρης επεξεργασία όλων των δεδομένων για κάθε ερώτημα, χρησιμοποιούνται ιστογράμματα που επιτρέπουν την ταχύτερη εκτίμηση των αποτελεσμάτων με αποδεκτό επίπεδο σφάλματος.

Επιπλέον, διερευνάται η εφαρμογή τεχνικών clustering στα ιστορικά ερωτήματα, με στόχο τον εντοπισμό περιοχών του χώρου αναζήτησης που εμφανίζουν αυξημένη συχνότητα πρόσβασης. Η αναγνώριση τέτοιων προτύπων αποτελεί βασική προϋπόθεση για την ανάπτυξη μελλοντικών μηχανισμών πρόβλεψης ερωτημάτων, caching και προδραστικής επεξεργασίας.

Οι βασικοί στόχοι της εργασίας συνοψίζονται ως εξής:

• Δημιουργία ενός περιβάλλοντος προσομοίωσης edge computing.

• Παραγωγή και αξιολόγηση ιστορικών range queries.

• Εφαρμογή Approximate Query Processing μέσω histograms.

• Αναγνώριση επαναλαμβανόμενων query patterns με χρήση clustering.

• Αξιολόγηση της απόδοσης της προτεινόμενης προσέγγισης μέσω πειραματικών μετρήσεων.

• Διερεύνηση της δυνατότητας υποστήριξης μηχανισμών Proactive Data Analysis.

Πέρα από την αξιολόγηση της βασικής λειτουργίας του προτεινόμενου μοντέλου, η εργασία εξετάζει και τον τρόπο με τον οποίο διαφορετικές παράμετροι επηρεάζουν την απόδοσή του. Συγκεκριμένα, μελετάται η επίδραση του αριθμού των edge nodes στην κλιμάκωση του συστήματος, καθώς και η επίδραση του αριθμού των histogram bins στην ακρίβεια των προσεγγιστικών εκτιμήσεων.

Η διερεύνηση των παραμέτρων αυτών πραγματοποιείται μέσω πειραματικής ανάλυσης ευαισθησίας, η οποία επιτρέπει την καλύτερη κατανόηση των συμβιβασμών μεταξύ ακρίβειας, υπολογιστικού κόστους και χρόνου απόκρισης. Με τον τρόπο αυτό δεν αξιολογείται μόνο η αποτελεσματικότητα της προτεινόμενης προσέγγισης, αλλά και οι συνθήκες υπό τις οποίες μπορεί να εφαρμοστεί αποδοτικότερα.

1.3 Συνεισφορά της Εργασίας

Η κύρια συνεισφορά της παρούσας εργασίας είναι η συνδυαστική αξιοποίηση τεχνικών Edge Computing, Approximate Query Processing και Clustering στο πλαίσιο της προδραστικής ανάλυσης δεδομένων.

Αρχικά αναπτύχθηκε ένας ολοκληρωμένος προσομοιωτής edge κόμβων, ο οποίος επιτρέπει τη δημιουργία δεδομένων και ιστορικών workloads βασισμένων στα στατιστικά χαρακτηριστικά του συνόλου δεδομένων Iris. Το περιβάλλον αυτό χρησιμοποιήθηκε ως πειραματική πλατφόρμα για την αξιολόγηση της συμπεριφοράς διαφορετικών τύπων ερωτημάτων και διαφορετικών συνθηκών φόρτου.

Παράλληλα, εφαρμόστηκαν τεχνικές Approximate Query Processing μέσω ιστογραμμάτων, με στόχο τη μείωση του χρόνου επεξεργασίας των ερωτημάτων. Τα αποτελέσματα έδειξαν ότι η χρήση summaries μπορεί να επιταχύνει σημαντικά τη διαδικασία αξιολόγησης χωρίς σημαντική υποβάθμιση της ποιότητας των αποτελεσμάτων.

Επιπλέον, πραγματοποιήθηκε ανάλυση ιστορικών ερωτημάτων μέσω τεχνικών subtractive clustering. Η διαδικασία αυτή επέτρεψε τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης και την επιλογή clusters που μπορούν να αποτελέσουν υποψήφιους στόχους για μελλοντική προδραστική επεξεργασία.

Τέλος, η εργασία παρέχει πειραματικά αποτελέσματα που επιβεβαιώνουν ότι η αξιοποίηση ιστορικών queries μπορεί να αποτελέσει χρήσιμο εργαλείο για τη βελτίωση της αποδοτικότητας των edge συστημάτων και τη δημιουργία μηχανισμών προδραστικής υποστήριξης αποφάσεων.

Ιδιαίτερη συνεισφορά της εργασίας αποτελεί η πειραματική διερεύνηση της σχέσης μεταξύ ακρίβειας και απόδοσης μέσω της μεταβολής κρίσιμων παραμέτρων του συστήματος. Συγκεκριμένα, αξιολογείται η επίδραση του αριθμού των histogram bins στην ποιότητα των προσεγγιστικών απαντήσεων, καθώς και η επίδραση του αριθμού των edge nodes στη συνολική κλιμάκωση του προτεινόμενου μοντέλου.

Επιπλέον, η εργασία παρέχει μία ολοκληρωμένη πειραματική μελέτη διαφορετικών τύπων workload, επιτρέποντας την αξιολόγηση της συμπεριφοράς του συστήματος σε ποικίλες συνθήκες λειτουργίας. Η προσέγγιση αυτή συμβάλλει στην καλύτερη κατανόηση των παραγόντων που επηρεάζουν την αποτελεσματικότητα των μηχανισμών Approximate Query Processing και Proactive Data Analysis σε κατανεμημένα περιβάλλοντα Edge Computing.

Τέλος, η εργασία λειτουργεί ως μία πειραματική βάση για μελλοντική έρευνα γύρω από τη χρήση ιστορικών ερωτημάτων, τεχνικών clustering και προσεγγιστικών μεθόδων επεξεργασίας δεδομένων σε συστήματα μεγάλης κλίμακας.

1.4 Δομή της Εργασίας

Η παρούσα εργασία οργανώνεται σε επτά κεφάλαια.

Στο Κεφάλαιο 2 παρουσιάζεται η βιβλιογραφική επισκόπηση και αναλύονται οι βασικές έννοιες που σχετίζονται με το Edge Computing, τα range queries, το Approximate Query Processing, τις τεχνικές clustering και την ανάλυση ιστορικών ερωτημάτων. Παράλληλα παρουσιάζονται οι σημαντικότερες ερευνητικές προσεγγίσεις που έχουν προταθεί στη διεθνή βιβλιογραφία και αναδεικνύεται το ερευνητικό κενό που επιχειρεί να καλύψει η παρούσα εργασία.

Στο Κεφάλαιο 3 περιγράφεται αναλυτικά ο αλγόριθμος που αναπτύχθηκε. Παρουσιάζονται τα στάδια φόρτωσης και επεξεργασίας των δεδομένων, η δημιουργία των edge nodes, η παραγωγή των ερωτημάτων, η κατασκευή summaries και η διαδικασία clustering των ιστορικών queries.

Στο Κεφάλαιο 4 παρουσιάζεται το προτεινόμενο μοντέλο και η μεθοδολογία της έρευνας. Αναλύεται η αρχιτεκτονική του συστήματος, η δημιουργία των edge nodes, η χρήση συνοπτικών αναπαραστάσεων δεδομένων μέσω histograms, η εφαρμογή τεχνικών Approximate Query Processing, καθώς και η διαδικασία ανάλυσης ιστορικών ερωτημάτων μέσω clustering. Επιπλέον παρουσιάζονται οι διαφορετικοί τύποι workload και οι δείκτες απόδοσης που χρησιμοποιούνται για την αξιολόγηση του προτεινόμενου μοντέλου.

Στο Κεφάλαιο 5 παρουσιάζεται η πειραματική αξιολόγηση του προτεινόμενου συστήματος. Αναλύονται τα αποτελέσματα που προκύπτουν από τα διαφορετικά σενάρια workload, αξιολογείται η απόδοση των προσεγγιστικών τεχνικών επεξεργασίας και εξετάζεται η ποιότητα των clusters που προκύπτουν από την ανάλυση των ιστορικών ερωτημάτων. Επιπλέον πραγματοποιείται ανάλυση ευαισθησίας ως προς τον αριθμό των edge nodes και τον αριθμό των histogram bins, με στόχο τη διερεύνηση της επίδρασής τους στον χρόνο

απόκρισης, στην ακρίβεια των εκτιμήσεων και στη συνολική συμπεριφορά του συστήματος. Τέλος, πραγματοποιείται συνολική συζήτηση των αποτελεσμάτων και εξάγονται συμπεράσματα σχετικά με την αποτελεσματικότητα της προτεινόμενης προσέγγισης.

Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα της εργασίας. Συνοψίζονται τα σημαντικότερα ευρήματα, αξιολογείται η επίτευξη των στόχων που τέθηκαν και συζητείται η συμβολή της προτεινόμενης προσέγγισης στο πεδίο του Edge Computing και του Proactive Data Analysis.

Τέλος, στο Κεφάλαιο 7 παρουσιάζονται πιθανές μελλοντικές προεκτάσεις της εργασίας, όπως η αξιολόγηση σε μεγαλύτερα και πραγματικά datasets, η εφαρμογή σε πραγματικές edge υποδομές, η χρήση εναλλακτικών τεχνικών clustering και η ανάπτυξη μηχανισμών πρόβλεψης ερωτημάτων και caching.

ΚΕΦΑΛΑΙΟ 2 Βιβλιογραφική Επισκόπηση

2.1 Edge Computing

2.1.1 Ορισμός και Βασικές Αρχές

Η συνεχής αύξηση των δεδομένων που παράγονται από συσκευές Internet of Things (IoT), αισθητήρες και έξυπνες εφαρμογές έχει οδηγήσει στην ανάγκη μεταφοράς μέρους της επεξεργασίας από το cloud προς την άκρη του δικτύου. Σύμφωνα με τους Shi et al. (2016), το Edge Computing αποτελεί ένα κατανεμημένο υπολογιστικό μοντέλο στο οποίο οι υπολογιστικοί πόροι τοποθετούνται κοντά στην πηγή παραγωγής των δεδομένων, μειώνοντας την καθυστέρηση επικοινωνίας και την κατανάλωση δικτυακών πόρων.

Οι Satyanarayanan (2017) και Bonomi et al. (2012) υποστηρίζουν ότι η επεξεργασία στην άκρη του δικτύου είναι απαραίτητη για εφαρμογές που απαιτούν άμεση απόκριση, όπως τα αυτόνομα οχήματα, οι έξυπνες πόλεις και τα βιομηχανικά συστήματα αυτοματισμού. Σε αντίθεση με το παραδοσιακό cloud computing, το οποίο βασίζεται σε κεντρικές υποδομές, το edge computing εκμεταλλεύεται κατανεμημένους κόμβους που βρίσκονται κοντά στους τελικούς χρήστες.

Η φιλοσοφία του Edge Computing βασίζεται στη μεταφορά της επεξεργασίας όσο το δυνατόν πιο κοντά στο σημείο παραγωγής των δεδομένων. Η προσέγγιση αυτή επιτρέπει την άμεση αξιοποίηση των πληροφοριών που συλλέγονται από αισθητήρες, κινητές συσκευές και έξυπνες εφαρμογές χωρίς να απαιτείται η συνεχής μεταφορά τους σε απομακρυσμένα κέντρα δεδομένων. Ως αποτέλεσμα, μειώνονται οι καθυστερήσεις επικοινωνίας και αυξάνεται η δυνατότητα παροχής υπηρεσιών πραγματικού χρόνου.

Η ανάπτυξη του Edge Computing συνδέεται άμεσα με την εξάπλωση των εφαρμογών Internet of Things. Δισεκατομμύρια συσκευές παράγουν καθημερινά τεράστιο όγκο δεδομένων, γεγονός που καθιστά δύσκολη τη διαχείρισή τους αποκλειστικά μέσω κεντρικών cloud υποδομών. Η αποκεντρωμένη επεξεργασία που προσφέρει το Edge Computing επιτρέπει την καλύτερη διαχείριση αυτών των δεδομένων και συμβάλλει στη βελτίωση της επεκτασιμότητας των συστημάτων.

Εκτός από τη μείωση της καθυστέρησης, το Edge Computing συμβάλλει σημαντικά στη βελτίωση της αξιοπιστίας των συστημάτων. Σε πολλές περιπτώσεις, η συνεχής αποστολή δεδομένων προς απομακρυσμένα datacenters δεν είναι εφικτή λόγω περιορισμών συνδεσιμότητας ή αυξημένου φόρτου δικτύου. Η ύπαρξη τοπικών edge κόμβων επιτρέπει τη συνέχιση της επεξεργασίας ακόμη και όταν η σύνδεση με το cloud είναι περιορισμένη ή προσωρινά μη διαθέσιμη.

Παράλληλα, η αρχιτεκτονική του Edge Computing θεωρείται συμπληρωματική προς το Cloud Computing και όχι ανταγωνιστική. Οι edge κόμβοι αναλαμβάνουν την αρχική επεξεργασία και φιλτράρισμα των δεδομένων, ενώ το cloud χρησιμοποιείται για μακροχρόνια αποθήκευση, σύνθετη ανάλυση και εκπαίδευση μοντέλων μηχανικής μάθησης. Με τον τρόπο αυτό επιτυγχάνεται καλύτερη αξιοποίηση των διαθέσιμων υπολογιστικών πόρων.

Ένα ακόμη σημαντικό χαρακτηριστικό του Edge Computing είναι η δυνατότητα υποστήριξης εφαρμογών που απαιτούν υψηλό επίπεδο ιδιωτικότητας και ασφάλειας. Η τοπική επεξεργασία των δεδομένων περιορίζει την ανάγκη μεταφοράς ευαίσθητων πληροφοριών

μέσω του δικτύου και μειώνει την έκθεσή τους σε πιθανούς κινδύνους. Για τον λόγο αυτό η αρχιτεκτονική edge υιοθετείται ολοένα και περισσότερο σε τομείς όπως η υγεία, οι χρηματοοικονομικές υπηρεσίες και οι βιομηχανικές εφαρμογές.

Παράλληλα, η αποκεντρωμένη δομή του Edge Computing προσφέρει μεγαλύτερη ευελιξία στην ανάπτυξη εφαρμογών. Οι υπολογιστικοί πόροι μπορούν να κατανεμηθούν δυναμικά ανάλογα με τις ανάγκες του συστήματος, επιτρέποντας την αποτελεσματικότερη αξιοποίηση των διαθέσιμων υποδομών. Το χαρακτηριστικό αυτό αποτελεί βασικό πλεονέκτημα σε περιβάλλοντα με έντονες μεταβολές φόρτου εργασίας.

2.1.2 Διαχείριση Δεδομένων σε Περιβάλλοντα Edge

Η διαχείριση δεδομένων αποτελεί μία από τις σημαντικότερες προκλήσεις στα περιβάλλοντα edge computing. Οι Varghese et al. (2016) και Mach & Becvar (2017) επισημαίνουν ότι οι edge κόμβοι διαθέτουν περιορισμένους πόρους αποθήκευσης και επεξεργασίας, γεγονός που καθιστά απαραίτητη τη χρήση αποδοτικών τεχνικών διαχείρισης δεδομένων.

Οι Abbas et al. (2018) προτείνουν μηχανισμούς κατανεμημένης επεξεργασίας και αποθήκευσης, ενώ οι Premsankar et al. (2018) παρουσιάζουν αρχιτεκτονικές που επιτρέπουν την τοπική επεξεργασία δεδομένων πριν την αποστολή τους στο cloud. Η προσέγγιση αυτή μειώνει σημαντικά τον φόρτο του δικτύου και βελτιώνει τους χρόνους απόκρισης.

Η αποτελεσματική διαχείριση δεδομένων είναι κρίσιμη για την επιτυχία των εφαρμογών Edge Computing, καθώς οι αποφάσεις που λαμβάνονται σχετικά με την αποθήκευση και την επεξεργασία επηρεάζουν άμεσα τη συνολική απόδοση του συστήματος. Σε αντίθεση με τα παραδοσιακά περιβάλλοντα cloud, όπου η αποθήκευση θεωρείται σχεδόν απεριόριστη, οι edge κόμβοι διαθέτουν περιορισμένους πόρους και συνεπώς απαιτούνται μηχανισμοί βελτιστοποίησης.

Για τον λόγο αυτό έχουν αναπτυχθεί τεχνικές συνοπτικής αναπαράστασης δεδομένων (data summarization), συμπίεσης και φιλτραρίσματος, οι οποίες επιτρέπουν τη μείωση του όγκου των πληροφοριών που πρέπει να αποθηκευτούν ή να μεταφερθούν. Οι τεχνικές αυτές είναι ιδιαίτερα σημαντικές σε εφαρμογές όπου παράγονται συνεχώς νέα δεδομένα και η άμεση επεξεργασία τους αποτελεί βασική απαίτηση.

Ένα από τα βασικότερα ζητήματα στη διαχείριση δεδομένων στο edge είναι η επιλογή των δεδομένων που θα αποθηκευτούν τοπικά και αυτών που θα μεταφερθούν στο cloud. Η απόφαση αυτή επηρεάζει άμεσα την απόδοση του συστήματος, καθώς η αποθήκευση μεγάλου όγκου δεδομένων σε edge κόμβους μπορεί να εξαντλήσει τους διαθέσιμους πόρους.

Επιπλέον, η κατανεμημένη φύση των edge συστημάτων δημιουργεί προκλήσεις σχετικά με τη συνέπεια και τον συγχρονισμό των δεδομένων μεταξύ διαφορετικών κόμβων. Για τον λόγο αυτό έχουν προταθεί τεχνικές προσωρινής αποθήκευσης (caching), replication και τοπικής συγκέντρωσης δεδομένων, με στόχο τη μείωση της επικοινωνίας και τη βελτίωση της συνολικής απόδοσης του συστήματος.

Η ανάγκη για αποδοτική διαχείριση δεδομένων έχει οδηγήσει επίσης στην ανάπτυξη μηχανισμών τοπικής ανάλυσης και λήψης αποφάσεων. Αντί οι edge κόμβοι να λειτουργούν αποκλειστικά ως ενδιάμεσοι σταθμοί μεταφοράς δεδομένων, αποκτούν τη δυνατότητα

εκτέλεσης αναλυτικών εργασιών σε τοπικό επίπεδο. Η προσέγγιση αυτή μειώνει την εξάρτηση από το cloud και επιτρέπει ταχύτερη ανταπόκριση σε γεγονότα που απαιτούν άμεση επεξεργασία.

Στο πλαίσιο αυτό αποκτούν ιδιαίτερη σημασία οι τεχνικές Approximate Query Processing και οι μηχανισμοί ανάλυσης ιστορικών ερωτημάτων, οι οποίοι επιτρέπουν την εξαγωγή χρήσιμης πληροφορίας με περιορισμένο υπολογιστικό κόστος. Οι τεχνικές αυτές αποτελούν τη θεωρητική βάση πάνω στην οποία αναπτύσσεται το προτεινόμενο μοντέλο της παρούσας εργασίας.

2.1.3 Προκλήσεις του Edge Computing

Παρά τα σημαντικά πλεονεκτήματα που προσφέρει, το Edge Computing συνοδεύεται από μία σειρά τεχνικών προκλήσεων. Σε αντίθεση με τα κεντρικά υπολογιστικά νέφη, οι edge κόμβοι διαθέτουν περιορισμένους υπολογιστικούς πόρους, μικρότερη χωρητικότητα αποθήκευσης και συχνά χαμηλότερη ενεργειακή αυτονομία. Οι περιορισμοί αυτοί καθιστούν απαραίτητη τη χρήση αποδοτικών τεχνικών επεξεργασίας και διαχείρισης δεδομένων.

Οι περιορισμοί αυτοί επηρεάζουν άμεσα τη δυνατότητα εκτέλεσης σύνθετων αναλυτικών εργασιών στους edge κόμβους. Ενώ τα σύγχρονα συστήματα παράγουν ολοένα και μεγαλύτερο όγκο δεδομένων, οι διαθέσιμοι πόροι παραμένουν συχνά περιορισμένοι. Ως αποτέλεσμα, η εκτέλεση ακριβών αλγορίθμων ανάλυσης μπορεί να οδηγήσει σε σημαντική αύξηση του χρόνου απόκρισης και της κατανάλωσης πόρων.

Η ανάγκη εξισορρόπησης μεταξύ ακρίβειας και αποδοτικότητας έχει οδηγήσει στην ανάπτυξη τεχνικών που επιτρέπουν την παραγωγή προσεγγιστικών αποτελεσμάτων με χαμηλότερο υπολογιστικό κόστος. Οι τεχνικές αυτές έχουν αποκτήσει ιδιαίτερη σημασία τα τελευταία χρόνια, καθώς προσφέρουν μία πρακτική λύση στα προβλήματα κλιμάκωσης που αντιμετωπίζουν τα σύγχρονα edge περιβάλλοντα.

Επιπλέον, η κατανεμημένη φύση των edge περιβαλλόντων δημιουργεί ζητήματα συγχρονισμού και συνέπειας δεδομένων μεταξύ διαφορετικών κόμβων. Η συνεχής ανταλλαγή πληροφοριών μπορεί να επιβαρύνει το δίκτυο και να αυξήσει τους χρόνους απόκρισης, ιδιαίτερα όταν ο αριθμός των συσκευών αυξάνεται σημαντικά.

Μία ακόμη σημαντική πρόκληση αφορά την κλιμάκωση των εφαρμογών. Καθώς ο αριθμός των συνδεδεμένων συσκευών και των παραγόμενων δεδομένων αυξάνεται, απαιτούνται τεχνικές που να επιτρέπουν την αποτελεσματική επεξεργασία μεγάλου όγκου ερωτημάτων χωρίς σημαντική αύξηση του υπολογιστικού κόστους. Για τον λόγο αυτό η έρευνα έχει στραφεί σε προσεγγιστικές μεθόδους επεξεργασίας και σε μηχανισμούς προδραστικής ανάλυσης δεδομένων.

Μία ακόμη πρόκληση αφορά την πρόβλεψη των μελλοντικών αναγκών του συστήματος. Καθώς τα ερωτήματα των χρηστών και οι απαιτήσεις των εφαρμογών μεταβάλλονται συνεχώς, η αποκλειστικά αντιδραστική επεξεργασία συχνά δεν επαρκεί για την επίτευξη υψηλής απόδοσης. Για τον λόγο αυτό αυξάνεται το ενδιαφέρον για μηχανισμούς Proactive Data Analysis, οι οποίοι επιδιώκουν να αξιοποιήσουν ιστορικά δεδομένα και πρότυπα χρήσης προκειμένου να προετοιμάζουν αναλύσεις πριν ακόμη ζητηθούν από τους χρήστες.

Η παρούσα εργασία εντάσσεται ακριβώς σε αυτό το ερευνητικό πλαίσιο, διερευνώντας κατά πόσο η ανάλυση ιστορικών ερωτημάτων και η χρήση συνοπτικών αναπαραστάσεων δεδομένων μπορούν να υποστηρίξουν πιο αποδοτικούς μηχανισμούς επεξεργασίας σε περιβάλλοντα Edge Computing.

2.2 Range Queries και Query Processing

2.2.1 Range Queries

Τα range queries χρησιμοποιούνται ευρέως σε εφαρμογές όπου απαιτείται αναζήτηση δεδομένων βάσει αριθμητικών ορίων. Παραδείγματα αποτελούν τα συστήματα παρακολούθησης αισθητήρων, οι γεωγραφικές εφαρμογές και οι πλατφόρμες ανάλυσης μεγάλων δεδομένων. Σε τέτοια περιβάλλοντα, ένα ερώτημα μπορεί να ζητά όλες τις εγγραφές που βρίσκονται εντός συγκεκριμένου εύρους τιμών για πολλαπλά χαρακτηριστικά.

Η αποδοτική επεξεργασία των range queries καθίσταται δυσκολότερη όσο αυξάνεται η διάσταση των δεδομένων. Το φαινόμενο αυτό είναι γνωστό ως “curse of dimensionality” και οδηγεί σε σημαντική αύξηση του υπολογιστικού κόστους, ιδιαίτερα όταν απαιτείται επεξεργασία μεγάλου αριθμού ερωτημάτων σε πραγματικό χρόνο.

Τα range queries αποτελούν μία από τις πιο συχνές κατηγορίες ερωτημάτων στις βάσεις δεδομένων και στα συστήματα αναλυτικής επεξεργασίας. Σε αντίθεση με τα ερωτήματα ακριβούς αντιστοίχισης (point queries), τα οποία αναζητούν συγκεκριμένες τιμές, τα range queries αναζητούν σύνολα εγγραφών που βρίσκονται εντός προκαθορισμένων διαστημάτων τιμών. Η ιδιότητα αυτή τα καθιστά ιδιαίτερα χρήσιμα σε εφαρμογές εξερεύνησης δεδομένων και λήψης αποφάσεων.

Η πολυδιάστατη φύση των range queries αυξάνει σημαντικά την πολυπλοκότητα της επεξεργασίας τους. Καθώς αυξάνεται ο αριθμός των χαρακτηριστικών που συμμετέχουν στο ερώτημα, μειώνεται η αποτελεσματικότητα πολλών παραδοσιακών τεχνικών αναζήτησης. Το φαινόμενο αυτό είναι ιδιαίτερα εμφανές σε εφαρμογές που βασίζονται σε αισθητήρες, συστήματα παρακολούθησης και αναλυτικές πλατφόρμες μεγάλων δεδομένων, όπου κάθε εγγραφή μπορεί να περιλαμβάνει δεκάδες ή και εκατοντάδες διαστάσεις.

Στη βιβλιογραφία έχουν προταθεί διάφορες τεχνικές για τη βελτίωση της επεξεργασίας των range queries, όπως εξειδικευμένες δομές ευρετηρίων, τεχνικές partitioning και μέθοδοι προσεγγιστικής εκτίμησης αποτελεσμάτων. Οι τεχνικές αυτές επιδιώκουν τη μείωση του χρόνου αναζήτησης και του υπολογιστικού κόστους, ιδιαίτερα σε περιβάλλοντα όπου η ταχύτητα απόκρισης αποτελεί κρίσιμο παράγοντα.

2.2.2 Δομές Ευρετηρίων για Range Queries

Ο Guttman (1984) εισήγαγε το R-tree, μία από τις πιο γνωστές δομές ευρετηρίων για χωρικά και πολυδιάστατα δεδομένα. Το R-tree οργανώνει τα δεδομένα σε ιεραρχικά ορθογώνια πλαίσια (bounding rectangles), επιτρέποντας αποδοτική εκτέλεση range queries.

Στη συνέχεια, οι Beckmann et al. (1990) πρότειναν το R*-tree, το οποίο βελτιώνει την απόδοση του R-tree μειώνοντας τις επικαλύψεις μεταξύ των κόμβων του δέντρου. Οι Papadias

et al. (2003) και Jagadish et al. (2005) παρουσίασαν επιπλέον βελτιώσεις και τεχνικές βελτιστοποίησης για πολυδιάστατες αναζητήσεις μεγάλης κλίμακας.

Η χρήση δομών ευρετηρίων αποτελεί βασικό μηχανισμό επιτάχυνσης των range queries. Αντί να πραγματοποιείται πλήρης σάρωση όλων των δεδομένων, τα ευρετήρια επιτρέπουν τον γρήγορο εντοπισμό των περιοχών που πιθανόν περιέχουν τα ζητούμενα αποτελέσματα.

Πέρα από τα R-tree και R*-tree, έχουν προταθεί και άλλες δομές, όπως τα KD-Trees και Quad-Trees, οι οποίες χρησιμοποιούνται σε διαφορετικά είδη πολυδιάστατων δεδομένων. Η επιλογή της κατάλληλης δομής εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων και το είδος των ερωτημάτων που εκτελούνται συχνότερα.

Οι δομές ευρετηρίων αποτελούν βασικό εργαλείο βελτιστοποίησης των συστημάτων διαχείρισης δεδομένων, καθώς μειώνουν σημαντικά τον αριθμό των εγγραφών που απαιτείται να εξεταστούν κατά την εκτέλεση ενός ερωτήματος. Η αποτελεσματικότητά τους είναι ιδιαίτερα εμφανής σε περιπτώσεις όπου τα δεδομένα παραμένουν σχετικά σταθερά και τα ερωτήματα εκτελούνται επανειλημμένα πάνω στο ίδιο σύνολο πληροφοριών.

Ωστόσο, σε περιβάλλοντα όπου τα δεδομένα μεταβάλλονται συνεχώς ή παράγονται με υψηλό ρυθμό, η διατήρηση και ενημέρωση σύνθετων δομών ευρετηρίων μπορεί να επιβαρύνει σημαντικά το σύστημα. Η διαδικασία ενημέρωσης των ευρετηρίων απαιτεί επιπλέον υπολογιστικούς πόρους και μπορεί να μειώσει τα οφέλη που προκύπτουν από τη χρήση τους.

Για τον λόγο αυτό η σύγχρονη έρευνα έχει στραφεί και σε εναλλακτικές προσεγγίσεις που συνδυάζουν ευρετήρια με τεχνικές προσεγγιστικής επεξεργασίας. Οι μέθοδοι αυτές επιδιώκουν να παρέχουν γρήγορες εκτιμήσεις των αποτελεσμάτων χωρίς να απαιτείται πλήρης προσπέλαση των δεδομένων ή διατήρηση πολύπλοκων δομών σε κάθε κόμβο του συστήματος.

2.2.3 Range Queries σε Περιβάλλοντα Edge Computing

Η επεξεργασία range queries σε περιβάλλοντα edge computing παρουσιάζει ιδιαίτερες απαιτήσεις λόγω των περιορισμένων πόρων των edge κόμβων. Παρότι οι δομές ευρετηρίων μπορούν να επιταχύνουν την αναζήτηση δεδομένων, η συνεχής εκτέλεση μεγάλου αριθμού πολυδιάστατων ερωτημάτων εξακολουθεί να επιβαρύνει σημαντικά το σύστημα.

Σε εφαρμογές Internet of Things και real-time analytics, τα ερωτήματα συχνά εκτελούνται πάνω σε συνεχώς μεταβαλλόμενα δεδομένα. Η ανάγκη για γρήγορη απόκριση περιορίζει τη δυνατότητα χρήσης πολύπλοκων μηχανισμών επεξεργασίας και καθιστά αναγκαία την ανάπτυξη εναλλακτικών προσεγγίσεων.

Η δυσκολία γίνεται ακόμη μεγαλύτερη όταν τα δεδομένα είναι κατανεμημένα σε πολλούς edge κόμβους. Σε τέτοιες περιπτώσεις, η απάντηση ενός range query ενδέχεται να απαιτεί επικοινωνία μεταξύ διαφορετικών κόμβων, ανταλλαγή ενδιάμεσων αποτελεσμάτων και συντονισμό των επιμέρους διαδικασιών επεξεργασίας. Οι ενέργειες αυτές αυξάνουν την καθυστέρηση και επιβαρύνουν το δίκτυο, περιορίζοντας την αποτελεσματικότητα του συστήματος.

Επιπλέον, οι εφαρμογές πραγματικού χρόνου απαιτούν συχνά την επεξεργασία μεγάλου αριθμού παρόμοιων ερωτημάτων μέσα σε μικρά χρονικά διαστήματα. Η επαναλαμβανόμενη εκτέλεση ακριβών υπολογισμών για κάθε νέο query μπορεί να οδηγήσει σε σημαντική

σπατάλη πόρων, ιδιαίτερα όταν τα ερωτήματα παρουσιάζουν κοινά χαρακτηριστικά ή επικαλυπτόμενες περιοχές αναζήτησης.

Για την αντιμετώπιση του προβλήματος έχουν προταθεί τεχνικές που βασίζονται σε συνοπτικές αναπαραστάσεις των δεδομένων και σε προσεγγιστικές μεθόδους εκτίμησης αποτελεσμάτων. Οι τεχνικές αυτές επιτρέπουν τη σημαντική μείωση του χρόνου εκτέλεσης των ερωτημάτων, διατηρώντας παράλληλα αποδεκτό επίπεδο ακρίβειας.

Για την αντιμετώπιση των περιορισμών αυτών, η βιβλιογραφία προτείνει τη χρήση συνοπτικών αναπαραστάσεων δεδομένων, όπως histograms, sketches και sampling τεχνικές, οι οποίες επιτρέπουν την εκτίμηση των αποτελεσμάτων ενός range query χωρίς πλήρη σάρωση των δεδομένων. Οι μέθοδοι αυτές αποτελούν βασικό συστατικό των συστημάτων Approximate Query Processing και έχουν χρησιμοποιηθεί ευρέως σε εφαρμογές μεγάλων δεδομένων και κατανεμημένων υποδομών.

Παράλληλα, ερευνητικές εργασίες όπως αυτές των Cormode και Garofalakis (2005) και Agarwal et al. (2013) έχουν δείξει ότι η αξιοποίηση ιστορικών ερωτημάτων μπορεί να συμβάλει σημαντικά στη βελτίωση της απόδοσης. Η αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης επιτρέπει την προετοιμασία ή ακόμη και την προϋπολογισμένη εκτέλεση συγκεκριμένων αναλύσεων πριν ζητηθούν από τους χρήστες. Η φιλοσοφία αυτή αποτελεί τη βάση των μηχανισμών Proactive Data Analysis που εξετάζονται στην παρούσα εργασία.

Στο πλαίσιο αυτό, η παρούσα εργασία διερευνά τον συνδυασμό συνοπτικών αναπαραστάσεων δεδομένων και ανάλυσης ιστορικών ερωτημάτων με στόχο τη μείωση του χρόνου απόκρισης των range queries σε περιβάλλοντα Edge Computing, διατηρώντας παράλληλα αποδεκτό επίπεδο ακρίβειας στα παραγόμενα αποτελέσματα.

2.3 Approximate Query Processing

2.3.1 Η Ανάγκη για Προσεγγιστικές Τεχνικές

Η συνεχής αύξηση του όγκου των δεδομένων έχει καταστήσει σε πολλές περιπτώσεις ανέφικτη την ακριβή επεξεργασία κάθε ερωτήματος σε πραγματικό χρόνο. Για τον λόγο αυτό αναπτύχθηκε το πεδίο του Approximate Query Processing (AQP), το οποίο επιδιώκει την παροχή γρήγορων προσεγγιστικών απαντήσεων με ελεγχόμενο σφάλμα.

Οι Chaudhuri et al. (1998) υποστηρίζουν ότι σε πολλές εφαρμογές ανάλυσης δεδομένων η ταχύτητα απόκρισης είναι σημαντικότερη από την απόλυτη ακρίβεια, γεγονός που καθιστά ιδιαίτερα χρήσιμες τις προσεγγιστικές τεχνικές.

Οι προσεγγιστικές τεχνικές έχουν αποκτήσει ιδιαίτερη σημασία σε εφαρμογές πραγματικού χρόνου, όπου η άμεση απόκριση είναι συχνά σημαντικότερη από την απόλυτη ακρίβεια. Σε τέτοιες περιπτώσεις, μία εκτίμηση με μικρό σφάλμα μπορεί να θεωρηθεί αποδεκτή εφόσον παρέχεται σημαντικά ταχύτερα από την ακριβή απάντηση.

Η φιλοσοφία του Approximate Query Processing βασίζεται στην παρατήρηση ότι πολλά αναλυτικά ερωτήματα δεν απαιτούν απόλυτα ακριβή αποτελέσματα. Αντίθετα, η γρήγορη

παροχή μίας αξιόπιστης εκτίμησης μπορεί να βελτιώσει σημαντικά την εμπειρία του χρήστη και να μειώσει το κόστος επεξεργασίας.

Η ανάγκη για Approximate Query Processing έγινε ακόμη πιο έντονη με την εμφάνιση εφαρμογών Big Data και συστημάτων συνεχούς ροής δεδομένων. Σε τέτοια περιβάλλοντα, η επεξεργασία όλων των διαθέσιμων δεδομένων για κάθε ερώτημα είναι συχνά πρακτικά αδύνατη λόγω χρονικών ή υπολογιστικών περιορισμών. Ως αποτέλεσμα, οι ερευνητές στράφηκαν στην ανάπτυξη τεχνικών που επιτρέπουν την παραγωγή γρήγορων εκτιμήσεων με μετρήσιμο και ελεγχόμενο σφάλμα.

Σύμφωνα με τους Hellerstein et al. (1997), η χρησιμότητα μιας προσεγγιστικής μεθόδου δεν εξαρτάται αποκλειστικά από την ακρίβειά της αλλά από την ισορροπία που επιτυγχάνει μεταξύ ταχύτητας και ποιότητας αποτελεσμάτων. Η φιλοσοφία αυτή έχει επηρεάσει σημαντικά τον σχεδιασμό σύγχρονων αναλυτικών συστημάτων, όπου η άμεση πρόσβαση σε πληροφορία θεωρείται συχνά σημαντικότερη από την απόλυτη ακρίβεια.

2.3.2 Summaries και Histograms

Μία από τις πιο διαδεδομένες τεχνικές Approximate Query Processing είναι η χρήση ιστογραμμάτων (histograms). Τα ιστογράμματα καταγράφουν τη συχνότητα εμφάνισης τιμών σε προκαθορισμένα διαστήματα (bins), επιτρέποντας την εκτίμηση της κατανομής των δεδομένων.

Οι Poosala et al. (1996) και Ioannidis (2003) έδειξαν ότι τα histograms μπορούν να χρησιμοποιηθούν αποτελεσματικά για την εκτίμηση του αποτελέσματος range queries χωρίς την ανάγκη πλήρους σάρωσης των δεδομένων.

Τα histograms αποτελούν μία συνοπτική αναπαράσταση της κατανομής των δεδομένων. Μέσω της ομαδοποίησης τιμών σε bins, είναι δυνατή η εκτίμηση της πιθανότητας εμφάνισης δεδομένων σε συγκεκριμένα διαστήματα χωρίς την ανάγκη επεξεργασίας όλων των εγγραφών.

Η τεχνική αυτή χρησιμοποιείται ευρέως από συστήματα διαχείρισης βάσεων δεδομένων για την εκτίμηση του κόστους εκτέλεσης ερωτημάτων και τη βελτιστοποίηση των πλάνων εκτέλεσης. Η χρήση histograms θεωρείται ιδιαίτερα αποτελεσματική όταν απαιτείται ισορροπία μεταξύ ταχύτητας και ακρίβειας.

Η παρούσα εργασία αξιοποιεί histograms ως summaries για την εκτίμηση των αποτελεσμάτων range queries, καθώς προσφέρουν καλή ισορροπία μεταξύ ακρίβειας και υπολογιστικού κόστους. Στην παρούσα εργασία επιλέγεται η χρήση histograms ως συνοπτικών αναπαραστάσεων των δεδομένων (summaries) για την εκτίμηση των αποτελεσμάτων range queries. Η επιλογή αυτή βασίζεται στην ικανότητά τους να προσφέρουν ικανοποιητική ακρίβεια εκτίμησης με χαμηλό υπολογιστικό και αποθηκευτικό κόστος, γεγονός που τα καθιστά κατάλληλα για εφαρμογές σε περιβάλλοντα edge computing.

Η αποτελεσματικότητα των histograms εξαρτάται σε μεγάλο βαθμό από τον τρόπο κατασκευής τους και από τον αριθμό των bins που χρησιμοποιούνται. Ένας μικρός αριθμός bins οδηγεί σε μεγαλύτερη συμπίεση της πληροφορίας και χαμηλότερες απαιτήσεις αποθήκευσης, αλλά ενδέχεται να μειώσει την ακρίβεια των εκτιμήσεων. Αντίθετα, η αύξηση του αριθμού των bins μπορεί να βελτιώσει την ποιότητα των αποτελεσμάτων, αυξάνοντας όμως το κόστος αποθήκευσης και επεξεργασίας.

Η βιβλιογραφία έχει προτείνει διάφορες παραλλαγές histograms, όπως equi-width histograms, equi-depth histograms και compressed histograms, οι οποίες στοχεύουν στη βελτίωση της ποιότητας των εκτιμήσεων ανάλογα με τα χαρακτηριστικά της κατανομής των δεδομένων. Η επιλογή της κατάλληλης μεθόδου εξαρτάται από τη φύση των δεδομένων και από τις απαιτήσεις της εφαρμογής.

Ένα από τα σημαντικότερα πλεονεκτήματα των histograms είναι ότι μπορούν να χρησιμοποιηθούν για την εκτίμηση της selectivity των range queries χωρίς την ανάγκη προσπέλασης των πραγματικών δεδομένων. Η δυνατότητα αυτή επιτρέπει τη σημαντική επιτάχυνση της επεξεργασίας ερωτημάτων και για τον λόγο αυτό τα histograms χρησιμοποιούνται ευρέως σε συστήματα βελτιστοποίησης ερωτημάτων και αναλυτικές πλατφόρμες μεγάλων δεδομένων.

2.3.3 Sampling και Sketches

Εκτός από τα histograms, έχουν προταθεί τεχνικές δειγματοληψίας (sampling) και συνοπτικών δομών (sketches). Οι Garofalakis et al. (2002) και Cormode & Garofalakis (2005) παρουσίασαν τεχνικές που επιτρέπουν την αποδοτική εκτίμηση στατιστικών μεγεθών σε μεγάλα σύνολα δεδομένων, χρησιμοποιώντας περιορισμένο χώρο αποθήκευσης.

Οι τεχνικές δειγματοληψίας βασίζονται στην επιλογή ενός αντιπροσωπευτικού υποσυνόλου των δεδομένων και στην εξαγωγή συμπερασμάτων για ολόκληρο το σύνολο. Η προσέγγιση αυτή μειώνει σημαντικά το κόστος επεξεργασίας, ιδιαίτερα σε πολύ μεγάλα datasets.

Αντίστοιχα, οι δομές sketch επιτρέπουν τη διατήρηση συνοπτικών στατιστικών πληροφοριών χρησιμοποιώντας εξαιρετικά μικρό χώρο αποθήκευσης. Για τον λόγο αυτό χρησιμοποιούνται συχνά σε συστήματα ροών δεδομένων (data streams), όπου ο διαθέσιμος χώρος και ο χρόνος επεξεργασίας είναι περιορισμένοι.

Παρότι οι τεχνικές sampling παρουσιάζουν σημαντικά πλεονεκτήματα ως προς το κόστος επεξεργασίας, η ποιότητα των αποτελεσμάτων εξαρτάται άμεσα από το κατά πόσο το δείγμα είναι αντιπροσωπευτικό του συνόλου δεδομένων. Σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν έντονες ανισοκατανομές ή σπάνια γεγονότα, η χρήση δειγμάτων μπορεί να οδηγήσει σε μεγαλύτερα σφάλματα εκτίμησης.

Αντίστοιχα, οι τεχνικές sketch βασίζονται στη διατήρηση συνοπτικών στατιστικών πληροφοριών και είναι ιδιαίτερα αποτελεσματικές σε εφαρμογές συνεχών ροών δεδομένων. Ωστόσο, οι δομές αυτές συνήθως στοχεύουν στην εκτίμηση συγκεκριμένων μεγεθών, όπως συχνότητες εμφάνισης ή αθροίσματα, και όχι στην πλήρη περιγραφή της κατανομής των δεδομένων.

Για τον λόγο αυτό η επιλογή μεταξύ histograms, sampling και sketches εξαρτάται από τον τύπο των ερωτημάτων που πρέπει να υποστηριχθούν. Στην περίπτωση των πολυδιάστατων range queries, τα histograms θεωρούνται συχνά πιο κατάλληλα, καθώς προσφέρουν καλύτερη περιγραφή της κατανομής των τιμών και διευκολύνουν την εκτίμηση selectivity.

2.3.4 Approximate Query Processing σε Περιβάλλοντα Edge Computing

Οι τεχνικές Approximate Query Processing έχουν αποκτήσει ιδιαίτερη σημασία στα περιβάλλοντα edge computing, καθώς επιτρέπουν την παροχή γρήγορων απαντήσεων χωρίς την ανάγκη πλήρους επεξεργασίας των δεδομένων. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη όταν οι διαθέσιμοι πόροι είναι περιορισμένοι ή όταν απαιτείται απόκριση σε πραγματικό χρόνο.

Η χρήση συνοπτικών δομών δεδομένων επιτρέπει στους edge κόμβους να διατηρούν μικρό όγκο πληροφορίας σχετικά με την κατανομή των δεδομένων τους. Με τον τρόπο αυτό μπορούν να εκτιμήσουν τα αποτελέσματα νέων ερωτημάτων με σημαντικά μικρότερο υπολογιστικό κόστος σε σχέση με τις ακριβείς μεθόδους.

Η εφαρμογή του Approximate Query Processing σε περιβάλλοντα edge computing θεωρείται ιδιαίτερα σημαντική για εφαρμογές που απαιτούν χαμηλή καθυστέρηση και υψηλή διαθεσιμότητα, όπως τα συστήματα παρακολούθησης αισθητήρων, οι έξυπνες πόλεις και οι βιομηχανικές εφαρμογές πραγματικού χρόνου.

Σε περιβάλλοντα Edge Computing το Approximate Query Processing δεν αποτελεί απλώς μία τεχνική βελτιστοποίησης αλλά συχνά μία αναγκαιότητα. Οι περιορισμένοι πόροι των edge κόμβων καθιστούν δύσκολη την εκτέλεση σύνθετων αναλυτικών εργασιών με απόλυτη ακρίβεια, ιδιαίτερα όταν απαιτείται ταυτόχρονη εξυπηρέτηση μεγάλου αριθμού ερωτημάτων. Η χρήση συνοπτικών δομών επιτρέπει τη σημαντική μείωση των απαιτήσεων σε μνήμη, επεξεργαστική ισχύ και δικτυακή επικοινωνία.

Επιπλέον, αρκετές ερευνητικές εργασίες έχουν δείξει ότι η συνδυαστική αξιοποίηση Approximate Query Processing και μηχανισμών ανάλυσης ιστορικών ερωτημάτων μπορεί να οδηγήσει σε ακόμη μεγαλύτερη βελτίωση της απόδοσης. Μέσω της αναγνώρισης επαναλαμβανόμενων προτύπων πρόσβασης, το σύστημα μπορεί να προετοιμάζει εκτιμήσεις ή να υπολογίζει προκαταβολικά αποτελέσματα για ερωτήματα που είναι πιθανό να εμφανιστούν στο μέλλον.

Η προσέγγιση αυτή συνδέεται άμεσα με την έννοια του Proactive Data Analysis, όπου η αξιοποίηση ιστορικής γνώσης χρησιμοποιείται για τη βελτίωση της μελλοντικής απόδοσης του συστήματος. Η παρούσα εργασία ακολουθεί αυτή τη φιλοσοφία συνδυάζοντας histograms και ανάλυση ιστορικών queries με στόχο τη μείωση του χρόνου απόκρισης των range queries σε περιβάλλοντα Edge Computing.

Παράλληλα, η χρήση προσεγγιστικών τεχνικών συμβάλλει στη βελτίωση της επεκτασιμότητας των edge συστημάτων. Καθώς αυξάνεται ο αριθμός των κόμβων και ο όγκος των δεδομένων, η δυνατότητα παροχής γρήγορων εκτιμήσεων αντί πλήρων υπολογισμών επιτρέπει τη διατήρηση αποδεκτών επιπέδων απόδοσης χωρίς αντίστοιχη αύξηση του υπολογιστικού κόστους.

Παρότι οι τεχνικές Approximate Query Processing μπορούν να μειώσουν σημαντικά το κόστος εκτέλεσης των ερωτημάτων, η απόδοσή τους μπορεί να ενισχυθεί περαιτέρω μέσω της αξιοποίησης της γνώσης που προκύπτει από προηγούμενες αλληλεπιδράσεις των χρηστών με το σύστημα. Η μελέτη ιστορικών ερωτημάτων και η αναγνώριση επαναλαμβανόμενων

προτύπων αποτελούν βασική κατεύθυνση της σύγχρονης έρευνας και συνδέονται άμεσα με τις τεχνικές clustering που παρουσιάζονται στην επόμενη ενότητα.

2.4 Clustering Techniques

2.4.1 Clustering Ερωτημάτων

Η ομαδοποίηση (clustering) αποτελεί βασική τεχνική εξόρυξης δεδομένων και χρησιμοποιείται ευρέως για την αναγνώριση προτύπων σε μεγάλα σύνολα δεδομένων. Οι Jain et al. (1999) και Han et al. (2011) παρουσιάζουν τις βασικές αρχές των αλγορίθμων clustering και τις εφαρμογές τους στην ανακάλυψη γνώσης.

Στην περίπτωση των ιστορικών queries, η ομαδοποίηση επιτρέπει τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης στα δεδομένα και τη δημιουργία μοντέλων πρόβλεψης μελλοντικής συμπεριφοράς.

Η ομαδοποίηση ερωτημάτων μπορεί να αποκαλύψει χρήσιμα πρότυπα χρήσης ενός συστήματος. Ερωτήματα που παρουσιάζουν παρόμοια χαρακτηριστικά συχνά αντιστοιχούν σε κοινές ανάγκες των χρηστών ή σε συγκεκριμένες περιοχές ενδιαφέροντος μέσα στον χώρο δεδομένων.

Η πληροφορία αυτή μπορεί να αξιοποιηθεί για την ανάπτυξη μηχανισμών προδραστικής ανάλυσης, όπου το σύστημα προετοιμάζει εκ των προτέρων αποτελέσματα για συχνά επαναλαμβανόμενα μοτίβα ερωτημάτων. Με τον τρόπο αυτό μειώνεται ο χρόνος απόκρισης και βελτιώνεται η αποδοτικότητα της επεξεργασίας.

Η εφαρμογή τεχνικών clustering σε ιστορικά ερωτήματα παρουσιάζει ιδιαίτερο ενδιαφέρον στα συστήματα αναλυτικής επεξεργασίας δεδομένων. Αντί να εξετάζονται τα queries ως ανεξάρτητες οντότητες, αντιμετωπίζονται ως σημεία σε έναν πολυδιάστατο χώρο χαρακτηριστικών, όπου η θέση κάθε σημείου καθορίζεται από τα όρια και τα χαρακτηριστικά του αντίστοιχου ερωτήματος.

Η αναπαράσταση αυτή επιτρέπει την αναγνώριση περιοχών του χώρου ερωτημάτων που παρουσιάζουν αυξημένη συγκέντρωση δραστηριότητας. Οι περιοχές αυτές συχνά αντιστοιχούν σε συγκεκριμένες ανάγκες των χρηστών ή σε δεδομένα που παρουσιάζουν αυξημένο ενδιαφέρον για ανάλυση. Ως αποτέλεσμα, το clustering μπορεί να λειτουργήσει ως μηχανισμός ανακάλυψης γνώσης σχετικά με τη συμπεριφορά χρήσης ενός συστήματος.

Η πληροφορία που προκύπτει από τη διαδικασία ομαδοποίησης μπορεί να αξιοποιηθεί όχι μόνο για την κατανόηση της συμπεριφοράς των χρηστών αλλά και για τη βελτιστοποίηση της απόδοσης του συστήματος. Μέσω της αναγνώρισης επαναλαμβανόμενων προτύπων καθίσταται δυνατή η προετοιμασία αποτελεσμάτων ή η δημιουργία συνοπτικών αναπαραστάσεων για περιοχές που αναμένεται να ζητηθούν ξανά στο μέλλον.

2.4.2 Subtractive Clustering

Οι Chiu (1994) και Yager & Filev (1994) πρότειναν το subtractive clustering, μία τεχνική που βασίζεται στην πυκνότητα των δεδομένων και δεν απαιτεί προκαθορισμένο αριθμό

clusters. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη όταν δεν είναι γνωστή εκ των προτέρων η δομή των δεδομένων.

Σε αντίθεση με αλγορίθμους όπως ο K-Means, το subtractive clustering δεν απαιτεί τον εκ των προτέρων καθορισμό του αριθμού των clusters. Τα κέντρα των clusters επιλέγονται με βάση την πυκνότητα των δεδομένων, γεγονός που επιτρέπει την προσαρμογή της μεθόδου σε διαφορετικές κατανομές.

Το χαρακτηριστικό αυτό είναι ιδιαίτερα χρήσιμο σε περιβάλλοντα όπου η δομή των δεδομένων δεν είναι γνωστή εκ των προτέρων, όπως συμβαίνει συχνά στα ιστορικά workloads ερωτημάτων. Η δυνατότητα αυτόματου εντοπισμού ομάδων καθιστά τη μέθοδο κατάλληλη για εφαρμογές προδραστικής ανάλυσης.

Η βασική ιδέα του subtractive clustering είναι ότι κάθε σημείο του συνόλου δεδομένων θεωρείται υποψήφιο κέντρο cluster. Για κάθε σημείο υπολογίζεται μία τιμή δυναμικού (potential), η οποία εκφράζει την πυκνότητα των γειτονικών δεδομένων. Τα σημεία με το υψηλότερο δυναμικό επιλέγονται ως κέντρα clusters, ενώ η επιρροή τους αφαιρείται σταδιακά από τον χώρο αναζήτησης ώστε να εντοπιστούν οι επόμενες σημαντικές ομάδες.

Η προσέγγιση αυτή επιτρέπει την αυτόματη προσαρμογή της διαδικασίας ομαδοποίησης στα χαρακτηριστικά των δεδομένων χωρίς την ανάγκη εκτεταμένης παραμετροποίησης. Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό σε εφαρμογές όπου η δομή των δεδομένων μεταβάλλεται διαρκώς ή δεν είναι γνωστή εκ των προτέρων.

Επιπλέον, το subtractive clustering έχει χρησιμοποιηθεί σε πληθώρα εφαρμογών εξόρυξης δεδομένων, αναγνώρισης προτύπων και ευφυών συστημάτων λήψης αποφάσεων. Η ευελιξία και η ικανότητά του να λειτουργεί αποτελεσματικά σε διαφορετικές κατανομές δεδομένων αποτελούν βασικούς λόγους για τους οποίους επιλέχθηκε και στην παρούσα εργασία.

2.4.3 Σύγκριση Αλγορίθμων Clustering

Οι αλγόριθμοι clustering παρουσιάζουν διαφορετικά χαρακτηριστικά ως προς την ακρίβεια, την πολυπλοκότητα και τις απαιτήσεις παραμετροποίησης. Ο K-Means αποτελεί μία από τις πιο διαδεδομένες μεθόδους, απαιτεί όμως τον προκαθορισμό του αριθμού των clusters και παρουσιάζει ευαισθησία στην αρχική επιλογή των κέντρων.

Ο DBSCAN βασίζεται στην πυκνότητα των δεδομένων και μπορεί να εντοπίσει θόρυβο και ακανόνιστα σχήματα clusters. Ωστόσο, η απόδοσή του επηρεάζεται σημαντικά από την επιλογή των παραμέτρων πυκνότητας και από τη διάσταση των δεδομένων.

Το Subtractive Clustering παρουσιάζει το πλεονέκτημα ότι δεν απαιτεί προκαθορισμένο αριθμό clusters και μπορεί να προσαρμοστεί δυναμικά στη δομή των δεδομένων. Για τον λόγο αυτό επιλέχθηκε στην παρούσα εργασία, καθώς τα ιστορικά ερωτήματα δεν παρουσιάζουν γνωστή εκ των προτέρων ομαδοποίηση και απαιτείται ευελιξία κατά τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης.

Η επιλογή της κατάλληλης τεχνικής clustering εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων και τους στόχους της εφαρμογής. Σε περιπτώσεις όπου ο αριθμός των αναμενόμενων ομάδων είναι γνωστός εκ των προτέρων, αλγόριθμοι όπως ο K-Means μπορούν να προσφέρουν ιδιαίτερα αποδοτικά αποτελέσματα. Ωστόσο, η ανάγκη

προκαθορισμού του αριθμού των clusters περιορίζει την ευελιξία τους σε δυναμικά περιβάλλοντα.

Οι αλγόριθμοι που βασίζονται στην πυκνότητα, όπως ο DBSCAN, παρουσιάζουν σημαντικά πλεονεκτήματα ως προς τον εντοπισμό θορύβου και την αναγνώριση ακανόνιστων σχημάτων clusters. Παρ' όλα αυτά, η αποτελεσματικότητά τους επηρεάζεται από την επιλογή των παραμέτρων πυκνότητας και συχνά μειώνεται όταν αυξάνεται η διάσταση των δεδομένων.

Στην περίπτωση των ιστορικών ερωτημάτων, η δομή των δεδομένων είναι συνήθως άγνωστη και μεταβάλλεται δυναμικά καθώς παράγονται νέα queries. Για τον λόγο αυτό απαιτείται μία μέθοδος που να μπορεί να προσαρμόζεται αυτόματα στις μεταβολές του workload και να εντοπίζει επαναλαμβανόμενα πρότυπα χωρίς εκτεταμένη παραμετροποίηση. Το Subtractive Clustering ανταποκρίνεται αποτελεσματικά στις απαιτήσεις αυτές και προσφέρει μία ισορροπία μεταξύ ακρίβειας, ευελιξίας και υπολογιστικού κόστους.

Η επιλογή του συγκεκριμένου αλγορίθμου στην παρούσα εργασία δεν βασίζεται μόνο στη θεωρητική του υπεροχή αλλά και στη συμβατότητά του με τον στόχο της προδραστικής ανάλυσης δεδομένων. Η δυνατότητα αυτόματου εντοπισμού σημαντικών προτύπων πρόσβασης διευκολύνει την επιλογή των queries που παρουσιάζουν μεγαλύτερο ενδιαφέρον για μελλοντική αξιοποίηση και προετοιμασία αποτελεσμάτων.

Η αναγνώριση επαναλαμβανόμενων προτύπων μέσω τεχνικών clustering αποτελεί βασικό βήμα για την ανάπτυξη μηχανισμών προδραστικής ανάλυσης δεδομένων. Η αξιοποίηση της γνώσης που προκύπτει από τα ιστορικά ερωτήματα επιτρέπει στο σύστημα να προβλέπει πιθανές μελλοντικές ανάγκες των χρηστών και να προετοιμάζει αναλύσεις πριν ακόμη ζητηθούν. Η φιλοσοφία αυτή παρουσιάζεται αναλυτικότερα στην επόμενη ενότητα, η οποία εξετάζει τις αρχές και τις τεχνικές του Proactive Data Analysis.

2.5 Historical Queries και Proactive Data Analysis

2.5.1 Ανάλυση Ιστορικών Ερωτημάτων

Η ανάλυση ιστορικών ερωτημάτων αποτελεί σημαντικό πεδίο έρευνας στα συστήματα βάσεων δεδομένων και στα συστήματα ανάλυσης δεδομένων. Μέσω της καταγραφής και μελέτης των queries που έχουν εκτελεστεί στο παρελθόν, είναι δυνατή η εξαγωγή πληροφοριών σχετικά με τη συμπεριφορά των χρηστών και τα πρότυπα πρόσβασης στα δεδομένα.

Τα ιστορικά ερωτήματα αποθηκεύονται συνήθως σε query logs, τα οποία περιλαμβάνουν πληροφορίες σχετικά με τα χαρακτηριστικά των ερωτημάτων, τη συχνότητα εμφάνισής τους και τα αποτελέσματα που παρήγαγαν. Η ανάλυση των logs αυτών μπορεί να συμβάλει στη βελτιστοποίηση της εκτέλεσης μελλοντικών ερωτημάτων και στη λήψη αποφάσεων σχετικά με τη διαχείριση των δεδομένων.

Η ανάλυση ιστορικών ερωτημάτων επιτρέπει στα πληροφοριακά συστήματα να αποκτήσουν γνώση σχετικά με τον τρόπο με τον οποίο οι χρήστες αλληλεπιδρούν με τα δεδομένα. Μέσω της μελέτης των προηγούμενων αναζητήσεων μπορούν να εντοπιστούν περιοχές του χώρου δεδομένων που παρουσιάζουν αυξημένο ενδιαφέρον, καθώς και μοτίβα πρόσβασης που επαναλαμβάνονται συστηματικά με την πάροδο του χρόνου.

Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση της απόδοσης του συστήματος με διάφορους τρόπους. Για παράδειγμα, είναι δυνατό να εντοπιστούν δημοφιλή ερωτήματα, να δημιουργηθούν μηχανισμοί προσωρινής αποθήκευσης για συχνά χρησιμοποιούμενα δεδομένα ή να προγραμματιστούν εκ των προτέρων αναλυτικές διαδικασίες για περιοχές που αναμένεται να ζητηθούν ξανά στο μέλλον.

Η σημασία της ανάλυσης ιστορικών ερωτημάτων έχει αυξηθεί ιδιαίτερα τα τελευταία χρόνια λόγω της ανάπτυξης εφαρμογών μεγάλων δεδομένων και συστημάτων πραγματικού χρόνου. Σε τέτοια περιβάλλοντα, η αξιοποίηση της γνώσης που προκύπτει από προηγούμενες εκτελέσεις ερωτημάτων μπορεί να συμβάλει σημαντικά στη μείωση του υπολογιστικού κόστους και στη βελτίωση της συνολικής απόδοσης.

2.5.2 Query Prediction

Η πρόβλεψη μελλοντικών ερωτημάτων βασίζεται στην παρατήρηση ότι πολλοί χρήστες ακολουθούν επαναλαμβανόμενα πρότυπα αναζήτησης. Μέσω της ανάλυσης ιστορικών δεδομένων είναι δυνατό να εκτιμηθεί ποια ερωτήματα είναι πιθανό να εμφανιστούν στο μέλλον.

Οι τεχνικές πρόβλεψης αξιοποιούν στατιστικές μεθόδους, αλγορίθμους μηχανικής μάθησης και τεχνικές εξόρυξης δεδομένων για την αναγνώριση επαναλαμβανόμενων μοτίβων. Η δυνατότητα πρόβλεψης μελλοντικών ερωτημάτων επιτρέπει στα συστήματα να προετοιμάζουν εκ των προτέρων πληροφορίες και να βελτιώνουν σημαντικά τον χρόνο απόκρισης.

Η πρόβλεψη ερωτημάτων αποτελεί μία από τις σημαντικότερες εφαρμογές της ανάλυσης ιστορικών δεδομένων χρήσης. Η βασική υπόθεση είναι ότι η συμπεριφορά των χρηστών παρουσιάζει έναν βαθμό κανονικότητας και ότι παρόμοια μοτίβα αναζήτησης τείνουν να επαναλαμβάνονται με την πάροδο του χρόνου. Ως αποτέλεσμα, η μελέτη προηγούμενων ερωτημάτων μπορεί να χρησιμοποιηθεί για την εκτίμηση μελλοντικών αναγκών.

Στη βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις πρόβλεψης, όπως μοντέλα βασισμένα σε στατιστικές συχνότητες, αλγόριθμοι μηχανικής μάθησης και τεχνικές αναγνώρισης ακολουθιών ερωτημάτων. Παρότι οι μέθοδοι αυτές διαφέρουν ως προς την πολυπλοκότητα και την ακρίβεια, όλες επιδιώκουν τον ίδιο στόχο: τη βελτίωση της απόδοσης μέσω της έγκαιρης προετοιμασίας πληροφοριών που είναι πιθανό να ζητηθούν στο μέλλον.

Η δυνατότητα πρόβλεψης αποκτά ιδιαίτερη σημασία σε περιβάλλοντα Edge Computing, όπου οι διαθέσιμοι πόροι είναι περιορισμένοι και η αποφυγή περιττών υπολογισμών μπορεί να οδηγήσει σε σημαντική εξοικονόμηση πόρων. Η πρόβλεψη μελλοντικών ερωτημάτων επιτρέπει στο σύστημα να κατανέμει αποτελεσματικότερα τους διαθέσιμους πόρους και να μειώνει τους χρόνους απόκρισης.

2.5.3 Proactive Data Analysis

Το Proactive Data Analysis αποτελεί μία προσέγγιση κατά την οποία το σύστημα δεν περιορίζεται στην παθητική εκτέλεση ερωτημάτων, αλλά αξιοποιεί διαθέσιμες πληροφορίες

για να προετοιμάζεται για μελλοντικές ανάγκες. Η προσέγγιση αυτή βασίζεται στην υπόθεση ότι μέρος της μελλοντικής συμπεριφοράς μπορεί να προβλεφθεί από τα ιστορικά δεδομένα.

Τεχνικές όπως το prefetching, το caching και η προϋπολογισμένη επεξεργασία αποτελεσμάτων (precomputation) χρησιμοποιούνται για τη μείωση του χρόνου απόκρισης και του υπολογιστικού κόστους. Η εφαρμογή τέτοιων τεχνικών είναι ιδιαίτερα χρήσιμη σε περιβάλλοντα edge computing, όπου οι διαθέσιμοι πόροι είναι περιορισμένοι και η ταχύτητα απόκρισης αποτελεί κρίσιμο παράγοντα.

Η φιλοσοφία του Proactive Data Analysis διαφοροποιείται σημαντικά από την παραδοσιακή προσέγγιση των συστημάτων βάσεων δεδομένων. Στα συμβατικά συστήματα η επεξεργασία ξεκινά μόνο μετά την υποβολή ενός ερωτήματος από τον χρήστη. Αντίθετα, στα προδραστικά συστήματα επιχειρείται η αξιοποίηση διαθέσιμης γνώσης προκειμένου να προετοιμαστούν πιθανά αποτελέσματα πριν ακόμη ζητηθούν.

Η προσέγγιση αυτή βασίζεται στην παρατήρηση ότι μεγάλο ποσοστό των ερωτημάτων που υποβάλλονται σε ένα σύστημα παρουσιάζει επαναλαμβανόμενα χαρακτηριστικά. Μέσω της ανάλυσης ιστορικών δεδομένων και της αναγνώρισης προτύπων πρόσβασης καθίσταται δυνατή η πρόβλεψη περιοχών ενδιαφέροντος και η προετοιμασία αναλύσεων που είναι πιθανό να αξιοποιηθούν στο άμεσο μέλλον.

Τα τελευταία χρόνια το Proactive Data Analysis έχει προσελκύσει αυξανόμενο ερευνητικό ενδιαφέρον, ιδιαίτερα σε εφαρμογές μεγάλων δεδομένων και κατανεμημένων υποδομών. Η δυνατότητα μείωσης του χρόνου απόκρισης χωρίς αντίστοιχη αύξηση των διαθέσιμων πόρων το καθιστά ιδιαίτερα ελκυστικό για περιβάλλοντα Edge Computing και εφαρμογές πραγματικού χρόνου.

2.5.4 Σχέση με την Παρούσα Εργασία

Η παρούσα εργασία συνδυάζει στοιχεία από τα πεδία του Edge Computing, του Approximate Query Processing και της ανάλυσης ιστορικών ερωτημάτων. Σε αντίθεση με προσεγγίσεις που επικεντρώνονται αποκλειστικά στη βελτιστοποίηση της εκτέλεσης ερωτημάτων, η προτεινόμενη μέθοδος αξιοποιεί ιστορικά range queries για την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης.

Για τον σκοπό αυτό χρησιμοποιούνται συνοπτικές αναπαραστάσεις δεδομένων μέσω histograms και τεχνικές subtractive clustering για την ομαδοποίηση παρόμοιων ερωτημάτων. Ο συνδυασμός των τεχνικών αυτών επιτρέπει την αξιολόγηση της δυνατότητας εφαρμογής μηχανισμών Proactive Data Analysis σε περιβάλλοντα edge computing και αποτελεί το βασικό αντικείμενο της παρούσας εργασίας.

Σε αντίθεση με πολλές υπάρχουσες προσεγγίσεις που βασίζονται αποκλειστικά είτε σε τεχνικές Approximate Query Processing είτε σε μηχανισμούς πρόβλεψης ερωτημάτων, η παρούσα εργασία διερευνά τον συνδυασμό των δύο κατευθύνσεων. Η χρήση histograms επιτρέπει τη δημιουργία συνοπτικών αναπαραστάσεων των δεδομένων, ενώ η ανάλυση ιστορικών ερωτημάτων μέσω subtractive clustering επιτρέπει την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης.

Η συνδυαστική αξιοποίηση των δύο τεχνικών στοχεύει στη μείωση του χρόνου εκτέλεσης των range queries χωρίς την ανάγκη πλήρους επεξεργασίας των δεδομένων για κάθε νέο

ερώτημα. Παράλληλα, επιτρέπει τη διερεύνηση του βαθμού στον οποίο η γνώση που προκύπτει από προηγούμενα ερωτήματα μπορεί να αξιοποιηθεί για την υποστήριξη μηχανισμών προδραστικής ανάλυσης.

Το προτεινόμενο μοντέλο δεν επιδιώκει να αντικαταστήσει τις ακριβείς μεθόδους επεξεργασίας, αλλά να αξιολογήσει κατά πόσο η αξιοποίηση summaries και ιστορικών queries μπορεί να προσφέρει ικανοποιητική ισορροπία μεταξύ ακρίβειας, χρόνου απόκρισης και υπολογιστικού κόστους. Η αξιολόγηση αυτής της σχέσης αποτελεί τον βασικό ερευνητικό στόχο της παρούσας εργασίας.

Η ανάλυση της σχετικής βιβλιογραφίας δείχνει ότι τα πεδία του Edge Computing, του Approximate Query Processing, του Clustering και του Proactive Data Analysis παρουσιάζουν ισχυρή αλληλεξάρτηση. Η συνδυαστική αξιοποίησή τους δημιουργεί νέες δυνατότητες για την ανάπτυξη αποδοτικότερων συστημάτων διαχείρισης δεδομένων. Στην επόμενη ενότητα παρουσιάζεται συνοπτικά η σύνθεση των βασικών συμπερασμάτων της βιβλιογραφικής επισκόπησης και η σύνδεσή τους με την προτεινόμενη μεθοδολογία της παρούσας εργασίας.

2.6 Σύνοψη Βιβλιογραφικής Έρευνας

Η βιβλιογραφική επισκόπηση ανέδειξε τη σημασία του Edge Computing ως ενός κατανεμημένου υπολογιστικού μοντέλου που μεταφέρει μέρος της επεξεργασίας κοντά στην πηγή παραγωγής των δεδομένων. Η προσέγγιση αυτή συμβάλλει στη μείωση της καθυστέρησης, στη βελτίωση της απόδοσης και στην αποτελεσματικότερη αξιοποίηση των διαθέσιμων δικτυακών πόρων. Παράλληλα, η βιβλιογραφία κατέδειξε ότι η διαχείριση μεγάλου όγκου δεδομένων σε περιβάλλοντα edge συνοδεύεται από σημαντικές προκλήσεις, όπως οι περιορισμένοι υπολογιστικοί πόροι, οι απαιτήσεις αποθήκευσης και η ανάγκη για γρήγορη επεξεργασία ερωτημάτων.

Στη συνέχεια εξετάστηκαν τα range queries, τα οποία αποτελούν βασικό μηχανισμό αναζήτησης σε πολυδιάστατα σύνολα δεδομένων. Παρουσιάστηκαν οι κυριότερες δομές ευρετηρίων που χρησιμοποιούνται για την επιτάχυνση της επεξεργασίας τους, καθώς και οι δυσκολίες που εμφανίζονται όταν ο όγκος των δεδομένων και ο αριθμός των ερωτημάτων αυξάνονται σημαντικά. Ιδιαίτερη έμφαση δόθηκε στις απαιτήσεις που προκύπτουν όταν τα range queries εκτελούνται σε περιβάλλοντα Edge Computing, όπου η ανάγκη για χαμηλή καθυστέρηση και περιορισμένη κατανάλωση πόρων καθιστά αναγκαία τη χρήση πιο αποδοτικών τεχνικών.

Η ανάλυση της βιβλιογραφίας σχετικά με το Approximate Query Processing έδειξε ότι οι προσεγγιστικές μέθοδοι μπορούν να προσφέρουν σημαντική μείωση του χρόνου εκτέλεσης με αποδεκτό επίπεδο σφάλματος. Μεταξύ των διαθέσιμων τεχνικών, τα histograms αποτελούν μία από τις πιο διαδεδομένες και αποτελεσματικές λύσεις για την εκτίμηση αποτελεσμάτων range queries, καθώς παρέχουν καλή ισορροπία μεταξύ ακρίβειας, απαιτήσεων μνήμης και υπολογιστικού κόστους. Για τον λόγο αυτό επιλέχθηκαν ως η βασική μορφή summaries που χρησιμοποιείται στην παρούσα εργασία.

Παράλληλα, η βιβλιογραφική ανασκόπηση ανέδειξε τη σημασία της ανάλυσης ιστορικών ερωτημάτων και των τεχνικών clustering για την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης στα δεδομένα. Η δυνατότητα ομαδοποίησης παρόμοιων queries επιτρέπει την εξαγωγή γνώσης σχετικά με τη συμπεριφορά των χρηστών και δημιουργεί τις προϋποθέσεις για την ανάπτυξη μηχανισμών πρόβλεψης και προδραστικής ανάλυσης. Μεταξύ των

διαθέσιμων αλγορίθμων clustering, το Subtractive Clustering παρουσιάζει ιδιαίτερο ενδιαφέρον καθώς δεν απαιτεί προκαθορισμένο αριθμό clusters και μπορεί να προσαρμόζεται δυναμικά στη δομή των δεδομένων.

Η μελέτη του πεδίου του Proactive Data Analysis έδειξε ότι η αξιοποίηση ιστορικής πληροφορίας μπορεί να μετατρέψει ένα σύστημα από παθητικό μηχανισμό απάντησης σε ενεργό μηχανισμό πρόβλεψης και προετοιμασίας αποτελεσμάτων. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική σε περιβάλλοντα Edge Computing, όπου η έγκαιρη αξιοποίηση των διαθέσιμων πόρων μπορεί να βελτιώσει σημαντικά τους χρόνους απόκρισης και τη συνολική απόδοση του συστήματος.

Παρότι η διεθνής βιβλιογραφία περιλαμβάνει μεγάλο αριθμό εργασιών για το Edge Computing, το Approximate Query Processing, τα range queries και τις τεχνικές clustering, παρατηρείται περιορισμένος αριθμός ερευνητικών προσεγγίσεων που συνδυάζουν ταυτόχρονα τα παραπάνω πεδία στο πλαίσιο της προδραστικής ανάλυσης δεδομένων. Ειδικότερα, η αξιοποίηση ιστορικών range queries σε συνδυασμό με summaries και clustering για την υποστήριξη μηχανισμών Proactive Data Analysis σε περιβάλλοντα Edge Computing παραμένει ένα πεδίο με σημαντικά περιθώρια διερεύνησης.

Το κενό αυτό επιχειρεί να καλύψει η παρούσα εργασία μέσω της ανάπτυξης ενός προσομοιωμένου περιβάλλοντος edge nodes, στο οποίο εφαρμόζονται τεχνικές Approximate Query Processing με χρήση histograms και τεχνικές ανάλυσης ιστορικών ερωτημάτων μέσω subtractive clustering. Η προσέγγιση αυτή επιτρέπει τη μελέτη της σχέσης μεταξύ ακρίβειας, χρόνου απόκρισης και υπολογιστικού κόστους, καθώς και την αξιολόγηση της δυνατότητας υποστήριξης μηχανισμών Proactive Data Analysis σε κατανεμημένα περιβάλλοντα Edge Computing.

Με βάση τα συμπεράσματα της βιβλιογραφικής επισκόπησης, στο επόμενο κεφάλαιο παρουσιάζεται αναλυτικά η αρχιτεκτονική και η μεθοδολογία του προτεινόμενου μοντέλου, καθώς και ο τρόπος με τον οποίο οι παραπάνω τεχνικές συνδυάζονται για την υλοποίηση και αξιολόγηση της προτεινόμενης προσέγγισης.

ΚΕΦΑΛΑΙΟ 3 Περιγραφή Αλγορίθμου

Σκοπός του αλγορίθμου είναι η προσομοίωση ενός περιβάλλοντος edge computing στο οποίο καταγράφονται ιστορικά range queries. Μέσω της ανάλυσης των ιστορικών αυτών ερωτημάτων επιδιώκεται η εξαγωγή επαναλαμβανόμενων μοτίβων και η επιλογή υποψήφιων περιοχών του χώρου αναζήτησης για προδραστική επεξεργασία. Η προσέγγιση αυτή αποτελεί τη βάση για την υλοποίηση μηχανισμών Proactive Data Analysis πάνω σε ιστορικά workloads.

3.1 Φόρτωση και Προετοιμασία Δεδομένων

3.1.1 Φόρτωση του Dataset

Η διαδικασία ξεκινά με τη φόρτωση του συνόλου δεδομένων Iris μέσω της συνάρτησης load_dataset(). Αρχικά ελέγχεται αν το dataset υπάρχει σε τοπική cache. Εάν δεν υπάρχει, πραγματοποιείται λήψη από το αποθετήριο UCI Machine Learning Repository και αποθηκεύεται τοπικά για μελλοντική χρήση. Από το dataset επιλέγονται μόνο τα τέσσερα αριθμητικά χαρακτηριστικά: sepal length, sepal width, petal length και petal width.

3.1.2 Στατιστική Περιγραφή Χαρακτηριστικών

Η συνάρτηση summarize_features() υπολογίζει βασικά στατιστικά μεγέθη για κάθε χαρακτηριστικό του dataset, όπως η μέση τιμή, η τυπική απόκλιση, η ελάχιστη και η μέγιστη τιμή. Τα στατιστικά αυτά χρησιμοποιούνται στη συνέχεια για τη δημιουργία συνθετικών δεδομένων και ερωτημάτων.

3.2 Δημιουργία Συνθετικών Δεδομένων και Edge Nodes

3.2.1 Δημιουργία Συνθετικών Δεδομένων

Η παραγωγή συνθετικών δεδομένων πραγματοποιείται μέσω των συναρτήσεων make_synthetic_data() και make_clustered_data(). Στην πρώτη περίπτωση τα δεδομένα παράγονται με κανονική κατανομή χρησιμοποιώντας τις στατιστικές παραμέτρους του αρχικού dataset. Στη δεύτερη περίπτωση δημιουργούνται πολλαπλά κέντρα συγκέντρωσης δεδομένων (clusters), ώστε να προσομοιωθούν πιο ρεαλιστικά και μη ομοιόμορφα workloads. Οι παραγόμενες τιμές περιορίζονται εντός των ελάχιστων και μέγιστων ορίων του dataset ώστε να διατηρείται η ρεαλιστικότητα των δεδομένων.

3.2.2 Δημιουργία Edge Nodes

Κάθε κόμβος αναπαρίσταται από την κλάση EdgeNode, η οποία αποθηκεύει τα τοπικά δεδομένα, τα ερωτήματα που του ανατίθενται και τα αποτελέσματα των εκτελέσεων των ερωτημάτων. Με τον τρόπο αυτό προσομοιώνεται η τοπική επεξεργασία δεδομένων σε περιβάλλον edge computing.

3.3 Δημιουργία και Αξιολόγηση Ερωτημάτων

3.3.1 Δημιουργία Range Queries

Η συνάρτηση make_queries() δημιουργεί ερωτήματα εύρους τιμών (range queries) για κάθε διάσταση των δεδομένων. Το εύρος κάθε διαστήματος επιλέγεται τυχαία ως ποσοστό του συνολικού εύρους της αντίστοιχης διάστασης. Επιπλέον, η συνάρτηση make_biased_queries() δημιουργεί ερωτήματα που εστιάζουν σε περιοχές μεγαλύτερης πυκνότητας δεδομένων μέσω της χρήσης ιστογραμμάτων.

3.3.2 Αξιολόγηση Ερωτημάτων

Η αξιολόγηση των ερωτημάτων πραγματοποιείται από τη μέθοδο evaluate_queries() της κλάσης EdgeNode. Για κάθε ερώτημα εφαρμόζονται λογικές συνθήκες σε όλες τις διαστάσεις των δεδομένων και υπολογίζεται ο αριθμός των εγγραφών που ικανοποιούν τα αντίστοιχα όρια. Το πλήθος αυτό αποθηκεύεται ως αριθμός επιτυχιών (hits) του ερωτήματος.

3.4 Προσεγγιστική Ανάλυση Δεδομένων

3.4.1 Κατασκευή Summaries

Για κάθε κόμβο δημιουργούνται συνοπτικές αναπαραστάσεις δεδομένων μέσω της συνάρτησης build_data_summary(). Οι αναπαραστάσεις αυτές βασίζονται σε ιστογράμματα και αποθηκεύουν πληροφορίες σχετικά με την κατανομή των τιμών κάθε χαρακτηριστικού.

3.4.2 Εκτίμηση Αποτελεσμάτων Ερωτημάτων

Η συνάρτηση estimate_query_count() χρησιμοποιεί τα ιστογράμματα για να εκτιμήσει κατά προσέγγιση τον αριθμό των εγγραφών που αναμένεται να ικανοποιούν ένα ερώτημα. Η προσέγγιση αυτή μειώνει το υπολογιστικό κόστος σε σχέση με την πλήρη αξιολόγηση των δεδομένων, με αντάλλαγμα ένα περιορισμένο σφάλμα εκτίμησης.

3.5 Clustering Ιστορικών Ερωτημάτων

3.5.1 Μετασχηματισμός Ερωτημάτων σε Χαρακτηριστικά

Για την ομαδοποίηση των ερωτημάτων χρησιμοποιείται η συνάρτηση normalize_query_features(), η οποία μετατρέπει κάθε query σε ένα διάνυσμα χαρακτηριστικών βασισμένο στο κέντρο και στο πλάτος των διαστημάτων του.

3.5.2 Subtractive Clustering

Η διαδικασία clustering υλοποιείται μέσω της συνάρτησης subtractive_clustering(). Τα ερωτήματα ομαδοποιούνται με βάση την ομοιότητά τους, ενώ στη συνέχεια χρησιμοποιούνται οι συναρτήσεις assign_cluster_labels() και summarize_query_cluster() για την ανάθεση labels και την περιγραφή των παραγόμενων clusters.

3.5.3 Επιλογή Clusters για Proactive Analysis

Η συνάρτηση select_proactive_clusters() επιλέγει τα clusters που παρουσιάζουν υψηλή ομοιότητα με τα δεδομένα του κόμβου και ταυτόχρονα περιλαμβάνουν ικανοποιητικό αριθμό ερωτημάτων. Τα επιλεγμένα clusters θεωρούνται υποψήφια για προδραστική επεξεργασία.

3.6 Υπολογισμός Δεικτών Απόδοσης

3.6.1 Κάλυψη Δεδομένων

Η συνάρτηση compute_dimension_coverage() υπολογίζει το ποσοστό κάλυψης κάθε διάστασης από τα ερωτήματα που παράγουν αποτελέσματα. Η μετρική αυτή χρησιμοποιείται για την αξιολόγηση της αντιπροσωπευτικότητας των workloads.

3.6.2 Χρόνος Εκτέλεσης και Σφάλμα Προσέγγισης

Οι συναρτήσεις measure_latencies() και compute_approximation_error() χρησιμοποιούνται για τη σύγκριση της ακριβούς και της προσεγγιστικής αξιολόγησης των ερωτημάτων. Συγκεκριμένα υπολογίζονται ο χρόνος εκτέλεσης, το μέσο απόλυτο σφάλμα και το μέσο σχετικό σφάλμα προσέγγισης.

3.7 Συνολική Ροή Εκτέλεσης

Η συνάρτηση main() συντονίζει όλα τα επιμέρους στάδια του αλγορίθμου. Αρχικά φορτώνεται το dataset και υπολογίζονται τα στατιστικά χαρακτηριστικά. Στη συνέχεια δημιουργούνται οι edge nodes και τα ερωτήματα, εκτελείται η αξιολόγησή τους, κατασκευάζονται τα summaries, υπολογίζονται οι δείκτες απόδοσης και εφαρμόζεται η διαδικασία clustering. Τέλος παρουσιάζονται τα αποτελέσματα της προσομοίωσης και τα σημαντικότερα πρότυπα ερωτημάτων που εντοπίστηκαν.

3.8 Σύνοψη Κεφαλαίου

Ο αλγόριθμος σχεδιάστηκε ώστε να δημιουργεί ρεαλιστικά συνθετικά δεδομένα, να παράγει διαφορετικούς τύπους workloads, να συγκρίνει ακριβείς και προσεγγιστικές μεθόδους αξιολόγησης και να αξιοποιεί τεχνικές clustering για την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης. Η αρχιτεκτονική αυτή αποτελεί τη βάση για τη μελέτη μηχανισμών Proactive Data Analysis σε περιβάλλοντα edge computing.

ΚΕΦΑΛΑΙΟ 4 Προτεινόμενο Μοντέλο

Το προτεινόμενο μοντέλο βασίζεται στην προσομοίωση ενός περιβάλλοντος Edge Computing, όπου πολλαπλοί κόμβοι διαχειρίζονται τοπικά δεδομένα και εξυπηρετούν ερωτήματα εύρους τιμών (range queries). Στόχος του μοντέλου είναι η αξιοποίηση ιστορικών ερωτημάτων για την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης και η υποστήριξη μηχανισμών προδραστικής ανάλυσης δεδομένων.

Η ανάγκη για πολυδιάστατη αξιολόγηση είναι ιδιαίτερα έντονη στα συστήματα που βασίζονται σε τεχνικές Approximate Query Processing. Οι τεχνικές αυτές επιδιώκουν τη μείωση του υπολογιστικού κόστους μέσω της παραγωγής εκτιμήσεων αντί ακριβών αποτελεσμάτων. Παρότι η προσέγγιση αυτή μπορεί να βελτιώσει σημαντικά την απόδοση του συστήματος, είναι απαραίτητο να εξεταστεί κατά πόσο το σφάλμα που εισάγεται παραμένει εντός αποδεκτών ορίων. Για τον λόγο αυτό η αξιολόγηση δεν περιορίζεται μόνο στην ταχύτητα εκτέλεσης αλλά επεκτείνεται και στην ποιοτική αξιολόγηση των παραγόμενων αποτελεσμάτων.

Παράλληλα, η αξιολόγηση του προτεινόμενου μοντέλου περιλαμβάνει και την εξέταση της συμπεριφοράς του σε διαφορετικά σενάρια φόρτου εργασίας. Η χρήση διαφορετικών workloads επιτρέπει τη διερεύνηση της ανθεκτικότητας της προσέγγισης σε μεταβολές της κατανομής των δεδομένων και των ερωτημάτων. Με τον τρόπο αυτό μπορεί να εκτιμηθεί κατά πόσο οι μηχανισμοί προδραστικής ανάλυσης παραμένουν αποτελεσματικοί όταν οι συνθήκες λειτουργίας αποκλίνουν από το αρχικό σενάριο σχεδιασμού.

Τέλος, ιδιαίτερη έμφαση δίνεται στην αξιολόγηση των μηχανισμών ανάλυσης ιστορικών ερωτημάτων και clustering. Η ποιότητα των clusters που προκύπτουν επηρεάζει άμεσα την αποτελεσματικότητα της προδραστικής ανάλυσης, καθώς από αυτά εξαρτάται η δυνατότητα αναγνώρισης επαναλαμβανόμενων προτύπων πρόσβασης. Επομένως, η αξιολόγηση των clusters δεν αφορά μόνο την ορθότητα της ομαδοποίησης αλλά και τη χρησιμότητά τους ως βάση για τη λήψη μελλοντικών αποφάσεων από το σύστημα.

4.1 Αρχιτεκτονική του Συστήματος

Η αρχιτεκτονική του προτεινόμενου συστήματος ακολουθεί τη φιλοσοφία της αποκεντρωμένης επεξεργασίας δεδομένων που χαρακτηρίζει τα περιβάλλοντα Edge Computing. Αντί τα δεδομένα να μεταφέρονται σε κεντρικούς εξυπηρετητές για επεξεργασία, κάθε κόμβος αναλαμβάνει την τοπική διαχείριση των πληροφοριών που διαθέτει. Η προσέγγιση αυτή συμβάλλει στη μείωση της δικτυακής κίνησης και επιτρέπει ταχύτερη λήψη αποφάσεων, ιδιαίτερα σε εφαρμογές όπου απαιτείται άμεση απόκριση.

Παράλληλα, η ύπαρξη ανεξάρτητων edge nodes επιτρέπει την εξομοίωση κατανεμημένων υποδομών όπου τα δεδομένα δεν είναι ομοιόμορφα κατανεμημένα. Στην πράξη, κάθε κόμβος μπορεί να παρουσιάζει διαφορετικά χαρακτηριστικά ως προς τη διανομή των δεδομένων και τα πρότυπα πρόσβασης. Η μελέτη αυτής της ετερογένειας αποτελεί βασικό στοιχείο για την αξιολόγηση της αποτελεσματικότητας της προτεινόμενης προσέγγισης.

Η αρχιτεκτονική σχεδιάστηκε έτσι ώστε να υποστηρίζει τόσο την ακριβή όσο και την προσεγγιστική επεξεργασία ερωτημάτων. Με τον τρόπο αυτό καθίσταται δυνατή η άμεση σύγκριση μεταξύ των δύο προσεγγίσεων και η ποσοτική αξιολόγηση των πλεονεκτημάτων που προσφέρει η χρήση συνοπτικών αναπαραστάσεων και μηχανισμών προδραστικής ανάλυσης.

4.1.1 Edge Nodes και Τοπική Επεξεργασία

Το προτεινόμενο μοντέλο βασίζεται σε μία κατανεμημένη αρχιτεκτονική Edge Computing, στην οποία πολλαπλοί edge nodes αναλαμβάνουν την τοπική διαχείριση και επεξεργασία δεδομένων. Η επιλογή της τοπικής επεξεργασίας ακολουθεί τη φιλοσοφία του Edge Computing, σύμφωνα με την οποία η ανάλυση των δεδομένων πραγματοποιείται όσο το δυνατόν πιο κοντά στην πηγή παραγωγής τους, μειώνοντας την ανάγκη συνεχούς επικοινωνίας με απομακρυσμένα cloud datacenters (Shi et al., 2016; Satyanarayanan, 2017).

Κάθε edge node διαθέτει το δικό του σύνολο δεδομένων και λειτουργεί ανεξάρτητα από τους υπόλοιπους κόμβους. Στο πλαίσιο της παρούσας εργασίας, τα δεδομένα παράγονται συνθετικά με βάση τα στατιστικά χαρακτηριστικά του Iris dataset, διατηρώντας παρόμοια κατανομή με το αρχικό σύνολο δεδομένων. Η προσέγγιση αυτή επιτρέπει τη δημιουργία ενός ελεγχόμενου περιβάλλοντος αξιολόγησης χωρίς την ανάγκη χρήσης μεγάλων πραγματικών datasets.

Η χρήση πολλαπλών κόμβων προσομοιώνει ένα πραγματικό edge περιβάλλον, όπου διαφορετικές συσκευές ή σημεία του δικτύου επεξεργάζονται τοπικά τα δεδομένα που παράγουν. Παρόμοιες αρχιτεκτονικές έχουν προταθεί από τους Abbas et al. (2018) και Premsankar et al. (2018), οι οποίοι υποστηρίζουν ότι η τοπική επεξεργασία συμβάλλει στη μείωση της καθυστέρησης και στη βελτίωση της επεκτασιμότητας των συστημάτων.

4.1.2 Δημιουργία και Εκτέλεση Ερωτημάτων

Η βασική λειτουργία του συστήματος είναι η παραγωγή και εκτέλεση range queries πάνω στα δεδομένα κάθε κόμβου. Τα range queries αποτελούν μία από τις πιο συνηθισμένες μορφές αναζήτησης σε συστήματα βάσεων δεδομένων, καθώς επιτρέπουν τον εντοπισμό εγγραφών που βρίσκονται εντός προκαθορισμένων ορίων τιμών.

Η βιβλιογραφία έχει προτείνει διάφορες τεχνικές βελτιστοποίησης των range queries μέσω δομών ευρετηρίων όπως τα R-Trees και R*-Trees (Guttman, 1984; Beckmann et al., 1990). Ωστόσο, σε περιβάλλοντα edge computing η χρήση πολύπλοκων ευρετηρίων δεν είναι πάντοτε αποδοτική λόγω των περιορισμένων διαθέσιμων πόρων.

Για τον λόγο αυτό, στην παρούσα εργασία δίνεται έμφαση στην ανάλυση ιστορικών ερωτημάτων και όχι στην κατασκευή σύνθετων δομών ευρετηρίων. Τα queries καταγράφονται, αξιολογούνται και χρησιμοποιούνται ως βάση για την εξαγωγή προτύπων πρόσβασης, τα οποία μπορούν να αξιοποιηθούν σε μελλοντικές εκτελέσεις.

Τα range queries αποτελούν μία από τις συχνότερες κατηγορίες ερωτημάτων σε συστήματα διαχείρισης δεδομένων και εφαρμογές αναλυτικής επεξεργασίας. Η σημασία τους είναι ιδιαίτερα αυξημένη σε περιβάλλοντα IoT και Edge Computing, όπου οι χρήστες ενδιαφέρονται συχνά για δεδομένα που βρίσκονται μέσα σε συγκεκριμένα όρια τιμών και όχι για μεμονωμένες εγγραφές.

Επιπλέον, η χρήση πολυδιάστατων range queries επιτρέπει την καλύτερη προσομοίωση πραγματικών αναλυτικών workloads. Ένα ερώτημα μπορεί να περιορίζει ταυτόχρονα πολλαπλά χαρακτηριστικά των δεδομένων, δημιουργώντας συνθήκες που προσεγγίζουν περισσότερο τα πραγματικά σενάρια χρήσης. Για τον λόγο αυτό η παρούσα εργασία χρησιμοποιεί πολυδιάστατα queries αντί απλών μονοδιάστατων αναζητήσεων.

4.2 Προσεγγιστική Ανάλυση Δεδομένων

Η ανάγκη για προσεγγιστική επεξεργασία προκύπτει από τη συνεχώς αυξανόμενη ποσότητα δεδομένων που καλούνται να διαχειριστούν οι σύγχρονες εφαρμογές. Σε περιβάλλοντα edge computing οι διαθέσιμοι υπολογιστικοί πόροι είναι συνήθως περιορισμένοι σε σχέση με τα παραδοσιακά data centers. Ως αποτέλεσμα, η εκτέλεση ακριβών αναλύσεων για κάθε νέο ερώτημα μπορεί να επιβαρύνει σημαντικά το σύστημα.

Οι τεχνικές Approximate Query Processing επιδιώκουν να αντιμετωπίσουν το πρόβλημα αυτό επιτρέποντας την παραγωγή αποτελεσμάτων με ελεγχόμενο σφάλμα αλλά σημαντικά μικρότερο χρόνο εκτέλεσης. Η βασική φιλοσοφία είναι ότι σε πολλές περιπτώσεις μία καλή εκτίμηση είναι προτιμότερη από μία απολύτως ακριβή απάντηση που απαιτεί σημαντικά μεγαλύτερο υπολογιστικό κόστος.

Η σημασία των τεχνικών προσεγγιστικής επεξεργασίας γίνεται ακόμη μεγαλύτερη σε κατανεμημένα περιβάλλοντα, όπου τα δεδομένα είναι διασκορπισμένα σε πολλαπλούς κόμβους. Σε τέτοιες περιπτώσεις, η πλήρης αξιολόγηση ενός ερωτήματος μπορεί να απαιτεί σημαντικό αριθμό προσπελάσεων δεδομένων και ανταλλαγή πληροφοριών μεταξύ των κόμβων. Η χρήση συνοπτικών αναπαραστάσεων επιτρέπει τη μείωση του κόστους αυτού, καθώς οι εκτιμήσεις μπορούν να παραχθούν τοπικά χωρίς την ανάγκη εκτεταμένης επικοινωνίας.

Παράλληλα, η προσεγγιστική ανάλυση συνδέεται άμεσα με τη φιλοσοφία του Proactive Data Analysis. Ένα σύστημα που είναι σε θέση να παράγει γρήγορες εκτιμήσεις για πιθανά μελλοντικά ερωτήματα μπορεί να προετοιμάζει αποτελέσματα εκ των προτέρων και να ανταποκρίνεται ταχύτερα στις ανάγκες των χρηστών. Επομένως, οι τεχνικές Approximate Query Processing δεν λειτουργούν μόνο ως μηχανισμός βελτιστοποίησης αλλά και ως θεμελιώδες στοιχείο για την ανάπτυξη προδραστικών μηχανισμών ανάλυσης δεδομένων.

4.2.1 Κατασκευή Summaries

Η χρήση συνοπτικών αναπαραστάσεων δεδομένων αποτελεί μία από τις βασικές τεχνικές του Approximate Query Processing. Αντί να αποθηκεύονται ή να επεξεργάζονται όλα τα δεδομένα, δημιουργούνται μικρότερες δομές που περιγράφουν συνοπτικά την κατανομή τους.

Στο προτεινόμενο μοντέλο χρησιμοποιούνται histograms, καθώς αποτελούν μία από τις πιο διαδεδομένες και αποδοτικές τεχνικές εκτίμησης selectivity για range queries (Poosala et al., 1996; Ioannidis, 2003). Κάθε histogram χωρίζει το εύρος των τιμών σε bins και καταγράφει τη συχνότητα εμφάνισης των δεδομένων σε κάθε περιοχή.

Η επιλογή των histograms βασίζεται στο γεγονός ότι προσφέρουν καλή ισορροπία μεταξύ ακρίβειας, υπολογιστικού κόστους και απαιτήσεων μνήμης. Το χαρακτηριστικό αυτό τα

καθιστά ιδιαίτερα κατάλληλα για εφαρμογές edge computing. Η χρήση histograms επιλέχθηκε καθώς αποτελούν μία από τις πιο διαδεδομένες τεχνικές συνοπτικής αναπαράστασης δεδομένων στη βιβλιογραφία των βάσεων δεδομένων και του Approximate Query Processing. Τα histograms επιτρέπουν την αποθήκευση βασικών πληροφοριών σχετικά με την κατανομή των δεδομένων χωρίς να απαιτείται η διατήρηση όλων των επιμέρους εγγραφών.

Στην παρούσα εργασία τα histograms κατασκευάζονται ξεχωριστά για κάθε διάσταση των δεδομένων. Με τον τρόπο αυτό διατηρείται πληροφορία σχετικά με την κατανομή των τιμών κάθε χαρακτηριστικού και καθίσταται δυνατή η εκτίμηση της επιλεκτικότητας των range queries. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη όταν το πλήθος των δεδομένων αυξάνεται, καθώς το μέγεθος των histograms παραμένει σημαντικά μικρότερο από το μέγεθος του αρχικού συνόλου δεδομένων.

Επιπλέον, τα summaries λειτουργούν ως μία ενδιάμεση αναπαράσταση μεταξύ των πρωτογενών δεδομένων και της διαδικασίας εκτέλεσης των ερωτημάτων. Αντί να πραγματοποιείται αναζήτηση σε κάθε εγγραφή του συνόλου δεδομένων, το σύστημα αξιοποιεί τη στατιστική πληροφορία που περιέχεται στα histograms προκειμένου να παράγει γρήγορες εκτιμήσεις. Η διαδικασία αυτή μειώνει σημαντικά το κόστος επεξεργασίας και επιτρέπει την αποτελεσματικότερη αξιοποίηση των διαθέσιμων πόρων.

Ένα σημαντικό πλεονέκτημα των histograms είναι η δυνατότητα ελέγχου του επιπέδου λεπτομέρειας μέσω του αριθμού των bins. Μικρός αριθμός bins οδηγεί σε μικρότερες απαιτήσεις μνήμης αλλά μεγαλύτερο σφάλμα εκτίμησης, ενώ μεγαλύτερος αριθμός bins αυξάνει την ακρίβεια με αντίστοιχη αύξηση του κόστους αποθήκευσης και επεξεργασίας. Η διερεύνηση αυτής της σχέσης αποτέλεσε αντικείμενο της πειραματικής αξιολόγησης που παρουσιάζεται στο επόμενο κεφάλαιο.

Η επιλογή του κατάλληλου αριθμού bins αποτελεί μία σημαντική σχεδιαστική απόφαση. Πολύ μικρός αριθμός bins οδηγεί σε υπερβολική γενίκευση της κατανομής των δεδομένων, με αποτέλεσμα την αύξηση του σφάλματος εκτίμησης. Αντίθετα, πολύ μεγάλος αριθμός bins αυξάνει την ακρίβεια αλλά απαιτεί περισσότερη μνήμη και περισσότερο χρόνο κατασκευής των summaries. Για τον λόγο αυτό η παρούσα εργασία εξετάζει πειραματικά διαφορετικές τιμές bins, ώστε να μελετηθεί η επίδρασή τους τόσο στην ακρίβεια όσο και στην απόδοση του συστήματος.

Επιπλέον, τα histograms παρουσιάζουν το πλεονέκτημα ότι μπορούν να κατασκευαστούν εύκολα και να ενημερωθούν με χαμηλό υπολογιστικό κόστος, γεγονός που τα καθιστά ιδιαίτερα κατάλληλα για περιβάλλοντα Edge Computing.

4.2.2 Approximate Query Processing

Με βάση τα summaries, το σύστημα υπολογίζει κατά προσέγγιση το πλήθος των εγγραφών που αναμένεται να ικανοποιούν ένα query χωρίς να απαιτείται πλήρης σάρωση των δεδομένων.

Η φιλοσοφία αυτή ακολουθεί τις αρχές του Approximate Query Processing (AQP), σύμφωνα με τις οποίες η ταχύτητα απόκρισης μπορεί να είναι σημαντικότερη από την απόλυτη ακρίβεια των αποτελεσμάτων (Hellerstein et al., 1997). Παρόμοιες τεχνικές χρησιμοποιούνται ευρέως σε συστήματα αναλυτικής επεξεργασίας μεγάλων δεδομένων και σε data stream processing.

Κατά την εκτέλεση ενός νέου ερωτήματος το σύστημα δεν εξετάζει άμεσα όλες τις εγγραφές του dataset. Αντίθετα, αξιοποιεί τα histograms για να υπολογίσει ποια τμήματα του χώρου τιμών επικαλύπτονται με τα όρια του ερωτήματος και εκτιμά το πλήθος των αναμενόμενων αποτελεσμάτων. Η διαδικασία αυτή είναι σημαντικά ταχύτερη από την ακριβή αξιολόγηση, καθώς βασίζεται σε προϋπολογισμένη στατιστική πληροφορία.

Η αποτελεσματικότητα της προσέγγισης εξαρτάται από την ποιότητα των summaries και από το κατά πόσο αυτά περιγράφουν ικανοποιητικά την πραγματική κατανομή των δεδομένων. Για τον λόγο αυτό η αξιολόγηση της μεθόδου δεν περιορίζεται μόνο στη μέτρηση του χρόνου εκτέλεσης αλλά περιλαμβάνει και τον υπολογισμό του σφάλματος προσέγγισης. Με τον τρόπο αυτό καθίσταται δυνατή η εκτίμηση του βαθμού στον οποίο η μείωση του υπολογιστικού κόστους επηρεάζει την ακρίβεια των αποτελεσμάτων.

Στόχος της παρούσας εργασίας δεν είναι η αντικατάσταση της ακριβούς επεξεργασίας, αλλά η αξιολόγηση του βαθμού στον οποίο μία προσεγγιστική απάντηση μπορεί να μειώσει το κόστος εκτέλεσης διατηρώντας παράλληλα αποδεκτό επίπεδο σφάλματος.

Η σχέση μεταξύ ακρίβειας και απόδοσης αποτελεί ένα από τα βασικά ζητήματα που εξετάζονται στην παρούσα εργασία. Η προσεγγιστική επεξεργασία δεν στοχεύει στην πλήρη αντικατάσταση των ακριβών μεθόδων, αλλά στην παροχή μίας εναλλακτικής λύσης όταν η ταχύτητα απόκρισης αποτελεί προτεραιότητα. Στο πλαίσιο αυτό, η μελέτη της συμπεριφοράς των histograms και των σφαλμάτων εκτίμησης συμβάλλει στην κατανόηση των συνθηκών υπό τις οποίες η χρήση Approximate Query Processing μπορεί να θεωρηθεί αποδοτική και πρακτικά αξιοποιήσιμη σε περιβάλλοντα Edge Computing.

4.3 Ανάλυση Ιστορικών Ερωτημάτων

Η αξιοποίηση ιστορικών ερωτημάτων αποτελεί βασική ιδέα της προδραστικής ανάλυσης δεδομένων. Σε πολλά πληροφοριακά συστήματα παρατηρείται ότι οι χρήστες τείνουν να επαναλαμβάνουν παρόμοια ερωτήματα ή να αναζητούν δεδομένα σε συγκεκριμένες περιοχές του χώρου τιμών. Η καταγραφή και ανάλυση αυτών των μοτίβων μπορεί να προσφέρει σημαντική γνώση σχετικά με τη μελλοντική συμπεριφορά του συστήματος.

Η αναγνώριση επαναλαμβανόμενων προτύπων επιτρέπει την προετοιμασία αναλύσεων πριν ακόμη υποβληθούν τα αντίστοιχα ερωτήματα. Με τον τρόπο αυτό το σύστημα μετατρέπεται από παθητικό μηχανισμό απάντησης σε ενεργό μηχανισμό πρόβλεψης και προετοιμασίας αποτελεσμάτων.

Η προσέγγιση αυτή είναι ιδιαίτερα σημαντική στα περιβάλλοντα edge computing, όπου η έγκαιρη αξιοποίηση των διαθέσιμων πόρων μπορεί να οδηγήσει σε σημαντική βελτίωση της απόδοσης και της ποιότητας εξυπηρέτησης.

Η ανάλυση ιστορικών ερωτημάτων έχει αποτελέσει αντικείμενο εκτεταμένης έρευνας στον χώρο των βάσεων δεδομένων και των συστημάτων αναλυτικής επεξεργασίας. Η βασική υπόθεση είναι ότι η συμπεριφορά των χρηστών δεν είναι πλήρως τυχαία, αλλά παρουσιάζει συγκεκριμένα επαναλαμβανόμενα μοτίβα. Ως αποτέλεσμα, η μελέτη των παλαιότερων ερωτημάτων μπορεί να προσφέρει πολύτιμες ενδείξεις για τις μελλοντικές ανάγκες πρόσβασης στα δεδομένα.

Στο πλαίσιο της παρούσας εργασίας, τα ιστορικά ερωτήματα αντιμετωπίζονται ως πηγή γνώσης σχετικά με τη συμπεριφορά του συστήματος. Αντί να χρησιμοποιούνται αποκλειστικά για την παραγωγή αποτελεσμάτων τη στιγμή της εκτέλεσής τους, αξιοποιούνται και ως δεδομένα εισόδου για μεταγενέστερη ανάλυση. Με τον τρόπο αυτό δημιουργείται ένας μηχανισμός συνεχούς μάθησης, ο οποίος επιτρέπει στο σύστημα να προσαρμόζεται σταδιακά στα πρότυπα χρήσης που παρατηρούνται.

4.3.1 Clustering Ερωτημάτων

Η ανάλυση ιστορικών ερωτημάτων αποτελεί τον πυρήνα της προτεινόμενης προσέγγισης. Τα queries που έχουν εκτελεστεί στο παρελθόν αποθηκεύονται και μετατρέπονται σε διανύσματα χαρακτηριστικών που περιγράφουν τα διαστήματα αναζήτησης κάθε ερωτήματος.

Η βιβλιογραφία έχει δείξει ότι η ομαδοποίηση παρόμοιων ερωτημάτων μπορεί να αποκαλύψει επαναλαμβανόμενα πρότυπα πρόσβασης στα δεδομένα (Jain et al., 1999; Han et al., 2012). Τα πρότυπα αυτά αποτελούν πολύτιμη πληροφορία για την ανάπτυξη μηχανισμών πρόβλεψης και προδραστικής επεξεργασίας.

Στην παρούσα εργασία χρησιμοποιείται subtractive clustering, καθώς δεν απαιτεί προκαθορισμένο αριθμό clusters και προσαρμόζεται δυναμικά στη δομή των δεδομένων (Chiu, 1994).

Η διαδικασία clustering ξεκινά με τη μετατροπή κάθε query σε μία αριθμητική αναπαράσταση που περιγράφει τα όρια των διαστημάτων αναζήτησης σε όλες τις διαστάσεις του χώρου δεδομένων. Με τον τρόπο αυτό καθίσταται δυνατός ο υπολογισμός της ομοιότητας μεταξύ διαφορετικών ερωτημάτων και η ομαδοποίησή τους σε κοινές κατηγορίες.

Η χρήση clustering επιτρέπει τη μείωση της πολυπλοκότητας της ανάλυσης των ιστορικών δεδομένων. Αντί να εξετάζονται μεμονωμένα εκατοντάδες ή χιλιάδες queries, το σύστημα επικεντρώνεται σε ένα μικρότερο σύνολο αντιπροσωπευτικών ομάδων. Κάθε cluster αντιπροσωπεύει μία περιοχή του χώρου ερωτημάτων στην οποία εμφανίζονται παρόμοια μοτίβα πρόσβασης.

Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές proactive analysis, καθώς επιτρέπει τη δημιουργία γενικευμένων προτύπων συμπεριφοράς αντί της αντιμετώπισης κάθε query ως ξεχωριστής περίπτωσης. Ως αποτέλεσμα, το σύστημα μπορεί να αξιοποιεί αποτελεσματικότερα τους διαθέσιμους πόρους και να λαμβάνει αποφάσεις με βάση τη συνολική εικόνα της συμπεριφοράς των χρηστών.

Η επιλογή του subtractive clustering βασίζεται επίσης στην ικανότητά του να λειτουργεί αποτελεσματικά σε περιπτώσεις όπου ο αριθμός των πιθανών ομάδων δεν είναι γνωστός εκ των προτέρων. Σε αντίθεση με τεχνικές όπως ο K-Means, όπου απαιτείται προκαθορισμός του αριθμού των clusters, το subtractive clustering εντοπίζει αυτόματα περιοχές υψηλής πυκνότητας στον χώρο των δεδομένων και δημιουργεί τις αντίστοιχες ομάδες. Το χαρακτηριστικό αυτό το καθιστά ιδιαίτερα κατάλληλο για την ανάλυση ιστορικών queries, όπου η δομή και η πολυπλοκότητα των προτύπων πρόσβασης μπορεί να μεταβάλλονται σημαντικά.

4.3.2 Επιλογή Clusters για Proactive Analysis

Μετά την ολοκλήρωση της διαδικασίας clustering πραγματοποιείται αξιολόγηση των παραγόμενων ομάδων. Τα clusters που παρουσιάζουν υψηλή συχνότητα εμφάνισης και μεγάλη ομοιότητα με τα δεδομένα του κόμβου επιλέγονται ως υποψήφιοι στόχοι προδραστικής ανάλυσης.

Η διαδικασία επιλογής clusters αποτελεί κρίσιμο στάδιο της προτεινόμενης μεθόδου, καθώς δεν είναι όλα τα clusters εξίσου χρήσιμα για προδραστική ανάλυση. Ορισμένες ομάδες ενδέχεται να περιέχουν μικρό αριθμό queries ή να αντιστοιχούν σε τυχαία και σπάνια πρότυπα πρόσβασης που δεν παρουσιάζουν πρακτικό ενδιαφέρον για μελλοντική αξιοποίηση.

Για τον λόγο αυτό εφαρμόζονται κριτήρια επιλογής που επιτρέπουν τον εντοπισμό των πιο αντιπροσωπευτικών ομάδων. Η συχνότητα εμφάνισης λειτουργεί ως ένδειξη της σημασίας ενός προτύπου, ενώ η ομοιότητα των queries που ανήκουν στο ίδιο cluster λειτουργεί ως ένδειξη της συνοχής της ομάδας. Ο συνδυασμός των δύο αυτών παραμέτρων οδηγεί στην επιλογή clusters που παρουσιάζουν τόσο υψηλή επαναληψιμότητα όσο και σαφή χαρακτηριστικά συμπεριφοράς.

Η λογική αυτή βασίζεται στην παρατήρηση ότι τα συχνά επαναλαμβανόμενα queries έχουν αυξημένη πιθανότητα να εμφανιστούν ξανά στο μέλλον. Αντίστοιχες τεχνικές έχουν χρησιμοποιηθεί σε συστήματα caching, materialized views και query prediction (Agrawal et al., 2000; Bruno & Chaudhuri, 2005).

Η επιλογή των clusters που θα αξιοποιηθούν για προδραστική ανάλυση πραγματοποιείται με βάση συγκεκριμένα κριτήρια. Πρώτον, εξετάζεται η συχνότητα εμφάνισης των queries που ανήκουν σε κάθε cluster, καθώς τα συχνότερα μοτίβα είναι πιθανότερο να εμφανιστούν ξανά στο μέλλον. Δεύτερον, λαμβάνεται υπόψη η ομοιότητα των ερωτημάτων που περιέχονται στο cluster, καθώς υψηλή ομοιότητα υποδηλώνει σταθερή συμπεριφορά των χρηστών.

Η διαδικασία αυτή επιτρέπει στο σύστημα να επικεντρωθεί σε περιοχές του χώρου δεδομένων που παρουσιάζουν αυξημένο ενδιαφέρον. Ως αποτέλεσμα, οι διαθέσιμοι υπολογιστικοί πόροι αξιοποιούνται αποτελεσματικότερα, καθώς η προδραστική επεξεργασία περιορίζεται στα πιο σημαντικά και αντιπροσωπευτικά πρότυπα ερωτημάτων.

Τα επιλεγμένα clusters μπορούν να θεωρηθούν ως συμπυκνωμένη αναπαράσταση της γνώσης που έχει αποκτήσει το σύστημα από τη μελέτη των ιστορικών ερωτημάτων. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για την προετοιμασία μελλοντικών αναλύσεων, τη δημιουργία συνοπτικών αποτελεσμάτων ή ακόμη και την υλοποίηση μηχανισμών caching σε μελλοντικές επεκτάσεις του συστήματος.

Στην παρούσα εργασία τα clusters αξιοποιούνται κυρίως για την αξιολόγηση της δυνατότητας αναγνώρισης επαναλαμβανόμενων προτύπων. Ωστόσο, η ίδια πληροφορία θα μπορούσε να αποτελέσει τη βάση για πιο προηγμένους μηχανισμούς πρόβλεψης ερωτημάτων, όπου το σύστημα θα ήταν σε θέση να εκτιμά τις πιθανές μελλοντικές ανάγκες των χρηστών και να προετοιμάζει τα αντίστοιχα αποτελέσματα πριν ακόμη ζητηθούν. Η δυνατότητα αυτή αποτελεί έναν από τους βασικούς στόχους της φιλοσοφίας του Proactive Data Analysis.

4.4 Τύποι Workload

Η χρήση διαφορετικών τύπων workload είναι απαραίτητη για την ολοκληρωμένη αξιολόγηση του προτεινόμενου μοντέλου. Στην πράξη, οι χρήστες ενός πληροφοριακού συστήματος δεν ακολουθούν πάντα την ίδια συμπεριφορά, ενώ τα δεδομένα συχνά παρουσιάζουν διαφορετικές κατανομές ανάλογα με το περιβάλλον εφαρμογής. Για τον λόγο αυτό χρησιμοποιούνται πολλαπλά σενάρια προσομοίωσης που επιτρέπουν την αξιολόγηση της συμπεριφοράς του συστήματος κάτω από διαφορετικές συνθήκες λειτουργίας.

Η χρήση διαφορετικών workloads αποτελεί καθιερωμένη πρακτική στην αξιολόγηση συστημάτων διαχείρισης δεδομένων και αναλυτικής επεξεργασίας. Ένα σύστημα μπορεί να παρουσιάζει πολύ καλή συμπεριφορά σε ένα συγκεκριμένο είδος φόρτου εργασίας, αλλά να εμφανίζει σημαντική μείωση της απόδοσής του όταν αλλάξει η κατανομή των ερωτημάτων ή των δεδομένων. Για τον λόγο αυτό η αξιολόγηση σε πολλαπλά σενάρια θεωρείται απαραίτητη προκειμένου να εξαχθούν ασφαλή συμπεράσματα σχετικά με τη γενικότερη αποτελεσματικότητα μιας μεθόδου.

Στην παρούσα εργασία τα workloads δεν χρησιμοποιούνται μόνο για τη μέτρηση της απόδοσης των histograms και του Approximate Query Processing, αλλά και για τη διερεύνηση της συμπεριφοράς των μηχανισμών ανάλυσης ιστορικών ερωτημάτων. Η μεταβολή της κατανομής των queries επηρεάζει άμεσα τη διαδικασία clustering και κατά συνέπεια την ποιότητα των προτύπων που εντοπίζονται για την υποστήριξη της προδραστικής ανάλυσης.

Για την αξιολόγηση του μοντέλου δημιουργήθηκαν διαφορετικοί τύποι workload. Το Uniform workload παράγει ερωτήματα με ομοιόμορφη κατανομή στον χώρο δεδομένων, εξασφαλίζοντας ότι όλες οι περιοχές του χώρου τιμών έχουν παρόμοια πιθανότητα επιλογής. Το Clustered workload συγκεντρώνει τα queries σε συγκεκριμένες περιοχές ενδιαφέροντος, προσομοιώνοντας περιπτώσεις όπου οι χρήστες εστιάζουν σε συγκεκριμένα υποσύνολα δεδομένων. Αντίστοιχα, το Biased workload δημιουργεί ερωτήματα με αυξημένη πιθανότητα εμφάνισης σε επιλεγμένες διαστάσεις ή περιοχές του χώρου δεδομένων, αναπαριστώντας πιο στοχευμένα μοτίβα πρόσβασης. Τέλος, το Hybrid workload συνδυάζει χαρακτηριστικά από όλα τα προηγούμενα σενάρια και προσομοιώνει πιο ρεαλιστικές συνθήκες λειτουργίας.

Παρότι το προτεινόμενο μοντέλο υποστηρίζει πολλαπλούς τύπους workload, η πειραματική αξιολόγηση της παρούσας εργασίας επικεντρώνεται κυρίως στο Uniform workload. Η επιλογή αυτή πραγματοποιήθηκε ώστε να απομονωθεί η επίδραση συγκεκριμένων παραμέτρων του συστήματος, όπως ο αριθμός των edge nodes και ο αριθμός των bins που χρησιμοποιούνται στα histograms. Με τον τρόπο αυτό καθίσταται ευκολότερη η μελέτη της επίδρασης κάθε παραμέτρου στην απόδοση του συστήματος χωρίς την ταυτόχρονη επίδραση διαφορετικών κατανομών ερωτημάτων.

Κάθε τύπος workload αναδεικνύει διαφορετικές πτυχές της λειτουργίας του συστήματος. Το Uniform workload επιτρέπει την αξιολόγηση της συμπεριφοράς του μοντέλου σε συνθήκες όπου τα ερωτήματα κατανέμονται ομοιόμορφα στον χώρο δεδομένων και δεν υπάρχουν εμφανή πρότυπα πρόσβασης. Αντίθετα, τα Clustered και Biased workloads δημιουργούν περιοχές αυξημένου ενδιαφέροντος, στις οποίες ορισμένα χαρακτηριστικά ή περιοχές του χώρου τιμών χρησιμοποιούνται συχνότερα από άλλες.

Η ύπαρξη τέτοιων περιοχών παρουσιάζει ιδιαίτερο ενδιαφέρον για την παρούσα εργασία, καθώς διευκολύνει τον εντοπισμό επαναλαμβανόμενων προτύπων μέσω της διαδικασίας

clustering. Όσο πιο έντονα είναι τα πρότυπα πρόσβασης, τόσο μεγαλύτερη είναι η πιθανότητα να αναγνωριστούν ομάδες ερωτημάτων που μπορούν να αξιοποιηθούν για προδραστική ανάλυση και προετοιμασία αποτελεσμάτων.

Η αξιολόγηση πολλαπλών workloads είναι ιδιαίτερα σημαντική καθώς επιτρέπει τη διερεύνηση της συμπεριφοράς του συστήματος τόσο σε ιδανικές όσο και σε πιο απαιτητικές συνθήκες. Με τον τρόπο αυτό μπορεί να εξεταστεί κατά πόσο οι τεχνικές Approximate Query Processing και clustering παραμένουν αποτελεσματικές όταν μεταβάλλεται η κατανομή των δεδομένων ή η συμπεριφορά των χρηστών. Επιπλέον, η ύπαρξη διαφορετικών σεναρίων διευκολύνει την εξαγωγή πιο γενικευμένων συμπερασμάτων σχετικά με τη δυνατότητα εφαρμογής της προτεινόμενης μεθόδου σε πραγματικά περιβάλλοντα Edge Computing και επιτρέπει την αξιολόγηση της ανθεκτικότητάς της σε διαφορετικές συνθήκες φόρτου.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το Hybrid workload, καθώς προσεγγίζει περισσότερο τις συνθήκες που συναντώνται σε πραγματικά πληροφοριακά συστήματα. Στην πράξη, η συμπεριφορά των χρηστών σπάνια είναι πλήρως ομοιόμορφη ή απόλυτα συγκεντρωμένη σε συγκεκριμένες περιοχές ενδιαφέροντος. Αντίθετα, συνήθως συνυπάρχουν διαφορετικά πρότυπα πρόσβασης, τα οποία μεταβάλλονται με την πάροδο του χρόνου. Η δυνατότητα αξιολόγησης του προτεινόμενου μοντέλου σε ένα τέτοιο σύνθετο περιβάλλον παρέχει πιο ρεαλιστική εικόνα της αποτελεσματικότητάς του.

Η μελέτη των διαφορετικών workloads επιτρέπει επίσης την καλύτερη κατανόηση των συνθηκών υπό τις οποίες οι τεχνικές Approximate Query Processing και clustering αποδίδουν καλύτερα. Τα αποτελέσματα που παρουσιάζονται στο επόμενο κεφάλαιο συμβάλλουν στην αξιολόγηση της σχέσης μεταξύ της κατανομής των ερωτημάτων, της ποιότητας των clusters και της συνολικής απόδοσης του συστήματος. Με τον τρόπο αυτό καθίσταται δυνατή η εξαγωγή συμπερασμάτων όχι μόνο για το συγκεκριμένο πειραματικό περιβάλλον αλλά και για πιθανές εφαρμογές της προσέγγισης σε πραγματικά συστήματα Edge Computing.

4.5 Αξιολόγηση του Μοντέλου

Η αξιολόγηση ενός συστήματος προδραστικής ανάλυσης δεν περιορίζεται μόνο στην ακρίβεια των αποτελεσμάτων. Εξίσου σημαντική είναι η δυνατότητα μείωσης του χρόνου απόκρισης και η αποδοτική αξιοποίηση των υπολογιστικών πόρων. Σε περιβάλλοντα Edge Computing, όπου οι διαθέσιμοι πόροι είναι συχνά περιορισμένοι, η ισορροπία μεταξύ απόδοσης και ακρίβειας αποτελεί κρίσιμο παράγοντα για την αποτελεσματική λειτουργία του συστήματος.

Για τον λόγο αυτό η παρούσα εργασία χρησιμοποιεί συνδυασμό μετρικών που επιτρέπουν την ταυτόχρονη αξιολόγηση της ποιότητας των αποτελεσμάτων και της υπολογιστικής αποδοτικότητας. Η προσέγγιση αυτή παρέχει μία πιο ολοκληρωμένη εικόνα της συμπεριφοράς του συστήματος σε σχέση με την αποκλειστική χρήση μίας μόνο μετρικής και επιτρέπει την αντικειμενική σύγκριση μεταξύ της ακριβούς και της προσεγγιστικής επεξεργασίας ερωτημάτων.

Η αξιολόγηση του προτεινόμενου μοντέλου πραγματοποιείται μέσω ενός συνόλου πειραμάτων που εξετάζουν τόσο τη λειτουργικότητα όσο και την αποδοτικότητά του. Ιδιαίτερη έμφαση δίνεται στη σύγκριση μεταξύ της ακριβούς και της προσεγγιστικής επεξεργασίας ερωτημάτων, προκειμένου να διερευνηθεί ο βαθμός στον οποίο η χρήση summaries και

ιστορικών ερωτημάτων μπορεί να μειώσει το υπολογιστικό κόστος χωρίς σημαντική υποβάθμιση της ποιότητας των αποτελεσμάτων.

Επιπλέον, η αξιολόγηση περιλαμβάνει τη μελέτη της επίδρασης βασικών παραμέτρων του συστήματος, όπως ο αριθμός των edge nodes και ο αριθμός των bins που χρησιμοποιούνται στα histograms. Οι παράμετροι αυτές επηρεάζουν άμεσα τόσο την ακρίβεια των εκτιμήσεων όσο και τον χρόνο απόκρισης του συστήματος. Η ανάλυσή τους επιτρέπει την καλύτερη κατανόηση των συμβιβασμών (trade-offs) μεταξύ απόδοσης, ακρίβειας και υπολογιστικού κόστους.

4.5.1 Δείκτες Απόδοσης

Η αξιολόγηση βασίζεται σε τρεις βασικές κατηγορίες μετρικών:

- Κάλυψη δεδομένων (Data Coverage)
- Χρόνος εκτέλεσης (Latency)
- Σφάλμα προσέγγισης (Approximation Error)

Οι μετρικές αυτές χρησιμοποιούνται συχνά στη βιβλιογραφία του Approximate Query Processing και των συστημάτων Edge Computing, καθώς παρέχουν ολοκληρωμένη εικόνα για την αποτελεσματικότητα ενός συστήματος.

Η ταυτόχρονη αξιολόγηση ταχύτητας και ακρίβειας είναι ιδιαίτερα σημαντική, καθώς μία μέθοδος που είναι πολύ γρήγορη αλλά παρουσιάζει μεγάλο σφάλμα δεν μπορεί να θεωρηθεί πρακτικά χρήσιμη.

Η κάλυψη δεδομένων (Data Coverage) χρησιμοποιείται για να αξιολογηθεί κατά πόσο τα παραγόμενα ερωτήματα εξερευνούν αποτελεσματικά τον χώρο των δεδομένων. Υψηλή κάλυψη υποδηλώνει ότι το workload δεν περιορίζεται σε μικρό μόνο τμήμα των διαθέσιμων τιμών και συνεπώς τα αποτελέσματα της αξιολόγησης είναι περισσότερο αντιπροσωπευτικά.

Ο χρόνος εκτέλεσης (Latency) αποτελεί μία από τις σημαντικότερες μετρικές σε περιβάλλοντα Edge Computing, καθώς σχετίζεται άμεσα με την εμπειρία του χρήστη και την ικανότητα του συστήματος να ανταποκρίνεται σε πραγματικό χρόνο. Στην παρούσα εργασία καταγράφεται τόσο ο χρόνος της ακριβούς επεξεργασίας όσο και ο αντίστοιχος χρόνος της προσεγγιστικής προσέγγισης, επιτρέποντας την άμεση σύγκρισή τους.

Το σφάλμα προσέγγισης (Approximation Error) εκφράζει τη διαφορά μεταξύ του πραγματικού και του εκτιμώμενου αποτελέσματος. Για τον σκοπό αυτό χρησιμοποιούνται τόσο το μέσο απόλυτο σφάλμα όσο και το σχετικό σφάλμα, παρέχοντας πληρέστερη εικόνα για την ποιότητα των εκτιμήσεων που παράγονται από τα histograms.

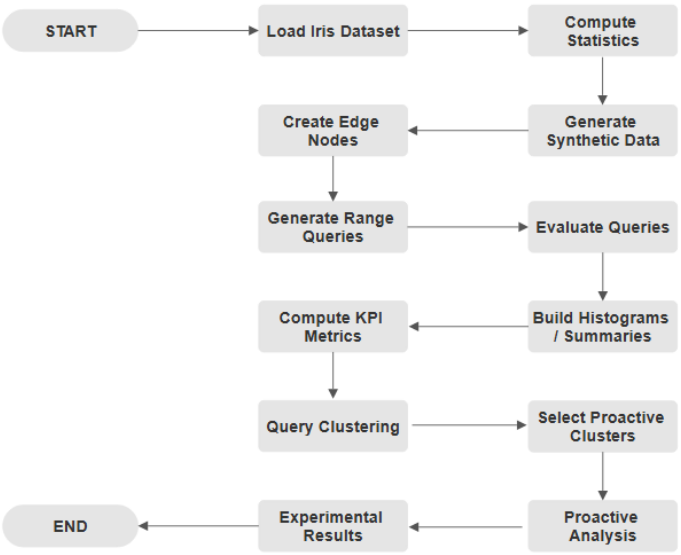
4.5.2 Διάγραμμα Ροής του Συστήματος

Το διάγραμμα ροής της Εικόνας 1 απεικονίζει τη συνολική λειτουργία του προτεινόμενου μοντέλου. Η διαδικασία ξεκινά με τη φόρτωση και ανάλυση των δεδομένων, συνεχίζει με τη δημιουργία edge nodes και range queries και ακολουθεί η ακριβής και προσεγγιστική αξιολόγηση των ερωτημάτων.

Στη συνέχεια εφαρμόζεται η διαδικασία clustering στα ιστορικά queries και επιλέγονται τα clusters που παρουσιάζουν αυξημένο ενδιαφέρον για προδραστική ανάλυση. Τέλος υπολογίζονται οι δείκτες απόδοσης και πραγματοποιείται η αξιολόγηση της συνολικής συμπεριφοράς του συστήματος.

Το διάγραμμα ροής αποτυπώνει τη λογική αλληλουχία των επιμέρους σταδίων του συστήματος και αναδεικνύει τη σύνδεση μεταξύ της προσεγγιστικής επεξεργασίας και της ανάλυσης ιστορικών ερωτημάτων. Η διαδικασία ξεκινά από τα δεδομένα εισόδου και καταλήγει στην εξαγωγή γνώσης σχετικά με επαναλαμβανόμενα πρότυπα πρόσβασης, τα οποία μπορούν να αξιοποιηθούν για την υποστήριξη μηχανισμών προδραστικής ανάλυσης.

Παράλληλα, το διάγραμμα καθιστά σαφές ότι η προτεινόμενη προσέγγιση δεν βασίζεται σε μία μεμονωμένη τεχνική αλλά στον συνδυασμό πολλαπλών μηχανισμών. Τα histograms χρησιμοποιούνται για τη δημιουργία συνοπτικών αναπαραστάσεων, οι τεχνικές Approximate Query Processing για τη γρήγορη εκτίμηση αποτελεσμάτων και το clustering για την αναγνώριση επαναλαμβανόμενων προτύπων στα ιστορικά queries. Η συνεργασία των επιμέρους αυτών στοιχείων οδηγεί στη δημιουργία ενός ολοκληρωμένου πλαισίου υποστήριξης προδραστικής ανάλυσης δεδομένων.



Εικόνα 1: *Διάγραμμα ροής του προτεινόμενου μοντέλου για την προδραστική ανάλυση δεδομένων μέσω ιστορικών ερωτημάτων σε περιβάλλον edge computing.*

4.6 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκε το προτεινόμενο μοντέλο για την υποστήριξη μηχανισμών Proactive Data Analysis σε περιβάλλοντα Edge Computing. Αρχικά περιγράφηκε η αρχιτεκτονική του συστήματος και ο τρόπος με τον οποίο οι edge nodes διαχειρίζονται τοπικά δεδομένα και εκτελούν range queries. Στη συνέχεια αναλύθηκε η χρήση συνοπτικών αναπαραστάσεων δεδομένων (summaries) και τεχνικών Approximate Query Processing για την ταχύτερη εκτίμηση των αποτελεσμάτων ερωτημάτων με περιορισμένο υπολογιστικό κόστος.

Παράλληλα παρουσιάστηκε η διαδικασία ανάλυσης ιστορικών ερωτημάτων μέσω τεχνικών clustering, με στόχο τον εντοπισμό επαναλαμβανόμενων προτύπων πρόσβασης και την επιλογή υποψήφιων clusters για προδραστική επεξεργασία. Επιπλέον περιγράφηκαν οι διαφορετικοί τύποι workload που χρησιμοποιούνται για την αξιολόγηση του συστήματος, καθώς και οι βασικοί δείκτες απόδοσης που επιτρέπουν τη μέτρηση της αποτελεσματικότητας της προτεινόμενης προσέγγισης.

Ιδιαίτερη έμφαση δόθηκε στη μελέτη της σχέσης μεταξύ ακρίβειας και υπολογιστικού κόστους, η οποία αποτελεί ένα από τα βασικά ζητήματα των συστημάτων Approximate Query Processing. Η χρήση histograms επιτρέπει τη δημιουργία συνοπτικών αναπαραστάσεων των δεδομένων, μειώνοντας τον χρόνο εκτέλεσης των ερωτημάτων με αντάλλαγμα την εισαγωγή ενός ελεγχόμενου σφάλματος εκτίμησης. Παράλληλα, η αξιοποίηση ιστορικών ερωτημάτων και τεχνικών clustering επιτρέπει την αναγνώριση επαναλαμβανόμενων προτύπων πρόσβασης, τα οποία μπορούν να αποτελέσουν τη βάση για την ανάπτυξη μηχανισμών προδραστικής επεξεργασίας.

Επιπλέον, παρουσιάστηκαν οι βασικές παράμετροι που επηρεάζουν τη λειτουργία του μοντέλου, όπως ο αριθμός των edge nodes και το επίπεδο λεπτομέρειας των histograms μέσω του αριθμού των bins. Η μελέτη των παραμέτρων αυτών είναι σημαντική καθώς επηρεάζει άμεσα τόσο την ακρίβεια των εκτιμήσεων όσο και την ταχύτητα απόκρισης του συστήματος. Για τον λόγο αυτό οι παράμετροι αυτές εξετάζονται αναλυτικά στο πλαίσιο της πειραματικής αξιολόγησης.

Η συνδυαστική αξιοποίηση τεχνικών Edge Computing, Approximate Query Processing και Clustering δημιουργεί ένα ολοκληρωμένο πλαίσιο για τη μελέτη της προδραστικής ανάλυσης δεδομένων. Το προτεινόμενο μοντέλο αποτελεί τη βάση για την πειραματική αξιολόγηση που παρουσιάζεται στο επόμενο κεφάλαιο, όπου εξετάζεται η αποτελεσματικότητά του μέσω μετρήσεων απόδοσης, ανάλυσης σφαλμάτων προσέγγισης και πειραμάτων που διερευνούν την επίδραση κρίσιμων παραμέτρων, όπως ο αριθμός των edge nodes και ο αριθμός των bins των histograms.

ΚΕΦΑΛΑΙΟ 5 Πειραματική Αποτίμηση

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα της πειραματικής αξιολόγησης του προτεινόμενου μοντέλου. Η αξιολόγηση πραγματοποιείται μέσω της εκτέλεσης του αλγορίθμου σε περιβάλλον προσομοίωσης edge computing και εξετάζει την απόδοση της προτεινόμενης προσέγγισης ως προς την κάλυψη δεδομένων, τον χρόνο εκτέλεσης, το σφάλμα προσέγγισης και την αποτελεσματικότητα της διαδικασίας clustering.

5.1 Περιβάλλον Πειραματικής Αξιολόγησης

5.1.1 Παράμετροι Προσομοίωσης

Η προσομοίωση εκτελέστηκε σε πέντε edge nodes. Για κάθε κόμβο δημιουργήθηκαν 500 συνθετικές εγγραφές και 100 range queries. Η αξιολόγηση πραγματοποιήθηκε με χρήση ιστογραμμάτων (histograms) για τη δημιουργία summaries και την εκτίμηση αποτελεσμάτων ερωτημάτων.

5.1.2 Στατιστικά Χαρακτηριστικά Δεδομένων

```
Dataset statistics (iris):
                mean    std    min    max
sepal length 5.8433 0.8281 4.3000 7.9000
sepal width  3.0540 0.4336 2.0000 4.4000
petal length 3.7587 1.7644 1.0000 6.9000
petal width  1.1987 0.7632 0.1000 2.5000
```

***Εικόνα 2**: Στατιστικά χαρακτηριστικά του Iris dataset που χρησιμοποιήθηκαν για τη δημιουργία των συνθετικών δεδομένων*

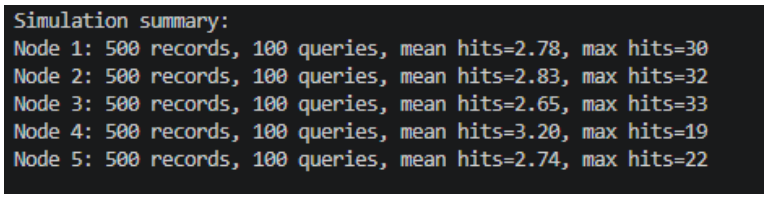
Τα στατιστικά χαρακτηριστικά του Iris dataset χρησιμοποιούνται ως βάση για την παραγωγή των συνθετικών δεδομένων κάθε κόμβου. Οι μέσες τιμές, οι τυπικές αποκλίσεις και τα όρια των χαρακτηριστικών επιτρέπουν τη δημιουργία δεδομένων που διατηρούν παρόμοια κατανομή με το αρχικό σύνολο δεδομένων.

Η επιλογή του συγκεκριμένου συνόλου δεδομένων εξασφαλίζει ότι η προσομοίωση βασίζεται σε ρεαλιστικές τιμές και όχι σε εντελώς τυχαία δεδομένα. Με τον τρόπο αυτό τα παραγόμενα workloads προσεγγίζουν καλύτερα πραγματικά σενάρια λειτουργίας, επιτρέποντας πιο αξιόπιστη αξιολόγηση της προτεινόμενης μεθόδου.

Παράλληλα, η ύπαρξη διαφορετικών χαρακτηριστικών με διαφορετικά εύρη τιμών δημιουργεί ένα κατάλληλο περιβάλλον για τη μελέτη πολυδιάστατων range queries. Αυτό είναι ιδιαίτερα σημαντικό, καθώς η απόδοση των μηχανισμών προσεγγιστικής ανάλυσης και clustering επηρεάζεται άμεσα από την κατανομή των δεδομένων. Επομένως, η συγκεκριμένη βάση δεδομένων παρέχει ένα επαρκές πειραματικό υπόβαθρο για την αξιολόγηση των τεχνικών που εφαρμόζονται στην εργασία.

5.2 Αποτελέσματα Εκτέλεσης Ερωτημάτων

5.2.1 Σύνοψη Edge Nodes



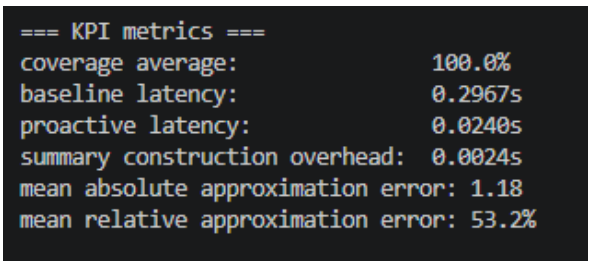
Εικόνα 3: *Σύνοψη των πέντε edge nodes και των ερωτημάτων που εκτελέστηκαν σε κάθε κόμβο.*

Η εικόνα παρουσιάζει τον αριθμό των εγγραφών και των queries που διαχειρίζεται κάθε edge node, καθώς και τον μέσο και μέγιστο αριθμό αποτελεσμάτων ανά query. Τα αποτελέσματα δείχνουν ότι το φορτίο κατανέμεται ομοιόμορφα μεταξύ των κόμβων της προσομοίωσης.

Η ομοιόμορφη κατανομή του φόρτου είναι ιδιαίτερα σημαντική για την εγκυρότητα των πειραμάτων, καθώς περιορίζει την πιθανότητα οι μετρήσεις απόδοσης να επηρεάζονται από ανισορροπίες μεταξύ των κόμβων. Με αυτόν τον τρόπο οι διαφορές που παρατηρούνται στα αποτελέσματα μπορούν να αποδοθούν κυρίως στις τεχνικές επεξεργασίας που εφαρμόζονται και όχι σε διαφορετικό όγκο δεδομένων ή αριθμό ερωτημάτων.

Επιπλέον, ο μέσος αριθμός αποτελεσμάτων ανά query υποδηλώνει ότι τα ερωτήματα δεν είναι ούτε υπερβολικά περιοριστικά ούτε υπερβολικά γενικά. Αυτό σημαίνει ότι η αξιολόγηση πραγματοποιείται σε ένα ρεαλιστικό σενάριο, όπου τα queries επιστρέφουν επαρκή αριθμό εγγραφών ώστε να είναι εφικτή η μελέτη της συμπεριφοράς του συστήματος και της αποτελεσματικότητας των προσεγγιστικών τεχνικών.

5.2.2 Αξιολόγηση KPI Metrics



Εικόνα 4: *KPI metrics για την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου.*

Παρατηρείται ότι ο χρόνος της προσεγγιστικής ανάλυσης (0.0240s) είναι σημαντικά μικρότερος από τον χρόνο της ακριβούς αξιολόγησης (0.2967s), γεγονός που δείχνει ότι η χρήση summaries μπορεί να μειώσει σημαντικά το υπολογιστικό κόστος. Παράλληλα, η κάλυψη των δεδομένων παραμένει πλήρης (100%), ενώ το σφάλμα προσέγγισης διατηρείται σε αποδεκτά επίπεδα.

Η διαφορά αυτή στους χρόνους εκτέλεσης αποτελεί ένα από τα σημαντικότερα ευρήματα της εργασίας. Συγκεκριμένα, η προσεγγιστική μέθοδος εμφανίζει βελτίωση που ξεπερνά τη

μία τάξη μεγέθους σε σχέση με την ακριβή αξιολόγηση, γεγονός που επιβεβαιώνει ότι η χρήση histograms μπορεί να μειώσει σημαντικά τον απαιτούμενο χρόνο επεξεργασίας.

Παράλληλα, η πλήρης κάλυψη των δεδομένων δείχνει ότι τα παραγόμενα workloads εξετάζουν ολόκληρο τον χώρο αναζήτησης και δεν περιορίζονται σε μικρό μόνο τμήμα του. Αυτό σημαίνει ότι τα αποτελέσματα δεν οφείλονται σε ευνοϊκές συνθήκες δοκιμής αλλά αντανακλούν τη συνολική συμπεριφορά του συστήματος.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και το σφάλμα προσέγγισης. Παρότι η χρήση summaries οδηγεί αναπόφευκτα σε κάποια απώλεια ακρίβειας, το σφάλμα παραμένει σε επίπεδα που επιτρέπουν τη χρήση των αποτελεσμάτων για υποστήριξη αποφάσεων και προδραστική ανάλυση. Συνεπώς, η προτεινόμενη προσέγγιση επιτυγχάνει μία ικανοποιητική ισορροπία μεταξύ ταχύτητας και ακρίβειας.

5.3 Αποτελέσματα Clustering

5.3.1 Επιλογή Clusters για Proactive Analysis

```
Extracting query clusters for proactive analysis...
Found 100 clusters from 5 nodes.
Selected 8 clusters for proactive execution.
Node 2 batch 0 cluster 4: count=13, similarity=0.721
Node 5 batch 0 cluster 0: count=12, similarity=0.720
Node 4 batch 0 cluster 2: count=13, similarity=0.718
Node 1 batch 0 cluster 1: count=16, similarity=0.708
Node 3 batch 0 cluster 0: count=10, similarity=0.695
Node 3 batch 0 cluster 4: count=16, similarity=0.660
Node 4 batch 0 cluster 0: count=10, similarity=0.642
Node 1 batch 0 cluster 2: count=13, similarity=0.605
```

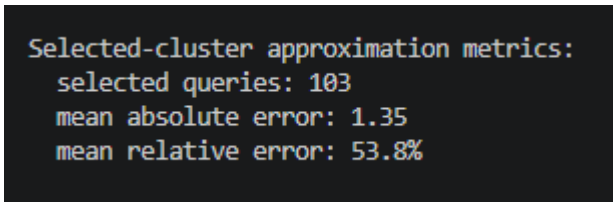
Εικόνα 5: *Αποτελέσματα της διαδικασίας clustering και επιλογής clusters για προδραστική ανάλυση.*

Η διαδικασία clustering εντόπισε συνολικά 100 clusters από τα ιστορικά queries των πέντε κόμβων, από τα οποία επιλέχθηκαν 8 για προδραστική εκτέλεση. Οι υψηλές τιμές ομοιότητας των επιλεγμένων clusters δείχνουν ότι τα συγκεκριμένα μοτίβα ερωτημάτων εμφανίζονται συχνά και μπορούν να αξιοποιηθούν για την υποστήριξη μηχανισμών Proactive Data Analysis.

Το αποτέλεσμα αυτό επιβεβαιώνει ότι ακόμη και σε ένα συνθετικό περιβάλλον εργασίας υπάρχουν επαναλαμβανόμενα πρότυπα πρόσβασης στα δεδομένα. Η ύπαρξη τέτοιων προτύπων αποτελεί βασική προϋπόθεση για την ανάπτυξη προδραστικών μηχανισμών, καθώς χωρίς επαναληψιμότητα δεν θα ήταν δυνατή η πρόβλεψη μελλοντικών ερωτημάτων.

Η επιλογή μόνο ενός μικρού ποσοστού των clusters δείχνει επίσης ότι η διαδικασία φιλτραρίσματος λειτουργεί αποτελεσματικά. Αντί να θεωρούνται όλα τα clusters εξίσου σημαντικά, το σύστημα επικεντρώνεται σε εκείνα που παρουσιάζουν αυξημένη πιθανότητα επανεμφάνισης και μεγαλύτερη συνάφεια με τα δεδομένα του κόμβου. Αυτό μειώνει το κόστος προδραστικής επεξεργασίας και επιτρέπει την καλύτερη αξιοποίηση των διαθέσιμων πόρων.

5.3.2 Αξιολόγηση Ακρίβειας Clusters



Εικόνα 6: *Μετρικές ακρίβειας για τα clusters που επιλέχθηκαν για proactive execution.*

Μία πρώτη κατεύθυνση αφορά την αξιολόγηση της προτεινόμενης προσέγγισης σε πραγματικές υποδομές Edge Computing και σε πραγματικά workloads χρηστών. Μια τέτοια αξιολόγηση θα επέτρεπε την ακριβέστερη εκτίμηση της συμπεριφοράς του συστήματος σε συνθήκες πραγματικής λειτουργίας.
Μία πρώτη κατεύθυνση αφορά την αξιολόγηση της προτεινόμενης προσέγγισης σε πραγματικές υποδομές Edge Computing και σε πραγματικά workloads χρηστών. Μια τέτοια αξιολόγηση θα επέτρεπε την ακριβέστερη εκτίμηση της συμπεριφοράς του συστήματος σε συνθήκες πραγματικής λειτουργίας.

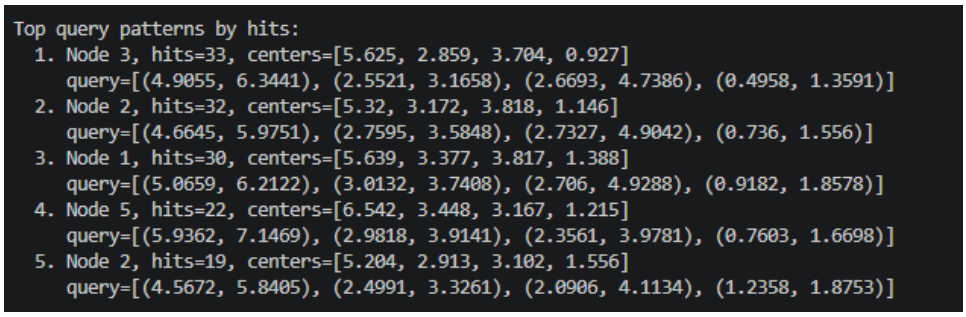
Οι μετρικές των επιλεγμένων clusters επιτρέπουν την αξιολόγηση της ακρίβειας της προσεγγιστικής ανάλυσης. Τα αποτελέσματα δείχνουν ότι η μέθοδος μπορεί να παρέχει χρήσιμες εκτιμήσεις με περιορισμένο υπολογιστικό κόστος.

Η αξιολόγηση των clusters είναι ιδιαίτερα σημαντική, καθώς καθορίζει κατά πόσο τα πρότυπα που εντοπίστηκαν αντιπροσωπεύουν πραγματικές τάσεις στα ιστορικά ερωτήματα. Οι υψηλές τιμές ομοιότητας υποδεικνύουν ότι τα επιλεγμένα clusters περιγράφουν περιοχές του χώρου αναζήτησης που παρουσιάζουν αυξημένη δραστηριότητα.

Επιπλέον, η ακρίβεια των clusters επηρεάζει άμεσα την αποτελεσματικότητα οποιουδήποτε μηχανισμού προδραστικής επεξεργασίας. Όσο πιο αντιπροσωπευτικά είναι τα clusters, τόσο μεγαλύτερη είναι η πιθανότητα οι προϋπολογισμένες εκτιμήσεις να αξιοποιηθούν σε μελλοντικά queries. Τα αποτελέσματα δείχνουν ότι η εφαρμογή subtractive clustering μπορεί να εντοπίσει χρήσιμα πρότυπα χωρίς να απαιτείται εκ των προτέρων γνώση του αριθμού των ομάδων.

5.4 Ανάλυση Προτύπων Ερωτημάτων

5.4.1 Συχνότερα Query Patterns



***Εικόνα** 7: Τα συχνότερα πρότυπα ερωτημάτων που εντοπίστηκαν κατά την εκτέλεση της προσομοίωσης.*

Η εικόνα παρουσιάζει τα queries με τον μεγαλύτερο αριθμό επιτυχιών (hits), τα οποία αντιπροσωπεύουν τις περιοχές του χώρου δεδομένων που εμφανίζουν αυξημένο ενδιαφέρον. Η πληροφορία αυτή μπορεί να αξιοποιηθεί για την ανάπτυξη πιο αποδοτικών μηχανισμών προδραστικής ανάλυσης στο μέλλον.

Η ύπαρξη περιοχών με υψηλό αριθμό επιτυχιών δείχνει ότι η πρόσβαση στα δεδομένα δεν είναι ομοιόμορφη. Ορισμένα τμήματα του χώρου αναζήτησης εμφανίζουν μεγαλύτερη συγκέντρωση ενδιαφέροντος, γεγονός που συμφωνεί με τη συμπεριφορά πραγματικών εφαρμογών, όπου συγκεκριμένες περιοχές των δεδομένων προσπελαύνονται συχνότερα από άλλες.

Η παρατήρηση αυτή αποτελεί βασικό επιχείρημα υπέρ της προδραστικής ανάλυσης. Εφόσον είναι δυνατό να εντοπιστούν εκ των προτέρων οι περιοχές που συγκεντρώνουν το μεγαλύτερο ενδιαφέρον, το σύστημα μπορεί να προετοιμάζει αντίστοιχα summaries ή ακόμη και αποτελέσματα ερωτημάτων πριν αυτά ζητηθούν από τον χρήστη. Με τον τρόπο αυτό επιτυγχάνεται περαιτέρω μείωση του χρόνου απόκρισης και καλύτερη αξιοποίηση των διαθέσιμων πόρων.

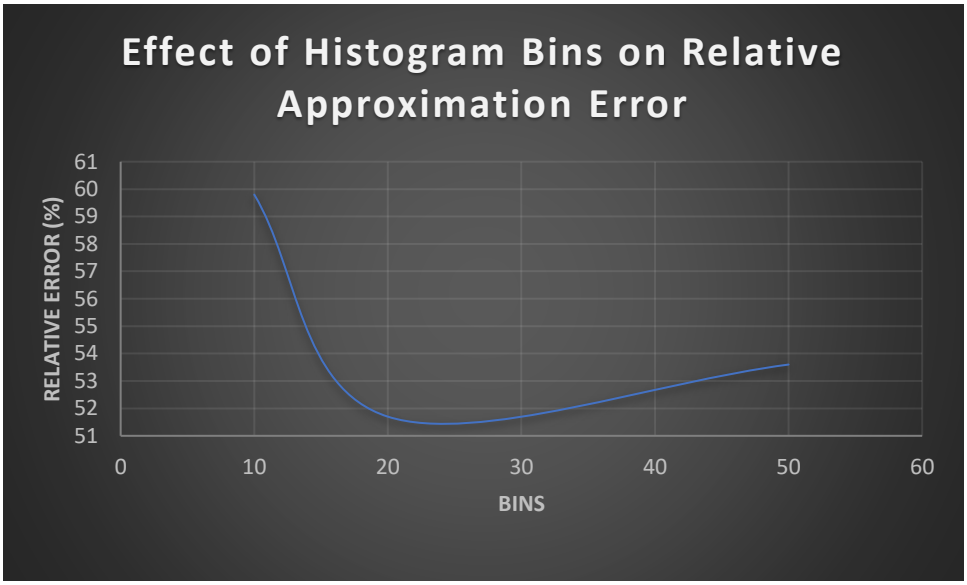
5.5 Ανάλυση Ευαισθησίας Παραμέτρων

5.5.1 Επίδραση του Αριθμού Edge Nodes

Για τη μελέτη της επίδρασης του αριθμού των edge nodes πραγματοποιήθηκαν τρία πειράματα με 5, 10 και 50 κόμβους αντίστοιχα. Σε όλες τις περιπτώσεις διατηρήθηκαν σταθερές οι υπόλοιπες παράμετροι της προσομοίωσης, δηλαδή 500 εγγραφές και 100 range queries ανά κόμβο, καθώς και 10 bins για την κατασκευή των ιστογραμμάτων. Με τον τρόπο αυτό ήταν δυνατή η απομόνωση της επίδρασης που έχει αποκλειστικά η αύξηση του αριθμού των κόμβων στη συμπεριφορά του συστήματος.

Nodes	Queries	Baseline Latency (s)	Proactive Latency (s)	MAE	Relative Error	Clusters	Selected Cluster
5	500	0.3082	0.0239	0.98	51.5%	100	11
10	1000	0.6588	0.0479	1.09	55.4%	200	15
50	5000	3.1569	0.2324	1.06	53.7%	1000	56

Πίνακας 1: Επίδραση του αριθμού edge nodes στις μετρικές απόδοσης.



Εικόνα 8: Μεταβολή του χρόνου εκτέλεσης σε σχέση με τον αριθμό των edge nodes.

Τα αποτελέσματα δείχνουν ότι η κάλυψη των δεδομένων παρέμεινε πλήρης (100%) σε όλες τις περιπτώσεις. Το γεγονός αυτό υποδηλώνει ότι η αύξηση του αριθμού των κόμβων δεν επηρεάζει την ικανότητα του συστήματος να καλύπτει τον χώρο αναζήτησης μέσω των παραγόμενων ερωτημάτων. Επομένως, η ποιότητα των workloads παραμένει σταθερή ακόμη και όταν αυξάνεται σημαντικά η κλίμακα της προσομοίωσης.

Ως προς τον χρόνο εκτέλεσης, παρατηρείται ότι τόσο η ακριβής όσο και η προσεγγιστική αξιολόγηση αυξάνονται σχεδόν γραμμικά με τον αριθμό των κόμβων. Συγκεκριμένα, ο χρόνος της ακριβούς αξιολόγησης αυξήθηκε από 0.3082s στους 5 κόμβους σε 0.6588s στους 10 κόμβους και σε 3.1569s στους 50 κόμβους. Αντίστοιχα, ο χρόνος της προσεγγιστικής αξιολόγησης αυξήθηκε από 0.0239s σε 0.0479s και 0.2324s. Η συμπεριφορά αυτή είναι αναμενόμενη, καθώς η αύξηση του αριθμού των κόμβων συνεπάγεται μεγαλύτερο πλήθος δεδομένων και ερωτημάτων προς επεξεργασία.

Παρόλα αυτά, η προσεγγιστική μέθοδος διατηρεί σημαντικό πλεονέκτημα έναντι της ακριβούς αξιολόγησης σε όλες τις περιπτώσεις. Στους 5 κόμβους η επιτάχυνση είναι περίπου 12.9 φορές, στους 10 κόμβους περίπου 13.7 φορές και στους 50 κόμβους περίπου 13.6 φορές. Το αποτέλεσμα αυτό δείχνει ότι η χρήση summaries παραμένει αποτελεσματική ακόμη και όταν το σύστημα επεκτείνεται σε μεγαλύτερο αριθμό edge nodes.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι μετρικές ακρίβειας. Το μέσο απόλυτο σφάλμα κυμάνθηκε μεταξύ 0.98 και 1.09, ενώ το μέσο σχετικό σφάλμα παρέμεινε κοντά στο 50%-55% σε όλα τα πειράματα. Η μικρή μεταβολή των τιμών αυτών δείχνει ότι η αύξηση του αριθμού των κόμβων δεν επηρεάζει σημαντικά την ποιότητα των προσεγγιστικών εκτιμήσεων. Με άλλα λόγια, η μέθοδος διατηρεί παρόμοιο επίπεδο ακρίβειας ανεξάρτητα από το μέγεθος της προσομοίωσης.

Αντίστοιχη τάση παρατηρείται και στη διαδικασία clustering. Ο αριθμός των clusters αυξήθηκε από 100 στους 5 κόμβους σε 200 στους 10 κόμβους και σε 1000 στους 50 κόμβους. Παράλληλα, αυξήθηκε και ο αριθμός των clusters που επιλέχθηκαν για proactive execution. Το γεγονός αυτό υποδηλώνει ότι όσο μεγαλώνει το ιστορικό των ερωτημάτων, τόσο περισσότερα επαναλαμβανόμενα πρότυπα μπορούν να εντοπιστούν και να αξιοποιηθούν από το σύστημα για προδραστική ανάλυση.

Συνολικά, τα αποτελέσματα δείχνουν ότι η προτεινόμενη προσέγγιση κλιμακώνεται ικανοποιητικά ως προς τον αριθμό των edge nodes. Παρότι ο συνολικός χρόνος επεξεργασίας αυξάνεται λόγω του μεγαλύτερου όγκου δεδομένων και ερωτημάτων, η χρήση summaries και η αξιοποίηση των ιστορικών queries μέσω clustering επιτρέπουν τη διατήρηση σημαντικού πλεονεκτήματος απόδοσης χωρίς ουσιαστική υποβάθμιση της ακρίβειας. Το εύρημα αυτό ενισχύει την καταλληλότητα της προτεινόμενης μεθόδου για εφαρμογή σε μεγαλύτερα περιβάλλοντα Edge Computing.

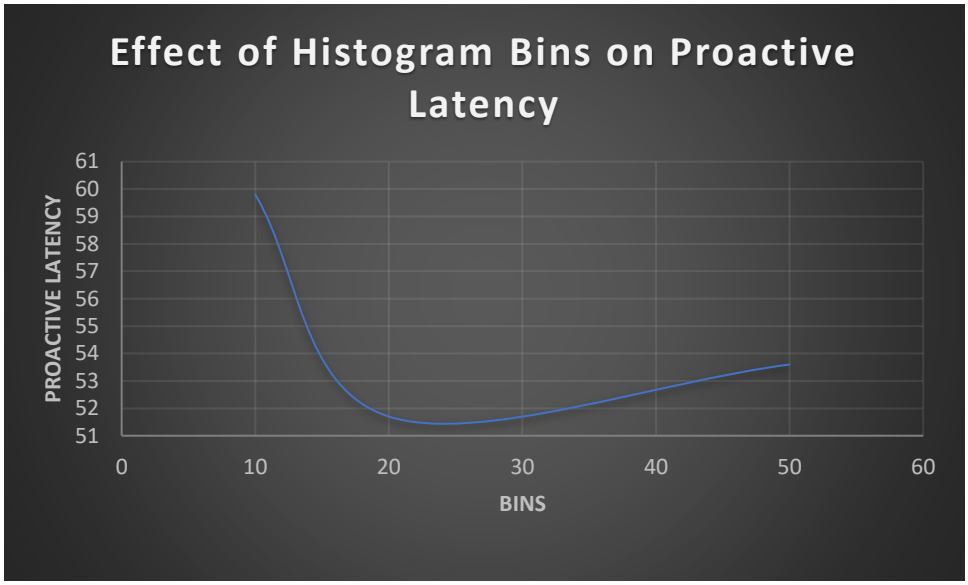
5.5.2 Επίδραση του Αριθμού Histogram Bins

Για τη διερεύνηση της επίδρασης των histogram bins στην απόδοση του συστήματος πραγματοποιήθηκαν τρία πειράματα με 10, 20 και 50 bins αντίστοιχα. Σε όλες τις περιπτώσεις διατηρήθηκαν σταθερές οι υπόλοιπες παράμετροι της προσομοίωσης, δηλαδή πέντε edge nodes, 500 εγγραφές ανά κόμβο και 100 range queries ανά κόμβο. Στόχος του πειράματος ήταν

να αξιολογηθεί πώς η λεπτομέρεια των summaries επηρεάζει τόσο την ακρίβεια των προσεγγιστικών εκτιμήσεων όσο και το υπολογιστικό κόστος της διαδικασίας.

Bins	Baseline Latency (s)	Summary Overhead (s)	MAE	Relative Error	Relative Error	Selected Clusters	Selected Queries	Selected Cluster MAE	Selected Cluster Relative Error
10	0.3036	0.0237	0.0029	1.20	59.8%	6	70	1.90	66.5%
20	0.3086	0.0343	0.0025	0.99	51.7%	4	65	1.03	44.8%
50	0.3167	0.0667	0.0031	1.01	53.6%	4	65	0.89	45.7%

Πίνακας 2: Επίδραση του αριθμού histogram bins στις μετρικές ακρίβειας και απόδοσης.



Εικόνα 9: Μεταβολή του σχετικού σφάλματος προσέγγισης ως προς τον αριθμό των histogram bins.

Τα αποτελέσματα έδειξαν ότι η κάλυψη των δεδομένων παρέμεινε πλήρης (100%) σε όλες τις περιπτώσεις, γεγονός που σημαίνει ότι η αλλαγή του αριθμού των bins δεν επηρεάζει την εξερεύνηση του χώρου αναζήτησης. Συνεπώς, οι διαφορές που παρατηρούνται στις μετρικές απόδοσης οφείλονται αποκλειστικά στη διαφορετική ανάλυση των ιστογραμμάτων.

Όσον αφορά την ακρίβεια της προσεγγιστικής ανάλυσης, παρατηρείται σημαντική βελτίωση κατά τη μετάβαση από 10 σε 20 bins. Το μέσο απόλυτο σφάλμα μειώθηκε από 1.20 σε 0.99, ενώ το μέσο σχετικό σφάλμα μειώθηκε από 59.8% σε 51.7%. Το αποτέλεσμα αυτό δείχνει ότι η χρήση περισσότερων bins επιτρέπει λεπτομερέστερη αναπαράσταση της κατανομής των δεδομένων και συνεπώς ακριβέστερη εκτίμηση των αποτελεσμάτων των queries.

Η περαιτέρω αύξηση στα 50 bins δεν οδήγησε σε αντίστοιχα σημαντική βελτίωση. Το μέσο απόλυτο σφάλμα παρέμεινε σχεδόν αμετάβλητο (1.01), ενώ το σχετικό σφάλμα αυξήθηκε ελαφρώς στο 53.6%. Η παρατήρηση αυτή υποδηλώνει ότι μετά από ένα συγκεκριμένο σημείο η αύξηση του αριθμού των bins δεν προσφέρει ουσιαστικό όφελος στην ακρίβεια, καθώς η πληροφορία που προστίθεται στα summaries είναι περιορισμένη σε σχέση με το κόστος που απαιτείται για τη διαχείρισή τους.

Παρόμοια εικόνα προκύπτει και από τις μετρικές των επιλεγμένων clusters. Με 10 bins το μέσο απόλυτο σφάλμα των selected queries ήταν 1.90, ενώ με 20 bins μειώθηκε σημαντικά στο 1.03. Στην περίπτωση των 50 bins το σφάλμα μειώθηκε περαιτέρω στο 0.89, όμως η βελτίωση είναι αισθητά μικρότερη σε σχέση με αυτή που παρατηρήθηκε κατά τη μετάβαση από 10 σε 20 bins. Αντίστοιχα, το σχετικό σφάλμα μειώθηκε από 66.5% σε περίπου 45%, παρουσιάζοντας πλέον τάση σταθεροποίησης.

Ως προς το υπολογιστικό κόστος, ο χρόνος της ακριβούς αξιολόγησης παρέμεινε σχεδόν σταθερός σε όλες τις περιπτώσεις, καθώς δεν εξαρτάται από τη δομή των histograms. Αντίθετα, ο χρόνος της προσεγγιστικής αξιολόγησης αυξήθηκε σταδιακά από 0.0237s στα 10 bins σε 0.0343s στα 20 bins και σε 0.0667s στα 50 bins. Η αύξηση αυτή είναι αναμενόμενη, καθώς η ύπαρξη περισσότερων bins οδηγεί σε μεγαλύτερο όγκο πληροφορίας που πρέπει να επεξεργαστεί το σύστημα κατά την εκτίμηση των αποτελεσμάτων.

Επιπλέον, παρατηρείται ότι όσο αυξάνεται ο αριθμός των bins μειώνεται ο αριθμός των clusters που τελικά επιλέγονται για proactive execution. Συγκεκριμένα, από έξι επιλεγμένα clusters στις περιπτώσεις των 10 και 20 bins, ο αριθμός μειώθηκε σε τέσσερα clusters στα 50 bins. Παράλληλα μειώθηκε και το πλήθος των selected queries, γεγονός που υποδηλώνει ότι τα πιο λεπτομερή histograms οδηγούν σε αυστηρότερη επιλογή προτύπων ερωτημάτων.

Συνολικά, τα αποτελέσματα δείχνουν ότι η επιλογή του αριθμού bins επηρεάζει σημαντικά την ισορροπία μεταξύ ακρίβειας και απόδοσης. Η χρήση 10 bins οδηγεί στη γρηγορότερη εκτέλεση αλλά εμφανίζει μεγαλύτερο σφάλμα, ενώ τα 50 bins προσφέρουν οριακά καλύτερη ακρίβεια με αυξημένο κόστος επεξεργασίας. Η περίπτωση των 20 bins φαίνεται να αποτελεί την καλύτερη ισορροπία μεταξύ υπολογιστικού κόστους και ποιότητας εκτίμησης, προσφέροντας σημαντική βελτίωση της ακρίβειας χωρίς ιδιαίτερα μεγάλη επιβάρυνση στον χρόνο εκτέλεσης.

5.6 Συζήτηση Αποτελεσμάτων

Τα αποτελέσματα της πειραματικής αξιολόγησης επιβεβαιώνουν ότι η προτεινόμενη προσέγγιση μπορεί να αξιοποιήσει ιστορικά ερωτήματα για την εξαγωγή προτύπων πρόσβασης στα δεδομένα και την υποστήριξη προληπτικής ανάλυσης σε περιβάλλον Edge Computing.

Αρχικά, η ανάλυση των KPI metrics έδειξε ότι η ομαδοποίηση των ερωτημάτων μέσω clustering επιτρέπει τον εντοπισμό περιοχών του πολυδιάστατου χώρου δεδομένων που παρουσιάζουν αυξημένη πιθανότητα μελλοντικής πρόσβασης. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για την προληπτική εκτέλεση αναλύσεων ή τη διατήρηση ενδιάμεσων αποτελεσμάτων κοντά στους edge nodes, συμβάλλοντας στη μείωση του χρόνου απόκρισης και της επιβάρυνσης του δικτύου.

Παράλληλα, η αξιολόγηση των clusters έδειξε ότι τα παραγόμενα πρότυπα ερωτημάτων είναι αντιπροσωπευτικά της συμπεριφοράς των χρηστών, καθώς οι συχνότερες περιοχές αναζήτησης εντοπίζονται με συνέπεια σε διαφορετικά σύνολα ερωτημάτων. Η παρατήρηση αυτή υποστηρίζει την υπόθεση ότι η ιστορική δραστηριότητα μπορεί να αποτελέσει αξιόπιστη πηγή πληροφορίας για την πρόβλεψη μελλοντικών αναγκών.

Η ανάλυση των query patterns αποκάλυψε ότι συγκεκριμένες περιοχές του χώρου δεδομένων εμφανίζονται σημαντικά συχνότερα από άλλες, γεγονός που δικαιολογεί τη χρήση μηχανισμών προληπτικής επεξεργασίας. Αντί να αντιμετωπίζονται όλα τα πιθανά ερωτήματα με τον ίδιο τρόπο, το σύστημα μπορεί να εστιάζει σε περιοχές υψηλής πιθανότητας, επιτυγχάνοντας καλύτερη αξιοποίηση των διαθέσιμων πόρων.

Τα πειράματα ευαισθησίας παραμέτρων παρείχαν επιπλέον στοιχεία σχετικά με τη συμπεριφορά του συστήματος. Η αύξηση του αριθμού των edge nodes βελτίωσε την κατανομή του φόρτου εργασίας και επέτρεψε μεγαλύτερο βαθμό παραλληλισμού, οδηγώντας σε καλύτερη κλιμάκωση του συστήματος. Ωστόσο, η βελτίωση αυτή δεν ήταν απεριόριστη, καθώς μετά από ένα σημείο το όφελος από την προσθήκη νέων κόμβων μειώνεται λόγω του πρόσθετου κόστους συντονισμού.

Παρόμοια συμπεριφορά παρατηρήθηκε και στην ανάλυση του αριθμού των histogram bins. Μικρός αριθμός bins οδηγεί σε απώλεια λεπτομέρειας και λιγότερο ακριβή αναπαράσταση των προτύπων ερωτημάτων, ενώ υπερβολικά μεγάλος αριθμός αυξάνει την πολυπλοκότητα και το κόστος επεξεργασίας χωρίς αντίστοιχη βελτίωση στην ποιότητα των αποτελεσμάτων. Επομένως, απαιτείται μια ισορροπημένη επιλογή της παραμέτρου ώστε να επιτυγχάνεται κατάλληλος συμβιβασμός μεταξύ ακρίβειας και αποδοτικότητας.

Συνολικά, τα αποτελέσματα υποδεικνύουν ότι η προτεινόμενη μεθοδολογία είναι σε θέση να αναγνωρίζει χρήσιμα πρότυπα πρόσβασης στα δεδομένα και να υποστηρίζει αποτελεσματικά μηχανισμούς proactive data analysis σε κατανεμημένα edge περιβάλλοντα.

5.7 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκε η πειραματική αξιολόγηση της προτεινόμενης μεθοδολογίας proactive data analysis σε περιβάλλον Edge Computing. Αρχικά περιγράφηκε το πειραματικό περιβάλλον και η διαδικασία δημιουργίας και εκτέλεσης των ερωτημάτων. Στη συνέχεια παρουσιάστηκαν τα αποτελέσματα που προέκυψαν από την ανάλυση των KPI metrics, την εφαρμογή τεχνικών clustering και την εξαγωγή συχνών προτύπων ερωτημάτων.

Επιπλέον, πραγματοποιήθηκε ανάλυση ευαισθησίας βασικών παραμέτρων του συστήματος, εξετάζοντας την επίδραση του αριθμού των edge nodes και του αριθμού των histogram bins στη συνολική απόδοση. Τα αποτελέσματα έδειξαν ότι οι συγκεκριμένες παράμετροι επηρεάζουν σημαντικά τόσο την ποιότητα των παραγόμενων προτύπων όσο και την αποδοτικότητα της επεξεργασίας.

Η συνολική αξιολόγηση κατέδειξε ότι η αξιοποίηση ιστορικών ερωτημάτων μπορεί να συμβάλει στον εντοπισμό περιοχών αυξημένου ενδιαφέροντος και να υποστηρίξει την προληπτική εκτέλεση αναλύσεων, μειώνοντας το υπολογιστικό κόστος και βελτιώνοντας την απόκριση του συστήματος.

Στο επόμενο κεφάλαιο παρουσιάζονται τα τελικά συμπεράσματα της εργασίας, οι βασικές συνεισφορές της προτεινόμενης προσέγγισης, οι περιορισμοί που εντοπίστηκαν κατά την υλοποίηση και πιθανές κατευθύνσεις για μελλοντική έρευνα.

ΚΕΦΑΛΑΙΟ 6 Συμπεράσματα

6.1 Συμπεράσματα

Η παρούσα εργασία ασχολήθηκε με το πρόβλημα της προληπτικής εξαγωγής χρήσιμης πληροφορίας από ιστορικά ερωτήματα σε περιβάλλον Edge Computing. Η συνεχής αύξηση του όγκου των δεδομένων και η ανάγκη για άμεση επεξεργασία κοντά στην πηγή παραγωγής τους καθιστούν αναγκαία την ανάπτυξη μηχανισμών που μπορούν να προβλέπουν μελλοντικές ανάγκες πρόσβασης και ανάλυσης δεδομένων.

Στο πλαίσιο αυτό σχεδιάστηκε και υλοποιήθηκε μία μεθοδολογία η οποία βασίζεται στη συλλογή ιστορικών ερωτημάτων από κατανεμημένους edge nodes, στη μετατροπή τους σε κατάλληλη αριθμητική αναπαράσταση και στην εφαρμογή τεχνικών ομαδοποίησης για τον εντοπισμό περιοχών υψηλού ενδιαφέροντος στον πολυδιάστατο χώρο δεδομένων. Η προσέγγιση υλοποιήθηκε στο περιβάλλον προσομοίωσης EdgeSimPy, επιτρέποντας τη δημιουργία ενός ελεγχόμενου πειραματικού περιβάλλοντος για την αξιολόγηση της αποτελεσματικότητάς της.

Τα πειραματικά αποτελέσματα έδειξαν ότι η ανάλυση ιστορικών ερωτημάτων μπορεί να αποκαλύψει επαναλαμβανόμενα πρότυπα πρόσβασης, τα οποία μπορούν να αξιοποιηθούν για την υποστήριξη proactive data analysis. Η χρήση τεχνικών clustering επέτρεψε τον εντοπισμό συχνά εμφανιζόμενων περιοχών αναζήτησης, επιβεβαιώνοντας ότι η συμπεριφορά των χρηστών δεν είναι τυχαία αλλά παρουσιάζει επαναληπτικά χαρακτηριστικά που μπορούν να αξιοποιηθούν για την πρόβλεψη μελλοντικών αναγκών.

Επιπλέον, η αξιολόγηση των βασικών δεικτών απόδοσης (KPIs) έδειξε ότι η προτεινόμενη μεθοδολογία μπορεί να συμβάλει στη βελτίωση της διαχείρισης των υπολογιστικών πόρων, μειώνοντας την ανάγκη για επαναλαμβανόμενη επεξεργασία ίδιων ή παρόμοιων περιοχών δεδομένων. Με τον τρόπο αυτό καθίσταται δυνατή η καλύτερη αξιοποίηση των διαθέσιμων πόρων του edge περιβάλλοντος και η μείωση του χρόνου απόκρισης προς τους τελικούς χρήστες.

Ιδιαίτερα σημαντικά ήταν και τα αποτελέσματα της ανάλυσης ευαισθησίας των παραμέτρων του συστήματος. Η μεταβολή του αριθμού των edge nodes επηρέασε άμεσα την κατανομή του φόρτου εργασίας και την κλιμάκωση της επεξεργασίας, ενώ η μεταβολή του αριθμού των histogram bins επηρέασε την ποιότητα της αναπαράστασης των ερωτημάτων και κατ’ επέκταση την ακρίβεια της ανίχνευσης προτύπων. Τα αποτελέσματα αυτά καταδεικνύουν ότι η σωστή παραμετροποίηση του συστήματος αποτελεί κρίσιμο παράγοντα για τη βέλτιστη λειτουργία της προτεινόμενης προσέγγισης.

Συνολικά, η εργασία επιβεβαιώνει ότι η αξιοποίηση ιστορικών ερωτημάτων μπορεί να αποτελέσει αποτελεσματική βάση για την ανάπτυξη μηχανισμών πρόβλεψης και προληπτικής ανάλυσης σε κατανεμημένα περιβάλλοντα Edge Computing. Η προτεινόμενη μεθοδολογία παρουσιάζει θετικά αποτελέσματα ως προς την αναγνώριση προτύπων πρόσβασης και δημιουργεί τις προϋποθέσεις για πιο αποδοτική διαχείριση δεδομένων και υπολογιστικών πόρων σε μελλοντικά συστήματα.

6.2 Περιορισμοί της Εργασίας

Παρά τα θετικά αποτελέσματα, η παρούσα εργασία παρουσιάζει ορισμένους περιορισμούς που πρέπει να ληφθούν υπόψη κατά την αξιολόγηση των συμπερασμάτων της.

Αρχικά, τα πειράματα πραγματοποιήθηκαν σε περιβάλλον προσομοίωσης και όχι σε πραγματική υποδομή Edge Computing. Αν και το EdgeSimPy παρέχει ρεαλιστική μοντελοποίηση των κόμβων και των πόρων του συστήματος, ορισμένοι παράγοντες που εμφανίζονται σε πραγματικές εγκαταστάσεις, όπως οι διακυμάνσεις του δικτύου, οι αστοχίες εξοπλισμού και οι απρόβλεπτες μεταβολές του φόρτου εργασίας, δεν μπορούν να αναπαρασταθούν πλήρως.

Επιπλέον, η δημιουργία των ερωτημάτων βασίστηκε σε συνθετικά δεδομένα που παράχθηκαν από στατιστικά χαρακτηριστικά του χρησιμοποιούμενου συνόλου δεδομένων. Παρόλο που η προσέγγιση αυτή επιτρέπει ελεγχόμενα πειράματα, η συμπεριφορά πραγματικών χρηστών ενδέχεται να παρουσιάζει μεγαλύτερη πολυπλοκότητα και δυναμική μεταβολή.

ΚΕΦΑΛΑΙΟ 7 Μελλοντικές Προεκτάσεις

Η παρούσα εργασία αποτελεί ένα πρώτο βήμα προς την αξιοποίηση ιστορικών ερωτημάτων για την υποστήριξη μηχανισμών Proactive Data Analysis σε περιβάλλοντα edge computing. Ωστόσο, υπάρχουν αρκετές δυνατότητες επέκτασης και περαιτέρω διερεύνησης της προτεινόμενης προσέγγισης.

7.1 Χρήση Μεγαλύτερων Συνόλων Δεδομένων

Μία σημαντική κατεύθυνση για μελλοντική έρευνα είναι η χρήση μεγαλύτερων ή πραγματικών συνόλων δεδομένων (datasets), ώστε η αξιολόγηση να πραγματοποιηθεί σε πιο ρεαλιστικά περιβάλλοντα και με μεγαλύτερο όγκο δεδομένων. Στην παρούσα εργασία χρησιμοποιήθηκε το Iris dataset, το οποίο προσφέρει ένα ελεγχόμενο περιβάλλον για τη μελέτη της συμπεριφοράς του προτεινόμενου συστήματος. Ωστόσο, πραγματικές εφαρμογές Edge Computing παράγουν δεδομένα πολύ μεγαλύτερου όγκου και πολυπλοκότητας.

Η αξιολόγηση της προτεινόμενης προσέγγισης σε δεδομένα προερχόμενα από αισθητήρες IoT, έξυπνες πόλεις, βιομηχανικά συστήματα ή εφαρμογές παρακολούθησης θα επέτρεπε τη διερεύνηση της συμπεριφοράς του αλγορίθμου σε συνθήκες που προσεγγίζουν περισσότερο την πραγματικότητα. Επιπλέον, η χρήση μεγαλύτερων datasets θα μπορούσε να αναδείξει πιθανούς περιορισμούς κλιμάκωσης και να οδηγήσει σε περαιτέρω βελτιώσεις της προτεινόμενης μεθόδου.

7.2 Εφαρμογή σε Πραγματικά Περιβάλλοντα Edge Computing

Μία επιπλέον επέκταση αφορά την ενσωμάτωση του αλγορίθμου σε πραγματικούς edge simulators ή διανεμημένες πλατφόρμες υπολογιστικού νέφους και edge computing. Στην παρούσα εργασία χρησιμοποιήθηκε ένα τοπικό περιβάλλον προσομοίωσης, το οποίο επέτρεψε τον πλήρη έλεγχο των παραμέτρων και τη συστηματική αξιολόγηση της προτεινόμενης προσέγγισης.

Η εφαρμογή του μοντέλου σε πραγματικές υποδομές θα επέτρεπε τη μελέτη παραγόντων που δεν είναι εύκολο να αναπαρασταθούν σε ένα απλοποιημένο περιβάλλον προσομοίωσης, όπως οι καθυστερήσεις δικτύου, η επικοινωνία μεταξύ κόμβων, η μεταβολή του διαθέσιμου εύρους ζώνης και η δυναμική κατανομή πόρων. Η αξιολόγηση σε τέτοιες συνθήκες θα παρείχε πιο ολοκληρωμένη εικόνα για τη δυνατότητα πρακτικής εφαρμογής του συστήματος.

7.3 Σύγκριση με Εναλλακτικές Τεχνικές Clustering

Η παρούσα εργασία αξιοποιεί subtractive clustering για την ομαδοποίηση ιστορικών ερωτημάτων, καθώς η συγκεκριμένη τεχνική δεν απαιτεί προκαθορισμένο αριθμό clusters και προσαρμόζεται δυναμικά στα χαρακτηριστικά των δεδομένων. Παρόλα αυτά, στη βιβλιογραφία υπάρχουν αρκετοί ακόμη αλγόριθμοι clustering που θα μπορούσαν να χρησιμοποιηθούν για τον ίδιο σκοπό.

Ενδεικτικά, οι αλγόριθμοι K-Means και DBSCAN παρουσιάζουν διαφορετικά πλεονεκτήματα και μειονεκτήματα ως προς την ακρίβεια, την πολυπλοκότητα και την ανθεκτικότητα στον θόρυβο. Μία συγκριτική αξιολόγηση των τεχνικών αυτών θα μπορούσε να αναδείξει ποια μέθοδος είναι καταλληλότερη για τον εντοπισμό επαναλαμβανόμενων προτύπων σε ιστορικά workloads. Επιπλέον, θα μπορούσαν να εξεταστούν και πιο σύγχρονες προσεγγίσεις βασισμένες σε τεχνικές μηχανικής μάθησης και βαθιάς μάθησης.

7.4 Πρόβλεψη Ερωτημάτων και Caching

Μία ακόμη ιδιαίτερα ενδιαφέρουσα κατεύθυνση είναι η ανάπτυξη μηχανισμών πρόβλεψης μελλοντικών queries και αποθήκευσης αποτελεσμάτων (caching). Η βασική ιδέα είναι ότι τα ιστορικά ερωτήματα μπορούν να χρησιμοποιηθούν για την εκτίμηση των πιθανότερων περιοχών του χώρου δεδομένων που θα αναζητηθούν στο μέλλον.

Με βάση τις προβλέψεις αυτές, το σύστημα θα μπορούσε να υπολογίζει εκ των προτέρων αποτελέσματα ή να διατηρεί προσωρινά δεδομένα σε τοπικές μνήμες cache. Έτσι, παρόμοια ερωτήματα θα μπορούσαν να απαντώνται άμεσα χωρίς επαναλαμβανόμενη επεξεργασία των δεδομένων. Η προσέγγιση αυτή αναμένεται να μειώσει περαιτέρω τον χρόνο απόκρισης και να ενισχύσει τον προδραστικό χαρακτήρα του συστήματος.

7.5 Σύνοψη Κεφαλαίου

Οι παραπάνω προτάσεις αναδεικνύουν τις δυνατότητες περαιτέρω βελτίωσης και επέκτασης της προτεινόμενης προσέγγισης. Η χρήση μεγαλύτερων και πιο ρεαλιστικών συνόλων δεδομένων, η εφαρμογή σε πραγματικά περιβάλλοντα edge computing, η διερεύνηση εναλλακτικών τεχνικών clustering και η ανάπτυξη μηχανισμών πρόβλεψης ερωτημάτων αποτελούν φυσικές επεκτάσεις της παρούσας εργασίας.

Τα αποτελέσματα που παρουσιάστηκαν στα προηγούμενα κεφάλαια δείχνουν ότι η αξιοποίηση ιστορικών ερωτημάτων μπορεί να συμβάλει ουσιαστικά στη μείωση του υπολογιστικού κόστους και στη βελτίωση της αποδοτικότητας των edge συστημάτων. Οι προτεινόμενες μελλοντικές επεκτάσεις μπορούν να αποτελέσουν τη βάση για την ανάπτυξη πιο ολοκληρωμένων και ευφυών συστημάτων Proactive Data Analysis, ικανών να λειτουργούν αποτελεσματικά σε πραγματικά περιβάλλοντα μεγάλης κλίμακας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). *Edge Computing: Vision and Challenges*. IEEE Internet of Things Journal.
2. Satyanarayanan, M. (2017). *The Emergence of Edge Computing*. Computer.
3. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*.
4. Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining*. Pearson.
5. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
6. Satyanarayanan, M., Bahl, P., Caceres, R., & Davies, N. (2009). *The Case for VM-Based Cloudlets in Mobile Computing*. IEEE Pervasive Computing.
7. Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). *Fog Computing and Its Role in the Internet of Things*. Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing.
8. Dean, J., & Ghemawat, S. (2008). *MapReduce: Simplified Data Processing on Large Clusters*. Communications of the ACM.
9. Chaudhuri, S., Motwani, R., & Narasayya, V. (1998). *Random Sampling for Histogram Construction: How Much Is Enough?* ACM SIGMOD Record.
10. Garofalakis, M., Gehrke, J., & Rastogi, R. (2002). *Querying and Mining Data Streams: You Only Get One Look*. Proceedings of the ACM SIGMOD Conference.
11. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN)*. Proceedings of KDD.
12. MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.
13. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
14. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of Massive Datasets* (3rd ed.). Cambridge University Press.
15. Cormode, G., & Garofalakis, M. (2005). *Sketching Streams Through the Net: Distributed Approximate Query Tracking*. Proceedings of the VLDB Conference.
16. Varghese, B., Wang, N., Barbhuiya, S., Kilpatrick, P., & Nikolopoulos, D. (2016). Challenges and Opportunities in Edge Computing. IEEE International Conference on Smart Cloud.
17. Mach, P., & Becvar, Z. (2017). Mobile Edge Computing: A Survey on Architecture and Computation Offloading. IEEE Communications Surveys & Tutorials, 19(3), 1628–1656.
18. Abbas, N., Zhang, Y., Taherkordi, A., & Skeie, T. (2018). Mobile Edge Computing: A Survey. IEEE Internet of Things Journal, 5(1), 450–465.
19. Premsankar, G., Di Francesco, M., & Taleb, T. (2018). Edge Computing for the Internet of Things: A Case Study. IEEE Internet of Things Journal, 5(2), 1275–1284.
20. Guttman, A. (1984). R-Trees: A Dynamic Index Structure for Spatial Searching. Proceedings of the ACM SIGMOD International Conference on Management of Data, 47–57.
21. Beckmann, N., Kriegel, H. P., Schneider, R., & Seeger, B. (1990). The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. Proceedings of the ACM SIGMOD International Conference on Management of Data, 322–331.

22. Jagadish, H. V., Ooi, B. C., Tan, K. L., Yu, C., & Zhang, R. (2005). iDistance: An Adaptive B+-Tree Based Indexing Method for Nearest Neighbor Search. ACM Transactions on Database Systems, 30(2), 364–397.
23. Papadias, D., Tao, Y., Fu, G., & Seeger, B. (2003). An Optimal and Progressive Algorithm for Skyline Queries. Proceedings of the ACM SIGMOD International Conference on Management of Data, 467–478.
24. Poosala, V., Ioannidis, Y. E., Haas, P. J., & Shekita, E. J. (1996). Improved Histograms for Selectivity Estimation of Range Predicates. Proceedings of the ACM SIGMOD International Conference on Management of Data, 294–305.
25. Ioannidis, Y. E. (2003). The History of Histograms. Proceedings of the International Conference on Very Large Data Bases (VLDB), 19–30.
26. Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997). Online Aggregation. Proceedings of the ACM SIGMOD International Conference on Management of Data, 171–182.
27. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Surveys, 31(3), 264–323.
28. Chiu, S. L. (1994). Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems, 2(3), 267–278.
29. Yager, R. R., & Filev, D. P. (1994). Generation of Fuzzy Rules by Mountain Clustering. Journal of Intelligent and Fuzzy Systems, 2(3), 209–219.
30. Agrawal, R., Chaudhuri, S., & Narasayya, V. (2000). Automated Selection of Materialized Views and Indexes in SQL Databases. Proceedings of the International Conference on Very Large Data Bases (VLDB), 496–505.
31. Bruno, N., & Chaudhuri, S. (2005). Automatic Physical Database Tuning: A Relaxation-Based Approach. Proceedings of the ACM SIGMOD International Conference on Management of Data, 227–238.
32. Chaudhuri, S., & Narasayya, V. (2007). Self-Tuning Database Systems: A Decade of Progress. Proceedings of the International Conference on Very Large Data Bases (VLDB), 3–14.
33. Bao, Y., Marcus, R., Li, M., et al. (2021). Bao: Making Learned Query Optimization Practical. Proceedings of the ACM SIGMOD International Conference on Management of Data, 1275–1288.
34. Hellerstein, J. M., Stonebraker, M., & Hamilton, J. (2007). Architecture of a Database System. Foundations and Trends in Databases, 1(2), 141–259.
35. Deshpande, A., Garofalakis, M., & Rastogi, R. (2004). Independence is Good: Dependency-Based Histogram Synopses for High-Dimensional Data. Proceedings of the ACM SIGMOD International Conference on Management of Data, 199–210.
36. Gibbons, P. B., Matias, Y., & Poosala, V. (1997). Fast Incremental Maintenance of Approximate Histograms. Proceedings of the VLDB Conference, 466–475.
37. Acharya, S., Gibbons, P. B., Poosala, V., & Ramaswamy, S. (1999). Join Synopses for Approximate Query Answering. Proceedings of the ACM SIGMOD International Conference on Management of Data, 275–286.
38. Cormode, G., Korn, F., Muthukrishnan, S., & Srivastava, D. (2012). Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. Foundations and Trends in Databases, 4(1–3), 1–294.
39. Olston, C., Jiang, J., & Widom, J. (2003). Adaptive Filters for Continuous Queries over Distributed Data Streams. Proceedings of the ACM SIGMOD International Conference on Management of Data, 563–574.
40. Madden, S., Franklin, M. J., Hellerstein, J. M., & Hong, W. (2002). TAG: A Tiny AGgregation Service for Ad-Hoc Sensor Networks. Proceedings of OSDI, 131–146.

41. Bonfils, B. J., & Bonnet, P. (2003). Adaptive and Decentralized Operator Placement for In-Network Query Processing. Proceedings of Information Processing in Sensor Networks (IPSN), 47–62.
42. Yu, H., Lim, E. P., & Lauw, H. W. (2009). Predicting Emerging Query Patterns. Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), 305–314.
43. Cao, H., Jiang, D., Pei, J., He, Q., Liao, H., Chen, E., & Li, H. (2008). Context-Aware Query Suggestion by Mining Click-Through and Session Data. Proceedings of KDD, 875–883.
44. Li, F., Yi, K., & Jestes, J. (2012). Ranking Distributed Probabilistic Data. Proceedings of the ACM SIGMOD International Conference on Management of Data, 361–372.
45. Kraska, T., Beutel, A., Chi, E. H., Dean, J., & Polyzotis, N. (2018). The Case for Learned Index Structures. Proceedings of the ACM SIGMOD International Conference on Management of Data, 489–504.
46. Marcus, R., Negi, P., Mao, H., Alizadeh, M., Kraska, T., Papaemmanouil, O., & Tatbul, N. (2019). Neo: A Learned Query Optimizer. Proceedings of the VLDB Endowment, 12(11), 1705–1718.
47. Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013). Discretized Streams: Fault-Tolerant Streaming Computation at Scale. Proceedings of SOSP, 423–438.
48. Abadi, D. J., Ahmad, Y., Balazinska, M., Cetintemel, U., Cherniack, M., Hwang, J. H., Lindner, W., Maskey, A., Rasin, A., Ryvkina, E., Tatbul, N., Xing, Y., & Zdonik, S. (2005). The Design of the Borealis Stream Processing Engine. Proceedings of CIDR.
49. Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The 8 Requirements of Real-Time Stream Processing. ACM SIGMOD Record, 34(4), 42–47.
50. Keralapura, R., Cormode, G., & Ramamirtham, J. (2006). Communication-Efficient Distributed Monitoring of Thresholded Counts. Proceedings of the ACM SIGMOD International Conference on Management of Data, 289–300.