Ατομική Διπλωματική Εργασία

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ ΠΛΑΙΣΙΟΥ ΓΙΑ ΑΜΝΗΣΙΑ ΜΕΓΑΛΩΝ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ

Ανδρέας Χαραλάμπους

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ ΠΛΑΙΣΙΟΥ ΓΙΑ ΑΜΝΗΣΙΑ ΜΕΓΑΛΩΝ ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ

Ανδρέας Χαραλάμπους

Επιβλέπων Καθηγητής
Δημήτρης Ζεϊναλιπούρ

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2018

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον Δρ. Δημήτρη Ζεϊναλιπούρ, έναν αξιόλογο καθηγητή με τεράστιο όγκο γνώσεων σε πολλούς τομείς της πληροφορικής, που μου έδωσε την ευκαιρία και μου πρόσφερε την συγκεκριμένη διπλωματική εργασία η οποία ήταν πολύ ενδιαφέρουσα και ελκυστική. Θα ήθελα επίσης να τον ευχαριστήσω για την τεράστια βοήθεια και υποστήριξη καθώς και την εμπιστοσύνη που έδειξε προς το πρόσωπό μου.

Επίσης θα ήθελα να ευχαριστήσω τον Κωνσταντίνο Κώστα, διδακτορικό φοιτητή του Πανεπιστημίου Κύπρου για την τεράστια βοήθεια, υποστήριξη, αφοσίωση, υπομονή και γενικώς για την όλη καθοδήγηση που μου έδωσε καθ' όλη την διάρκεια της υλοποίησης της διπλωματικής εργασίας καθώς και για την εξαιρετική συνεργασία που είχαμε αυτό τον καιρό.

Επιπλέον ευχαριστώ το Τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου αλλά και το Πανεπιστήμιο Κύπρου για την διάθεση όλων όσων χρειάστηκα για την εκπόνηση της διπλωματικής εργασίας.

Τέλος οφείλω ένα τεράστιο ευχαριστώ στους γονείς μου Κυριάκο και Έλλη για την οικονομική και ψυχολογική στήριξη τους καθ' όλη την διάρκεια των σπουδών μου αλλά και στους φίλους μου για την συμπαράσταση και την βοήθεια τους.

Αναγνώριση

Η παρούσα διπλωματική εργασία στηρίζεται σε γνώσεις που αποκομίστηκαν στα πλαίσια των δύο πιο κάτω επιστημονικών εργασιών τα οποία βρίσκονται υπό δημοσίευση.

1. "Decaying Telco Big Data with Data Postdiction", Constantinos Costa, Andreas Charalampous, Andreas Konstantinidis, Mohamed F. Mokbel, Proceedings of the 19th IEEE International Conference on Mobile Data Management (MDM '18), IEEE Computer Society, 10 pages, June 25 - June 28, 2018, AAU, Aalborg, Denmark, **2018**

2. "TBD-DP: Telco Big Data Visual Analytics with Data Postdiction", Constantinos Costa, Andreas Charalampous, Andreas Konstantinidis, Mohamed F. Mokbel, Proceedings of the 19th IEEE International Conference on Mobile Data Management (MDM '18), IEEE Computer Society, 2 pages, June 25 - June 28, 2018, AAU, Aalborg, Denmark, **2018**.

Περίληψη

Στις μέρες μας οι εταιρίες τηλεπικοινωνιών (telcos) δίνουν ιδιαίτερη προσοχή στην αποσύνθεση των μεγάλων δεδομένων (Big Data Decaying) και αυτό οφείλεται κατά κύριο λόγο στην τεράστια εξάπλωση και "έκρηξη" των δεδομένων που παράγονται μέσω του Διαδικτύου αλλά και μέσω των "έξυπνων συσκευών" (smart devices). Όπως είναι ήδη γνωστό τα συστήματα διαχείρισης μεγάλων δεδομένων, όπως για παράδειγμα οι κατανεμημένοι χώροι αποθήκευσης επικεντρώνονται κυρίως στην ελαχιστοποίηση του χρόνου απόκρισης ενός επερωτήματος και ταυτόχρονα στην μεγιστοποίηση της ακρίβειας (accuracy) της απόκρισης αγνοώντας κατά κύριο λόγο το κόστος αποθήκευσης και τις επιπτώσεις μια μακροχρόνιας διαχείρισης μεγάλων δεδομένων.

Η εν λόγω διπλωματική εργασία ασχολείται εν μέρει με την μελέτη, την υλοποίηση του τελεστή DP (Data-Postdiction operator) ο οποίος βελτιστοποιεί τον χώρο αποθήκευσης με την αποσύνθεση των μεγάλων τηλεπικοινωνιακών δεδομένων διατηρώντας μια βέλτιστη ακρίβεια. Ο DP χρησιμοποιεί "postdiction", με βάσει ένα LSTM-based αλγόριθμο εκμάθησης μηχανών (machine learning) με σκοπό την μείωση του χώρου αποθήκευσης, με την πάροδο του χρόνου, και μεγιστοποίηση του χρόνου απόκρισης και ακρίβειας των αποτελεσμάτων χρησιμοποιώντας τελευταίας τεχνολογίας τεχνικές αποσύνθεσης δεδομένων.

Με την διεξαγωγή πειραμάτων και την χρήση συγκεκριμένης μεθοδολογίας αποτιμάται η αποδοτικότητα του προαναφερθέντα τελεστή χρησιμοποιώντας ανώνυμα πραγματικά δεδομένα δικτύων. Με τα αποτελέσματα των πειραμάτων μπορεί να διαφανεί ότι ο προτεινόμενος τελεστής DP εξασφαλίζει την εξοικονόμηση του χώρου αποθήκευσης τάξης μεγέθους ενώ διατηρεί μια υψηλή ακρίβεια στις αποκρίσεις των διαφόρων επερωτημάτων.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Στα επερχόμενα υποκεφάλαια θα γίνει μια αναφορά στην υποκίνηση της εν λόγω διπλωματικής εργασίας δηλαδή ποιοι είναι οι κυριότεροι λόγοι που μας ώθησαν να ερευνήσουμε αυτό το τομέα όπου θα γίνει και μια εισαγωγή στην Αμνησία Δεδομένων. Επιπλέον θα γίνει μια αναφορά στο περίγραμμα εργασίας όπου θα παρουσιαστεί η δομή αυτής της διπλωματικής εργασίας. Τέλος, γίνεται μια παρουσίαση της συνεισφοράς της εργασίας.

1.2 Υποκίνηση

Η τεράστια και παράλληλα ταχεία εξάπλωση του διαδικτύου (Internet of Things) αλλά και των έξυπνων κινητών συσκευών δημιούργησαν μια συνεχώς αυξανόμενη ζήτηση όσον αφορά την γρήγορη επεξεργασία και αποθήκευση μεγάλων τηλεπικοινωνιακών δεδομένων (Telco Big Data). Συγκεκριμένα, λεπτό με λεπτό παράγονται τεράστιες ποσότητες τηλεπικοινωνιακών δεδομένων από τα τηλεπικοινωνιακά δίκτυα [1]. Για παράδειγμα όπως αναφέρουν ερευνητές [2] ένα τηλεπικοινωνιακό δίκτυο συλλέγει κατά μέσο όρο 5TBs δεδομένα την ημέρα που αφορούν 10 εκατομμύρια χρήστες του δικτύου.

Οι παραδοσιακοί χώροι αποθήκευσης και συστήματα διαχείρισης βάσεων δεδομένων δεν μπορούν να ανταποκριθούν σε αυτές τις υψηλές απαιτήσεις γι' αυτό και η αποδοτικότητα τους μειώνεται χρόνο με το χρόνο. Σε αντίθεση με τα κατανεμημένα συστήματα χώρου αποθήκευσης τα οποία μπορούν να διαχειριστούν αυτά τα δεδομένα καθώς αναπτύσσεται ο όγκος τους. Παρόλα αυτά ο απαιτούμενος χώρος αποθήκευσης και συνεπώς το κόστος αυτών για ένα κατανεμημένο σύστημα μεγάλων δεδομένων αυξάνεται συστηματικά με την αύξηση των δεδομένων. Για παράδειγμα ένα petabyte Hadoop συγκρότημα (cluster) κοστίζει κατά προσέγγιση 1 εκατομμύριο δολάρια χρησιμοποιώντας μεταξύ 125 και 250 κόμβους. Επιπλέον η ποσότητα αποθήκευσης αυξάνεται κάθε χρόνο κατά το διπλάσιο και το κόστος των μέσων αποθήκευσης μειώνεται σε ένα ποσοστό περίπου 20% τον χρόνο [3].

Αυτό αποτελεί ένα μεγάλο κίνητρο για την κοινότητα που ασχολείται με τα μεγάλα δεδομένα. Πιο συγκεκριμένα, η αποδοτική αποθήκευση μεγάλων δεδομένων με την χρήση ελάχιστου χώρου αποθήκευσης, παρέχοντας παράλληλα μια χαμηλή απόκριση σε χρόνο, υψηλή απόκριση σε ακρίβεια (accuracy) και υψηλή αξιοπιστία δια μέσου ανοχής σφάλματος μέσω της αναπαραγωγής δεδομένων [4].

Προηγούμενες ερευνητικές μελέτες εστιάζονται στην ελαχιστοποίηση του χρόνου απόκρισης χρησιμοποιώντας συνάθροιση σε μεγάλα δεδομένα (aggregation). Για παράδειγμα η δειγματοληψία δεδομένων [5] τεχνική στην οποία επιλέγονται ένα μικρό κομμάτι των δεδομένων για επεξεργασία και επιστρέφεται ένα κατά προσέγγιση αποτέλεσμα με κάποιο σφάλμα. Τα συστήματα βάσεων δεδομένων προσαρμόστηκαν στην τεχνική δειγματοληψίας δεδομένων στο να αναπτύξουν νέες προσεγγίσεις στις οποίες εστιάζονται στην ακρίβεια παρά το χρόνο απόκρισης και την χωρητικότητα του χώρου αποθήκευσης[6]. Η σύγχρονη λύση για την συνάθροιση δεδομένων [7] περιλαμβάνει προσαρμοστική στρωματοποιημένη και δυναμική δειγματοληψία με σκοπό την υψηλή ακρίβεια αγνοώντας όλους τους άλλους στόχους. Ο κυβικός λειτουργός (Data Cube Operator) [8] πέτυχε υψηλές λειτουργίες συνάθροισης δεδομένων με την πολυδιάστατη γενίκευση (multidimensional generalization) και χρησιμοποιήθηκε για την ελαχιστοποίηση της εκτέλεσης των ερωτημάτων στα μεγάλα δεδομένα. Άλλες τεχνικές συνάθροισης περιλαμβάνουν την γενικευμένη συνάθροιση [9] και τις μικρές δομές δεδομένων στην μνήμη (ονομάζεται sketches,summaries ή synopsis) [10]. Τέλος για την αύξηση της αποτελεσματικότητας χρησιμοποιούνται προϋπολογισμένες προβολές με οριακό σφάλμα για την αύξηση της

αποδοτικότητας των συστημάτων αποθήκευσης δεδομένων χρησιμοποιώντας μεροληπτική δειγματοληψία [11]. Παρόλα αυτά ελάχιστες ερευνητικές μελέτες [12], [13] εστιάζονται στις προσεγγίσεις αποσύνθεσης δεδομένων για την αντιμετώπιση του προβλήματος της ελαχιστοποίησης του αποθηκευτικού χώρου.

Όπως αναφέρθηκε και προηγουμένως, στις μέρες μας τα ποσά δεδομένων που συσσωρεύονται και αποθηκεύονται στις βάσεις δεδομένων είναι τεράστια. Τα περισσότερα από αυτά αποθηκεύονται και χρησιμοποιούνται ανάλογα με την χρήση τους και ενδεχομένως να μην εξεταστούν κατά πόσο πρέπει να παραμείνουν στις βάσεις δεδομένων ή να διαγραφούν επειδή υπάρχουν οι πιθανότητες να γίνει επαναχρησιμοποίηση τους. Αποτέλεσμα αυτού, οι αποθηκευτικοί χώροι να γεμίζουν και τα παλιά δεδομένα να καταλαμβάνουν μεγάλο και σημαντικό χώρο. Το κόστος διαχείρισης για την αποθήκευση τόσου μεγάλου όγκου δεδομένων γίνετε ήδη αντικείμενο μελέτης για πολλούς ερευνητές. Το κόστος αυτό αυξάνεται όχι μόνο για την αγορά του υλικού αλλά και επειδή η αξία χρησιμότητας των δεδομένων μειώνεται σημαντικά με την πάροδο του χρόνου. Με τον όρο αμνησία δεδομένων εννοούμε την σταδιακή μείωση των παλιών δεδομένων από τους αποθηκευτικούς χώρους είτε γιατί τα δεδομένα δεν χρησιμοποιούνται πια είτε γιατί θεωρούνται ξεπερασμένα.

1.3 Περίγραμμα εργασίας

Στο πρώτο κεφάλαιο αυτής της εργασίας γίνεται μια εισαγωγή σε αυτή την διπλωματική εργασία. Πιο συγκεκριμένα παρουσιάζεται η υποκίνηση της εργασίας και πώς οι κατανεμημένοι χώροι αποθήκευσης αντιμετωπίζουν προβλήματα στην συνεχή αποθήκευση μεγάλων τηλεπικοινωνιακών δεδομένων λόγω της ταχείας εξάπλωσης των έξυπνων κινητών συσκευών και γενικά του διαδικτύου. Επεξηγούνται κάποια παραδείγματα και οι λόγοι που είναι αναγκαία η προώθηση της τεχνολογίας που αφορά την αποσύνθεση μεγάλων δεδομένων. Ακολουθεί το περίγραμμα της εργασίας όπου θα δοθεί μια πλήρης επεξήγηση της δομής της εν λόγω διπλωματικής εργασίας ανά κεφάλαιο. Στο τέλος του πρώτου κεφαλαίου γίνεται μια αναφορά στην συνεισφορά που είχα σε αυτήν την εργασία.

Στο δεύτερο κεφάλαιο γίνεται μια αναφορά και επεξήγηση σχετικής δουλειάς και ερευνητικής μελέτης σύμφωνα με την βιβλιογραφία. Σε αυτό το κεφάλαιο αναλύονται κάποιες τεχνικές συμπίεσης οι οποίες προτάθηκαν στο παρελθόν οι οποίες αφορούν μεγάλα χωροχρονικά δεδομένα. Επίσης γίνεται μια αναφορά στις κατηγορίες έρευνας σε τηλεπικοινωνιακά δεδομένα. Επιπλέον, κάνουμε λόγο για την αποθήκευση μεγάλων δεδομένων και στα διάφορα συστήματα. Ακολουθεί μια αναφορά σε τεχνικές δειγματοληψίας και μια συνοπτική περιγραφή του συστήματος SPATE, μια ερευνητική δουλεία που στοχεύει στην ελαχιστοποίηση του χρόνου απόκρισης και χώρου αποθήκευσης. Στο τέλος του δεύτερου κεφαλαίου γίνεται μια αναφορά σε σχετική δουλειά που αφορά μηχανική μάθηση για τη πρόβλεψη δεδομένων.

Στο τρίτο κεφάλαιο θα γίνει μια αναφορά στο τεχνολογικό και θεωρητικό υπόβαθρο. Στην υποενότητα τεχνολογικό υπόβαθρο θα αναφερθούν οι τεχνολογίες που χρησιμοποιήθηκαν για την εκπόνηση αυτής της διπλωματικής εργασίας όπως για παράδειγμα το Tensorflow που χρησιμοποιήθηκε για την υλοποίηση του αλγόριθμου DP, για την δημιουργία των μοντέλων και την μετέπειτα επαναφορά τους για την πρόβλεψη διαφόρων στοιχείων. Όλα αυτά στην γλώσσα προγραμματισμού Python (καθώς επίσης και τα scripts που υλοποιήθηκαν για την πειραματική αποτίμηση) θα γίνει μια αναφορά στο Play Framework και στην γλώσσα προγραμματισμού SCALA. Όσον αφορά το θεωρητικό υπόβαθρο θα γίνει μια εισαγωγή στην τεχνική μάθηση (machine learning) στα Νευρωνικά δίκτυα και πιο συγκεκριμένα στα LSTM δίκτυα. Στο τέλος θα γίνει μια εισαγωγή στο HDFS.

Στο τέταρτο κεφάλαιο θα γίνει μια εκτενής ανάλυση του προβλήματος που μελετάμε αλλά και του τελεστή DP για την κατασκευή των μοντέλων για την "εκπαίδευση τους" χρησιμοποιώντας τεχνικές μάθησης (deep learning) για την αποθήκευση του στο δίσκο και την μετέπειτα επαναφορά τους για πρόβλεψη χωροχρονικών δεδομένων τα οποία δεδομένα χρησιμοποιούνται στη γραφική διαπροσωπεία που επεκτείνεται.

Στο πέμπτο κεφάλαιο θα γίνει μια πλήρης περιγραφή του πρωτότυπου καθώς και περιγραφή της αρχιτεκτονικής του αναλύοντας τα διάφορα επίπεδα καθώς και την περιγραφή της διεπαφής χρήστη.

Στο έκτο κεφάλαιο αναλύεται εις βάθος η μεθοδολογία που ακολουθήσαμε για την πειραματική αποτίμηση και αξιολόγηση των αλγορίθμων που προτείνουμε. Παρουσιάζονται τα αποτελέσματα των δύο πειραματικών σειρών σε γραφικές παραστάσεις. Επιπλέον γίνεται

η σύγκριση του τελεστή DP που προτείνουμε με άλλους αλγόριθμους και ανάλυση και επεξήγηση όλων των δεικτών - μετρητών που χρησιμοποιούνται στις πειραματικές διαδικασίες.

Στο έβδομο και τελευταίο κεφάλαιο γίνεται ο επίλογος και αναφορά σε μελλοντική δουλειά που μπορεί να γίνει. Αναφέρονται τα συμπεράσματα που εξάγονται από την πειραματική αποτίμηση των αλγορίθμων συνοψίζοντας τα αποτελέσματα που πήραμε από τις σειρές πειραμάτων. Στην υποενότητα μελλοντική δουλειά γίνονται κάποιες σκέψεις για την μετέπειτα επέκταση των προτεινόμενων αλγορίθμων και γραφικών διαπροσωπειών.

1.4 Συνεισφορά

Η συνεισφορά αυτής της διπλωματικής εργασίας χωρίζεται σε 3 βασικές ενότητες οι οποίες αναλύονται σε βάθος στα επόμενα κεφάλαια.

- **Τελεστής DP:** Υλοποίηση του τελεστή DP χρησιμοποιώντας αλγόριθμους τεχνικής μάθησης και με βάση τους αλγόριθμους της κατασκευής και επαναφοράς των μοντέλων, για μεγάλα τηλεπικοινωνιακά δεδομένα, τους οποίους προτείναμε χρησιμοποιώντας Data Postdiction.

- **Γραφική Διαπροσωπεία:** Υλοποίηση Γραφικής Διαπροσωπείας όπου ο χρήστης μπορεί να αλληλοεπιδράσει με το σύστημα, να δει στην πράξη γιατί είναι χρήσιμη η εφαρμογή του τελεστή DP σύμφωνα με κάποια σενάρια και να δει τα συμπεράσματα τόσο για την ακρίβεια όσο και για την χωρητικότητα αποθήκευσης.

- **Πειραματική Αποτίμηση:** Πειραματική αποτίμηση του τελεστή DP εκτελώντας δύο σειρές πειραμάτων. Για την αξιολόγηση του αλγόριθμου DP χρησιμοποιήθηκαν 3 επιπλέον αλγόριθμοι με σκοπό την σύγκριση μεταξύ τους. Οι τρείς αυτοί αλγόριθμοι είναι ο Data Sampling, Random και Compression. Συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος DEFLATE που χρησιμοποιείται από την βιβλιοθήκη GZIP. Στην πρώτη

πειραματική σειρά αξιολογούμε την απόδοση του DP αλγορίθμου, σε σχέση με τους υπόλοιπους τρείς αλγορίθμους, με βάση τον αποθηκευτικό χώρο και την ακρίβεια των δεδομένων που αποσυντέθηκαν. Στην δεύτερη πειραματική σειρά εξετάζουμε την επίδραση διαφόρων παραμέτρων ελέγχου στην απόδοση του DP σε space (s) και Accuracy (NRMSE). Πιο συγκεκριμένα αλλάζουμε τον παράγοντα αποσύνθεσης (Decaying Factor), τα μοντέλα και των αριθμό των νευρώνων του δικτύου LSTM.

Κεφάλαιο 2

Σχετική Δουλειά

2.1 Εισαγωγή

Στις πιο κάτω υποενότητες γίνεται μια εκτενής ανάλυση έργων και ερευνητικών μελετών που ασχολούνται με τον χώρο και τον χρόνο σε τηλεπικοινωνιακά δεδομένα μεγάλου όγκου. Αρχικά, αναλύονται κάποιες έρευνες όσον αφορά μεγάλα τηλεπικοινωνιακά δεδομένα, έρευνες για αποθήκευση και συμπίεση αρχείων. Μετέπειτα γίνεται μια συνοπτική περιγραφή της αρχιτεκτονικής του SPATE και παρουσίαση των στόχων του συστήματος. Στο τέλος του κεφαλαίου αναφέρονται σχετικές μελέτες που αφορούν σύνοψη δεδομένων και μηχανική μάθηση.

2.2 Έρευνα σε Μεγάλα Τηλεπικοινωνιακά Δεδομένα

Η έρευνα στα telcos εμπίπτει σε μία από τις ακόλουθες κατηγορίες ανάλογα με την περίπτωση:

i. Ανάλυση και ανίχνευση σε πραγματικό χρόνο.
ii. Προβλέποντας την συμπεριφορά των χρηστών.
iii. Ιδιωτικότητα (Privacy).

Αξιοσημείωτο είναι το γεγονός ότι υπάρχουν έρευνες σε telcos οι οποίες δεν σχετίζονται άμεσα ή έμμεσα με δεδομένα μεγάλου όγκου ούτε "έξυπνες" πόλεις αλλά αποτελούνται από εφαρμογές τις οποίες χρησιμοποιούν οι τηλεπικοινωνιακές εταιρίες για να βελτιώσουν τις παρεχόμενες υπηρεσίες τους και τα έσοδά τους.

Ανάλυση και ανίχνευση σε πραγματικό χρόνο: Η OceanRT [21] αναπτύχθηκε για την διαχείριση μεγάλων χωροχρονικών δεδομένων όπως τα στοιχεία OSS Telco που τρέχουν πάνω από την υποδομή του cloud. Η συγκεκριμένη ένα καινούριο πρόγραμμα αποθήκευσης που βελτιστοποιεί τα ερωτήματα (queries) με ενώσεις (joins) και πολυδιάστατες επιλογές (multi-dimensional selections). Στην συνέχεια παρουσιάστηκε η OceanST [2] που περιλαμβάνει τα ακόλουθα χαρακτηριστικά :

i. Έναν αποτελεσματικό μηχανισμό φόρτωσης για Telco MBB δεδομένα που συνεχώς αυξάνονται.
ii. Νέες χωροχρονικές δομές δεικτών για την ακριβή επεξεργασία και την κατά προσέγγιση χωροχρονική συνάθροιση ερωτημάτων με απώτερο σκοπό την αντιμετώπιση του τεράστιου όγκου MBB δεδομένων.

Επιπλέον, παρουσιάζεται το CellQ με σκοπό την βελτιστοποίηση ερωτημάτων όπως είναι τα "χωροχρονικά σημεία κυκλοφορίας" (spatiotemporal traffic hotspots) και "ακολουθίες μεταβιβάσεων (handoff sequences) με προβλήματα απόδοσης". Παρουσιάζει τα στιγμιότυπα (snapshots) των δεδομένων ενός κυψελοειδούς δικτύου (Cellular Network Data) σαν γράφους και τα χρησιμοποιεί στη χωροχρονική τοποθεσία των δεδομένων κυψελοειδούς δικτύου [22].

Μια σχετική ερευνητική δουλειά αφορά την ανάπτυξη ενός επεκτάσιμου κατανεμημένου συστήματος το οποίο επεξεργάζεται με αποδοτικό τρόπο μικτά φορτία εργασίας για την απόκριση σε ερωτήματα που αφορούν ροή συμβάντων και ερωτήματα ανάλυσης σε telco δεδομένα [23].

Επιπλέον σχετικές ερευνητικές δουλειές που αφορούν την ανάλυση τηλεπικοινωνιακών δεδομένων είναι η ανάπτυξη ενός συστήματος πάνω από το ενδιάμεσο λογισμικό (middleware) της IBM InfoSphere Streams η οποία αναλύει 6 δισεκατομμύρια CDR δεδομένα(Call Detail Record) την ημέρα σε πραγματικό χρόνο [24]. Τα cdr δεδομένα παράγονται από το τηλεπικοινωνιακό δίκτυο και καταγράφουν τις λεπτομέρειες μιας τηλεφωνικής κλήσης ή γενικά μιας τηλεπικοινωνιακής συναλλαγής. Ακόμα υπάρχει ένα σύστημα για την διατήρηση των προφίλ κλήσεων των πελατών σε μια ροή με την εφαρμογή επεξεργασία κατανεμημένων ροών [25].

2.3 Αποθήκευση Μεγάλων δεδομένων

Η αποθήκευση μεγάλων δεδομένων αναφέρεται σε τεχνολογίες αποθήκευσης οι οποίες αναπτύχθηκαν προκειμένου να ικανοποιήσουν τις απαιτήσεις που προκύπτουν από την αποθήκευση τεράστιων φορτίων με δεδομένα. Τυπικά, τα συστήματα αποθήκευσης μεγάλου όγκου δεδομένων είναι κατανεμημένα, όπως για παράδειγμα το Hadoop Distributed File System (HDFS), και δεν μοιράζονται αρχιτεκτονικές. Επιπρόσθετα οι NoSQL βάσεις δεδομένων εισάγουν στοιχεία τύπου κλειδί – τιμή (key – value) όπως για παράδειγμα την DynamoDB και Cassandra . Πέραν αυτού υπάρχουν NoSQL βάσεις δεδομένων οι οποίες εισάγουν στοιχεία στηλών (columns) (π.χ. HBase), στοιχεία εγγράφων όπως για παράδειγμα Couchbase, CouchDB, MongoDB και βάσεις δεδομένων με γράφους όπως η Neo4j. Το Hive και Impala είναι καινούριες πλατφόρμες ερωτημάτων (query platforms) που βρίσκονται πάνω από τα κατανεμημένα συστήματα αποθήκευσης δεδομένων, HDFS ή Hbase παρέχοντας την δυνατότητα σε αυτούς που αλληλοεπιδρούν με τέτοια συστήματα να γράφουν ερωτήματα σε κοινή SQL γλώσσα. Ιδιαίτερα, τα τελευταίας τεχνολογίας κατανεμημένα συστήματα αποθήκευσης όπως είναι το OctopusFS, διαχειρίζονται αυτά τα δεδομένα χρησιμοποιώντας ετερογενή μέσα αποθήκευσης.

2.4 Συμπίεση Αρχείων

Οι τεχνικές συμπίεσης του συγκεκριμένου τομέα προτάθηκαν στο παρελθόν. Για παράδειγμα, για την συμπίεση χωροχρονικού (spatiotemporal) δεδομένων κλίματος [15], συλλογές εγγράφων κειμένων [16], δεδομένα κινητής υποδιαστολής καθώς και τα κυμαινόμενα δεδομένα (floating point data streams) [17]. Καμία από αυτές τις ερευνητικές μελέτες δεν έχει προταθεί για κατανεμημένα συστήματα και δεν μπορούν να εφαρμοστούν άμεσα σε τηλεπικοινωνιακά δεδομένα. Τα τηλεπικοινωνιακά δεδομένα ως συνήθως περιέχουν συμβολοσειρές (strings) και ακέραιες τιμές (integers) .

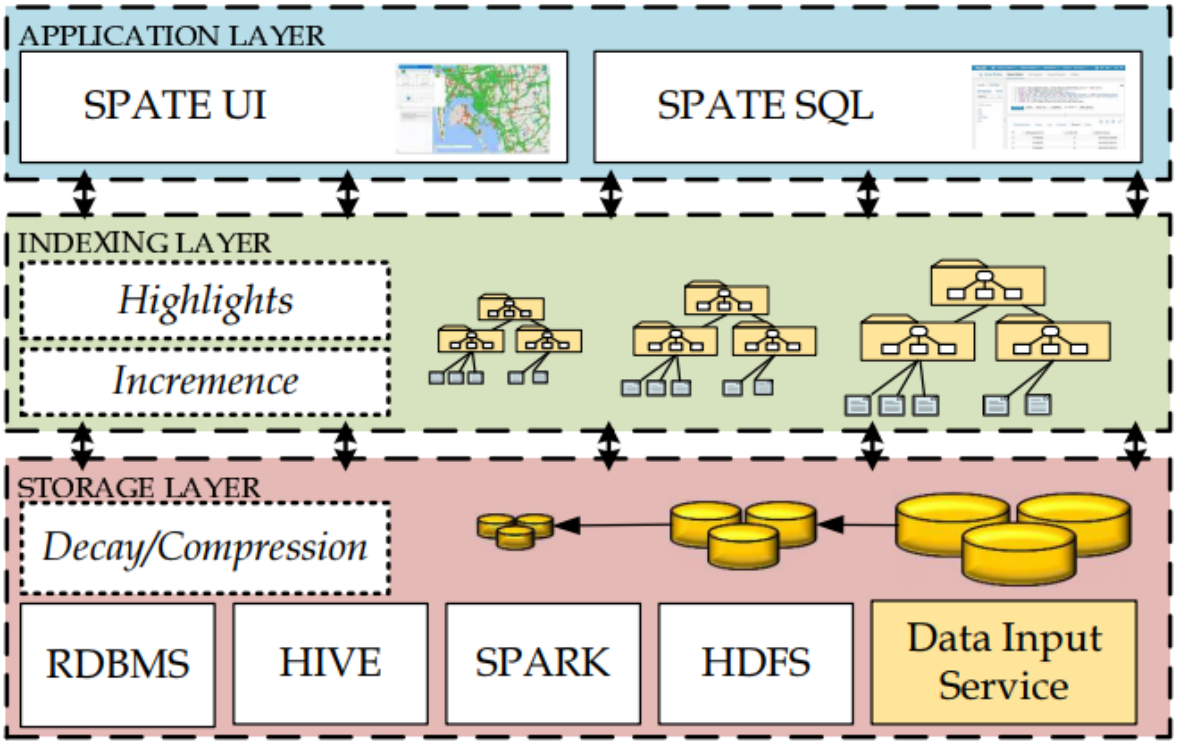
Στα έργα [18]-[20] σημαντικό στοιχείο της ερευνάς τους είναι οι τεχνικές διαφορικής συμπίεσης και η σχέση μεταξύ αναλογίας συμπίεσης και χρόνοι αποσυμπίεσης για αυξητικά αρχειακά δεδομένα . Αυτή η έρευνα σχετίζεται άμεσα με το αντικείμενο συμπίεσης και η λύση είναι βασισμένη στην μελέτη και τα ευρήματά τους.

2.5 Η Αρχιτεκτονική SPATE

Στο έργο SPATE [1] προτάθηκε ένα καινοτόμο πλαίσιο εξερεύνησης μεγάλων τηλεπικοινωνιακών δεδομένων το οποίο αποσκοπεί :

i. Στην ελαχιστοποίηση του χώρου αποθήκευσης που χρειάζεται για την αποθήκευση και διατήρηση των μεγάλων τηλεπικοινωνιακών δεδομένων με την πάροδο του χρόνου.
ii. Στην ελαχιστοποίηση του χρόνου απόκρισης για επερωτήματα που αφορούν χωροχρονικά δεδομένα.

Το SPATE χρησιμοποιεί αφενός συμπίεση δεδομένων χωρίς απώλειες (lossless) και αφετέρου αποσύνθεση δεδομένων με απώλειες (lossy) με σκοπό την αποθήκευση μεγάλων τηλεπικοινωνιακών δεδομένων με τον πιο συμπαγή τρόπο. Επιπλέον επιτρέπει στις εργασίες που εκτελούν την εξερεύνηση των μεγάλων δεδομένων να διατηρούνται σε υψηλό επίπεδο για προκαθορισμένα επερωτήματα σε πολύ μεγάλα χρονικά διαστήματα χωρίς να



Σχήμα 2.5.1 Επίπεδα Αρχιτεκτονικής SPATE[1]

καταναλώνουν τεράστιο χώρο αποθήκευσης. Για την απόδειξη των πιο πάνω η ερευνητές του προτεινόμενου πλαισίου μέτρησαν την αποτελεσματικότητα των χρησιμοποιώντας μια σειρά από εργασίες σχετιζόμενες με τηλεπικοινωνιακά δεδομένα όπως είναι εκτέλεση OLAP και OLTP επερωτημάτων, αναζήτηση, και έδειξαν ότι η χρήση αυτού του πλαισίου μπορεί να επιτύχει συγκρίσιμους χρόνους απόκρισης με λιγότερο χώρο αποθήκευσης. Το σχήμα 2.5.1 δείχνει τα επίπεδα αρχιτεκτονικής του SPATE. Το επίπεδο Storage αποθηκεύει τα νέα στιγμιότυπα του δικτύου μέσω μιας διαδικασίας συμπίεσης χωρίς απώλειες και αποθηκεύει τα αποτελέσματα σε ένα σύστημα μεγάλου αρχείου δεδομένων για διαθεσιμότητα και απόδοση. Αυτό το στοιχείο είναι υπεύθυνο για την ελαχιστοποίηση του απαιτούμενου χώρου αποθήκευσης με ελάχιστη επιβάρυνση στον χρόνο απόκρισης του ερωτήματος. Το επίπεδο Indexing είναι υπεύθυνη για την ελαχιστοποίηση του χρόνου απόκρισης ερωτήματος για ερωτήματα διερεύνησης δεδομένων. Το επίπεδο Application υλοποιεί την ενότητα επερωτήσεων και την διεπαφή χρήστη.

2.6 Σύνοψη Δεδομένων

Η δειγματοληψία (sampling) αναφέρεται στην διαδικασία τυχαίας επιλογής ενός υποσυνόλου στοιχείων από δεδομένα από ένα σχετικά μεγάλο σύνολο δεδομένων. Προηγμένες τεχνικές δειγματοληψίας όπως για παράδειγμα αυτή του Bernoulli ή του Poission επιλέγουν στοιχεία δεδομένων χρησιμοποιώντας σαν βάση τις πιθανότητες και την στατιστική.

Η διαστρωμάτωση η αλλιώς στρωματοποιημένη δειγματοληψία είναι μια τεχνική δειγματοληψίας πιθανότητας στην οποία ο ερευνητής διαιρεί ολόκληρο το σύνολο των δεδομένων σε διαφορετικά υποσύνολα τα οποία ονομάζονται στρώματα (strata) και στην συνέχεια επιλέγει τυχαία τα τελικά δεδομένα ανάλογα με τα διαφορετικά στρώματα. Υπάρχουν ερευνητικές μελέτες όπου προτάθηκε η διαστρωμάτωση όπου η πιθανότητα επιλογής είναι προμελετημένη [26]. Προκειμένου να αντιμετωπιστούν κάποια ζητήματα της δειγματοληψίας δεδομένων υπάρχει το G-OLA [9] το οποίο είναι ένα μοντέλο που γενικεύει το OLA για να υποστηρίξει γενικά OLAP (Online Analytical Processing) ερωτήματα που χρησιμοποιούν αλγόριθμους συντήρησης DELTA.

OLAP (Online Analytical Processing): Το OLAP είναι μια τεχνολογία βάσης δεδομένων που έχει βελτιστοποιηθεί για την αναζήτηση (querying) και αναφορά παρά την επεξεργασία συναλλαγών. Τα δεδομένα πηγής για το OLAP είναι βάσεις δεδομένων που αποθηκεύονται σε συνήθως σε αποθήκες δεδομένων και ονομάζονται Online Transactional Processing (OLTP).

Ερωτήματα OLAP: Τα ερωτήματα OLAP είναι περίπλοκα ερωτήματα για τα οποία ισχύουν τα ακόλουθα:

- Επεξεργάζονται μεγάλες ποσότητες δεδομένων.
- Ανακαλύπτουν πρότυπα και τάσεις στα δεδομένα.
- Σχετικά “ακριβές” ερωτήσεις που έχουν μεγάλο χρόνο απόκρισης.

- Ονομάζονται και ερωτήματα υποστήριξης λήψης αποφάσεων.

Ιδιαίτερα το BlinkDB [7] επιτρέπει στους χρήστες να επιλέγουν τα όρια σφάλματος και τον χρόνο απόκρισης ενός query χρησιμοποιώντας δυναμικούς αλγόριθμους δειγματοληψίας. Το SciBORQ [11] είναι ένα πλαίσιο εφαρμογών (framework) το οποίο επιτρέπει στους χρήστες να επιλέξουν την ποιότητα του ερωτήματος με βάσει πολλαπλά ενδιαφέροντα δείγματα δεδομένων τα οποία ονομάζονται impressions.

Αρκετά έργα έχουν προσαρμόσει τις διαδικασίες δειγματοληψίας προκειμένου να δημιουργήσουν σύνοψη δεδομένων ούτως ώστε να επιτευχθεί χαμηλός χρόνος απόκρισης για ερωτήματα adhoc queries [11].

Ad-Hoc queries: Ένα ερώτημα τύπου ad-hoc είναι ερώτημα το οποίο δεν μπορεί να προσδιοριστεί πριν από την έκδοση του ερωτήματος. Δημιουργείται προκειμένου να αποκτηθούν πληροφορίες όταν προκύψει ανάγκη και αποτελείται από δυναμικά κατασκευασμένη γλώσσα SQL η οποία ώς συνήθως κατασκευάζεται από εργαλεία ερωτημάτων. (Desktop-resident query tools).

Επιπρόσθετα, προτάθηκε ένα νέο online μοντέλο συνάθροισης [9] για την υποστήριξη ερωτημάτων OLAP με χρήση τεχνικών συντήρησης DELTA.

Σκίτσα Δεδομένων (Data Sketches): Τα σκίτσα δεδομένων αποθηκεύονται συνήθως στην μνήμη και είναι πολύ ευέλικτα λόγω της μικρής τροποποίησης που χρειάζεται να γίνει προκειμένου να έχουμε μια αλλαγή στα σκίτσα (δράση) όπως για παράδειγμα σε μια διαγραφή ή εισαγωγή μιας σειράς (εγγραφής δεδομένων) [10].

Επιπρόσθετα υπάρχουν στην βιβλιογραφία ερευνητικές μελέτες όπου προτάθηκαν τα επίμονα σκίτσα δεδομένων (persistent data sketches) που απαντούν σε ερωτήματα σε οποιαδήποτε προηγούμενη χρόνική στιγμή [27] και έχουν την δυνατότητα της συγχώνευσης για να απαντήσουν σε ένα γενικευμένο ερώτημα.

2.7 Μηχανική Μάθηση

Οι τηλεπικοινωνιακοί οργανισμοί (telcos) επικεντρώνονται στο να προβλέπουν την δραστηριότητα των χρηστών ούτως ώστε να βελτιστοποιήσουν το δίκτυο τους λαμβάνοντας υπόψιν την εξοικονόμηση του κόστους ανάπτυξης και την ελαχιστοποίηση όσον αφορά την κατανάλωση ενέργειας [28]. Επιπρόσθετα ερευνητικές μελέτες έδειξαν πώς οι διάφοροι αλγόριθμοι μηχανικής μάθησης (machine learning) όπως Linear Regression, Classification and Regression Trees και K-Nearest Neighbors μπορούν να εφαρμοστούν στα τηλεπικοινωνιακά δεδομένα μεγάλου όγκου για να προβλέπουν ποιοι πελάτες ενδεχομένως να σταματήσουν να χρησιμοποιούν τις τηλεπικοινωνιακές υπηρεσίες που παρέχονται [29]. Σε επιπλέον έρευνες προέβλεψαν ποιοι πελάτες ενδεχομένως να παύσουν κάποια υπηρεσία χρησιμοποιώντας δύο νέες παραλλαγές του επαναλαμβανόμενου νευρωνικού δικτύου (RNN) τα οποία ονομάζονται Elman και Jordan δίκτυα [30]. Στο έργο [31] προτάθηκε ένα πλαίσιο εφαρμογών για την πρόβλεψη της συμπεριφοράς των χρηστών, περιλαμβάνοντας περισσότερο από ένα εκατομμύριο χρήστες τηλεπικοινωνιών. Το πλαίσιο εφαρμογών αυτό χρησιμοποίησε έγγραφα τα οποία περιέχουν συλλογές από μεταβαλλόμενες χωροχρονικές “λέξεις” που περιγράφει τον χρήστη και την συμπεριφορά του. Εξάγοντας πληροφορίες πρόσβασής του χώρου και χρόνου από δεδομένα MBB (Mobile BroadBand) το μοντέλο μηχανικής μάθησης βρισκόταν σε θέση να μάθει ειδικά χαρακτηριστικά του χρήστη επιτυγχάνοντας με αυτό το τρόπο υψηλού επιπέδου πρόβλεψης της δραστηριότητας του χρήστη. Επιπρόσθετα, χρησιμοποιήθηκε ένα δίκτυο LSTM (Long Short Term Memory) για την ανίχνευση μακροπρόθεσμων συσχετίσεων στα δεδομένα ανθρώπινης δραστηριότητας [14]. Οι συγγραφείς έδειξαν ότι το LSTM δίκτυο έχει πολύ καλύτερη απόδοση εάν το συγκρίνουμε με άλλες παραδοσιακές προσεγγίσεις που βασίζονται στην εξόρυξη αλληλουχίας και στα κρυμμένα μοντέλα του Markov.

Κεφάλαιο 3

Υπόβαθρο

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιαστεί το τεχνολογικό και θεωρητικό υπόβαθρο. Αρχικά, στην υποενότητα Τεχνολογικό Υπόβαθρο, θα μιλήσουμε για τις τεχνολογίες που χρησιμοποιήθηκαν με σκοπό την υλοποίηση του αλγόριθμου Data Postdiction. Στην συνέχεια θα γίνει μια αναφορά στις τεχνολογίες που χρησιμοποιήθηκαν για την δημιουργία της γραφικής διαπροσωπείας. Στην υποενότητα Θεωρητικό Υπόβαθρο θα γίνει μια εκτενής αναφορά στην θεωρητική μελέτη που προηγήθηκε για την εκπόνηση της διπλωματικής εργασίας.

3.2 Τεχνολογικό Υπόβαθρο

Για την διεκπεραίωση αυτής της εργασίας χρησιμοποιήθηκαν κάποιες σύγχρονες τεχνολογίες. Για την δημιουργία των μοντέλων χρησιμοποιήθηκε το Tensorflow σε γλώσσα προγραμματισμού Python. Για την δημιουργία της γραφικής διαπροσωπείας χρησιμοποιήθηκε το PLAY framework σε Scala και επι μέρους τεχνολογίες όπως HTML/CSS, Angular. Τα δεδομένα βρίσκονται σε κατανεμημένο σύστημα αποθήκευσης (HDFS)

3.2.1 Python

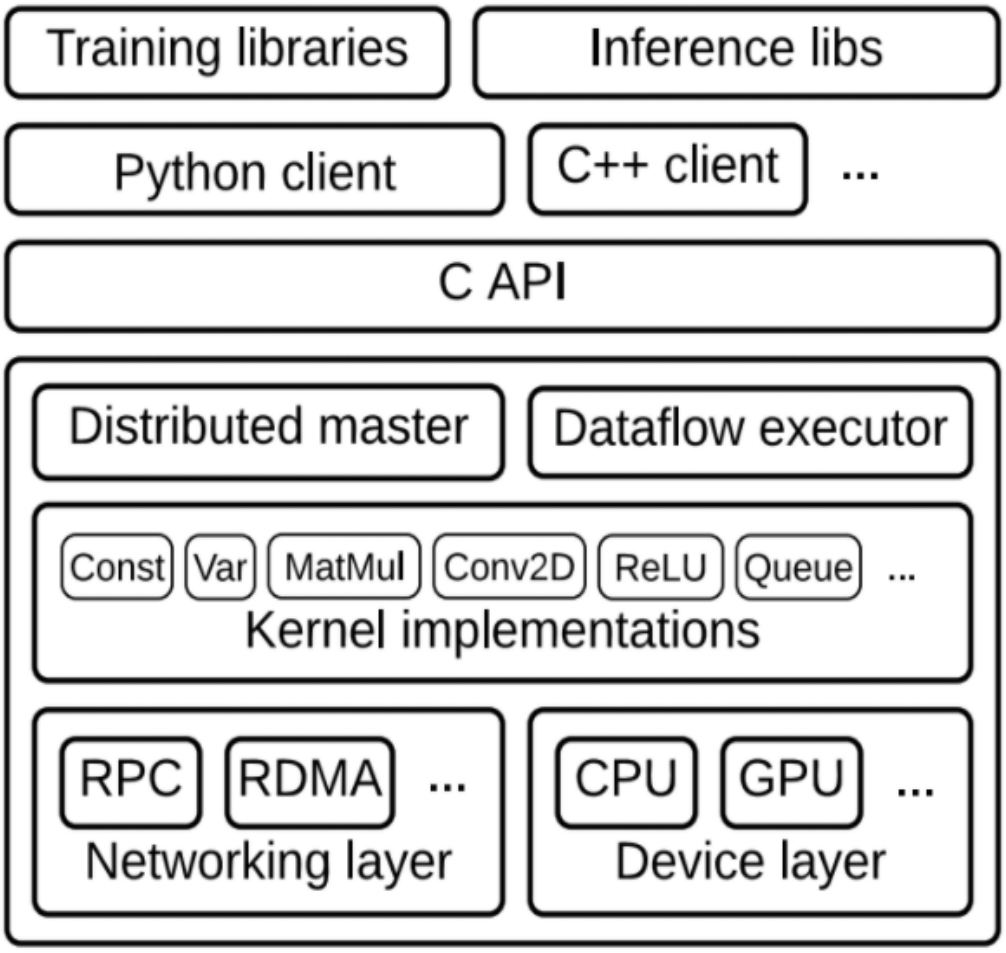
Η Python είναι μια ερμηνευμένη γλώσσα προγραμματισμού υψηλού επιπέδου. Δημιουργήθηκε από τον Guido van Rossum και το 1991 κυκλοφόρησε για πρώτη φορά. Η φιλοσοφία της Python δίνει έμφαση στην αναγνωσιμότητα κώδικα . Παρέχει κάποιες δομές οι οποίες καθιστούν εφικτό το σαφή προγραμματισμό σε μικρή ή μεγάλη κλίμακα.

3.2.2 Tensorflow

Το Tensorflow είναι μια βιβλιοθήκη λογισμικού ανοικτού κώδικα για αριθμητικούς υπολογισμούς υψηλής απόδοσης. Η ευέλικτη αρχιτεκτονική του Tensorflow επιτρέπει την ανάπτυξη υπολογισμών σε διάφορες πλατφόρμες, όπως CPU, GPU, TPU, και από επιτραπέζιους υπολογιστές σε συγκροτήματα εξυπηρετητών (clusters) και σε κινητές συσκευές. Αναπτύχθηκε από ερευνητές και μηχανικούς της Google Brain και υποστηρίζει σε μεγάλο βαθμό τις τεχνικές μηχανικής μάθησης και βαθιάς εκμάθησης και ο ευέλικτος πυρήνας του χρησιμοποιείται σε πολλούς άλλους επιστημονικούς τομείς. Ακολουθεί μια σύντομη ανάλυση της αρχιτεκτονικής του Tensorflow.
Η εκτέλεση (runtime) του Tensorflow είναι μια βιβλιοθήκη πολλαπλών πλατφορμών. Το API C διαχωρίζει τον κώδικα επιπέδου χρήστη σε διάφορες γλώσσες από τον πυρήνα εκτέλεσης. Το σχήμα 3.2.1 απεικονίζει την γενική αρχιτεκτονική του Tensorflow.

Client

- Ορίζει τους υπολογισμούς σαν γράφημα ροής δεδομένων
- Ξεκινά την εκτέλεση του γραφήματος χρησιμοποιώντας ένα session



Σχήμα 3.2.1 Η Αρχιτεκτονική του Tensorflow [35]

Distributed Master

- Ξεχωρίζει ένα συγκεκριμένο υπογράφημα από το γράφημα όπως ορίζεται από τις παραμέτρους του Session.run().
- Διαχωρίζει το υπογράφημα σε κομμάτια τα οποία εκτελούνται σε διαφορετικές διεργασίες και συσκευές.
- Διανέμει τα κομμάτια στις υπηρεσίες που θα εργαστούν.
- Ξεκινά την εκτέλεση του κάθε κομματιού από την υπηρεσία που θα εργαστεί.

Worker Services

- Προγραμματίζουν την εκτέλεση των λειτουργιών του γραφήματος χρησιμοποιώντας εφαρμογές πυρήνα κατάλληλες για το διαθέσιμο υλικό (CPU, GPU).
- Λαμβάνουν και αποστέλλουν τα αποτελέσματα της λειτουργίας από και προς τις υπηρεσίες εργαζομένων.

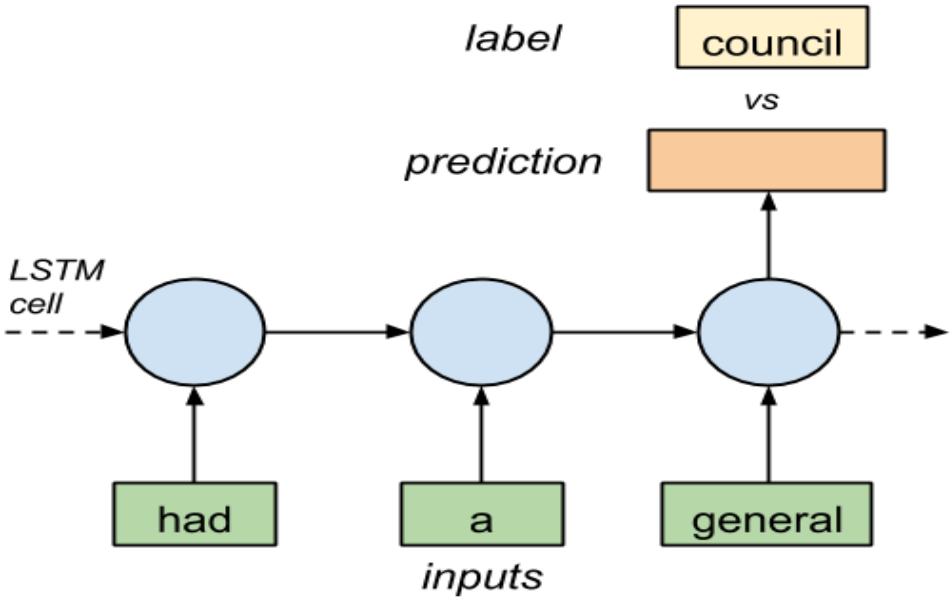
Kernel Implementations

- Εκτελούν τους υπολογισμούς για μεμονωμένες λειτουργίες γραφημάτων.

3.2.3 Tensorflow και LSTM

Στην μηχανική μάθηση τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) είναι μια οικογένεια νευρωνικών δικτύων τα οποία υπερέχουν στην μάθηση από διαδοχικά δεδομένα. Μια κατηγορία RNN στην οποία εφαρμόζεται είναι τα Long Short-Term Memory (LSTM) δίκτυα τα οποία εξηγήθηκε η λειτουργία τους στην υποενότητα Θεωρητικό Υπόβαθρο.

Υποθέστε ότι θέλουμε να εκπαιδεύσουμε ένα δίκτυο LSTM το οποίο θα προβλέπει την επόμενη λέξη χρησιμοποιώντας το Tensorflow. Εάν τροφοδοτήσουμε ένα LSTM με ακολουθίες από ένα κείμενο το οποίο περιέχει λέξεις και σύμβολα τότε το LSTM θα μπορεί να προβλέπει το επόμενο σύμβολο στην σειρά όπως φαίνεται στο σχήμα 3.2.2. Τεχνικά, οι εισόδοι στο LSTM πρέπει να αντιστοιχούν σε πραγματικούς αριθμούς οπόταν το μόνο που έχουμε να κάνουμε σ΄αυτό το σημείο είναι να αντιστοιχήσουμε ένα μοναδικό αριθμό σε κάθε σύμβολο με βάσει την συχνότητα εμφάνισης του. Αυτό μπορεί να γίνει χρησιμοποιώντας μια συνάρτηση για την κατασκευή ενός λεξιλογίου το οποίο περιέχει αριθμούς τα οποία αντιστοιχούν σε λέξεις. Η προβλεπόμενη τιμή στην έξοδο θα είναι ένας αριθμός ο οποίος θα αντιστοιχεί στην λέξη που βρίσκεται στο λεξιλόγιο.



Σχήμα 3.2.2 Ένα κελί LSTM με 3 σύμβολα σαν είσοδο και 1 σαν έξοδο [34]

Ο πυρήνας της εφαρμογής είναι το μοντέλο LSTM το οποίο υλοποιείται σχετικά εύκολα στο Tensorflow χρησιμοποιώντας την γλώσσα προγραμματισμού Python. Στο σχήμα 3.2.3 φαίνεται η υλοποίηση του LSTM.

Το πιο δύσκολο κομμάτι είναι η σωστή τροφοδότηση των εισόδων στο δίκτυο με την σωστή μορφή και ακολουθία. Σε αυτό το παράδειγμα στο LSTM τροφοδοτείται με μια ακολουθία από 3 ακέραιους αριθμούς. Στο σχήμα 3.2.4 φαίνεται ο κώδικας στον οποίο κάθε φορά παίρνουμε 3 σύμβολα από το κείμενο μετατρέπονται σε ακεραίους και μπαίνουν στο διάνυσμα εισόδου με το οποίο θα γίνει η τροφοδοσία καθώς και διάνυσμα εξόδου (symbols_out_onehot). Επίσης φαίνεται και ο ορισμός των biases και weights καθώς και τα hidden layers του LSTM.

```python
def RNN(x, weights, biases):

    # reshape to [1, n_input]
    x = tf.reshape(x, [-1, n_input])
    # Generate a n_input-element sequence of inputs
    # (eg. [had] [a] [general] -> [20] [6] [33])
    x = tf.split(x,n_input,1)
    # 1-layer LSTM with n_hidden units.
    rnn_cell = rnn.BasicLSTMCell(n_hidden)
    # generate prediction
    outputs, states = rnn.static_rnn(rnn_cell, x, dtype=tf.float32)
    # there are n_input outputs but
    # we only want the last output
    return tf.matmul(outputs[-1], weights['out']) + biases['out']
```

Σχήμα 3.2.3 Υλοποίηση ενός LSTM σε Python

```python
vocab_size = len(dictionary)
n_input = 3
# number of units in RNN cell
n_hidden = 512
# RNN output node weights and biases
weights = {
    'out': tf.Variable(tf.random_normal([n_hidden, vocab_size]))
}
biases = {
    'out': tf.Variable(tf.random_normal([vocab_size]))
}

symbols_in_keys = [ [dictionary[ str(training_data[i])]]
for i in range(offset, offset+n_input) ]

symbols_out_onehot = np.zeros([vocab_size], dtype=float)
symbols_out_onehot[dictionary[str(training_data[offset+n_input])]] = 1.0

_, acc, loss, onehot_pred = session.run([optimizer, accuracy, cost, pred],
    feed_dict={x: symbols_in_keys, y: symbols_out_onehot})
```

Σχήμα 3.2.4 Δημιουργία διανύσματος εισόδου, ορισμοί και optimizations

Στην εικόνα 3.2.4 και πιο συγκεκριμένα στην τελευταία εντολή γίνεται η βελτιστοποίηση των βημάτων εκπαίδευσης. Η ακρίβεια (accuracy) και απώλεια (loss) είναι

για την παρακολούθηση της προόδου της εκπαίδευσης. Με 50,000 επαναλήψεις στην εκπαίδευση μπορούμε να έχουμε μια αποδεκτή ακρίβεια όπως φαίνεται και στο σχήμα 3.2.5. Μπορούν να γίνουν αρκετές βελτιστοποιήσεις όσον αφορά το LSTM για παράδειγμα να χρησιμοποιήσεις επιπλέον επίπεδα. Με αυτό το τρόπο μπορείς να παίρνεις κάθε φορά το output να το δίνεις σαν είσοδο και να φτιάξεις μια ιστορία με βάσει ένα κείμενο που δόθηκε σαν είσοδο.

```
...
Iter= 49000, Average Loss= 0.528684, Average Accuracy= 88.50%
['could', 'easily', 'retire'] - [while] vs [while]
Iter= 50000, Average Loss= 0.415811, Average Accuracy= 91.20%
['this', 'means', 'we'] - [should] vs [should]
```

Σχήμα 3.2.5 Αποτέλεσμα εκπαίδευσης ενός LSTM με 50,000 επαναλήψεις

3.2.4 Play Framework

Το play framework είναι πλαίσιο εφαρμογών διαδικτύου ανοικτού κώδικα, γραμμένο σε γλώσσα προγραμματισμού Scala και μπορεί να χρησιμοποιηθεί σε πολλές άλλες γλώσσες προγραμματισμού οι οποίες συντάσσονται σε Bytecode π.χ JAVA, το οποίο ακολουθεί το αρχιτεκτονικό μοτίβο μοντέλο-προβολέα-ελεγκτή (MVC). Στόχος του συγκεκριμένου πλαισίου είναι να βελτιστοποιήσει την παραγωγικότητα των προγραμματιστών χρησιμοποιώντας συμβάσεις, επαναφόρτωση θερμού κώδικα και εμφάνιση σφαλμάτων στο πρόγραμμα περιήγησης(browser).

3.2.5 HTML/CSS/Javascript

Η γλώσσα σήμανσης υπερκειμένου (HTML) είναι η τυπική γλώσσα σήμανσης για την δημιουργία ιστοσελίδων και εφαρμογών διαδικτύου. Χρησιμοποιώντας Cascading Style

Sheets μπορείς να μορφοποιήσεις τα στοιχεία της ιστοσελίδας. Αυτά τα τρία μαζί αποτελούν τον ακρογωνιαίο λίθο του παγκόσμιου ιστού. Τα προγράμματα περιήγησης ιστού λαμβάνουν έγγραφα HTML από τους εξυπηρετητές ιστού και αποδίδουν τα έγγραφα σε ιστοσελίδες πολυμέσων. Η HTML περιγράφει την δομή μιας ιστοσελίδας σημασιολογικά και περιλαμβάνει στοιχεία για την εμφάνιση του εγγράφου.

3.2.6 Hadoop Distributed File System

Το HDFS χρησιμοποιείται για την κλιμάκωση ενός ενιαίου συμπλέγματος με εκατοντάδες (και ακόμη και χιλιάδες) κόμβους. Πρόκειται για ένα κατανεμημένο σύστημα αρχείων που χειρίζεται μεγάλα σύνολα δεδομένων που τρέχουν σε hardware.

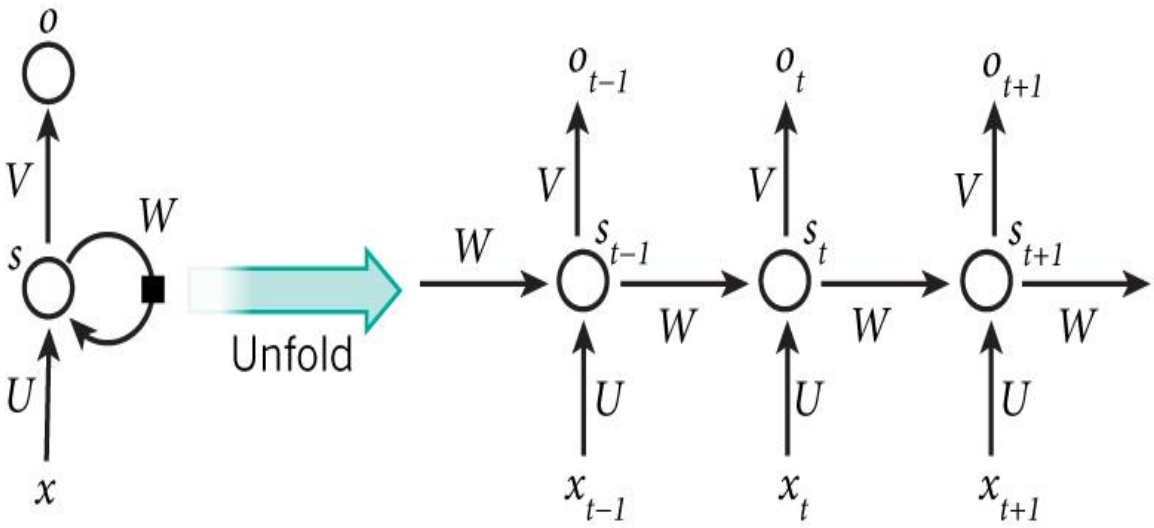
Ο στόχος του Hadoop είναι να χρησιμοποιεί τους servers σε ένα cluster όπου ο κάθε server έχει ένα σύνολο φθηνών δίσκων στο εσωτερικό του. Για υψηλότερες επιδόσεις , το MapReduce κατανέμει το φόρτο εργασίας στους servers στους οποίους πρόκειται να επεξεργαστούν και να αποθηκευτούν τα δεδομένα. Στο HDFS τα δεδομένα χωρίζονται σε blocks και αντίγραφα αυτών των block αποθηκεύονται στο Hadoop cluster. Δηλαδή ένα μεμονωμένο αρχείο αποθηκεύεται στην πραγματικότητα σαν μικρότερα blocks που αναπαράγονται σε πολλούς διακομιστές σε ολόκληρο το cluster.

3.3 Θεωρητικό Υπόβαθρο

Για την διεκπεραίωση της εν λόγω διπλωματικής εργασίας έγινε θεωρητική μελέτη όσον αφορά τα Αναδρομικά Νευρωνικά Δίκτυα (RNN) και Δίκτυα LSTM κα τις τεχνικές μάθησης τόσο για την υλοποίηση του αλγόριθμου DP αλλά και για την πειραματική αποτίμησή του.

3.3.1 Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN)

Η ιδέα πίσω από τα RNNs είναι η χρήση διαδοχικών πληροφοριών. Στα παραδοσιακά νευρωνικά δίκτυα υποθέτουμε ότι όλες οι εισόδοι και εξόδοι είναι ανεξάρτητες το ένα από



Σχήμα 3.3.1 Εξέλιξη του χρόνου του υπολογισμού ενός RNN [36]

το άλλο. Πολλές φορές αυτό δεν μπορεί να δώσει τα αναμενόμενα αποτελέσματα. Για παράδειγμα, αν θέλει κάποιος να προβλέψει τον επόμενο αριθμό σε μια αριθμητική έκφραση τότε πρέπει να γνωρίζει τους αριθμούς που ήρθαν προηγουμένως. Τα δίκτυα RNN ονομάζονται επαναλαμβανόμενα επειδή εκτελούν την ίδια εργασία για κάθε στοιχείο μιας ακολουθίας και η έξοδος εξαρτάται από τους προηγούμενους υπολογισμούς.

Το σχήμα 3.3.1 δείχνει ένα RNN που ξετυλίγεται σε ένα πλήρες δίκτυο. Με το ξεδίπλωμα εννοούμε την δημιουργία του δικτύου για την πλήρη ακολουθία. Για παράδειγμα εάν μας ενδιαφέρει μία ακολουθία 5 αριθμών τότε το δίκτυο θα ξεδιπλωθεί σέ 5 στρώματα ένα για κάθε αριθμό. Οι τύποι που διέπουν τον υπολογισμό που συμβαίνει σε ένα νευρωνικό δίκτυο είναι οι εξής:

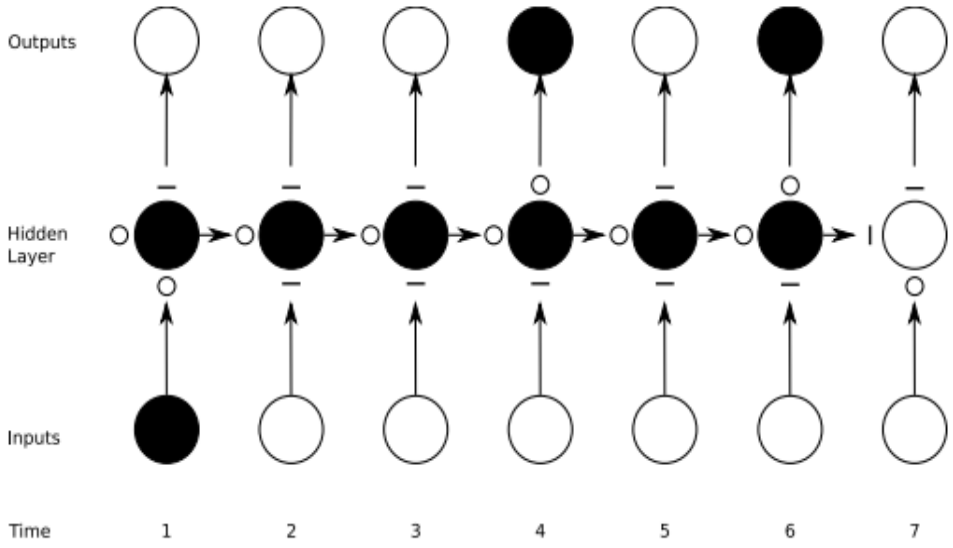
- Το x_t είναι η είσοδος το χρονικό βήμα t. Για παράδειγμα το x_1 θα μπορούσε να είναι ένα διάνυσμα που αντιστοιχεί στο δεύτερο στοιχείο μιας αριθμητικής έκφρασης.

- Το s_t είναι η κρυφή κατάσταση στο χρονικό βήμα t. Είναι η "μνήμη" του δικτύου. Το s_t υπολογίζεται με βάση την προηγούμενη κρυφή κατάσταση και την είσοδο στο τρέχον βήμα: $s_t = f(Ux_t + Ws_{t-1})$.

- Το o_t είναι η έξοδος στο βήμα t. Για παράδειγμα αν θέλαμε να προβλέψουμε τον επόμενο αριθμό σε μια έκφραση θα ήταν ένα διάνυσμα με πιθανότητες για το λεξιλόγιο μας.

3.3.2 Long Short-Term Memory (LSTM)

Τα LSTM δίκτυα είναι ο συνηθέστερος τύπος RNN που χρησιμοποιείται. Οι μονάδες μακράς βραχυπρόθεσμης μνήμης (LSTM) είναι μια δομική μονάδα για στρώματα επαναλαμβανόμενου νευρικού δικτύου (RNN). Ένα RNN δίκτυο που αποτελείται από μονάδες LSTM ονομάζεται LSTM δίκτυο. Μία μονάδα LSTM αποτελείται από ένα κελί (cell), μία πύλη εισόδου (input gate), μια πύλη εξόδου (output gate) και μια πύλη ξεχάσματος (forget gate). Το κελί είναι υπεύθυνο για τις τιμές απομνημόνευσης σε αυθαίρετα χρονικά διαστήματα, δηλαδή παίρνει τις αποφάσεις με το τι πρέπει να αποθηκεύσει και πότε πρέπει να επιτρέπεται η ανάγνωση, η εγγραφή και η διαγραφή μέσω της πύλης που ανοίγει και κλείνει. Κάθε μια από τις τρείς πύλες (εισόδου, εξόδου, ξεχάσματος) θεωρείται ένας νευρώνας όπως σε ένα νευρωνικό δίκτυο πολλαπλών στρωμάτων. Υπολογίζουν μια ενεργοποίηση ενός σταθμισμένου ποσού. Συγκεκριμένα, μπορούν να θεωρηθούν ως ρυθμιστές της ροής των τιμών που περνούν από τις συνδέσεις του LSTM γι' αυτό και χαρακτηρίζονται ως πύλες. Αυτές οι πύλες αντιδρούνε στα σήματα που λαμβάνουν και όπως γίνεται με τους κόμβους του νευρωνικού δικτύου μπλοκάρουν ή μεταδίδουν πληροφορίες τις οποίες φιλτράρουν με τα δικά τους βάρη. Τα βάρη αυτά, ρυθμίζονται μέσω τις διαδικασία εκμάθηση επαναλαμβανόμενων δικτύων. Δηλαδή τα κελιά μαθαίνουν πότε πρέπει να επιτρέπεται η είσοδος, η έξοδος ή η διαγραφή των δεδομένων μέσω της επαναληπτικής διαδικασίας υπολογισμού εικασιών, σφάλματος εκ νέου προώθησης και προσαρμογής των βαρών. Υπάρχουν διασυνδέσεις μεταξύ των πυλών και του κελιού.



Σχήμα 3.3.2 Αναπαράσταση των πυλών [38]

Στο σχήμα 3.3.2 βλέπουμε τις πύλες, με ευθείες γραμμές αναπαριστώνται οι κλειστές πύλες και με κενούς κύκλους οι ανοικτές πύλες. Οι γραμμές και οι κύκλοι που πηγαίνουν οριζόντια με το hidden layer είναι οι πύλες ξεχάσματος.

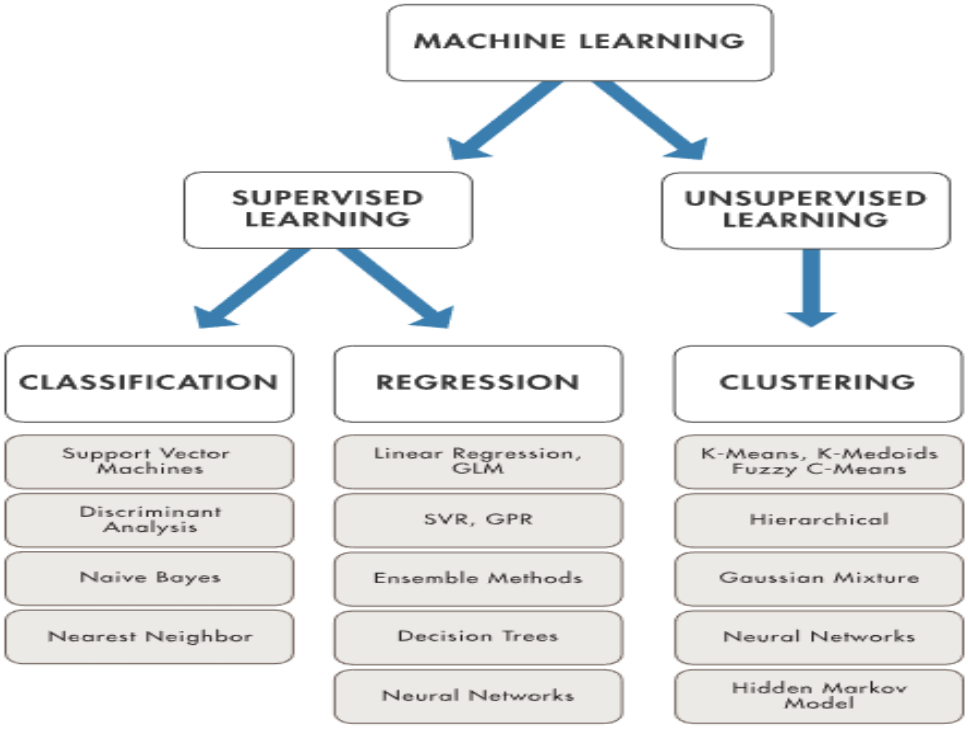
3.3.3 Μηχανική Μάθηση

Η μηχανική μάθηση είναι μια τεχνική ανάλυσης δεδομένων που διδάσκει στους υπολογιστές να κάνουν αυτά που έρχονται φυσικά στους ανθρώπους δηλαδή να μαθαίνουν από την εμπειρία. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν υπολογιστικές μεθόδους για να μαθαίνουν τις πληροφορίες απευθείας από τα δεδομένα χωρίς να στηρίζονται σε μια προκαθορισμένη εξίσωση ώς μοντέλο. Οι αλγόριθμοι βελτιώνουν προσαρμοστικά την απόδοση τους καθώς αυξάνεται ο αριθμός των διαθέσιμων δειγμάτων για μάθηση.

Με την άνοδο των μεγάλων δεδομένων στο κόσμο της πληροφορικής η μηχανική μάθηση έχει καταστεί βασική τεχνική για την επίλυση προβλημάτων σε τομείς όπως:

- Υπολογιστική οικονομικών, για αλγοριθμικές συναλλαγές.
- Επεξεργασία εικόνας και όρασης υπολογιστή για αναγνώριση προσώπου, ανίχνευση κίνησης και ανίχνευση αντικειμένων.
- Υπολογιστική βιολογία, για ανίχνευση όγκων, ανακάλυψη φαρμάκου και αλληλουχίας DNA.
- Επεξεργασία φυσικής γλώσσας για εφαρμογές κινητής αναγνώρισης.

Η μηχανική μάθηση χρησιμοποιεί δύο τύπους τεχνικών, την επιτηρούμενη μάθηση η οποία εκπαιδεύει ένα μοντέλο σε γνωστά δεδομένα εισόδου και εξόδου έτσι ώστε να μπορεί να προβλέψει μελλοντικές εξόδους, και μάθηση χωρίς επίβλεψη, η οποία βρίσκει μοτίβα στα δεδομένα εισόδου. Στο σχήμα 3.3.3 βλέπουμε τους 2 τύπους τεχνικής μάθησης καθώς και του αντίστοιχους αλγόριθμους για τον καθένα.



Σχήμα 3.3.3 Τεχνικές μηχανικής μάθησης [37]

i. Επιτηρούμενη μάθηση

Κατασκευάζει ένα μοντέλο που κάνει προβλέψεις βάσει αποδείξεων αλλά με παρουσία αβεβαιότητας. Ένας αλγόριθμος αυτού του είδους μάθησης παίρνει ένα γνωστό σύνολο δεδομένων εισόδου και γνωστών απαντήσεων στα δεδομένα (έξοδος) και εκπαιδεύει ένα μοντέλο για να παράγει λογικές προβλέψεις για την απάντηση σε νέα δεδομένα. Για παράδειγμα η πρόβλεψη της τιμής μιας μετοχής. Η επιτηρούμενη μάθηση χρησιμοποιεί τεχνικές ταξινόμησης και παλινδρόμησης για την ανάπτυξη προγνωστικών μοντέλων.

ii. Μάθηση χωρίς επιτήρηση

Η μη εποπτευόμενη μάθηση βρίσκει κρυμμένα μοτίβα η εγγενείς δομές στα δεδομένα. Χρησιμοποιείται κυρίως για την εξαγωγή συμπερασμάτων από σύνολα δεδομένων που αποτελούνται από δεδομένα εισόδου χωρίς επισημασμένες απαντήσεις. Για παράδειγμα την διάσπαση δεδομένων σε ομάδες (clusters).

Κεφάλαιο 4

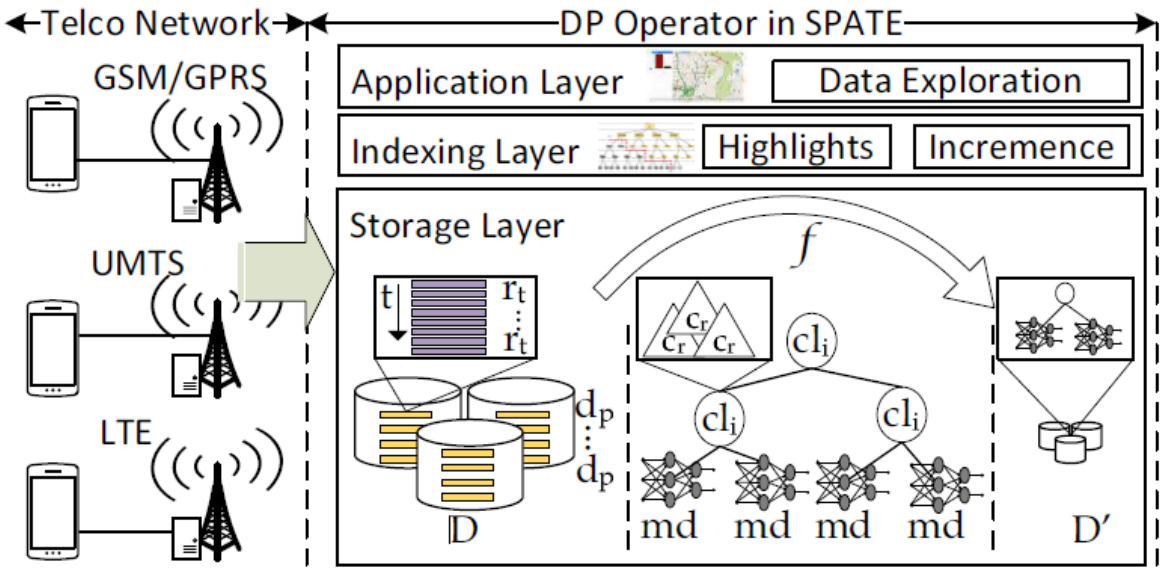
Τελεστής DP

4.1 Εισαγωγή

Αυτό το κεφάλαιο επισημοποιεί το μοντέλο του συστήματος, τις υποθέσεις και το πρόβλημα. Γίνεται μια αναφορά στο τηλεπικοινωνιακό σύστημα και το σύστημα διαχείρισης των τηλεπικοινωνιακών δεδομένων SPATE. Αναλύονται εις βάθος τα χαρακτηριστικά που απαρτίζουν ένα σύνολο τηλεπικοινωνιακών δεδομένων, το τι γίνεται με το που φτάσουν αυτά τα δεδομένα στα αντίστοιχα Data Centers. Στην συνέχεια κεφαλαίου περιγράφονται οι ερευνητικοί στόχοι καθώς και οι δύο μετρήσεις που χρησιμοποιούνται για την αξιολόγηση του συστήματος. Μετέπειτα θα γίνει επεξήγηση ενός νέου τελεστή για μεγάλα τηλεπικοινωνιακά δεδομένα που χρησιμοποιεί Data Postdiction. Αρχικά θα γίνει μία παρουσίαση του όρου Data Postdiction και στην συνέχεια θα γίνει αναφορά στους 2 αλγόριθμους του τελεστή DP για την κατασκευή των μοντέλων και την επαναφορά τους.

4.2 Μοντέλο Συστήματος

Ένα τυπικό σύστημα τηλεπικοινωνιών το οποίο απεικονίζεται στο σχήμα 4.2.1 αποτελείται από το τηλεπικοινωνιακό δίκτυο το οποίο είναι υπεύθυνο για την παροχή τηλεπικοινωνιακών υπηρεσιών στους διάφορους πελάτες και από ένα σύστημα διαχείρισης τηλεπικοινωνιακών δεδομένων όπως είναι το SPATE [1] του οποίου κύρια υπευθυνότητα του είναι η αποτελεσματική εξερεύνηση του συνόλου των δεδομένων που αποτελούνται από τα τηλεπικοινωνιακά δεδομένα (Telco datasets).



Σχήμα 4.2.1: Το σύστημα τηλεπικοινωνιών και τα κομμάτια που το απαρτίζουν [32]

Σημειογραφία	Περιγραφή
p, dp,D	Περίοδος λήψης, στιγμιότυπο δεδομένων ενός p, σύνολο όλων των dps
T,rt	Χρονοσήμανση εντός ενός κύκλου λήψης, καταγραφή στο t
C,cr,cli	Σύνολο όλων των Cell Towers, Cell of record r, cluster of records I=I,. . ., k

mdi,MD	LSTM μοντέλα του cluster cli, σύνολο των μοντέλων
f	Παράγοντας αποσύνθεσης δεδομένων (Decaying factor): ποσοστό των δεδομένων που θα αφαιρεθούν.

Πίνακας 4.2 Επεξήγηση Συμβόλων σχήματος 4.2.1

Ένα σύνολο δεδομένων αποτελείται κυρίως από :

- **Business Supporting Systems data (BSS):** Τα δεδομένα συστημάτων υποστήριξης επιχειρήσεων παράγονται από τα συστατικά IT του Telco και χρησιμοποιούνται για την εκτέλεση των επιχειρηματικών δραστηριοτήτων του προς τους πελάτες (π.χ Αρχεία βάσης χρηστών, συμβατά αρχεία, αρχεία χρέωσης, και voice/ message τα οποία ονομάζονται CDR- Call Detailed Records).

- **Operation Supporting Systems (OSS):** Τα δεδομένα συστημάτων υποστήριξης λειτουργίας παράγονται από τα υπολογιστικά συστήματα της telco που χρησιμοποιούνται για την διαχείριση των δικτύων της τα οποία δίκτυα περιλαμβάνουν στατιστικά στοιχεία ποιότητας μιας σύνδεσης κλήσης (call connection) όπως για παράδειγμα τον ρυθμό πτώσης κλήσεων (call drops) και τον ρυθμό σύνδεσης κλήσεων (call connection rate). Εκτός από τα αυτά τα στοιχεία περιλαμβάνουν επίσης και στοιχεία συμπεριφοράς του χρήστη στο δίκτυο κινητής τηλεφωνίας (πχ ταχύτητα web, ποσοστό επιτυχίας σύνδεσης, ερωτήματα αναζήτησης κινητής και δεδομένα που συλλέχθηκαν από ανιχνευτές χρησιμοποιώντας Deep Packet Inspection).

  Τα δεδομένα φτάνουν στο Data center σε παρτίδες, και από εκείνη την στιγμή και έπειτα ονομάζονται στιγμιότυπα δεδομένων με d_p, σε οριζόντια κατακερματισμένα (segmented) αρχεία με μια περίοδο λήψης p . Ένα στιγμιότυπο δεδομένων d_p περιέχει πολλαπλές εγγραφές rt που δημιουργήθηκαν σε μια συγκεκριμένη χρονική στιγμή, σε ένα timestamp δηλαδή t. Κάθε εγγραφή rt αποτελείται από ένα προκαθορισμένο σύνολο χαρακτηριστικών συμπεριλαμβανομένου και του αναγνωριστικού κυψέλης

cell ID cr το οποίο αντιπροσωπεύει τις χωρικές πληροφορίες σε ένα δίκτυο τηλεπικοινωνιών. Συγκεκριμένα κάθε αναγνωριστικό κυψέλης αντιστοιχεί σε ένα κελί που καλύπτει μια γεωγραφική κυψελοειδή περιοχή. Η περιοχή αυτή ως συνήθως εκτείνεται σε εκατοντάδες μέτρα. Τέλος, οι κυψέλες είναι χωρικά ομαδοποιημένες στα clusters cli, i = i . . . k διευκολύνοντας την διαδικασία του Postdiction με την δημιουργία ενός μοντέλου mdi, i = i . . . k για κάθε cli όπως θα εξηγηθεί εκτενέστερα στην επόμενη ενότητα.

4.3 Διατύπωση προβλήματος

Ο στόχος της έρευνας : Δεδομένης της σύνθεσης ενός Telco, η παρούσα διπλωματική εργασία στοχεύει κατά κύριο λόγο αφενός στην επίτευξη μίας καθορισμένης αποσύνθεσης τηλεπικοινωνιακών δεδομένων μεγάλου όγκου που αυτό μας οδηγεί στην ελαχιστοποίηση της επιπλέον χωρητικότητας του χώρου αποθήκευσης και αφετέρου στην ικανότητα για ανάκτηση των αποσυντεθημένων δεδομένων με ακρίβεια και αποτελεσματικότητα.

Η αποτελεσματικότητα των προτεινόμενων τεχνικών για την επίτευξη των προαναφερθέντων στόχων μετριέται χρησιμοποιώντας τα ακόλουθα στοιχεία όσον αφορά τα δεδομένα και τον χώρο αποθήκευσής τους.

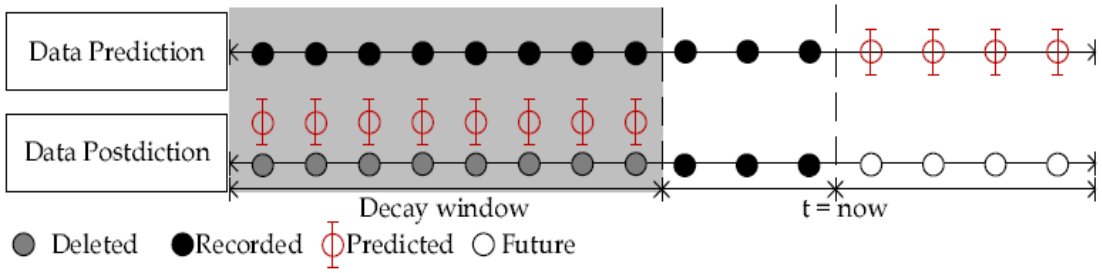
- **Χωρητικότητα Αποθήκευσης (S) :** Είναι ο συνολικός αποθηκευτικός χώρος που απαιτείται για να επιτευχθεί η αποσύνθεση των τηλεπικοινωνιακών δεδομένων με βάσει ένα προκαθορισμένο συντελεστή αποσύνθεσης f .

- **Ακρίβεια αποτελέσματος (NRMSE) :** Ακρίβεια (accuracy) είναι το ποσοστό των τηλεπικοινωνιακών αποσυντεθημένων δεδομένων που ανακτήθηκαν σωστά χρησιμοποιώντας ένα Postdiction μοντέλο. Αυτό αποτιμάται χρησιμοποιώντας το κανονικοποιημένο τετραγωνικό σφάλμα ρίζας (root mean squared error), το οποίο είναι η κανονικοποιημένη διαφορά μεταξύ των

προβλεπόμενων, δηλαδή των δεδομένων που υπολογίστηκαν χρησιμοποιώντας το Postdiction μοντέλο και των πραγματικών δεδομένων πριν αποσυντεθούν.

4.4 Data Postdiction

Σε αντίθεση με την πρόβλεψη δεδομένων (data prediction), η οποία στοχεύει στην πρόβλεψη μιας μελλοντικής τιμής σε μια πλειάδα, ο αλγόριθμος Data Postdiction στοχεύει στην πρόβλεψη μιας προηγούμενης τιμής κάποιας πλειάδας που δεν υπάρχει πια καθώς διαγράφηκε



Σχήμα 4.4.1 Data Prediction and Data Postdiction [32]

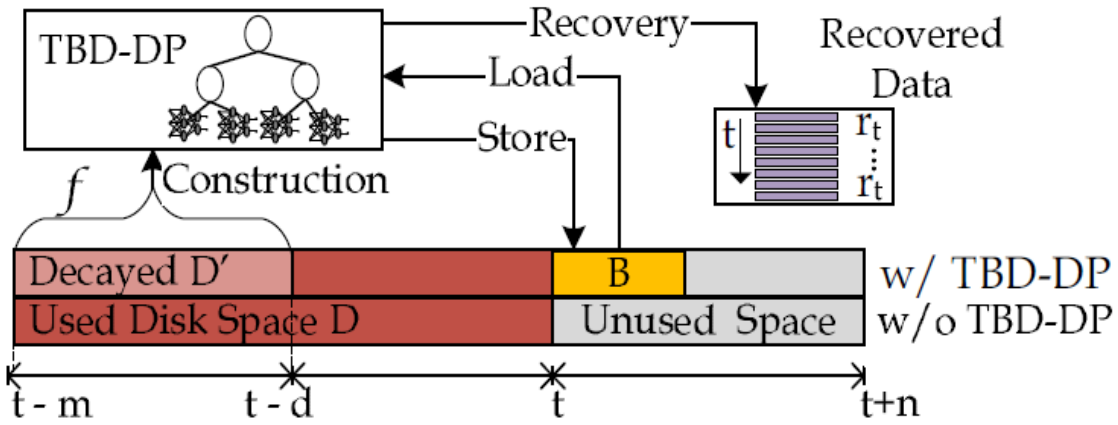
από τον χώρο αποθήκευσης για την ελευθέρωση χώρου με βάσει την αμνησία δεδομένων. Αυτό φαίνεται και στο σχήμα 4.4.1. Ο DP βασίζεται στους υπάρχοντες αλγόριθμους μηχανικής μάθησης

με τους οποίους μετατρέπουμε τα μεγάλα τηλεπικοινωνιακά δεδομένα σε συμπαγή μοντέλα που μπορούν να αποθηκευτούν στον χώρο καταλαμβάνοντας λιγότερη μνήμη και τα οποία μπορούν να φορτωθούν όποτε είναι απαραίτητο. Ο προτεινόμενος τελεστής DP περιλαμβάνει τις δύο ακόλουθες φάσεις:

i. Χρησιμοποιεί έναν ιεραρχικό αλγόριθμο τεχνική μάθησης βασισμένος σε LSTM για να μάθει ένα δέντρο με μοντέλα σε σχέση με το χώρο και χρόνο.

ii. Χρησιμοποιεί αυτό το δέντρο για την επαναφορά των μοντέλων για την ανάκτηση των δεδομένων με μια ορισμένη ακρίβεια.

Στο σχήμα 4.4.2 παρουσιάζονται πώς τα εισερχόμενα σήματα τηλεπικοινωνιακών δεδομένων απορροφώνται από την αρχιτεκτονική TBD και αποθηκεύονται στον χώρο αποθήκευσης (D). Αυτό βοηθά στην μετέπειτα εκτέλεση λειτουργικών διεργασιών με την ανάλυση των δεδομένων. Στην πρώτη φάση του τελεστή DP χρησιμοποιούμε ένα επαναλαμβανόμενο νευρωνικό δίκτυο που αποτελείται από μονάδες LSTM το οποίο έχει την δυνατότητα να ανιχνεύει μακροχρόνιες συσχετίσεις στα δεδομένα και το εκπαιδευμένο μοντέλο έχει μικρό αποτύπωμα χώρου στο δίσκο. Αυτό επιτρέπει στο τελεστή DP να χρησιμοποιεί την ελάχιστη χωρητικότητα αποθήκευσης των αποσυντεθημένων δεδομένων τα οποία αντιστοιχούν σε LSTM μοντέλα στο δίσκο (D') και



Σχήμα 4.4.2 Τελεστής DP για μεγάλα τηλεπικοινωνιακά δεδομένα [32]

παρέχοντας σε πραγματικό χρόνο προβλέψεις (postdictions) στην επόμενη φάση όπου γίνεται η επαναφορά των μοντέλων εφόσον μια διεργασία ζητήσει αυτά τα δεδομένα. Μια σφαιρική εικόνα του τελεστή DP στο SPATE φαίνεται στο σχήμα 4.2.1.

4.5 Κατασκευή Μοντέλων

Ο αλγόριθμος της Κατασκευής των μοντέλων μπορεί να τεθεί σε λειτουργία είτε από τον χρήστη η αυτόματα όταν η χωρητικότητα αποθήκευσης φτάσει σε ένα ορισμένο επίπεδο. Και στις δύο περιπτώσεις τα δεδομένα αρχικά διαχωρίζονται βάσει χωρικών χαρακτηριστικών και στην συνέχεια ταξινομούνται βάσει χρονικών χαρακτηριστικών. Για τον κάθε διαχωρισμό δημιουργούνται τα μοντέλα με βάσει την προσέγγιση της μηχανικής μάθησης για τα LSTM μοντέλα και τα πραγματικά δεδομένα αποσυντίθενται με βάσει κάποιο παράγοντα f (decaying factor).

Ακολουθεί ο αλγόριθμος για την κατασκευή των μοντέλων:

Είσοδος: Δίνουμε σαν είσοδο στο πρόγραμμα ένα σύνολο δεδομένων D, ένα σύνολο κυψελοειδών πύργων (cell towers) C και ένα αριθμό ομάδων k.

Έξοδος: Σαν έξοδο Β παίρνουμε το σύνολο των μοντέλων M D.

i. **Βήμα 1:** Επιλέγουμε την τιμή του Decaying factor f. Δηλαδή επιλέγουμε το ποσοστό των δεδομένων που θα "ξεχαστούν" .
ii. **Βήμα 2:** Σε αυτό το σημείο γίνεται ο χωρικός διαμοιρασμός των δεδομένων σε ομάδες χρησιμοποιώντας την τοποθεσία των κυψελοειδών πύργων και το πεδίο cell id της κάθε εγγραφής.
iii. **Βήμα 3:** Σε αυτό το σημείο γίνεται ο χρονική ταξινόμηση. Όλες οι εγγραφές της κάθε ομάδας ταξινομούνται βάσει του χρόνου που παράχθηκε η κάθε μια ξεχωριστά.
iv. **Βήμα 4:** Στο τελευταίο βήμα για κάθε ομάδα k δημιουργούνται τα postdiction μοντέλα χρησιμοποιώντας ένα εξειδικευμένο επαναλαμβανόμενο νευρωνικό δίκτυο που αναφέραμε πιο πάνω το LSTM.

Ο συγκεκριμένος αλγόριθμος κατασκευής των μοντέλων εξάγει ένα σύνολο από Postdiction μοντέλα σε ένα DP δέντρο για να διευκολύνει τον αλγόριθμο επαναφοράς των μοντέλων που ακολουθεί. Με το πέρας του αλγόριθμου κατασκευής των μοντέλων το σύνολο των δεδομένων που "ξεχάστηκαν" διαγράφονται με σκοπό την εξοικονόμηση χώρου αποθήκευσης και αντικαθίστανται από τα Postdiction μοντέλα που κατασκευάστηκαν.

4.6 Επαναφορά Μοντέλων

Ο αλγόριθμος επαναφοράς των μοντέλων χρησιμοποιεί την δομή του DP δέντρου που περιέχει τα Postdiction μοντέλα του αλγόριθμου κατασκευής για την ανάκτηση ενός επιλεγμένου υποσυνόλου δεδομένων από τα διαγραφόμενα δεδομένα.

Ο αλγόριθμος για την επαναφορά των μοντέλων έχει ως εξής :

Είσοδος: Σύνολο των δεδομένων καθώς και κάποιες χωροχρονικές πληροφορίες οι οποίες θα καθορίσουν την ποσότητα των αποσυνθεμένων δεδομένων που θα ανακτηθούν.

Έξοδος: Τα αποσυνθεμένα δεδομένα.

i. **Βήμα 1:** Χρησιμοποιώντας μια συνάρτηση LOOKUP ανακτούμε από το DP δέντρο το υποσύνολο των μοντέλων.
ii. **Βήμα 2:** Χρησιμοποιώντας τα LSTM μοντέλα ανακτούμε τα αποσυνθεμένα δεδομένα.

Κεφάλαιο 5

Περιγραφή Πρωτοτύπου

5.1 Εισαγωγή

Στο παρόν κεφάλαιο θα γίνει η περιγραφή του πρωτοτύπου που υλοποιήθηκε πάνω στο SPATE [1]. Παρουσιάζονται οι λειτουργίες που παρέχει το πρωτότυπο μέσω της γραφικής διαπροσωπείας που παρέχεται και πώς ο χρήστης μπορεί να αλληλοεπιδράσει με το σύστημα και να δει στην πράξη την σημασία του τελεστή DP πραγματοποιώντας προβλέψεις σε δεδομένα που ήδη έχουν καταστραφεί . Στο πρωτότυπο τα αποτελέσματα φαίνονται σε γραφικές παραστάσεις και οπτικά χρησιμοποιώντας χάρτες θερμότητας όπου φαίνονται τα πραγματικά δεδομένα και τα δεδομένα που υπολογίστηκαν χρησιμοποιώντας Data Postdiction.

5.2 Back-end

Η προτεινόμενη αρχιτεκτονική αποτελείται από 3 επίπεδα τα οποία ονομαστικά είναι το επίπεδο Αποθήκευσης (Storage Layer) το επίπεδο ευρετηρίου (Indexing Layer) και το επίπεδο εφαρμογής (Application Layer). Τα 3 επίπεδα εξηγούνται παρακάτω:

- **Storage Layer**

Το επίπεδο Storage αποθηκεύει τα νέα στιγμιότυπα του δικτύου μέσω μιας διαδικασίας συμπίεσης χωρίς απώλειες και αποθηκεύει τα αποτελέσματα σε ένα σύστημα αρχείου δεδομένων για διαθεσιμότητα και απόδοση. Το επίπεδο αυτό είναι υπεύθυνο για την ελαχιστοποίηση του απαιτούμενου χώρου αποθήκευσης με ελάχιστη επιβάρυνση στον χρόνο απόκρισης του ερωτήματος. Το επίπεδο Storage επιπλέον χρησιμοποιεί τον τελεστή DP για την παροχή των μεθόδων αποσύνθεσης δεδομένων για το επόμενο επίπεδο.

- **Indexing Layer**

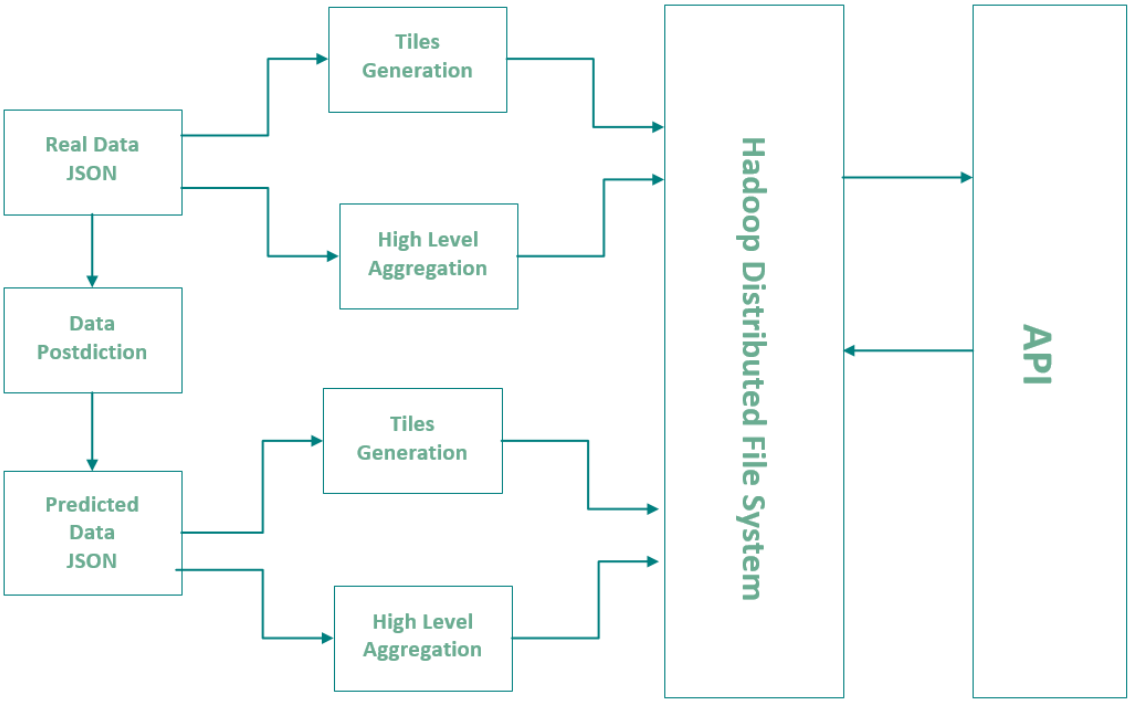
Το επίπεδο Indexing χρησιμοποιεί ένα πολλαπλών διαστάσεων χωροχρονικό δείκτη ο οποίος αυξάνεται με κάθε νέο στιγμιότυπο που καταφθάνει, δηλαδή κάθε 30 λεπτά.

- **Application Layer**

  Το επίπεδο Application υλοποιεί την ενότητα επερώτησης και τις γραφικές διαπροσωπείες διερεύνησης δεδομένων, οι οποίες λαμβάνουν τα επερωτήματα διερεύνησης δεδομένων σε οπτική λειτουργία και χρησιμοποιούν το Indexing επίπεδο για να συνδυάσουν τις απαραίτητες επισημάνσεις και στιγμιότυπα για την απάντηση του επερωτήματος. Το SPATE είναι εξοπλισμένο με ένα εύκολο στην χρήση χάρτη που κρύβει την πολυπλοκότητα του συστήματος με μια απλή διεπαφή ιστού.

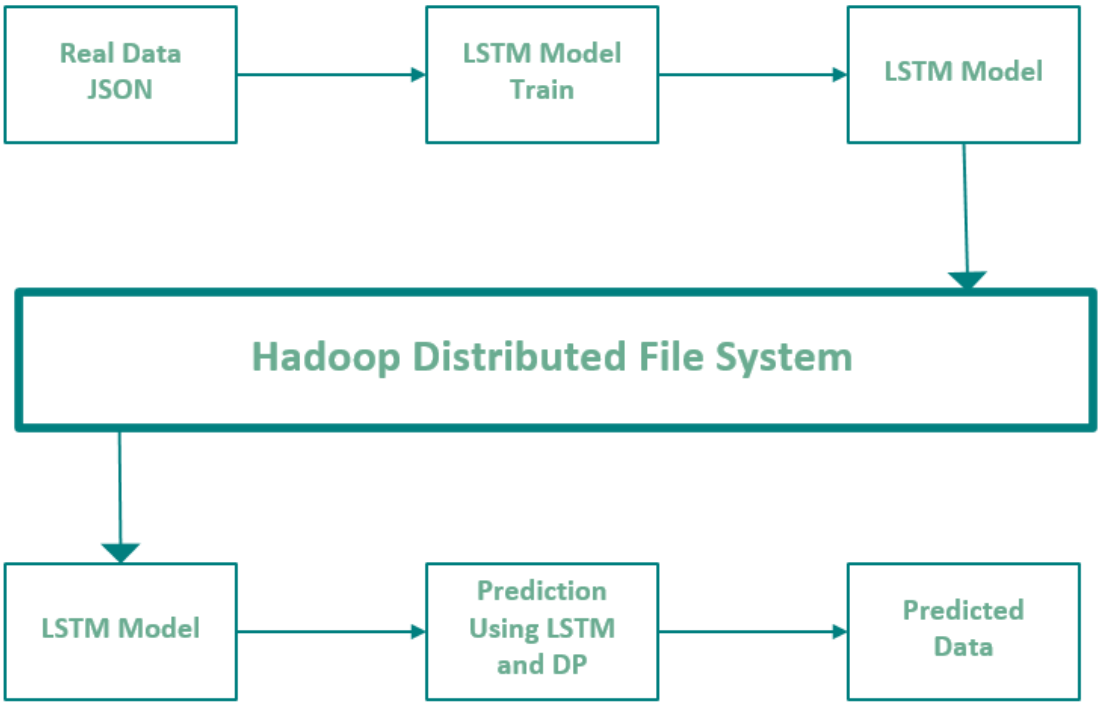
Το σχήμα 5.2.1 περιγράφει την διαδικασία με την οποία διαχωρίζονται τα δεδομένα μέσω μιας μηχανής σε tiles καθώς και την αποθήκευσης τους στο κατανεμημένο σύστημα HDFS για την φόρτωση τους στη διαπροσωπεία χωρίς να υπάρχει υπερφόρτωση δεδομένων στον φυλλομετρητή που θα οδηγήσει προβλήματα απόδοσης. Τα δεδομένα φορτώνονται από JSON αρχεία και μέσω ενός υποπρογράμματος γίνεται ο διαχωρισμός τους με βάσει την χρονοσφραγίδα της κάθε εγγραφής σε tiles (όσον αφορά το τελευταίο επίπεδο μεγέθυνσης του παραθύρου πλοήγησης). Για τα πιο πάνω επίπεδα μεγέθυνσης του παραθύρου πλοήγησης πραγματοποιείται high-level aggregation για σκοπούς απόδοσης. Τα δεδομένα

αφού χωριστούν (κάθε 30 λεπτά) αποθηκεύονται στο HDFS. Το ίδιο ισχύει και για τα δεδομένα που προβλέπονται χρησιμοποιώντας Data Postdiction.



Σχήμα 5.2.1 Διαχωρισμός των δεδομένων σε Tiles και Average καθώς και η αποθήκευση τους στο κατανεμημένο σύστημα HDFS

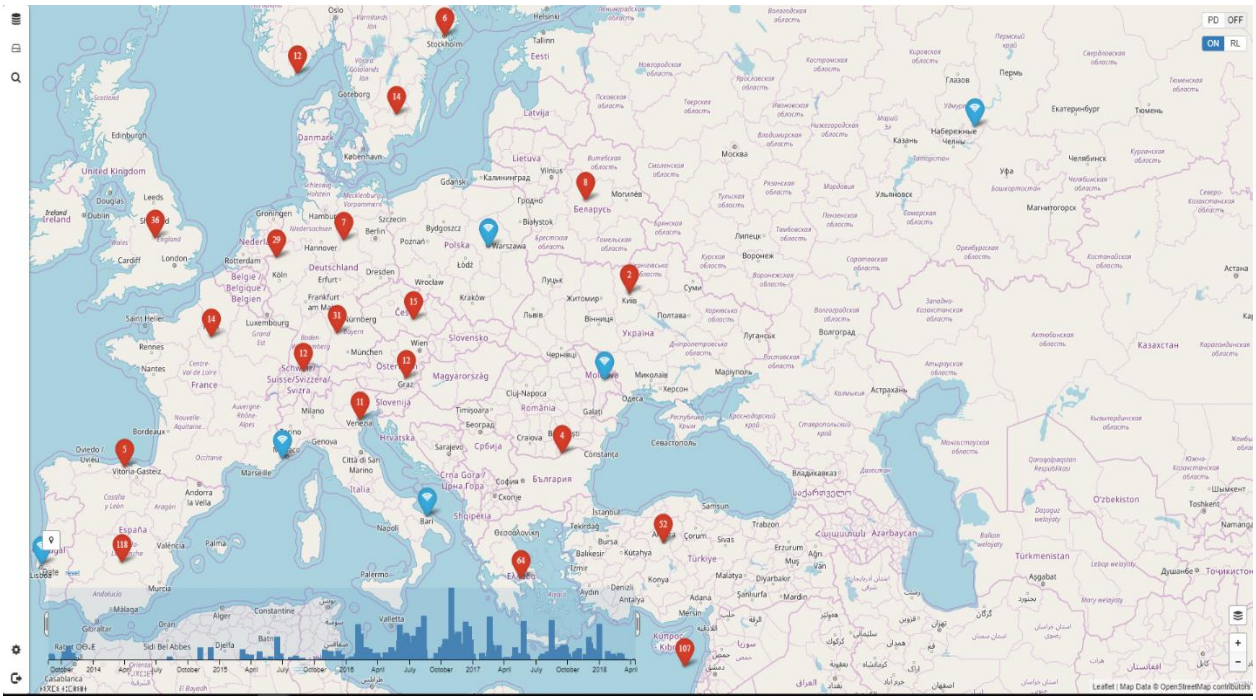
Τα δεδομένα πριν διαγραφούν από τον χώρο αποθήκευσης δίνονται σαν είσοδο στον τελεστή DP για την ανάλυσή τους. Το σχήμα 5.2.2 πιο συγκεκριμένα περιγράφει την διαδικασία που υλοποιεί ο τελεστής DP (δηλαδή το κομμάτι Data Postdiction του σχήματος 5.2.1). Το πρώτο κομμάτι αφορά την κατασκευή των μοντέλων και εκπαίδευση τους. και το δεύτερο αφορά την επαναφορά των μοντέλων για πρόβλεψη των τιμών. Τα δεδομένα προς αποσύνθεση δίνονται σαν δεδομένα εκπαίδευσης στο τελεστή DP ο οποίος κατασκευάζει ένα LSTM μοντέλο το οποίο το εκπαιδεύει χρησιμοποιώντας τις υπάρχουσες τιμές. Το μοντέλο LSTM αποθηκεύεται στο κατανεμημένο σύστημα αποθήκευσης HDFS το οποίο αντικαθιστά τα πραγματικά δεδομένα τα οποία διαγράφονται.



Σχήμα 5.2.2 Περιγράφεται το Data Postdiction κομμάτι του σχήματος 6.2.1 όπου πριν τα καταστροφή των δεδομένων έχουμε την κατασκευή του LSTM και την μετέπειτα επαναφορά τους.

Τα δεδομένα στην συνέχεια καταστρέφονται από τον χώρο αποθήκευσης για την εξοικονόμηση χώρου. Όταν και εφόσον ζητηθεί από τον χρήστη, είτε αυτόματα από το σύστημα τότε φορτώνεται στο σύστημα το εκπαιδευμένο πλέον μοντέλο και με βάσει κάποια χρονοσφραγίδα υπολογίζεται η τιμή. Τα προβλεπόμενα δεδομένα αποθηκεύονται στο HDFS για περαιτέρω ανάλυση τους στην συνέχεια και παρουσίαση τους στην γραφική διαπροσωπεία σαν δεδομένα υπολογισμένα χρησιμοποιώντας Data Postdiction. Πιο συγκεκριμένα το σχήμα 5.2.2 περιγράφει τις δύο φάσεις του τελεστή DP, κατασκευή μοντέλων και επαναφορά μοντέλων οι οποίες επεξηγήθηκαν στο προηγούμενο κεφάλαιο όπου αναλύθηκε η λειτουργία του τελεστή DP.

5.3 Front-end



Σχήμα 5.3.1 Η γραφική διεπαφή σε μηδενικό επίπεδο μεγέθυνσης

Ο τελεστής DP υλοποιήθηκε πάνω από την χωρο-χρονική αρχιτεκτονική του SPATE. Η γραφική διαπροσωπεία επιτρέπει στους χρήστες να πραγματοποιούν οπτικές αναλύσεις των δεδομένων χωρίς να καταναλώνουν τεράστια ποσά αποθήκευσης.

Όσον αφορά το χάρτη χρησιμοποιήθηκε ένας Open StreetMap παγκόσμιος χάρτης όπως φαίνεται στην εικόνα 5.3.1 όπου ο χρήστης μπορεί να περιηγηθεί στα διάφορα μέρη του κόσμου και να δει οπτικά τα δεδομένα. Στο κάτω μέρος του χάρτη υπάρχει ένα χρονικό ραβδόγραμμα με το οποίο ο χρήστη μπορεί να επιλέξει το παράθυρο χρόνου στο οποίο θέλει να κάνει τις οπτικές αναλύσεις των δεδομένων. Στο ραβδόγραμμα απεικονίζονται αναπαραστάσεις των δεδομένων τις αντίστοιχες χρονικές στιγμές. Σε σημεία όπου δεν υπάρχει αναπαράσταση στο ραβδόγραμμα τότε δεν υπάρχουν δεδομένα για εμφάνιση τους στο χάρτη. Επιπρόσθετα ο χρήστης έχει την επιλογή να αλλάξει την διαμόρφωση του χάρτη όπως φαίνεται και στο σχήμα 5.3.2 καθώς και να μεγεθύνει το χάρτη σε 18 επίπεδα. Όταν ο χρήστης βρίσκεται στο μηδενικό επίπεδο μεγέθυνσης μέχρι και το προτελευταίο επίπεδο

Σχήμα 5.3.2 Η γραφική διεπαφή σε μηδενικό επίπεδο μεγέθυνσης με αλλαγμένη διαμόρφωση του χάρτη

(δηλαδή το 17) φαίνονται τα σημεία τα οποία, στο χρονικό παράθυρο που είναι επιλεγμένο, περιέχουν κάποια δεδομένα. Τα δεδομένα στα πιο πάνω επίπεδα αναπαριστώνται με καρφίτσες (pins) όπως φαίνονται και στις εικόνες 5.3.1 και 5.3.2. Οι προαναφερθέντες καρφίτσες δείχνουν το high-level aggregation που έγινε για λόγους απόδοσης και εμφάνισης. Ο αριθμός που περιλαμβάνει η κάθε καρφίτσα δείχνει τα διαφορετικά timestamps που υπάρχουν δεδομένα στο επιλεγμένο χρονικό παράθυρο.

Στο σχήμα 5.3.3 φαίνεται το τελευταίο επίπεδο μεγέθυνσης όπου ο χρήστης μπορεί να δει σε εκτενή ανάλυση χρησιμοποιώντας χάρτες θερμότητας τα πραγματικά δεδομένα που υπάρχουν σε κτήρια σύμφωνα με μετρήσεις τα οποία απεικονίζουν το σήμα ενός χρήστη. Το κουμπί RL (Real) που βρίσκεται πάνω δεξιά δίνει την δυνατότητα στο χρήστη να εμφανίσει αυτά τα δεδομένα ή να τα κρύψει. Καθώς ο χρήστης μετακινείται στο χάρτη (tiles) τα

Σχήμα 5.3.3 Η γραφική διεπαφή στο τελευταίο επίπεδο μεγέθυνσης όπου ο χρήστης μπορεί να δει με χάρτες θερμότητας τα πραγματικά δεδομένα

δεδομένα φορτώνονται από τον αποθηκευτικό χώρο λαμβάνοντας πάντα υπόψιν το επιλεγμένο χρονικό παράθυρο.

Στο σχήμα 5.3.4 φαίνονται τα πραγματικά δεδομένα τα οποία αφορούν δικτυακή κάλυψη εντός των κτηρίων (στο παράδειγμα φαίνεται η Πανεπιστημιούπολη) όμως φαίνεται και η χρήση του τελεστή DP. Κάποια δεδομένα υπολογίστηκαν χρησιμοποιώντας τον τελεστή DP και διαγράφηκαν από τον χώρο αποθήκευσης για εξοικονόμηση χώρου. Τα προβλεπόμενα δεδομένα φαίνονται επίσης με χάρτες θερμότητας ο οποίο μπαίνει πάνω από το χάρτη θερμότητας που απεικονίζει τα πραγματικά δεδομένα. Με αυτό το τρόπο, ο χρήστης μπορεί να δει την ακρίβεια των πραγματικών δεδομένων συγκρίνοντας τα 2 heatmaps. Στο σχήμα 5.3.4 με πορτοκαλί χρώμα φαίνονται τα δεδομένα που προβλέφθηκαν χρησιμοποιώντας τον προτεινόμενο αλγόριθμο Data Postdiction. την χρήση του κουμπιού PD (Predicted) το οποίο βρίσκεται επίσης πάνω δεξιά. Τα δεδομένα του χάρτη θερμότητας που απεικονίζει τα προβλεπόμενα δεδομένα προβλέφθηκαν με την επαναφορά των μοντέλων LSTM από τον κατανεμημένο χώρο αποθήκευσης, αφού αντικατέστησαν τα πραγματικά δεδομένα, σε μια φάση εκτός σύνδεσης (offline).

Σχήμα 5.3.4 Η γραφική διεπαφή στο τελευταίο επίπεδο μεγέθυνσης όπου ο χρήστης μπορεί να δει πραγματικά δεδομένα και προβλεπόμενα δεδομένα με την χρήση του DP

Το σχήμα 5.3.5 δείχνει την αποτελεσματικότητα του τελεστή DP όσον αφορά την εξοικονόμηση του χώρου αποθήκευσης. Στα αριστερά βλέπουμε μια γραφική παράσταση η οποία απεικονίζει το μέγεθος (σε ποσοστό) των πραγματικών δεδομένων(RAW) στο κατανεμημένο σύστημα αποθήκευσης HDFS και το μέγεθος των μοντέλων που αντικατέστησαν τα αποσυνθεμένα δεδομένα με την χρήση του τελεστή DP (With DP). Παρατηρώντας την γραφική παράσταση τα μοντέλα των δεδομένων που αποσυντέθηκαν καταλαμβάνουν μόλις το ~20% σε σχέση με τα πραγματικά δεδομένα. Αξιοσημείωτο είναι να αναφέρουμε ότι το μέγεθος των μοντέλων είναι σταθερό για οποιοδήποτε σύνολο δεδομένων που αφορά ένα κτήριο.

Σχήμα 5.3.5 Τελεστής DP και χωρητικότητα αποθήκευσης

Χρησιμοποιώντας την γραφική αυτή διαπροσωπεία ο χρήστης μπορεί να εκτελέσει χωροχρονικά επερωτήματα για να αλληλοεπιδράσει με το σύστημα το οποίο δείχνει την ακρίβεια των προβλεπόμενων δεδομένων αλλά και την εξοικονόμηση χώρου αποθήκευσης με την χρήση του τελεστή DP συγκρίνοντας τους χάρτες θερμότητας των πραγματικών και προβλεπόμενων δεδομένων και παρατηρώντας το γράφημα.

Κεφάλαιο 6

Πειραματική Αποτίμηση και Αξιολόγηση

6.1 Εισαγωγή

Σε αυτή την ενότητα παρουσιάζεται μια πειραματική αξιολόγηση του προτεινόμενου τελεστή DP. Αρχικά θα δούμε την εγκατάσταση και την πειραματική μεθοδολογία που ακολουθήσαμε ακολουθούμενη από δυο σειρές πειραμάτων. Στην πρώτη σειρά πειραμάτων οι επιδόσεις του DP συγκρίνονται με δύο προσεγγίσεις βασικής γραμμής και δύο προσεγγίσεις με αποσύνθεση δεδομένων σε σχέση με διάφορες μετρήσεις σε ένα σύνολο ανώνυμων δεδομένων. Στην δεύτερη σειρά πειραμάτων εξετάζεται η επίδραση πολλών παραμέτρων ελέγχου στην απόδοση του τελεστή DP.

6.2 Μεθοδολογία

Στην υποενότητα αυτή παρουσιάζονται λεπτομέρειες σχετικά με τους αλγόριθμους τις μετρικές και τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση της προτεινόμενης προσέγγισης.

Testbed

Η αξιολόγηση μας πραγματοποιείται στο datacenter του DMSL Vcenter IaaS, ένα ιδιωτικό σύστημα νέφους, το οποίο περιλαμβάνει πέντε IBM System x3550 M3 και HP ProLiant DL 360 G7 που διαθέτουν ένα Socket (8 πυρήνες) η διπλό Socket (16 πυρήνες) Xeon CPU E5620 @ 2.40 GHz αντίστοιχα. Οι υπολογιστές αυτοί έχουν συλλογικά 300 GB κύριας μνήμης, 16TB αποθήκευσης RAID-5 σε ένα IBM 3512 και είναι διασυνδεδεμένοι μέσω ενός δικτύου Gigabit. Το datacenter διαχειρίζεται μέσω VMWare vCenter Server 5.1 που συνδέεται με τους αντίστοιχους κεντρικούς υπολογιστές VMWare ESXi 5.0.0. Το σύμπλεγμα υπολογιστών που λειτουργεί πάνω από το VCenter IaaS αποτελείται από τέσσερεις εικονικούς εξυπηρετητές Ubuntu 16.04 με τον καθένα να έχει 8GB μνήμης RAM και δύο εικονικούς επεξεργαστές στα 2.40 GHz. Χρησιμοποιούν τοπικούς γρήγορους δίσκους 10K RPM RAID-5 LSILogic SCSI διαμορφωμένους με VMFS 5.54 (μέγεθος μπλοκ 1 MB). Κάθε κόμβος χρησιμοποιεί το Hadoop v2.5.2.

Χρησιμοποιούμε ανώνυμες μετρήσεις από ένα πραγματικό παροχέα τηλεπικοινωνιακών υπηρεσιών που αποτελείται από 1192 πραγματικούς κυψελοειδής πύργους (3360 κελιά από 2G, 3G και LTE δίκτυα) οι οποίοι διανέμονται σε μια έκταση 5896 τετραγωνικών χιλιομέτρων. Τα κελιά συνδέονται μέσω ενός δικτύου Gigabit στο κέντρο δεδομένων. Κάθε κυψελοειδής πύργος διατηρεί αρκετά αρχεία καταγραφής δικτύου UMTS/GSM για την απόδοση του πύργου και προωθεί τις πληροφορίες μέσω του ελεγκτή του σταθμού βάσης (BSC) ή του ελεγκτή ασύρματου δικτύου (RNC) που πρόκειται να αποθηκευτεί. Ακόμα υπάρχει ένας διακομιστής CDR ο οποίος παράγει όλες τις εγγραφές που περιλαμβάνουν τις λεπτομέρειες κλήσεων (CDRs) για εισερχόμενες και εξερχόμενες κλήσεις στην επιχείρηση. Όταν δημιουργείται ένα CDR στον διακομιστή CDR, ο διακομιστής διαχείρισης και μια τρίτη εφαρμογή χρησιμοποιούν το πρωτόκολλο SFTP για να αποκτήσουν την CDR πληροφορία από τον διακομιστή CDR. Στην συνέχεια ο τηλεπικοινωνιακός παροχέας μπορεί να ζητήσει από τους CDR πληροφορείς σχετικά με κλήσεις/δεδομένα και να ελέγξει τις εξερχόμενες χρεώσεις κλήσεων δεδομένων με τον μεταφορέα.

Αλγόριθμοι

Ο προτεινόμενος τελεστής DP συγκρίνεται με τις ακόλουθες προσεγγίσεις:

- **RAW** : Δεν εφαρμόζει καμία φθορά στο σύνολο των δεδομένων.
- **ΣΥΜΠΙΕΣΗ (COMPRESSION) :** Το αποσυνθεμένο σύνολο δεδομένων συμπιέζεται χρησιμοποιώντας την βιβλιοθήκη GZIP και συγκεκριμένα τον αλγόριθμο DEFLATE η οποία παρουσιάστηκε στο [1] για να προσφέρει την καλύτερη ισορροπία ανάμεσα στις ταχύτητες συμπίεσης και αποσυμπίεσης, τις αναλογίες συμπίεσης και την συμβατότητα με τις βιβλιοθήκες ρευμάτων εισόδου και εξόδου.
- **ΔΕΙΓΜΑΤΟΛΗΨΙΑ (SAMPLING**): Χρησιμοποιήθηκε μια μέθοδος δειγματοληψίας που επιλέγει ένα κάθε δύο στοιχεία, αποδίδοντας ένα μέγεθος δείγματος 50 %.
- **ΤΥΧΑΙΟ (RANDOM):** Επιλέγει ομοιόμορφα τυχαία μια εγγραφή από το αποσυνθεμένο σύνολο δεδομένων.

Σε αυτό το σημείο πρέπει να σημειώσουμε ότι οι προσεγγίσεις RAW και RANDOM είναι οι βασικές προσεγγίσεις που χρησιμοποιούνται για να αποδειχθεί για να αποδειχθεί η σχέση μεταξύ της χωρητικότητας αποθήκευσης και του NRMSE.

Σύνολο Δεδομένων

Χρησιμοποιούμε ένα ανώνυμο σύνολο δεδομένων από τηλεπικοινωνιακά δεδομένα το οποίο περιλαμβάνει περίπου 100M εγγραφές μετρήσεων δικτύου (NMS) και 3660 κελιών που προέρχονται από κεραίες 2G, 3G και LTE. Η κίνηση δεδομένων δημιουργείται από περίπου 300K αντικείμενα και έχει συνολικό μέγεθος 10GB. Κατασκευάσαμε 6 ρεαλιστικά σύνολα δεδομένων από πραγματικά μεγάλα τηλεπικοινωνιακά δεδομένα που αποκτήθηκαν μέσω του SPATE [1] βάσει των βασικών δεικτών απόδοσης (KPIs).

Τα 6 ρεαλιστικά σύνολα δεδομένων που κατασκευάσαμε έχουν ώς εξής:

- **Calls (CS):** Ο αριθμός των κλήσεων που τερματίστηκαν κανονικά κατά την διάρκεια ενός στιγμιότυπου**.**
- **Call Drops (CSD):** Ο αριθμός των κλήσεων ο οποίος διακόπηκε για κάποιο λόγο κατά την διάρκεια ενός στιγμιότυπου.
- **Handover Attempts (HA):** Το ποσό των μεταβιβάσεων σε ή από τα κελιά που επιχειρήθηκαν κατά την διάρκεια ενός στιγμιότυπου.
- **Handovers (HS**): Ο αριθμός των επιτυχημένων μεταβιβάσεων σε ή από τα κελιά κατά τη διάρκεια ενός στιγμιότυπου.
- **Call Setup Attempts (CSA):** Το ποσό των διεργασιών ρύθμισης κλήσεων που επιχειρήθηκαν κατά την διάρκεια του στιγμιότυπου.
- **Call Setups (CE):** Το ποσό των επιτυχημένων διεργασιών ρύθμισης κλήσεων κατά την διάρκεια του στιγμιότυπου.

Μετρικές

Για την αξιολόγηση της απόδοσης του τελεστή DP χρησιμοποιήθηκαν μετρικές σε όλα τα πειράματα οι οποίες ορίζονται πιο κάτω:

Storage Capacity (S): Μετρά το συνολικό χώρο που καταλαμβάνουν τα δεδομένα μαζί με το DP δέντρο, ως ποσοστό της αποθήκευσης που απαιτείται από την μέθοδο RAW (χωρίς φθορά και συμπίεση).

Normalized Root Mean Square Error (NRMSE): Μετρά το σφάλμα των ανακτηθέντων δεδομένων χρησιμοποιώντας τον τύπο του NRMSE που παρουσιάζεται πιο πάνω . Μικρότερη τιμή της μετρικής NRMSE σημαίνει υψηλότερη ακρίβεια στην ανάκτηση των δεδομένων.

Παραμέτροι

Σε όλα τα πειράματα οι παραμέτροι προσομοίωσης διαμορφώθηκαν ώς εξής:

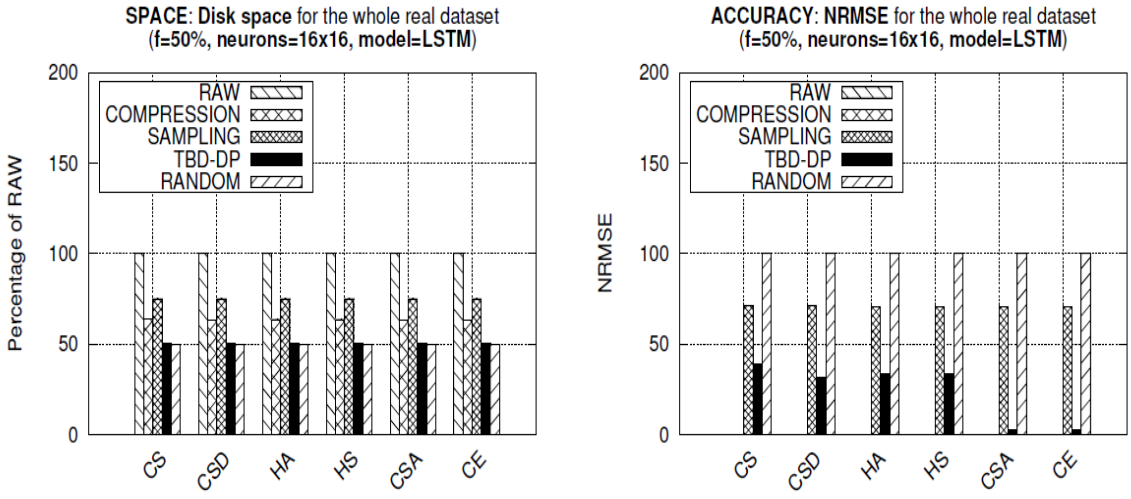
i. **Συντελεστής αποσύνθεσης** (decaying factor) f=50% . Η τιμή αυτή υποδεικνύει το ποσοστό επί του οποίου εκτελούμε το LSTM.
ii. Το πρότυπο ML είναι το **LSTM** και ο αριθμός των νευρώνων είναι 16 x 16.

Η επιρροή της κάθε μιας παραμέτρου στην προτεινόμενη προσέγγιση εξετάζεται ξεχωριστά στην δεύτερη πειραματική σειρά καθορίζοντας τις υπόλοιπες παραμέτρους αναλόγως.

6.3 Πειραματική Σειρά 1

Στην πρώτη σειρά πειραμάτων αξιολογείται η απόδοση του προτεινόμενου τελεστή DP ενάντια στους άλλους τέσσερις αλγόριθμους και σε όλα τα σύνολα δεδομένων που προαναφέραμε στην προηγούμενη υποενότητα, σε σχέση με την χωρητικότητα χώρου (ως ποσοστό δεδομένων RAW) και ακρίβεια (χρησιμοποιώντας τον όρο NRMSE στα αποσυνθεμένα σύνολα δεδομένων).

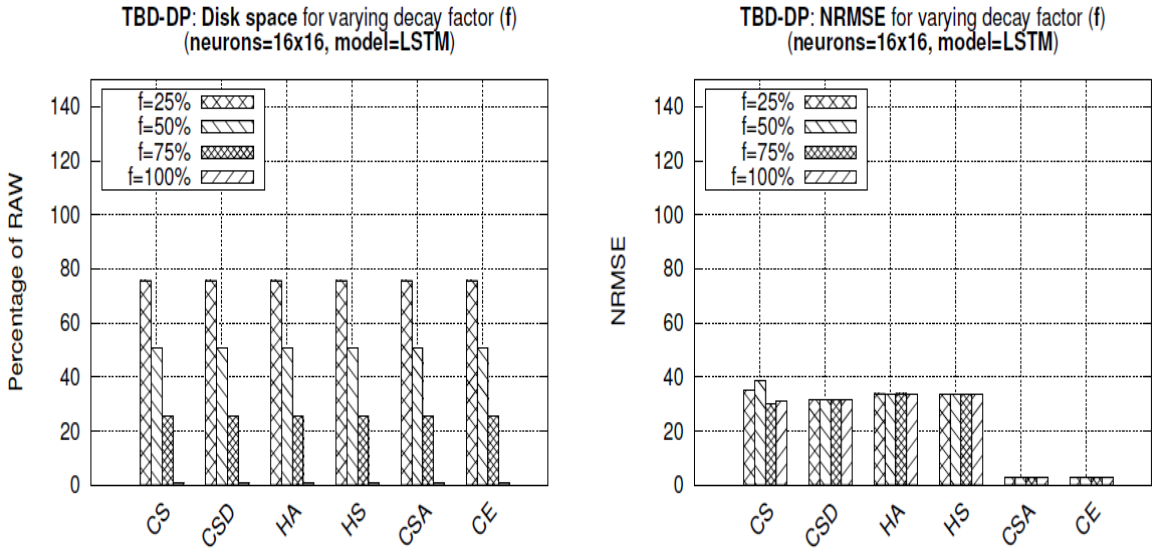
Στο σχήμα δείχνει την σχέση μεταξύ της χωρητικότητας S και των στόχων του NRMSE στα αποτελέσματα των προσεγγίσεων βασικής γραμμής, δεδομένου ότι η προσέγγιση RAW (χωρίς αποσύνθεση στο σύνολο δεδομένων) απέδωσε το χειρότερο δυνατό S που φτάνει το 100% του συνόλου δεδομένων και το χαμηλότερο NRMSE = 0, σε αντίθεση με την προσέγγιση RANDOM (σχεδόν όλα τα δεδομένα αποσυντέθηκαν) που απέδωσε S = 50% του συνόλου δεδομένων και το χειρότερο NRMSE = 100% για συντελεστή αποσύνθεσης f=50%. Τα αποτελέσματα των τριών άλλων προσεγγίσεων εμφανίζονται μεταξύ των αποτελεσμάτων των δύο βασικών προσεγγίσεων. Ωστόσο, ο προτεινόμενος τελεστής DP παρέχει περίπου 25% και 50% καλύτερη χωρητικότητα S σε σχέση με τον αλγόριθμο συμπίεσης (COMPRESSION) και δειγματοληψίας (SAMPLING) αντίστοιχα. Αυτό οφείλεται στο γεγονός ότι ο πρόσθετος χώρος που απαιτείται για το σύνολο των μοντέλων



Σχήμα 6.3.1 Αξιολόγηση Απόδοσης : Αξιολόγηση του DP όσον αφορά την χωρητικότητα αποθήκευσης S ώς ποσοστό δεδομένων RAW (αριστερά) και την ακρίβεια σε όρους NRMSE για το αποσυνθεμένο σύνολο δεδομένων. (Αφορά και τα 6 σύνολα δεδομένων)

LSTM είναι πολύ μικρότερος σε σχέση με το σύνολο δειγμάτων δειγματοληψίας και το συμπιεσμένο σύνολο δεδομένων συμπίεσης.

Όσον αφορά το NRMSE ο τελεστής DP υπερέχει της προσέγγισης SAMPLING κατά 50%, κατά μέσο όρο, σε όλα τα σύνολα δεδομένων. Η προσέγγιση COMPRESSION παρέχει ένα βέλτιστο NRMSE = 0, για τον λόγο ότι δεν εφαρμόζει καμία πρόβλεψη στα αποσυνθεμένα δεδομένα, αλλά τα ανακτά μέσω αποσυμπίεσης όταν και εφόσον ζητηθεί. Παρόλα αυτά η προσέγγιση COMPRESSION δεν μπορεί να προσαρμοστεί ώστε να επιτευχθεί ακόμα χαμηλότερη κατάληψη χωρητικότητας στο δίσκο. Από την άλλη ο τελεστής DP μπορεί να ρυθμιστεί κατάλληλα μέσω της παραμέτρου αποσύνθεσης f για να επιτύχει μια χωρητικότητα στο δίσκο η οποία θα ταιριάζει με τον προϋπολογισμό χώρου της εφαρμογής. Η συγκεκριμένη παράμετρος F θα μελετηθεί εκτενέστερα στην επόμενη πειραματική σειρά.

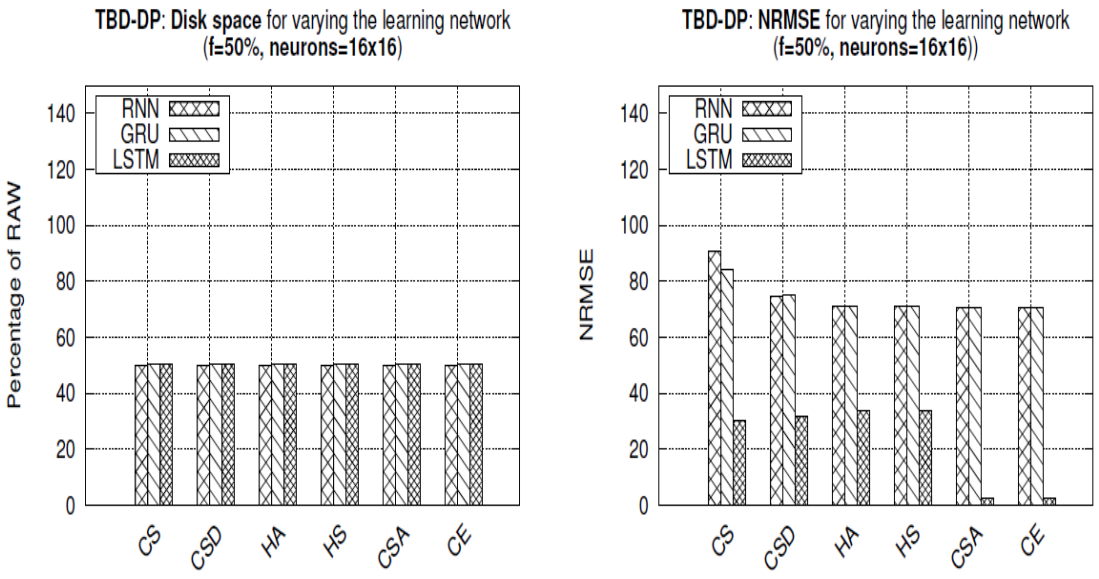


Σχήμα 6.4.1 Πείραμα ελέγχου-Συντελεστής αποσύνθεσης f: εξέταση της χωρητικότητας αποθήκευσης S και NRMSE της προτεινόμενης προσέγγισης DP ενώ το f μεταβάλλεται.

6.4 Πειραματική Σειρά 2

Στην δεύτερη πειραματική σειρά εξετάζουμε την επίδραση πολλών παραμέτρων ελέγχου στην απόδοση του προτεινόμενου DP σε όρους S και NRMSE. Συγκεκριμένα μεταβάλλουμε τον παράγοντα αποσύνθεσης f, τα μοντέλα ML και τον αριθμό των νευρώνων στο LSTM. Το σχήμα 6.4.1 δείχνει πώς ο συντελεστής αποσυνθέσεως f, και κατά συνέπεια όγκος των δεδομένων που θα αποσυντεθούν και αντιπροσωπεύονται στον χώρο αποθήκευσης από μοντέλα LSTM, θα επηρεάσουν το S και το NRMSE των προτεινόμενων. Τα αποτελέσματα δείχνουν ότι η χωρητικότητα αποθήκευσης που απαιτείται από τον τελεστή DP μειώνεται όσο αυξάνεται ο συντελεστής αποσυνθέσεως, πράγμα που είναι λογικό λόγω του γεγονότος ότι όσο υψηλότερο είναι το f τόσα περισσότερα δεδομένα πρέπει να υποστούν φθορά (να ξεχαστούν) και συνεπώς να απελευθερωθεί περισσότερος χώρος στο δίσκο. Ωστόσο, η ακρίβεια του προτεινόμενου τελεστή DP δεν επηρεάζεται δεδομένου ότι

το NRMSE παραμένει σχεδόν το ίδιο για όλους τους συντελεστές αποσύνθεσης στα περισσότερα σύνολα

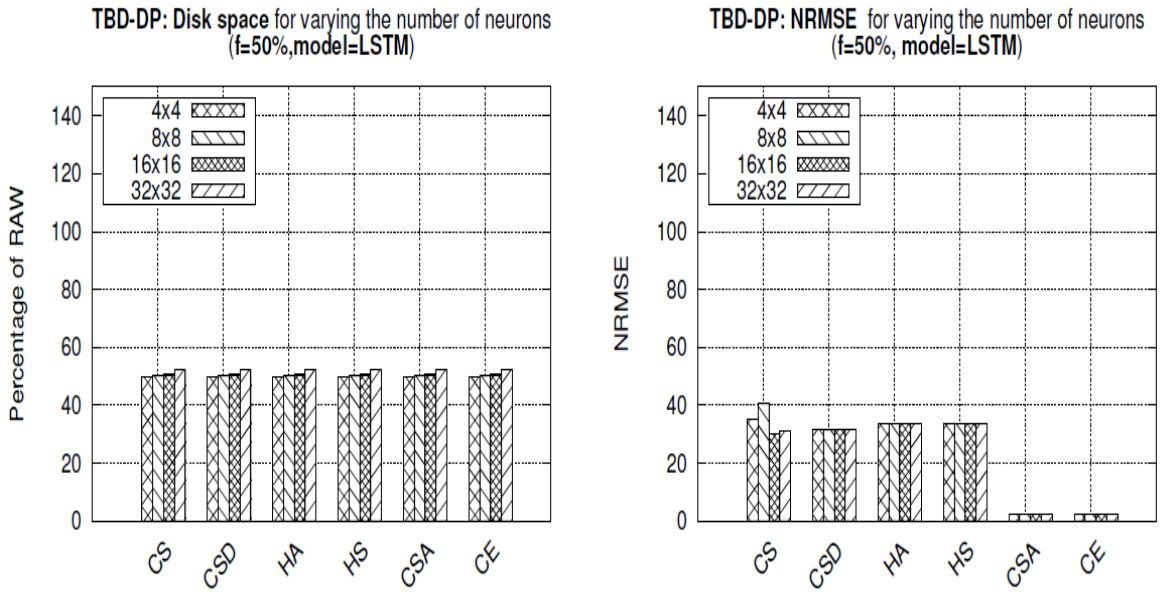


Σχήμα 6.4.2 Πείραμα ελέγχου - Μαθησιακά μοντέλα: Εξέταση της χωρητικότητας αποθήκευσης S και NRMSE της προτεινόμενης προσέγγισης DP ενώ συνδυάζεται με διάφορα μοντέλα τεχνικής μάθησης (ML)

δεδομένων. Αυτό δείχνει την επεκτασιμότητα (scalability) και τη γενικευσιμότητα (generalizability) της προτεινόμενης προσέγγισης, η οποία δεν επηρεάζεται από την αύξηση του μεγέθους του συνόλου δεδομένων. Είναι επίσης σημαντικό να σημειωθεί ότι οι μεταβολές της NRMSE που αποκτήθηκαν από το TBD-DP μεταξύ των συνόλων δεδομένων οφείλονται κυρίως στα διαφορετικά χαρακτηριστικά κάθε συνόλου δεδομένων.

Στο σχήμα 6.4.2 εξετάζεται η απόδοση του τελεστή DP σε όρους S και NRMSE καθώς συνδυάζεται με τρία διαφορετικά μοντέλα τεχνικής μάθησης, Αυτά τα μοντέλα είναι το παραδοσιακό Επαναλαμβανόμενο Νευρωνικό Δίκτυο (RNN), την Περιφραγμένη Επαναλαμβανόμενη Μονάδα (GRU) και μοντέλο LSTM το οποίο είναι αυτό που τελικά

υιοθετείται από την προτεινόμενη προσέγγιση. Τα αποτελέσματα δείχνουν ότι η προσέγγιση DP διατηρεί μια παρόμοια χωρητικότητα αποθήκευσης για διαφορετικά μοντέλα μάθησης,



Σχήμα 6.4.3 Πείραμα ελέγχου - Αριθμός νευρώνων στο LSTM: εξέταση της χωρητικότητας αποθήκευσης S και NRMSE της προτεινόμενης προσέγγισης DP, ενώ μεταβάλλουμε τον αριθμό των νευρώνων στο LSTM.

με ελαφρά αύξηση (περίπου 1%) όταν χρησιμοποιείται το μοντέλο LSTM. Από την άποψη του NRMSE, ωστόσο, ο συνδυασμός DP + LSTM ξεπερνά σαφώς τους άλλους δύο συνδυασμούς DP + RNN και DP + GRU, παρέχοντας κατά μέσο όρο περίπου 75% λιγότερα σφάλματα. Αυτός είναι και ένας από τους λόγους που επιλέχθηκε το μοντέλο LSTM για την αντικατάσταση των αποσυνθεμένων δεδομένων στον χώρο αποθήκευσης.

Τέλος το σχήμα 6.4.3 εξετάζει κατά πόσο ο αριθμός των νευρώνων του μοντέλου LSTM επηρεάζει την απόδοση του τελεστή DP. Τα αποτελέσματα υποστηρίζουν τις προηγούμενες παρατηρήσεις μας σχετικά με την επεκτασιμότητα και την γενικευσιμότητα της προτεινόμενης προσέγγισης DP. Η αύξηση στον αριθμό των νευρώνων επηρεάζει ελαφρώς τον τελεστή DP όσον αφορά την ικανότητα αποθήκευσης καθώς ο απαιτούμενος

χώρος αυξάνεται ελαφρά. Αυτό είναι λογικό καθώς η αύξηση στον αριθμό των νευρώνων οδηγεί σε “μεγαλύτερα” μοντέλα που απαιτούν περισσότερο χώρο στο δίσκο του συστήματος για την αποθήκευση τους. Ο πρόσθετος απαιτούμενος χώρος, ωστόσο, είναι σχεδόν αμελητέος σε σύγκριση με τον χώρο στο δίσκο που απαιτείται για την αποθήκευση των πραγματικών δεδομένων πριν από τη φθορά. Όσον αφορά το NRMSE, η αύξηση του αριθμού των νευρώνων δεν επηρεάζει την απόδοση του χειριστή του TBD-DP, καθώς το NRMSE παραμένει σχεδόν το ίδιο, ενώ μεταβάλλεται αυτή η παράμετρος ελέγχου σε όλα σχεδόν τα σύνολα δεδομένων.

Κεφάλαιο 7

Συμπεράσματα και Μελλοντική Δουλειά

7.1 Εισαγωγή

Σε αυτό το τελευταίο κεφάλαιο τις διπλωματικής εργασίας παρουσιάζουμε τα συμπεράσματα που εξάχθηκαν καθόλη την διάρκεια της εκπόνησης της εργασίας σύμφωνα με την πειραματική αποτίμηση που έγινε για τον προτεινόμενο αλγόριθμο DP. Στην συνέχεια αναφερόμαστε σε μελλοντική δουλειά που σχεδιάζεται να γίνει για επέκταση του DP, την βελτιστοποίηση των αλγορίθμων κατασκευής και επαναφοράς των μοντέλων, καθώς και την επέκταση της γραφικής διαπροσωπείας του πρωτοτύπου.

7.2 Συμπεράσματα

Σε αυτήν την εργασία παρουσιάζουμε ένα νέο τελεστή για μεγάλα τηλεπικοινωνιακά δεδομένα, τον DP, ο οποίος σχεδιάστηκε χρησιμοποιώντας Data Postdiction καθώς και την πειραματική αποτίμηση του αλγορίθμου αυτού με σκοπό την αξιολόγηση του. Ο τελεστής αυτός βασίζεται σε υπάρχοντες αλγόριθμους τεχνικής μάθησης (ML) για την μετατροπή των μεγάλων τηλεπικοινωνιακών δεδομένων σε συμπαγή μοντέλα τα οποία μπορούν να αποθηκευτούν στον αποθηκευτικό χώρο στην θέση των μεγάλων δεδομένων και να ερωτηθούν όταν είναι απαραίτητο από τον χρήστη ή αυτόματα από το σύστημα. Ο προτεινόμενος τελεστής DP έχει δύο εννοιολογικές φάσεις. Στην πρώτη φάση αναφερόμαστε

στην εκτός σύνδεσης φάση όπου ό τελεστής DP χρησιμοποιεί ένα ιεραρχικό αλγόριθμο τεχνικής μάθησης βασισμένο σε LSTM για να μάθει ένα δέντρο μοντέλων όσον αφορά χώρο και χρόνο. Στην δεύτερη φάση αναφερόμαστε σε μια απευθείας σύνδεση όπου ο τελεστής DP χρησιμοποιεί το δέντρο για την ανάκτηση των δεδομένων με μια ορισμένη ακρίβεια. Στην πειραματική μας εγκατάσταση μετράμε την αποτελεσματικότητα του προτεινόμενου τελεστή DP χρησιμοποιώντας ένα ανώνυμο πραγματικό τηλεπικοινωνιακό δίκτυο ίσο με 10GB και όσον αφορά την πειραματική αποτίμηση, τα αποτελέσματα στο Tensorflow πάνω από το HDFS είναι εξαιρετικά ενθαρρυντικά καθώς δείχνουν ότι ο προτεινόμενος αλγόριθμος εξοικονομεί χώρο αποθήκευσης διατηρώντας παράλληλα μια υψηλή ακρίβεια στα ανακτηθέντα δεδομένα.

7.2 Μελλοντική Δουλειά

Στο μέλλον στοχεύουμε στην γενίκευση των τελεστών που σταδιακά καταστρέφουν δεδομένα από τους αποθηκευτικούς χώρους πέρα από μεγάλα τηλεπικοινωνιακά δεδομένα και σε νέους τομείς (π.χ. σήματα από άλλο τύπο IoT). Αυτή η εργασία μπορεί να δώσει χώρο σε νέους αλγόριθμους τεχνικής μάθησης. Επιπρόσθετα ένας από τους κύριους στόχους μας είναι να αντλήσουμε θεωρητικά τα όρια της ακρίβεια και αποδοτικότητας του data postdiction πλαισίου μας. Τέλος σχεδιάζουμε να εκτελέσουμε μια πειραματική μελέτη και αποτίμηση που θα επικεντρωθεί αποκλειστικά στην αποσύνθεση μεγάλων δεδομένων.

Βιβλιογραφία

[1] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, and M. F. Mokbel, "Efficient exploration of telco big data with compression and decaying," in 2017 IEEE 33rd International Conference on Data Engineering (ICDE), April 2017, pp. 1332–1343.

[2] M. Yuan, K. Deng, J. Zeng, Y. Li, B. Ni, X. He, F. Wang, W. Dai, and Q. Yang, "Oceanst A distributed analytic system for large-scale spatiotemporal mobile broadband data," Proc. VLDB Endow., vol. 7, no. 13, pp. 1561–1564, Aug. 2014.

[3] C. LaChapelle, "The cost of data storage and management: where is the it headed in 2016?" 2016. [Online]. Available: http://www.datacenterjournal.com/cost-data-storage-managementheaded-2016/

[4] E. Kakoulli and H. Herodotou, "Octopusfs: A distributed file system with tiered storage management," in Proceedings of the 2017 ACM International Conference on Management of Data, ser. SIGMOD '17. New York, NY, USA: ACM, 2017, pp. 65–78. [Online]. Available: http://doi.acm.org/10.1145/3035918.3064023

[5] Y. Yan, L. J. Chen, and Z. Zhang, "Error-bounded sampling for analytics on big sparse data," Proc. VLDB Endow., vol. 7, no. 13, pp. 1508–1519, Aug. 2014. [Online]. Available: http://dx.doi.org/10.14778/2733004.2733022

[6] A. Klein, R. Gemulla, P. Rosch, and W. Lehner, "Derby/s: A dbms for sample-based query answering," in Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '06. New York, NY, USA: ACM, 2006, pp. 757–759. [Online]. Available: https://dl.acm.org/citation.cfm?id=1142579

[7] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica, "Blinkdb: Queries with bounded errors and bounded response times on very large data," in Proceedings of the 8th ACM European Conference on Computer Systems, ser. EuroSys '13. New York, NY, USA: ACM, 2013, pp. 29–42. [Online]. Available: http://doi.acm.org/10.1145/2465351.2465355

[8] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, "Data cube: A relational aggregation operator generalizing group-

by, cross-tab, and sub-totals," Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 29–53, Mar 1997. [Online]. Available: https://doi.org/10.1023/A:1009726021843

[9] K. Zeng, S. Agarwal, A. Dave, M. Armbrust, and I. Stoica, "G-ola: Generalized online aggregation for interactive analysis on big data," in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 913–918. [Online]. Available: http://doi.acm.org/10.1145/2723372.2735381

[10] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for massive data: Samples, histograms, wavelets, sketches," Found. Trends databases, vol. 4, no. 1–3, pp. 1–294, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1561/1900000004

[11] L. Sidirourgos, Martin, and P. Boncz, "Sciborq: Scientific data management with bounds on runtime and quality," in In Proc. of the Intl Conf. on Innovative Data Systems Research (CIDR, 2011, pp. 296–301.

[12] M. L. Kersten and L. Sidirourgos, "A database system with amnesia." in CIDR, 2017.

[13] M. L. Kersten, "Big data space fungus," in CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings, 2015.

[14] K. Krishna, D. Jain, S. V. Mehta, and S. Choudhary, "An lstm based system for prediction of human activities with durations," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, pp. 147:1–147:31, Jan. 2018. [Online]. Available: http://doi.acm.org/10.1145/3161201

[15] T. Bicer, J. Yin, D. Chiu, G. Agrawal, and K. Schuchardt, "Integrating online compression to accelerate large-scale data analytics applications," in Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on. IEEE, 2013, pp. 1205–1216.

[16] H. Yan, S. Ding, and T. Suel, "Inverted index compression and query processing with optimized document ordering," in Proceedings of the 18th international conference on World wide web. ACM, 2009, pp. 401–410. and Statistical Database Management, ser. SSDBM '14. New York, NY, USA: ACM, 2014, pp. 25:1–25:12. [Online]. Available http://doi.acm.org/10.1145/2618243.2618268

[17] M. Burtscher and P. Ratanaworabhan, “Fpc: A high-speed compressor for double-precision floating-point data,” IEEE Transactions on Computers, vol. 58, no. 1, pp. 18–31, 2009.

[18] F. Douglis and A. Iyengar, “Application-specific delta-encoding via resemblance detection.” in USENIX Annual Technical Conference, General Track, 2003, pp. 113126.

[19] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, “Extreme binning: Scalable, parallel deduplication for chunk-based file backup,” in 2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems. IEEE, 2009, pp. 1–9.

[20] L. L. You, K. T. Pollack, D. D. Long, and K. Gopinath, “Presidio: a framework for efficient archival data storage,” ACM Transactions on Storage (TOS), vol. 7, no. 2, p. 6, 2011. no. 12, pp. 1346–1357, 2015.

[21] S. Zhang, Y. Yang, W. Fan, L. Lan, and M. Yuan, “Oceanrt: Realtime analytics over large temporal data,” in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD ’14. New York, NY, USA: ACM, 2014, pp. 1099–1102.

[22] A. P. Iyer, L. E. Li, and I. Stoica, “Celliq: Real-time cellular network analytics at scale,” in Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation, ser. NSDI’15. Berkeley, CA, USA: USENIX Association, 2015, pp. 309–322.

[23] L. Braun, T. Etter, G. Gasparis, M. Kaufmann, D. Kossmann, D.Widmer, A. Avitzur, A. Iliopoulos, E. Levy, and N. Liang, “Analytics in motion: High performance event-processing and real-time analytics in the same database,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD ’15. New York, NY, USA: ACM, 2015, pp. 251–264.

[24] E. Bouillet, R. Kothari, V. Kumar, L. Mignet, S. Nathan, A. Ranganathan, D. S. Turaga, O. Udrea, and O. Verscheure, “Processing 6 billion cdrs/day: From research to production (experience report),” in Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems, ser. DEBS ’12. New York, NY, USA: ACM, 2012, pp. 264–267.

[25] M. A. Abbasoˇglu, B. Gedik, and H. Ferhatosmanoˇglu, “Aggregate profile clustering for telco analytics,” Proc. VLDB Endow., vol. 6, no. 12, pp. 1234–1237, Aug. 2013.

[26] S. Chaudhuri, G. Das, and V. Narasayya, “Optimized stratified sampling for approximate query processing,” ACM Trans. Database Syst., vol. 32, no. 2, Jun. 2007. [Online]. Available: http://doi.acm.org/10.1145/1242524.1242526

[27] Z. Wei, G. Luo, K. Yi, X. Du, and J.-R. Wen, “Persistent data sketching,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD ’15. New York, NY, USA: ACM, 2015, pp. 795–810. [Online]. Available: http://doi.acm.org/10.1145/2723372.274944

[28] C. Luo, J. Zeng, M. Yuan, W. Dai, and Q. Yang, “Telco user activity level prediction with massive mobile broadband data,” ACM Trans. Intell. Syst. Technol., vol. 7, no. 4, pp. 63:1–63:30, May 2016. [Online]. Available: http://doi.acm.org/10.1145/2856057

[29] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang, and J. Zeng, “Telco churn prediction with big data,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD. New York, NY, USA: ACM, 2015, pp. 60 618.

[30] Z. Kasiran, Z. Ibrahim, and M. S. M. Ribuan, “Mobile phone customers churn prediction using elman and jordan recurrent neural network,” in 2012 7th International Conference on Computing and Convergence Technology (ICCCT), Dec 2012, pp. 673–678.

[31] C. Luo, J. Zeng, M. Yuan, W. Dai, and Q. Yang, “Telco user activity level prediction with massive mobile broadband data,” ACM Trans. Intell. Syst. Technol., vol. 7, no. 4, pp. 63: 63:30, May 2016.

[32] "Decaying Telco Big Data with Data Postdiction", Constantinos Costa, Andrea Charalampous, Andreas Konstantinidis, Mohamed F. Mokbel, Proceedings of the 19th IEEE International Conference on Mobile Data Management (MDM '18), IEEE Computer Society, pp. 10 pages, June 25 - June 28, 2018, AAU, Aalborg, Denmark, **2018**

[33] Wikipedia.com, “Long Short-Term memory,2016. [Online]. Available: https://en.wikipedia.org/wiki/Long_short-term_memory

[34] Towardsdatascience.com, “LSTM by Example using Tensorflow”. [Online]. Available: https://towardsdatascience.com/lstm-by-example-using-tensorflow-feb0c1968537

[35] Tensorflow.com, “TensorFlow Architecture”. [Online]. Available: https://www.tensorflow.org/extend/architecture

[36] Wildml.com, “Recurrent Neural Networks Tutorial”. [Online]. Available: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[37] Mathworks.com, “Machine Learning” ”. [Online]. Available: https://www.mathworks.com/discovery/machine-learning.html

[38] Deeplearning4j.com, “Recurrent Networks and LSTMs”. [Online]. Available: https://deeplearning4j.org/lstm.html

Παράρτημα Α

Ο κώδικας βρίσκεται στο SVN αποθετήριο του Πανεπιστημίου Κύπρου στους εξής φακέλους.

Στο φάκελο **MDM18** περιλαμβάνεται ο κώδικας ο οποίος αφορά την υλοποίηση του τελεστή DP.

Στο φάκελο **spate-api** και **spate-ui** περιλαμβάνεται ο κώδικας που αφορά την υλοποίηση του πρωτοτύπου.