ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Σχολή Τεχνολογίας

Τμήμα Ψηφιακών Συστημάτων

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΠΕΞΕΡΓΑΣΙΑΣ ΠΛΗΡΟΦΟΡΙΩΝ ΓΙΑ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ (BIG DATA)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τουραλιά Θεοδώρα (ΑΜ: 3119035)

Επιβλέπων: ***Όμηρος Ιατρέλλης***

ΛΑΡΙΣΑ, Ιούνιος 2025

UNIVERSITY OF THESSALY

School of Technology

Digital Systems Department

OPTIMIZING INFORMATION PROCESSING ALGORITHMS FOR BIG DATA

BACHELOR THESIS

Touralia Theodora (RN: 3119035)

Supervisor: ***Omiros Iatrellis***

LARISSA, June 2025

Υπεύθυνη Δήλωση περί Ακαδημαϊκής Δεοντολογίας και Πνευματικών Δικαιωμάτων

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα πτυχιακή εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον, και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Η Δηλούσα

(Υπογραφή)

Θεοδώρα Τουραλιά

23/6/2025

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων	**Όμηρος Ιατρέλλης** Επίκουρος Καθηγητής Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Απόστολος Ξενάκης** Επίκουρος Καθηγητής Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Κωνσταντίνος Χαϊκάλης** Αναπληρωτής Καθηγητής Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 23/6/2025

Περίληψη

Η εκρηκτική αύξηση των μεγάλων δεδομένων (Big Data) έχει μεταβάλει ριζικά το πεδίο της ανάλυσης πληροφοριών, αναδεικνύοντας την ανάγκη για αποτελεσματικούς και βελτιστοποιημένους αλγόριθμους επεξεργασίας. Η παρούσα εργασία εστιάζει στη μελέτη τεχνικών και μεθοδολογιών που ενισχύουν την αποδοτικότητα αλγορίθμων όταν εφαρμόζονται σε περιβάλλοντα μεγάλων δεδομένων. Στόχος είναι η κατανόηση των βασικών αρχών παραλληλισμού, κατανεμημένης επεξεργασίας και στρατηγικών μείωσης της υπολογιστικής πολυπλοκότητας, καθώς και η αξιολόγηση της αποτελεσματικότητάς τους βάσει των χαρακτηριστικών των δεδομένων. Μέσα από την ανασκόπηση της σχετικής επιστημονικής βιβλιογραφίας, καταγράφονται οι πλέον σύγχρονες ακαδημαϊκές και τεχνολογικές προσεγγίσεις. Εξετάζονται αλγόριθμοι μηχανικής μάθησης, βαθιάς μάθησης, εξόρυξης δεδομένων και επεξεργασίας φυσικής γλώσσας, σε συνδυασμό με τεχνικές βελτιστοποίησης. Ιδιαίτερη έμφαση δίνεται σε υβριδικές λύσεις που ενσωματώνουν πολλαπλές τεχνικές για μεγαλύτερη προσαρμοστικότητα και ακρίβεια. Τα ευρήματα υπογραμμίζουν τη σημασία των τεχνικών βελτιστοποίησης για την αποδοτική και κλιμακούμενη επεξεργασία μεγάλων δεδομένων με μειωμένο κόστος και πόρους. Επιπλέον, αναδεικνύεται η συμβολή τους στη βελτίωση της αποδοτικότητας των υπολογιστικών συστημάτων, ενώ προτείνονται κατευθύνσεις για μελλοντική έρευνα.

Λέξεις-κλειδιά: *αλγόριθμοι, μεγάλα δεδομένα (big data), τεχνικές βελτιστοποίησης*

Abstract

The explosive growth of Big Data has radically changed the field of information analysis, highlighting the need for efficient and optimized processing algorithms. This paper focuses on the study of techniques and methodologies that enhance the efficiency of algorithms when applied in big data environments. The goal is to understand the basic principles of parallelism, distributed processing and computational complexity reduction strategies, as well as to evaluate their effectiveness based on the characteristics of the data. Through a review of the relevant scientific literature, the most modern academic and technological approaches are recorded. Machine learning, deep learning, data mining and natural language processing algorithms are examined, in combination with optimization techniques. Particular emphasis is given to hybrid solutions that integrate multiple techniques for greater adaptability and accuracy. The findings highlight the importance of optimization techniques for efficient and scalable processing of big data with reduced costs and resources. Furthermore, their contribution to improving the efficiency of computing systems is highlighted, while directions for future research are suggested.

Keywords*: algorithms, big data, optimization techniques*

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες προς τον επιβλέποντα καθηγητή μου, κ. Όμηρο Ιατρέλλη, για την αφοσίωση, την καθοδήγηση και τη στήριξη που μου πρόσφερε καθ' όλη τη διάρκεια της εκπόνησης αυτής της πτυχιακής εργασίας. Η επιστημονική του προσέγγιση και η πολύτιμη εμπειρία του ήταν καθοριστικές για την ολοκλήρωση του έργου με επιτυχία..

Επιπλέον, ευχαριστώ την οικογένειά μου και τους φίλους μου για τη συνεχή στήριξη και ενθάρρυνσή τους, καθώς και για την κατανόηση και την αγάπη τους που μου έδωσαν τη δύναμη να ολοκληρώσω αυτή την εργασία.

Τέλος, εκφράζω την ευγνωμοσύνη μου στο Πανεπιστήμιο Θεσσαλίας και το Τμήμα Ψηφιακών Συστημάτων για τις ευκαιρίες μάθησης και προσωπικής εξέλιξης που μου προσέφεραν όλα αυτά τα χρόνια.

Θεοδώρα Τουραλιά

23/6/2025

Περιεχόμενα

1 Εισαγωγή

Η ραγδαία αύξηση της παραγωγής και συλλογής δεδομένων στις σύγχρονες κοινωνίες, επιχειρήσεις και επιστήμες έχει οδηγήσει στην εμφάνιση των μεγάλων δεδομένων (Big Data), τα οποία χαρακτηρίζονται από τεράστιο όγκο, ποικιλία και ταχύτητα. Τα μεγάλα δεδομένα έχουν αναδειχθεί σε μία από τις πιο σημαντικές τεχνολογικές εξελίξεις της σύγχρονης εποχής, καθώς η ικανότητα συλλογής, αποθήκευσης και ανάλυσης τεράστιων ποσοτήτων δεδομένων έχει αλλάξει τον τρόπο λειτουργίας πολλών τομέων. Παρέχουν πρωτοφανείς ευκαιρίες για την ανάλυση και εξαγωγή γνώσεων που μπορούν να οδηγήσουν σε σημαντικές βελτιώσεις στη λήψη αποφάσεων. Προσφέρουν επίσης τη δυνατότητα κατανόησης πολύπλοκων προβλημάτων και επιτρέπουν στους οργανισμούς να προσαρμοστούν σε έναν συνεχώς μεταβαλλόμενο κόσμο, προσφέροντας καινοτόμες λύσεις και βελτιώνοντας την απόδοση και την αποτελεσματικότητα σε διάφορους τομείς [1]. Ωστόσο, η επεξεργασία και η ανάλυση αυτών των δεδομένων αποτελούν σημαντική πρόκληση για τα παραδοσιακά υπολογιστικά συστήματα και τους αλγορίθμους, καθώς απαιτούν σημαντικούς υπολογιστικούς πόρους και χρόνο [2].

Η βελτιστοποίηση αλγορίθμων για την επεξεργασία μεγάλων δεδομένων είναι ζωτικής σημασίας, καθώς επιτρέπει την αποτελεσματική διαχείριση των τεράστιων όγκων δεδομένων που παράγονται συνεχώς από διάφορες πηγές. Χωρίς κατάλληλα βελτιστοποιημένους αλγόριθμους, η επεξεργασία των δεδομένων μπορεί να γίνει αργή και αναποτελεσματική, αυξάνοντας τον χρόνο επεξεργασίας και την κατανάλωση πόρων, όπως η μνήμη και η υπολογιστική ισχύς [3]. Η βελτιστοποίηση βελτιώνει την ταχύτητα, την ακρίβεια και την απόδοση των συστημάτων, εξασφαλίζοντας ότι οι αλγόριθμοι μπορούν να χειριστούν μεγάλες κλίμακες δεδομένων με αποδοτικότητα, ακόμη και σε πραγματικό χρόνο. Επιπλέον, η βελτιστοποίηση αλγορίθμων επιτρέπει τη μείωση του κόστους των υποδομών, ενώ ενισχύει την κλιμακωσιμότητα των συστημάτων, καθιστώντας τα πιο προσαρμόσιμα σε αυξανόμενες απαιτήσεις[4].

Η παρούσα πτυχιακή εργασία στοχεύει στην ενδελεχή μελέτη και θεωρητική τεκμηρίωση των βασικών τεχνικών που σχετίζονται με τη βελτιστοποίηση αλγορίθμων για την επεξεργασία μεγάλων δεδομένων, ένα πεδίο το οποίο έχει αναδειχθεί σε κρίσιμη περιοχή έρευνας και εφαρμογής στο πλαίσιο της σύγχρονης υπολογιστικής επιστήμης. Η μελέτη

επιδιώκει να αναδείξει τη σημασία αυτών των τεχνικών όχι μόνο σε θεωρητικό επίπεδο, αλλά και ως πρακτικά εργαλεία που συμβάλλουν στην αποτελεσματική και κλιμακώσιμη διαχείριση δεδομένων υψηλού όγκου και πολυπλοκότητας, όπως αυτά προκύπτουν από ένα ευρύ φάσμα σύγχρονων πηγών πληροφορίας.

Η εργασία υιοθετεί ως κύριο μεθοδολογικό άξονα τη βιβλιογραφική ανασκόπηση, η οποία επιτρέπει την ενσωμάτωση και κριτική αποτίμηση υφιστάμενης επιστημονικής γνώσης και ερευνητικών αποτελεσμάτων αναφορικά με τις τεχνικές βελτιστοποίησης. Κεντρικό ζητούμενο της προσέγγισης αυτής αποτελεί η κατανόηση του τρόπου με τον οποίο η χρήση συγκεκριμένων τεχνικών, όπως η κατανομή του υπολογιστικού φόρτου, η υλοποίηση παραλληλισμού, η μείωση της πολυπλοκότητας των αλγοριθμικών διαδικασιών και η αξιοποίηση κατανεμημένων υπολογιστικών αρχιτεκτονικών, συντελούν στη βελτίωση της απόδοσης των αλγορίθμων επεξεργασίας πληροφοριών σε συνθήκες μεγάλης κλίμακας.

Η ερευνητική συνεισφορά της εργασίας εντοπίζεται στην προσπάθεια συγκρότησης ενός ενιαίου θεωρητικού πλαισίου που να καλύπτει τις βασικές αρχές, τα τεχνικά χαρακτηριστικά και τις εφαρμογές των αλγοριθμικών τεχνικών βελτιστοποίησης σε περιβάλλοντα Big Data. Η εργασία επιχειρεί, επίσης, να αναδείξει το πώς η χρήση των εν λόγω τεχνικών μπορεί να οδηγήσει σε συστήματα υψηλότερης αποτελεσματικότητας, μειώνοντας το απαιτούμενο υπολογιστικό κόστος και επιτρέποντας την καλύτερη αξιοποίηση των διαθέσιμων πόρων σε επίπεδο υποδομής. Με αυτόν τον τρόπο, συμβάλλει στην πληρέστερη κατανόηση των σύγχρονων προκλήσεων που αντιμετωπίζει η επιστήμη των δεδομένων και προτείνει, με βάση τεκμηριωμένα ευρήματα, εφικτούς και αποδοτικούς τρόπους αξιοποίησης βελτιστοποιημένων αλγορίθμων σε ένα ευρύ φάσμα εφαρμογών.

Η δομή της εργασίας οργανώνεται σε πέντε διακριτά αλλά αλληλένδετα κεφάλαια, τα οποία ακολουθούν μία λογική προοδευτικής εμβάθυνσης στο αντικείμενο μελέτης. Στο πρώτο κεφάλαιο διατυπώνεται η εισαγωγή, η οποία οριοθετεί το αντικείμενο της έρευνας, παρουσιάζει τη σημασία της θεματικής και θέτει το θεωρητικό υπόβαθρο στο οποίο βασίζεται η ανάλυση. Στο δεύτερο κεφάλαιο παρατίθεται η επισκόπηση της έννοιας των μεγάλων δεδομένων, με έμφαση στα χαρακτηριστικά, τις τεχνολογικές προκλήσεις και τις απαιτήσεις που δημιουργούν στον σχεδιασμό και την υλοποίηση αποδοτικών υπολογιστικών συστημάτων. Στο τρίτο κεφάλαιο, η ανάλυση επικεντρώνεται στους κύριους τύπους αλγορίθμων που χρησιμοποιούνται στην επεξεργασία πληροφοριών μεγάλου

όγκου, με ειδική αναφορά στους αλγόριθμους μηχανικής μάθησης, εξόρυξης δεδομένων, ανάλυσης κειμένου και αναλυτικής δικτύων.

Ακολουθεί το τέταρτο κεφάλαιο, το οποίο συνιστά τον πυρήνα της εργασίας, καθώς αναλύει τις τεχνικές βελτιστοποίησης που εφαρμόζονται σε αλγόριθμους για την αποτελεσματική διαχείριση μεγάλων δεδομένων. Τέλος, το πέμπτο κεφάλαιο συνοψίζει τα βασικά συμπεράσματα της μελέτης, αξιολογεί τα ερευνητικά ευρήματα σε σχέση με άλλες μελέτες στο ίδιο πεδίο και προτείνει κατευθύνσεις για μελλοντική έρευνα, υπογραμμίζοντας τις δυνατότητες περαιτέρω αξιοποίησης των τεχνικών αυτών σε πραγματικά περιβάλλοντα εφαρμογής.

Η εργασία φιλοδοξεί να αποτελέσει μία επιστημονικά τεκμηριωμένη συμβολή στην καλύτερη κατανόηση και αποτελεσματικότερη εφαρμογή των τεχνικών βελτιστοποίησης στον τομέα των Big Data, προσφέροντας ένα εργαλείο αναφοράς τόσο για μελλοντικούς ερευνητές όσο και για επαγγελματίες του κλάδου.

2 Θεωρητικό Υπόβαθρο

2.1 Έννοια και χαρακτηριστικά των μεγάλων δεδομένων

2.1.1 Έννοια και σημασία των μεγάλων δεδομένων

Τα μεγάλα δεδομένα (Big Data) αναφέρονται σε εξαιρετικά μεγάλους όγκους δεδομένων, οι οποίοι υπερβαίνουν τη δυνατότητα των παραδοσιακών εργαλείων επεξεργασίας να τα αποθηκεύσουν και να τα επεξεργαστούν αποτελεσματικά. Τα δεδομένα αυτά παράγονται καθημερινά από διάφορες πηγές, όπως κοινωνικά δίκτυα, συσκευές IoT, εφαρμογές επιχειρήσεων, συναλλαγές κλπ., και περιλαμβάνουν δομημένα, ημιδομημένα και αδόμητα δεδομένα [5].

Η διαχείριση αυτών των δεδομένων έχει γίνει ένα σημαντικό ζήτημα για πολλές βιομηχανίες και οργανισμούς, καθώς η εξαγωγή χρήσιμων πληροφοριών από αυτά μπορεί να προσφέρει σημαντικό ανταγωνιστικό πλεονέκτημα. Τα μεγάλα δεδομένα αναλύονται για την υποστήριξη της λήψης αποφάσεων, της δημιουργίας προγνωστικών μοντέλων και της κατανόησης τάσεων και μοτίβων, αλλάζοντας τον τρόπο λειτουργίας πολλών τομέων[6].

Στον επιχειρηματικό τομέα, τα big data χρησιμοποιούνται για τη βελτιστοποίηση των λειτουργιών και την αύξηση της κερδοφορίας. Επιχειρήσεις που συγκεντρώνουν και αναλύουν δεδομένα καταναλωτικής συμπεριφοράς μπορούν να κατανοήσουν καλύτερα τις ανάγκες των πελατών τους, να προσαρμόσουν τις στρατηγικές μάρκετινγκ και να προβλέψουν τις τάσεις της αγοράς. Μέσω της ανάλυσης αυτών των δεδομένων, οι εταιρείες μπορούν να προβλέπουν τη ζήτηση για προϊόντα και υπηρεσίες, να βελτιώνουν την εφοδιαστική αλυσίδα τους και να ενισχύουν τη λήψη στρατηγικών αποφάσεων [7].

Ο τομέας της υγείας είναι ένας άλλος κλάδος όπου οι τεράστιοι όγκοι πληροφορίας παίζουν καθοριστικό ρόλο. Η συλλογή και ανάλυση ιατρικών δεδομένων, όπως δεδομένα από ιατρικές συσκευές, αρχεία ασθενών και γενετικές πληροφορίες, βοηθά στη βελτίωση της διάγνωσης και της θεραπείας των ασθενών. Με τη χρήση μεγάλων δεδομένων, οι επαγγελματίες υγείας μπορούν να προβλέψουν επιδημίες, να παρακολουθούν την εξάπλωση ασθενειών και να εξατομικεύσουν θεραπείες για κάθε ασθενή [8].

Στον τομέα των επιστημών, τα μεγάλα δεδομένα έχουν μετασχηματίσει τον τρόπο με τον οποίο διεξάγεται η έρευνα. Οι ερευνητές χρησιμοποιούν μεγάλες συλλογές δεδομένων για την εξαγωγή νέων επιστημονικών γνώσεων και την επιβεβαίωση ή απόρριψη θεωριών. Για παράδειγμα, στον τομέα της αστρονομίας, οι παρατηρήσεις από διαστημικά τηλεσκόπια παράγουν δεδομένα τα οποία χρησιμοποιούνται για τη μελέτη των γαλαξιών και των αστέρων σε κλίμακα που δεν ήταν δυνατή πριν από μερικές δεκαετίες [9].

Οι κυβερνήσεις επίσης επωφελούνται από τα big data, αξιοποιώντας τα για την παρακολούθηση της κοινωνικής δραστηριότητας και τη βελτίωση των δημόσιων πολιτικών. Τα δεδομένα από κοινωνικές υπηρεσίες, μεταφορές και την υγεία μπορούν να χρησιμοποιηθούν για την ανάλυση και την πρόβλεψη τάσεων, όπως η κίνηση στους δρόμους ή οι ανάγκες για πόρους σε περιόδους κρίσης. Αυτό βοηθά στη βελτιστοποίηση των υπηρεσιών και την ορθότερη κατανομή των πόρων, προσφέροντας στους πολίτες πιο αποτελεσματικές και προσαρμοσμένες υπηρεσίες[10].

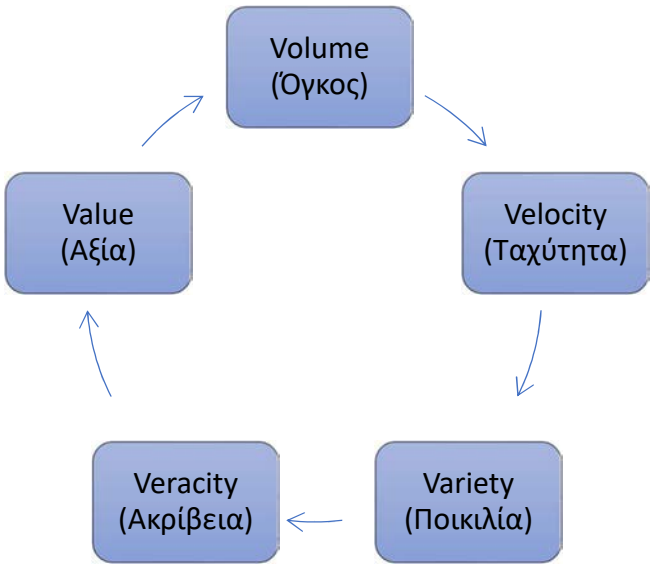
Γενικά, η σημασία των μεγάλων δεδομένων έγκειται στο γεγονός ότι μπορούν να οδηγήσουν σε καλύτερη λήψη αποφάσεων, καινοτομία και ενίσχυση της ανταγωνιστικότητας. Όταν οι οργανισμοί και οι κυβερνήσεις έχουν τη δυνατότητα να αναλύουν δεδομένα σε μεγάλη κλίμακα, μπορούν να αναγνωρίζουν ευκαιρίες που ήταν αόρατες ή αδύνατο να εκτιμηθούν με τις παραδοσιακές μεθόδους ανάλυσης.

2.1.2 Χαρακτηριστικά των μεγάλων δεδομένων

Τα big data διακρίνονται από τα ακόλουθα πέντε βασικά χαρακτηριστικά, που συνήθως περιγράφονται ως τα 5Vs (Εικόνα 1) [11]:

1. Volume (Όγκος): Ο όγκος είναι το πιο εμφανές χαρακτηριστικό των μεγάλων δεδομένων. Η προέλευσή τους είναι εξαιρετικά ετερογενής, περιλαμβάνοντας τόσο φυσικές όσο και ψηφιακές πηγές, όπως αισθητήρες, έξυπνες συσκευές, κοινωνικά μέσα και εφαρμογές. Η πρόκληση του όγκου σχετίζεται με την αποθήκευση και τη διαχείριση αυτών των τεράστιων ποσοτήτων δεδομένων [12].
2. Velocity (Ταχύτητα): Η ταχύτητα αναφέρεται στον γρήγορο ρυθμό με τον οποίο τα δεδομένα δημιουργούνται, μεταφέρονται και απαιτούν ανάλυση για άμεση αξιοποίηση. Για πολλές εφαρμογές, όπως η ανάλυση χρηματοοικονομικών δεδομένων ή η παρακολούθηση κοινωνικών μέσων, τα δεδομένα πρέπει να αναλυθούν σε πραγματικό χρόνο ή σε σχεδόν σε πραγματικό χρόνο. Αυτό δημιουργεί την ανάγκη για γρήγορες υποδομές επεξεργασίας και ανάλυσης.

3. Variety (Ποικιλία): Η ποικιλία αναφέρεται στη διαφορετικότητα των τύπων δεδομένων. Τα δεδομένα μπορεί να είναι δομημένα (π.χ. βάσεις δεδομένων), ημιδομημένα (π.χ. XML αρχεία) ή αδόμητα (π.χ. βίντεο, εικόνες, κείμενα από κοινωνικά δίκτυα). Η μεγάλη ποικιλία των δεδομένων απαιτεί την ανάπτυξη νέων τεχνικών για την ανάλυσή τους και την εξαγωγή χρήσιμων πληροφοριών.
4. Veracity (Αλήθεια/Ακρίβεια): Η αλήθεια αφορά την ποιότητα και την ακρίβεια των δεδομένων. Σε πολλούς τομείς, τα δεδομένα μπορεί να είναι ατελή, ανακριβή ή προκατειλημμένα, γεγονός που επηρεάζει την ακρίβεια των αποτελεσμάτων. Η πρόκληση είναι να εξασφαλιστεί η αξιοπιστία των δεδομένων, διαχωρίζοντας τα χρήσιμα δεδομένα από τον "θόρυβο" [12].
5. Value (Αξία): Η αξία είναι το τελευταίο και πιο σημαντικό χαρακτηριστικό. Τα μεγάλα δεδομένα από μόνα τους δεν έχουν αξία, εκτός και αν μπορούν να χρησιμοποιηθούν για να παράγουν χρήσιμες πληροφορίες. Η αξία έγκειται στην ικανότητα ανάλυσης και εξαγωγής πληροφοριών που οδηγούν σε καλύτερη λήψη αποφάσεων, καινοτομία και βελτίωση των υπηρεσιών [12].



Εικόνα 1. Τα 5V των μεγάλων δεδομένων

2.2 Προκλήσεις στην επεξεργασία μεγάλων δεδομένων

Η επεξεργασία μεγάλων δεδομένων παρουσιάζει σημαντικές προκλήσεις λόγω των παραπάνω χαρακτηριστικών τους. Οι προκλήσεις αυτές δεν σχετίζονται μόνο με την τεχνολογία, αλλά και με τη διαχείριση των πόρων, την ασφάλεια, την ιδιωτικότητα και την ικανότητα εξαγωγής χρήσιμων πληροφοριών [13]. Στη συνέχεια, αναλύονται αυτές οι βασικές προκλήσεις.

2.2.1 Όγκος δεδομένων

Το μέγεθος των παραγόμενων πληροφοριών αποτελεί μία από τις σημαντικότερες προκλήσεις στην επεξεργασία μεγάλων δεδομένων. Με την εκθετική αύξηση της παραγωγής δεδομένων, οργανισμοί και επιχειρήσεις καλούνται να διαχειριστούν και να επεξεργαστούν τεράστιες ποσότητες πληροφοριών που προέρχονται από διάφορες πηγές. Οι προκλήσεις που σχετίζονται με αυτόν τον τεράστιο όγκο είναι πολυεπίπεδες και επηρεάζουν κάθε στάδιο της διαδικασίας επεξεργασίας δεδομένων, από την αποθήκευση και την ανάλυση έως την εξαγωγή χρήσιμων πληροφοριών [14].

- Προκλήσεις αποθήκευσης

Η πρώτη και ίσως πιο άμεση πρόκληση που συνδέεται με τον όγκο δεδομένων είναι το ζήτημα της αποθήκευσής τους. Οι παραδοσιακές τεχνολογίες αποθήκευσης δεν μπορούν να ανταπεξέλθουν σε τόσο μεγάλα μεγέθη δεδομένων χωρίς υψηλό κόστος ή χωρίς να επηρεαστεί η απόδοση. Απαιτούνται νέες τεχνολογίες αποθήκευσης, όπως κατανεμημένα συστήματα αποθήκευσης (π.χ. Hadoop Distributed File System - HDFS), τα οποία μπορούν να αποθηκεύσουν και να διαχειριστούν δεδομένα κατανεμημένα σε πολλούς διακομιστές ή κόμβους [14]. Επιπλέον, η διαχείριση αυτών των δεδομένων σε κατανεμημένες υποδομές απαιτεί ειδικές γνώσεις και τεχνογνωσία, ενώ το κόστος αποθήκευσης δεδομένων σε τόσο μεγάλη κλίμακα αυξάνεται συνεχώς [20]. Οι επιχειρήσεις καλούνται να βρουν οικονομικά αποδοτικές λύσεις που να μπορούν να κλιμακωθούν με την αύξηση των δεδομένων τους.

- Προκλήσεις επεξεργασίας

Η τεράστια κλίμακα πληροφοριών δεν αφορά μόνο την αποθήκευσή τους αλλά και την επεξεργασία τους. Η ανάλυση τεράστιων ποσοτήτων δεδομένων σε μια εύλογη χρονική περίοδο είναι ιδιαίτερα απαιτητική, ειδικά όταν τα δεδομένα πρέπει να αναλυθούν σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο. Η επεξεργασία σε τέτοια κλίμακα

απαιτεί κατανεμημένα συστήματα υπολογισμού που μπορούν να χειριστούν μεγάλες ποσότητες δεδομένων ταυτόχρονα και να κατανεμηθούν σε πολλά συστήματα. Η παραδοσιακή επεξεργασία δεδομένων, όπως η εκτέλεση ερωτημάτων σε βάσεις δεδομένων ή η ανάλυση δεδομένων σε μηχανές επεξεργασίας, δεν επαρκεί πλέον για τον όγκο των δεδομένων που συγκεντρώνονται [13]. Ειδικά εργαλεία και πλατφόρμες, όπως το Apache Spark, το Apache Flink, και το Google BigQuery, έχουν αναπτυχθεί για να προσφέρουν δυνατότητες επεξεργασίας δεδομένων σε μεγάλη κλίμακα, επιτρέποντας την ανάλυση δεδομένων σε παράλληλους κόμβους, ενισχύοντας έτσι την ταχύτητα και την αποδοτικότητα της επεξεργασίας [14].

- Προκλήσεις στην εξαγωγή χρήσιμων πληροφοριών

Ακόμα μια πρόκληση που συνδέεται με τον όγκο των δεδομένων είναι η εξαγωγή χρήσιμων πληροφοριών. Όσο μεγαλύτερος είναι ο όγκος των δεδομένων, τόσο πιο δύσκολη είναι η ανάλυσή τους για την εξαγωγή αξιόπιστων και σημαντικών πληροφοριών. Τα δεδομένα που αποθηκεύονται περιλαμβάνουν συχνά μεγάλο "θόρυβο" και περιττές πληροφορίες, γεγονός που δυσκολεύει τον εντοπισμό των σημαντικών στοιχείων. Η χρήση τεχνικών μηχανικής μάθησης και τεχνητής νοημοσύνης μπορεί να βοηθήσει στην επεξεργασία και ανάλυση τεράστιων ποσοτήτων δεδομένων, ωστόσο αυτές οι τεχνικές απαιτούν υπολογιστικούς πόρους υψηλής απόδοσης. Η πρόκληση έγκειται στη διασφάλιση ότι η επεξεργασία μεγάλων δεδομένων μπορεί να οδηγήσει σε ακριβείς και χρήσιμες πληροφορίες που μπορούν να αξιοποιηθούν από τους οργανισμούς για τη λήψη αποφάσεων [18].

- Κόστος και πόροι

Ο όγκος δεδομένων συνδέεται άμεσα και με το κόστος διαχείρισης πόρων. Η ανάγκη για αυξημένη υπολογιστική ισχύ και αποθηκευτική ικανότητα σημαίνει υψηλότερα λειτουργικά έξοδα, τόσο σε υλικό όσο και σε λογισμικό. Ο παραδοσιακός εξοπλισμός δεν επαρκεί πλέον για τη διαχείριση τέτοιων ποσοτήτων δεδομένων, και αυτό μπορεί να οδηγήσει σε αυξημένο κόστος συντήρησης και αναβάθμισης υποδομών [16].

2.2.2 Ταχύτητα επεξεργασίας

Η ταχύτητα επεξεργασίας αποτελεί μία από τις πιο κρίσιμες προκλήσεις στην επεξεργασία μεγάλων δεδομένων. Καθώς οι ποσότητες των δεδομένων που παράγονται αυξάνονται εκθετικά, οι οργανισμοί καλούνται να επεξεργάζονται αυτά τα δεδομένα ταχύτερα από ποτέ, συχνά σε πραγματικό χρόνο, για να ανταποκριθούν στις σύγχρονες απαιτήσεις

των επιχειρήσεων και της τεχνολογίας [13]. Η ταχύτητα αφορά τόσο την ταχύτητα με την οποία τα δεδομένα δημιουργούνται και συλλέγονται όσο και την ταχύτητα με την οποία πρέπει να υποβληθούν σε ανάλυση και να παράγουν αποτελέσματα.

Η ταχύτητα επεξεργασίας των δεδομένων είναι κρίσιμη για εφαρμογές όπου η ανάλυση σε πραγματικό χρόνο είναι απαραίτητη. Για παράδειγμα, στην χρηματοπιστωτική αγορά οι συναλλαγές και οι αλγοριθμικές στρατηγικές βασίζονται σε δεδομένα που πρέπει να επεξεργαστούν και να αναλυθούν μέσα σε χιλιοστά του δευτερολέπτου. Ακόμα και μια μικρή καθυστέρηση μπορεί να οδηγήσει σε απώλειες εκατομμυρίων [17]. Στον κλάδο των IoT, οι αισθητήρες συλλέγουν και στέλνουν δεδομένα συνεχώς, τα οποία πρέπει να υποβληθούν σε ανάλυση σε πραγματικό χρόνο για τη λήψη κρίσιμων αποφάσεων. Για παράδειγμα, σε ένα έξυπνο εργοστάσιο, η καθυστερημένη επεξεργασία δεδομένων από μηχανήματα μπορεί να οδηγήσει σε δυσλειτουργίες και μειωμένη αποδοτικότητα. Στην υγειονομική περίθαλψη, τα δεδομένα ασθενών πρέπει να επεξεργάζονται άμεσα, ειδικά σε περιπτώσεις κρίσιμης φροντίδας. Η άμεση επεξεργασία βιομετρικών δεδομένων ή άλλων ιατρικών μετρήσεων μπορεί να σώσει ζωές [6].

Η πρόκληση της ταχύτητας σχετίζεται άμεσα με την ανάγκη για ισχυρές υποδομές που μπορούν να επεξεργαστούν μεγάλες ροές δεδομένων ταυτόχρονα και με ελάχιστη καθυστέρηση. Παραδοσιακές υποδομές πληροφορικής δεν μπορούν να αντεπεξέλθουν στην ανάγκη επεξεργασίας δεδομένων σε πραγματικό χρόνο, οδηγώντας στην ανάγκη για νέες τεχνολογίες και κατανεμημένα συστήματα υπολογιστικής ισχύος [18].

Συστήματα όπως το Apache Kafka, το Apache Storm, το Apache Flink και το Spark Streaming έχουν αναπτυχθεί για να διαχειρίζονται τη ροή και την επεξεργασία δεδομένων σε πραγματικό χρόνο, επιτρέποντας στους οργανισμούς να αναλύουν τεράστιους όγκους δεδομένων καθώς παράγονται, χωρίς καθυστερήσεις. Ωστόσο, η υλοποίηση αυτών των συστημάτων έρχεται με τις δικές της προκλήσεις, όπως το υψηλό κόστος υποδομών και η ανάγκη για εξειδικευμένες γνώσεις στη διαχείριση και παραμετροποίηση τέτοιων εργαλείων. Η δυνατότητα κλιμάκωσης αυτών των συστημάτων σε περιβάλλοντα όπου τα δεδομένα αυξάνονται εκθετικά παραμένει μια διαρκής πρόκληση [13].

Επιπλέον, η ταχύτητα επεξεργασίας big data απαιτεί την ανάπτυξη αλγορίθμων που να μπορούν να λειτουργούν αποδοτικά σε κατανεμημένα περιβάλλοντα και να υποστηρίζουν παράλληλη επεξεργασία. Πολλοί παραδοσιακοί αλγόριθμοι δεν έχουν σχεδιαστεί για να λειτουργούν σε περιβάλλοντα μεγάλων δεδομένων και απαιτείται ανασχεδιασμός

ή βελτιστοποίηση για να είναι συμβατοί με τις νέες απαιτήσεις ταχύτητας. Η ανάπτυξη λογισμικού που μπορεί να διαχειριστεί τέτοιες ταχύτητες επεξεργασίας απαιτεί εκτενή χρήση τεχνικών παραλληλισμού και βελτιστοποίησης της απόδοσης, κάτι που μπορεί να είναι εξαιρετικά περίπλοκο και να απαιτεί εξειδικευμένο προσωπικό. Η πολυπλοκότητα της υλοποίησης και διαχείρισης τέτοιων συστημάτων συχνά αποτελεί τροχοπέδη για πολλές εταιρείες που δεν διαθέτουν τους απαραίτητους πόρους [29].

Η επεξεργασία μεγάλων δεδομένων σε πραγματικό χρόνο αυξάνει την ανάγκη για άμεση ανταπόκριση και εξαιρετικά μικρούς χρόνους καθυστέρησης. Η πρόκληση είναι να εξασφαλιστεί ότι τα συστήματα μπορούν να επεξεργάζονται τα δεδομένα τόσο γρήγορα όσο παράγονται, χωρίς να υπάρχει στέρηση στην ανάλυση [16]. Εάν τα δεδομένα δεν επεξεργαστούν άμεσα, χάνεται η ευκαιρία για χρήσιμα συμπεράσματα και λήψη αποφάσεων.

2.2.3 Ποικιλία δεδομένων

Η ποικιλία δεδομένων είναι άλλη μία από τις κύριες προκλήσεις στην επεξεργασία Big Data, και αφορά τη διαφορετικότητα των τύπων δεδομένων που συλλέγονται από διάφορες πηγές και σε διάφορες μορφές. Σε αντίθεση με τις παραδοσιακές βάσεις δεδομένων, όπου τα δεδομένα είναι δομημένα και αποθηκεύονται σε μορφές που είναι εύκολα επεξεργάσιμες, τα μεγάλα δεδομένα περιλαμβάνουν μεγάλο εύρος μορφών [18].

Αναλυτικότερα:

- Δομημένα δεδομένα: Αυτά τα δεδομένα έχουν καθορισμένη μορφή, όπως οι παραδοσιακές βάσεις δεδομένων (π.χ. πίνακες SQL). Είναι εύκολα διαχειρίσιμα και αναλύσιμα με παραδοσιακά εργαλεία και αλγορίθμους.
- Ημιδομημένα δεδομένα: Τα δεδομένα αυτά έχουν κάποια δομή αλλά δεν είναι πλήρως καθορισμένα, όπως αρχεία XML ή JSON, όπου η δομή υπάρχει αλλά δεν είναι τόσο αυστηρή όσο στα δομημένα δεδομένα.
- Αδόμητα δεδομένα: Αυτή η κατηγορία περιλαμβάνει δεδομένα που δεν ακολουθούν καθορισμένη μορφή, όπως κείμενα από κοινωνικά δίκτυα, email, βίντεο, εικόνες και ηχητικά αρχεία. Αυτά τα δεδομένα είναι δύσκολο να επεξεργαστούν και να αναλυθούν με παραδοσιακές μεθόδους [19].

Η πρόκληση έγκειται στο ότι κάθε τύπος δεδομένων απαιτεί διαφορετικές τεχνικές επεξεργασίας, αποθήκευσης και ανάλυσης. Η ενσωμάτωση δεδομένων από διαφορετικές πηγές είναι μία από τις κύριες προκλήσεις της ποικιλίας δεδομένων. Η ετερογένεια στη

δομή και στο περιεχόμενο των δεδομένων απαιτεί εξειδικευμένες τεχνικές ενοποίησης και ανάλυσης. Τα δεδομένα από διαφορετικές πηγές μπορεί να περιλαμβάνουν αντικρουόμενες ή ασυνεπείς πληροφορίες, γεγονός που καθιστά δύσκολη την αποδοτική ενοποίηση [20]. Για παράδειγμα, μια επιχείρηση που συλλέγει δεδομένα από τα κοινωνικά δίκτυα, τα αρχεία συναλλαγών, και τους αισθητήρες IoT, πρέπει να βρει τρόπους να συνδυάσει δεδομένα από αδόμητα κείμενα και εικόνες με πιο παραδοσιακά δεδομένα, όπως αριθμητικά στοιχεία από βάσεις δεδομένων.

Επίσης, η ανάλυση των δεδομένων ποικίλης μορφής συνιστά μια σημαντική πρόκληση. Παραδοσιακές μέθοδοι και εργαλεία ανάλυσης, όπως οι βάσεις δεδομένων SQL, είναι καλά προσαρμοσμένα για τη διαχείριση δομημένων δεδομένων, αλλά είναι λιγότερο αποδοτικά για την επεξεργασία ημιδομημένων και αδόμητων δεδομένων. Η ανάλυση αδόμητων δεδομένων, όπως οι εικόνες, τα βίντεο ή τα κείμενα, απαιτεί ειδικές τεχνικές, όπως επεξεργασία φυσικής γλώσσας (NLP) για τα κείμενα, ή αλγόριθμους μηχανικής μάθησης για την αναγνώριση προτύπων στις εικόνες και τα βίντεο. Αυτό σημαίνει ότι οι οργανισμοί πρέπει να επενδύσουν σε νέες τεχνολογίες και υποδομές για να μπορούν να αντλήσουν χρήσιμες πληροφορίες από τα δεδομένα [18]. Οι πλατφόρμες ανάλυσης προσφέρουν δυνατότητες για την επεξεργασία διαφορετικών μορφών δεδομένων σε μεγάλο όγκο, αλλά η ανάπτυξη και διαχείριση αυτών των πλατφορμών απαιτεί εξειδικευμένες γνώσεις και σημαντικούς πόρους [13].

Μια άλλη πρόκληση που προκύπτει από την ποικιλία των δεδομένων είναι η διαχείριση ασυνεπών ή θορυβωδών δεδομένων. Δεδομένα από διαφορετικές πηγές μπορεί να περιλαμβάνουν αντικρουόμενες πληροφορίες ή δεδομένα χαμηλής ποιότητας που μπορούν να επηρεάσουν την ακρίβεια της ανάλυσης. Για παράδειγμα, τα δεδομένα από τα κοινωνικά δίκτυα μπορεί να είναι γεμάτα από ανακρίβειες, σφάλματα ή πληροφορίες που δεν είναι σχετικές με την ανάλυση που επιδιώκεται. Η διαδικασία καθαρισμού αυτών των δεδομένων, η εξάλειψη των θορύβων και η διασφάλιση της συνέπειας είναι πολύπλοκες και χρονοβόρες διαδικασίες [17].

2.2.4 Ποιότητα και ακρίβεια δεδομένων

Μια από τις σημαντικότερες προκλήσεις στην επεξεργασία μεγάλων δεδομένων αποτελεί η ποιότητα και η ακρίβεια των δεδομένων. Καθώς οι οργανισμοί συλλέγουν τεράστιες ποσότητες δεδομένων από πολλές διαφορετικές πηγές, η εξασφάλιση της ποιότητας και

της ακρίβειας αυτών των δεδομένων γίνεται κρίσιμη για την αξιόπιστη ανάλυση και την εξαγωγή χρήσιμων συμπερασμάτων.

Αναλυτικότερα, η ποιότητα των δεδομένων αναφέρεται στην ακεραιότητα, την πληρότητα, τη συνέπεια και την ακρίβεια των δεδομένων. Η ακρίβεια είναι το μέτρο που δείχνει το πόσο ακριβείς και αξιόπιστες είναι οι πληροφορίες που περιέχονται στα δεδομένα. Κατά την επεξεργασία μεγάλων δεδομένων, αυτά τα δύο χαρακτηριστικά είναι απαραίτητα για να διασφαλιστεί ότι τα δεδομένα που χρησιμοποιούνται για ανάλυση οδηγούν σε σωστά και αξιόπιστα αποτελέσματα [22]. Ωστόσο, η συλλογή δεδομένων από πολλές διαφορετικές πηγές, αυξάνει τον κίνδυνο να περιέχονται στα δεδομένα λάθη, ανακρίβειες ή θόρυβοι που επηρεάζουν την ποιότητα και την ακρίβειά τους [21].

Μια από τις βασικές προκλήσεις που συνδέονται με την ποιότητα των δεδομένων είναι η διαχείριση ελλιπών και ασυνεπών δεδομένων. Οι τεράστιοι όγκοι πληροφορίας συχνά περιλαμβάνουν αρχεία που μπορεί να είναι ατελή ή να περιέχουν αντικρουόμενες πληροφορίες. Για παράδειγμα, δεδομένα από διάφορες πηγές μπορεί να έχουν διαφορετικές μορφές, διαφορετικές μονάδες μέτρησης ή να περιέχουν κενά πεδία. Αυτή η ασυνέπεια δυσχεραίνει τη σωστή ενοποίηση των δεδομένων και την εξαγωγή αξιόπιστων αποτελεσμάτων. Η έλλειψη πληρότητας στα δεδομένα μπορεί επίσης να οδηγήσει σε λανθασμένες αναλύσεις, καθώς τα υπολογιστικά μοντέλα που χρησιμοποιούνται για την ανάλυση βασίζονται συχνά στην υπόθεση ότι όλα τα δεδομένα είναι διαθέσιμα και ακριβή [22]. Η αντιμετώπιση των ελλιπών δεδομένων απαιτεί διαδικασίες καθαρισμού δεδομένων, στις οποίες τα δεδομένα που λείπουν πρέπει να συμπληρώνονται ή να αφαιρούνται χωρίς να αλλοιώνονται τα αποτελέσματα της ανάλυσης [18].

Η παρουσία θορύβου στα δεδομένα είναι μια άλλη σοβαρή πρόκληση που συνδέεται με την ποιότητα και την ακρίβεια των δεδομένων. Όπως αναφέρθηκε πιο πάνω, ο θόρυβος αναφέρεται σε πληροφορίες που είναι άσχετες ή λανθασμένες και μπορούν να προκαλέσουν προβλήματα στην ανάλυση των δεδομένων. Θορυβώδη δεδομένα μπορεί να προκύψουν από σφάλματα στην εισαγωγή δεδομένων, δυσλειτουργίες συσκευών ή λάθη στη συλλογή τους. Τα ανακριβή δεδομένα μπορεί να περιλαμβάνουν διπλά εγγραφές, λάθη σε αριθμούς ή ημερομηνίες, ή ακόμη και εσκεμμένες ψευδείς πληροφορίες. Η ανάλυση αυτών των δεδομένων χωρίς κατάλληλο καθαρισμό μπορεί να οδηγήσει σε παραπλανητικά ή λανθασμένα συμπεράσματα [22].

Ένα άλλο ζήτημα που σχετίζεται με την ποιότητα των δεδομένων είναι η επαλήθευση και η εγκυρότητα των δεδομένων. Οι οργανισμοί πρέπει να διασφαλίσουν ότι τα δεδομένα που συλλέγουν είναι ακριβή, έγκυρα και προέρχονται από αξιόπιστες πηγές. Δεδομένα που δεν έχουν επαληθευτεί μπορεί να περιέχουν σφάλματα ή να είναι παρωχημένα, κάτι που μπορεί να επηρεάσει αρνητικά την ακρίβεια των αναλύσεων και τη λήψη αποφάσεων [20]. Ειδικά στον τομέα της υγειονομικής περίθαλψης ή των χρηματοπιστωτικών υπηρεσιών, όπου η ακρίβεια των δεδομένων είναι κρίσιμη, η επαλήθευση των δεδομένων είναι απαραίτητη για την εξασφάλιση της ποιότητας των αποτελεσμάτων. Οι οργανισμοί πρέπει να εφαρμόζουν μηχανισμούς που επαληθεύουν την προέλευση και την εγκυρότητα των δεδομένων πριν αυτά χρησιμοποιηθούν για ανάλυση.

2.2.5 Ασφάλεια και ιδιωτικότητα

Η ασφάλεια και η ιδιωτικότητα αποτελούν μείζονες προκλήσεις στην επεξεργασία big data. Η συλλογή, αποθήκευση και ανάλυση τεράστιων ποσοτήτων δεδομένων από ποικίλες πηγές περιλαμβάνει σημαντικά θέματα που σχετίζονται με την προστασία των δεδομένων από παραβιάσεις και την εξασφάλιση ότι η ιδιωτικότητα των χρηστών δεν παραβιάζεται. Καθώς οι οργανισμοί συλλέγουν όλο και περισσότερα δεδομένα, όπως προσωπικές πληροφορίες, δεδομένα από συσκευές IoT, δεδομένα από κοινωνικά δίκτυα και άλλες πηγές, η ανάγκη για αυξημένη ασφάλεια και προστασία της ιδιωτικότητας είναι πιο επείγουσα από ποτέ [23]. Η παραβίαση αυτών των δεδομένων μπορεί να οδηγήσει σε σοβαρές συνέπειες για τους χρήστες και τις επιχειρήσεις, όπως κλοπή ταυτότητας, απώλεια εμπιστοσύνης, και οικονομικές κυρώσεις.

Οι κυριότερες πτυχές της πρόκλησης σχετικά με την ασφάλεια των μεγάλων δεδομένων είναι:

- Διαχείριση της πρόσβασης

Καθώς οι οργανισμοί διαχειρίζονται μεγάλες ποσότητες δεδομένων, είναι απαραίτητο να ελέγχεται ποιος έχει πρόσβαση σε αυτά τα δεδομένα και σε ποιο επίπεδο. Αυτό απαιτεί την εφαρμογή αυστηρών πολιτικών πρόσβασης, που καθορίζουν ποιοι χρήστες μπορούν να δουν ή να επεξεργαστούν συγκεκριμένα δεδομένα. Οι πολιτικές αυτές πρέπει να εφαρμόζονται με τεχνολογίες όπως ο έλεγχος ταυτότητας, η κρυπτογράφηση και τα συστήματα διαχείρισης δικαιωμάτων. Η διαχείριση πρόσβασης είναι ιδιαίτερα δύσκολη σε περιβάλλοντα μεγάλων δεδομένων, όπου τα δεδομένα βρίσκονται διασκορπισμένα σε πολλαπλά συστήματα και τοπικές ή κατανεμημένες υποδομές [24].

- Κρυπτογράφηση δεδομένων

Η κρυπτογράφηση των δεδομένων είναι μία από τις πιο συνηθισμένες πρακτικές για την προστασία τους από μη εξουσιοδοτημένη πρόσβαση ή παραβιάσεις. Η κρυπτογράφηση εξασφαλίζει ότι τα δεδομένα είναι κατανοητά μόνο από εξουσιοδοτημένα άτομα ή συστήματα, ακόμη και αν αυτά τα δεδομένα αποκτηθούν από κακόβουλους χρήστες. Στο πλαίσιο των μεγάλων δεδομένων, η κρυπτογράφηση γίνεται πιο περίπλοκη λόγω της ποσότητας των δεδομένων και της ανάγκης για επεξεργασία τους σε πραγματικό χρόνο. Η κρυπτογράφηση πρέπει να εφαρμόζεται τόσο κατά τη μεταφορά (data in transit) όσο και κατά την αποθήκευση (data at rest), ενώ η κλιμάκωση αυτής της διαδικασίας μπορεί να έχει σημαντικές επιπτώσεις στην απόδοση των συστημάτων [24].

- Ασφάλεια σε κατανεμημένα συστήματα

Τα μεγάλα δεδομένα συνήθως επεξεργάζονται σε κατανεμημένα περιβάλλοντα, όπως το Hadoop και το Spark, όπου τα δεδομένα κατανέμονται σε πολλούς κόμβους ή ακόμα και σε διαφορετικές γεωγραφικές τοποθεσίες. Αυτό προσθέτει μια επιπλέον πολυπλοκότητα στην ασφάλεια, καθώς τα δεδομένα πρέπει να προστατεύονται όχι μόνο στον κεντρικό διακομιστή, αλλά και σε κάθε επιμέρους κόμβο του δικτύου. Η ασφάλεια σε κατανεμημένα συστήματα απαιτεί τη διασφάλιση της εμπιστευτικότητας, της ακεραιότητας και της διαθεσιμότητας των δεδομένων σε πολλαπλά σημεία. Οι οργανισμοί πρέπει να εφαρμόζουν πολιτικές ασφαλείας σε κάθε κόμβο του δικτύου, χρησιμοποιώντας εργαλεία κρυπτογράφησης, firewalls και συστήματα ανίχνευσης απειλών [23].

Όσον αφορά τις διαστάσεις της πρόκλησης σχετικά με την ιδιωτικότητα, οι βασικότερες είναι:

- Προστασία προσωπικών δεδομένων

Η διαχείριση μεγάλων δεδομένων συχνά περιλαμβάνει την επεξεργασία προσωπικών δεδομένων, όπως δεδομένα καταναλωτών, ιατρικά αρχεία ή πληροφορίες κοινωνικών μέσων. Οι κανονισμοί όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση απαιτούν από τους οργανισμούς να προστατεύουν τα προσωπικά δεδομένα των ατόμων και να εξασφαλίζουν ότι δεν παραβιάζεται η ιδιωτικότητα τους [25]. Η συμμόρφωση με κανονισμούς όπως ο GDPR απαιτεί αυστηρές διαδικασίες για την επεξεργασία των προσωπικών δεδομένων, την ανωνυμοποίηση και την ψευδωνυμοποίηση τους, καθώς και τη διασφάλιση ότι οι χρήστες έχουν τον έλεγχο των δεδομένων τους [24]. Οποιαδήποτε παραβίαση αυτών των δεδομένων μπορεί να έχει

σοβαρές νομικές συνέπειες για τους οργανισμούς, όπως υψηλά πρόστιμα και ζημιά στη φήμη τους.

- Ανωνυμοποίηση και ψευδωνυμοποίηση δεδομένων

Για να διασφαλιστεί η ιδιωτικότητα, πολλά δεδομένα πρέπει να υποβάλλονται σε διαδικασίες ανωνυμοποίησης ή ψευδωνυμοποίησης. Η ανώνυμοποίηση των δεδομένων αναφέρεται στη διαδικασία κατά την οποία τα δεδομένα τροποποιούνται με τέτοιο τρόπο ώστε να μην είναι δυνατόν να συνδεθούν με συγκεκριμένα άτομα. Η ψευδωνυμοποίηση είναι μια λιγότερο αυστηρή μέθοδος που επιτρέπει την απόκρυψη προσωπικών ταυτοτήτων, αλλά διατηρεί μια σύνδεση με τα δεδομένα σε περίπτωση που χρειαστεί να επανέλθουν. Ωστόσο, η ανώνυμοποίηση των δεδομένων δεν είναι πάντα εύκολη και συχνά μπορεί να αντιστραφεί με τη χρήση σύνθετων αλγορίθμων, κάτι που θέτει σε κίνδυνο την ιδιωτικότητα [25]. Αυτό αποτελεί μια σοβαρή πρόκληση για τους οργανισμούς που διαχειρίζονται δεδομένα ευαίσθητα ή προσωπικά, όπως δεδομένα υγείας ή οικονομικά στοιχεία.

- Ασφάλεια κατά την ανταλλαγή δεδομένων

Η ανταλλαγή δεδομένων μεταξύ οργανισμών ή εσωτερικών τμημάτων είναι συχνή στις εφαρμογές μεγάλων δεδομένων. Αυτή η ανταλλαγή δημιουργεί κινδύνους για την ιδιωτικότητα, καθώς δεδομένα που μοιράζονται μπορεί να περιέχουν προσωπικές ή εμπιστευτικές πληροφορίες που δεν πρέπει να εκτίθενται σε τρίτους. Ο έλεγχος της πρόσβασης σε αυτά τα δεδομένα και η διασφάλιση ότι ανταλλάσσονται με ασφαλή τρόπο είναι απαραίτητα για την προστασία της ιδιωτικότητας [16].

2.2.6 Κόστος υποδομών

Το κόστος υποδομών είναι ακόμα μια σημαντική πρόκληση στην επεξεργασία μεγάλων δεδομένων. Η υποδομή για την επεξεργασία μεγάλων δεδομένων δεν περιλαμβάνει μόνο την αποθήκευση μεγάλων ποσοτήτων δεδομένων, αλλά και την επεξεργασία και ανάλυση αυτών των δεδομένων σε πραγματικό ή σχεδόν πραγματικό χρόνο. Αυτό απαιτεί σημαντικούς πόρους σε επίπεδο υλικού (servers, δίκτυα, αποθηκευτικά συστήματα) και λογισμικού (πλατφόρμες ανάλυσης, διαχείριση δεδομένων, εργαλεία μηχανικής μάθησης κ.λπ.). Αυτές οι υποδομές πρέπει να είναι ευέλικτες, αξιόπιστες και σε θέση να διαχειρίζονται συνεχώς αυξανόμενους όγκους δεδομένων [26]

Ειδικότερα, ένα κρίσιμο ζήτημα σχετικά με το κόστος υποδομών είναι η αποθήκευση των δεδομένων. Τα big data απαιτούν τεράστιες ποσότητες αποθηκευτικού χώρου για να διατηρηθούν τα δεδομένα τόσο σε δομημένη όσο και σε αδόμητη μορφή. Η αγορά και συντήρηση εξειδικευμένων αποθηκευτικών συστημάτων, όπως οι συστοιχίες NAS (Network Attached Storage) ή τα SAN (Storage Area Networks), μπορεί να είναι εξαιρετικά δαπανηρή. Οι ανάγκες για αποθήκευση αυξάνονται συνεχώς, καθώς τα δεδομένα παράγονται με εκθετικό ρυθμό από αισθητήρες, εφαρμογές διαδικτύου, κοινωνικά δίκτυα και άλλες πηγές [27]. Επιπλέον, οι οργανισμοί χρειάζονται ασφαλή και αξιόπιστα αποθηκευτικά μέσα, τα οποία μπορεί να αυξήσουν ακόμη περισσότερο το κόστος, ειδικά όταν απαιτείται αδιάκοπη λειτουργία και μεγάλοι χρόνοι διατήρησης δεδομένων.

Η υπολογιστική ισχύς είναι ένας άλλος κρίσιμος παράγοντας που συμβάλλει στο κόστος υποδομών. Η ανάλυση μεγάλων δεδομένων απαιτεί τεράστιους υπολογιστικούς πόρους για να επεξεργαστεί τα δεδομένα με ταχύτητα και αποδοτικότητα. Οι οργανισμοί πρέπει να επενδύσουν σε ισχυρούς servers ή σε κατανεμημένα συστήματα, όπως το Hadoop και το Spark, που μπορούν να κατανεμηθούν σε πολλούς κόμβους και να εκτελούν παράλληλες διεργασίες για να επιταχύνουν την ανάλυση των δεδομένων. Η συντήρηση αυτών των συστημάτων απαιτεί συνεχή επένδυση σε υλικό, όπως επεξεργαστές υψηλής απόδοσης (CPU), μονάδες επεξεργασίας γραφικών (GPU), και μνήμες υψηλής ταχύτητας (RAM), τα οποία κοστίζουν σημαντικά [14]. Επιπλέον, η ανάπτυξη εξειδικευμένων υποδομών για την παράλληλη επεξεργασία δεδομένων και η ενσωμάτωση αλγορίθμων μηχανικής μάθησης απαιτούν εξειδικευμένες τεχνολογίες, οι οποίες είναι ακριβές τόσο στην απόκτηση όσο και στη λειτουργία τους.

Οι απαιτήσεις για υπολογιστική ισχύ και αποθήκευση οδηγούν σε υψηλή κατανάλωση ενέργειας και ανάγκες ψύξης. Οι datacenters, όπου αποθηκεύονται και επεξεργάζονται τα μεγάλα δεδομένα, καταναλώνουν τεράστιες ποσότητες ενέργειας, τόσο για τη λειτουργία του εξοπλισμού όσο και για την ψύξη του, προκειμένου να αποφεύγεται η υπερθέρμανση. Αυτή η αυξημένη κατανάλωση ενέργειας οδηγεί σε αυξημένα λειτουργικά κόστη, ειδικά όταν οι εγκαταστάσεις πρέπει να λειτουργούν 24/7 για να εξασφαλιστεί η συνεχής διαθεσιμότητα των δεδομένων και των συστημάτων [27].

Το κόστος του λογισμικού είναι επίσης σημαντικό μέρος της πρόκλησης των υποδομών για μεγάλους όγκους πληροφορίας. Τα εργαλεία ανάλυσης, η διαχείριση δεδομένων, οι πλατφόρμες επεξεργασίας και οι βάσεις δεδομένων μπορεί να έχουν υψηλά κόστη απόκτησης και άδειας χρήσης. Επιπλέον, τα συστήματα που απαιτούνται για τη λειτουργία

αυτών των υποδομών, όπως συστήματα διαχείρισης βάσεων δεδομένων (π.χ. Oracle, SQL Server) και πλατφόρμες ανάλυσης (π.χ. SAS, Tableau), συχνά απαιτούν συνδρομητικές υπηρεσίες ή αδειοδοτήσεις που κοστίζουν ακριβά σε βάθος χρόνου [28]. Ειδικά σε περιβάλλοντα μεγάλων δεδομένων, όπου τα εργαλεία πρέπει να είναι κλιμακώσιμα και να διαχειρίζονται μεγάλα φορτία εργασίας, το κόστος του λογισμικού μπορεί να αυξηθεί σημαντικά, ειδικά αν απαιτούνται προσαρμοσμένες λύσεις ή αν επενδυθούν σε λύσεις τεχνητής νοημοσύνης και μηχανικής μάθησης [26].

Η συντήρηση και η αναβάθμιση των υποδομών είναι ένα ακόμα σημαντικό κόστος που πρέπει να ληφθεί υπόψη. Καθώς οι τεχνολογίες εξελίσσονται και οι απαιτήσεις για ανάλυση δεδομένων αυξάνονται, οι υποδομές πρέπει να αναβαθμίζονται τακτικά για να παραμείνουν ανταγωνιστικές και αποδοτικές. Η αντικατάσταση παλαιότερων συστημάτων με νεότερες, ταχύτερες και πιο αποδοτικές λύσεις μπορεί να απαιτεί συνεχή επένδυση σε νέο εξοπλισμό, λογισμικό και τεχνολογίες. Η συντήρηση αυτών των συστημάτων περιλαμβάνει επίσης την επισκευή και την αντικατάσταση εξοπλισμού, καθώς και την παροχή τεχνικής υποστήριξης για την επίλυση προβλημάτων που μπορεί να προκύψουν κατά τη λειτουργία [28]. Αυτά τα κόστη μπορεί να αυξηθούν δραματικά όταν πρόκειται για μεγάλες εγκαταστάσεις ή συστήματα με σύνθετες ανάγκες επεξεργασίας.

Το ανθρώπινο δυναμικό είναι ένα ακόμη σημαντικό κομμάτι του κόστους των υποδομών. Οι οργανισμοί που επεξεργάζονται μαζικά σύνολα δεδομένων χρειάζονται εξειδικευμένο προσωπικό για την ανάπτυξη, την παρακολούθηση και τη διαχείριση αυτών των υποδομών. Οι ειδικοί σε τομείς όπως η διαχείριση βάσεων δεδομένων, η μηχανική μάθηση, η ανάλυση δεδομένων και η διαχείριση συστημάτων είναι εξαιρετικά περιζήτητοι, και η αμοιβή τους μπορεί να είναι υψηλή. Επιπλέον, η συνεχής εκπαίδευση και κατάρτιση του προσωπικού είναι απαραίτητη για την αντιμετώπιση των συνεχών τεχνολογικών εξελίξεων [6]. Το κόστος των προσλήψεων και της διατήρησης αυτού του εξειδικευμένου δυναμικού αποτελεί ένα σημαντικό κόστος που πρέπει να συνυπολογιστεί στις συνολικές δαπάνες υποδομών για μεγάλα δεδομένα.

2.2.7 Διαχείριση πόρων και παραλληλισμός

Ένα κρίσιμο στοιχείο στην επεξεργασία μεγάλων δεδομένων είναι και η διαχείριση πόρων και ο παραλληλισμός. Οι υπολογιστικοί πόροι που απαιτούνται για την αποδοτική ανάλυση και επεξεργασία τεράστιων ποσοτήτων δεδομένων είναι μεγάλοι, ενώ η κατανομή και ο συγχρονισμός αυτών των πόρων πρέπει να γίνονται με τρόπο που να

εξασφαλίζει την αποδοτικότητα και την ταχύτητα της ανάλυσης. Ο παραλληλισμός, δηλαδή η ταυτόχρονη εκτέλεση πολλαπλών διεργασιών σε κατανεμημένα συστήματα, είναι ένας από τους βασικούς μηχανισμούς για την επιτάχυνση της επεξεργασίας μεγάλων δεδομένων [29]. Ωστόσο, η αποτελεσματική διαχείριση πόρων και παραλληλισμός παρουσιάζουν σημαντικές τεχνικές προκλήσεις.

Αναλυτικότερα, η διαχείριση πόρων περιλαμβάνει την κατανομή των διαθέσιμων υπολογιστικών πόρων, όπως η CPU, η μνήμη, και η αποθήκευση, με τρόπο που να επιτυγχάνεται η μέγιστη αποδοτικότητα. Στην επεξεργασία μεγάλων δεδομένων, οι πόροι αυτοί πρέπει να κατανέμονται μεταξύ πολλών διεργασιών και κόμβων, ανάλογα με τις ανάγκες κάθε εργασίας. Ορισμένα είδη επεξεργασίας μπορεί να απαιτούν περισσότερη υπολογιστική ισχύ, ενώ άλλα μπορεί να χρειάζονται περισσότερη μνήμη ή αποθηκευτικό χώρο. Ένα από τα κύρια ζητήματα που προκύπτουν στη διαχείριση πόρων είναι η εξισορρόπηση του φορτίου μεταξύ των κόμβων ενός κατανεμημένου συστήματος. Εάν ορισμένοι κόμβοι είναι υπερφορτωμένοι ενώ άλλοι υποχρησιμοποιούνται, η συνολική απόδοση του συστήματος θα είναι χαμηλή. Η δυναμική κατανομή των πόρων, ανάλογα με τις ανάγκες των διεργασιών, είναι κρίσιμη για την αποφυγή συμφόρησης και την εξασφάλιση ότι οι πόροι χρησιμοποιούνται αποτελεσματικά [30].

Ο παραλληλισμός είναι η βασική τεχνική που χρησιμοποιείται για την επιτάχυνση της επεξεργασίας μεγάλων δεδομένων. Ουσιαστικά, ο παραλληλισμός επιτρέπει τη διάσπαση μιας μεγάλης εργασίας σε μικρότερα μέρη, τα οποία εκτελούνται ταυτόχρονα σε διαφορετικούς επεξεργαστές ή κόμβους του συστήματος. Ο στόχος είναι να αυξηθεί η αποδοτικότητα και να μειωθεί ο συνολικός χρόνος επεξεργασίας [29].

Ωστόσο, η επίτευξη αποτελεσματικού παραλληλισμού παρουσιάζει τεχνικές προκλήσεις, όπως για παράδειγμα η ανισομερής κατανομή εργασιών. Ορισμένες εργασίες μπορεί να είναι πιο περίπλοκες και να απαιτούν περισσότερους υπολογιστικούς πόρους από άλλες. Αν δεν κατανεμηθούν σωστά οι εργασίες, ορισμένοι κόμβοι μπορεί να επιβαρυνθούν περισσότερο, ενώ άλλοι παραμένουν αδρανείς [31]. Σε ένα κατανεμημένο σύστημα, εάν ένας κόμβος έχει πολύ περισσότερη εργασία από άλλους, μπορεί να οδηγήσει σε συμφόρηση που επηρεάζει ολόκληρο το σύστημα. Η συμφόρηση αυτή μπορεί να μειώσει την αποδοτικότητα του παραλληλισμού και να αυξήσει τους χρόνους επεξεργασίας. Επίσης, όταν πολλές διεργασίες εκτελούνται ταυτόχρονα, ο συγχρονισμός τους είναι απαραίτητος για να εξασφαλιστεί ότι τα αποτελέσματα από διαφορετικές διεργασίες θα συνδυαστούν σωστά και θα οδηγήσουν σε ακριβή τελικά αποτελέσματα. Ο συγχρονισμός

αυτός μπορεί να είναι περίπλοκος, ειδικά σε περιβάλλοντα κατανεμημένης επεξεργασίας, όπου οι διεργασίες εκτελούνται σε διαφορετικούς κόμβους και συστήματα [29]. Ο παραλληλισμός σε κατανεμημένα συστήματα εξαρτάται από την επικοινωνία μεταξύ των κόμβων μέσω δικτύου. Οποιαδήποτε καθυστέρηση ή συμφόρηση στο δίκτυο μπορεί να επηρεάσει την απόδοση του συστήματος και να αυξήσει τους χρόνους επεξεργασίας [31].

2.2.8 Εξαγωγή πληροφοριών και λήψη αποφάσεων

Η εξαγωγή πληροφοριών και η λήψη αποφάσεων είναι δύο από τις πιο κρίσιμες προκλήσεις στην επεξεργασία Big Data. Η επεξεργασία μεγάλων όγκων δεδομένων δεν έχει αξία αν δεν μπορεί να οδηγήσει σε ουσιαστικές πληροφορίες και να υποστηρίξει τη λήψη αποτελεσματικών αποφάσεων. Η διαδικασία εξαγωγής χρήσιμων πληροφοριών από τα μεγάλα δεδομένα και η αξιοποίηση αυτών για τη λήψη τεκμηριωμένων αποφάσεων είναι ιδιαίτερα πολύπλοκη και απαιτεί εξειδικευμένες τεχνικές, εργαλεία και υποδομές [32].

Οι προκλήσεις που σχετίζονται με την εξαγωγή πληροφοριών και τη λήψη αποφάσεων από μεγάλα δεδομένα συνδέονται με την πολυπλοκότητα και την ποικιλία των δεδομένων, την ταχύτητα με την οποία πρέπει να ληφθούν αποφάσεις, την ποιότητα των δεδομένων και την ικανότητα των οργανισμών να ερμηνεύουν τα αποτελέσματα της ανάλυσης με αξιόπιστο και αποδοτικό τρόπο. Το ζήτημα της πολυπλοκότητας, της ταχύτητας και της ποιότητας των δεδομένων αναλύθηκαν εκτενώς πιο πάνω. Επιπλέον, υπάρχει και το ζήτημα της μετατροπής των ακατέργαστων δεδομένων σε χρήσιμες πληροφορίες που μπορούν να αξιοποιηθούν για τη λήψη αποφάσεων [18]. Τα ακατέργαστα δεδομένα από μόνα τους μπορεί να μην έχουν νόημα έως ότου υποστούν ανάλυση, εξαγωγή προτύπων ή ενσωματωθούν σε μοντέλα πρόβλεψης. Η εξαγωγή πληροφοριών μπορεί να περιλαμβάνει την ανάλυση τάσεων, τη δημιουργία προγνωστικών μοντέλων, την εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων και την κατανόηση σύνθετων σχέσεων μεταξύ μεταβλητών. Η πρόκληση είναι να αναπτυχθούν και να εφαρμοστούν οι κατάλληλοι αλγόριθμοι που μπορούν να εξάγουν αξιόπιστες πληροφορίες από αυτά τα δεδομένα. Στην περίπτωση μεγάλων δεδομένων, η ανάλυση δεν είναι απλή, καθώς απαιτεί τεχνικές όπως η μηχανική μάθηση, η εξόρυξη δεδομένων και η ανάλυση φυσικής γλώσσας (NLP), οι οποίες μπορούν να είναι ιδιαίτερα απαιτητικές σε επίπεδο υπολογιστικών πόρων και τεχνογνωσίας [28].

Εκτός από την εξαγωγή πληροφοριών από τα δεδομένα, άλλη μια ακόμη σημαντική πρόκληση είναι η ερμηνεία των αποτελεσμάτων και η εφαρμογή τους στη λήψη

αποφάσεων. Τα δεδομένα και τα αποτελέσματα των αναλύσεων πρέπει να ερμηνευθούν σωστά για να μπορέσουν να ληφθούν οι σωστές αποφάσεις. Αυτό απαιτεί όχι μόνο τεχνογνωσία στον τομέα της ανάλυσης δεδομένων αλλά και βαθιά κατανόηση του επιχειρηματικού πλαισίου και των συγκεκριμένων προκλήσεων που αντιμετωπίζει ο οργανισμός. Είναι επίσης σημαντικό να κατανοηθεί ότι τα αποτελέσματα της ανάλυσης δεν παρέχουν πάντα απόλυτες απαντήσεις [22]. Οι αποφάσεις που βασίζονται σε μεγάλα δεδομένα ενδέχεται να περιλαμβάνουν κάποια αβεβαιότητα ή ρίσκο, ειδικά όταν τα δεδομένα είναι ελλιπή ή ατελή. Για αυτό, η σωστή ερμηνεία των αποτελεσμάτων και η αξιολόγηση του βαθμού αβεβαιότητας αποτελούν κρίσιμα στοιχεία για τη λήψη τεκμηριωμένων αποφάσεων [22].

Μια άλλη πρόκληση είναι η σύνδεση των δεδομένων με στρατηγικές επιχειρηματικές αποφάσεις. Τα μαζικά σύνολα δεδομένων μπορεί να προσφέρουν σημαντικές γνώσεις, αλλά αυτές οι γνώσεις πρέπει να συνδεθούν με τις στρατηγικές προτεραιότητες και τους στόχους του οργανισμού. Η μετατροπή των πληροφοριών σε δράση απαιτεί τη διασύνδεση της ανάλυσης με τα επιχειρηματικά αποτελέσματα, όπως η αύξηση της απόδοσης, η μείωση των δαπανών ή η βελτίωση της εμπειρίας των πελατών. Η εφαρμογή αυτών των πληροφοριών για τη λήψη αποφάσεων μπορεί να είναι πολύπλοκη, ειδικά σε μεγάλης κλίμακας οργανισμούς, όπου οι στρατηγικές αποφάσεις πρέπει να λαμβάνονται με βάση πολυεπίπεδη και σύνθετη ανάλυση [32].

Συνοψίζοντας, η επεξεργασία μεγάλων δεδομένων παρουσιάζει πολυάριθμες προκλήσεις που απαιτούν την ανάπτυξη καινοτόμων τεχνολογιών, τη σωστή διαχείριση πόρων και τη διασφάλιση της ποιότητας και της ασφάλειας των δεδομένων. Καθώς τα δεδομένα συνεχίζουν να αυξάνονται με εκθετικούς ρυθμούς, οι οργανισμοί καλούνται να βελτιστοποιήσουν τις διαδικασίες επεξεργασίας τους για να παραμείνουν ανταγωνιστικοί και αποτελεσματικοί.

3 Ανάλυση Βασικών Αλγορίθμων

3.1 Αλγόριθμοι μηχανικής μάθησης για μεγάλα δεδομένα

Οι αλγόριθμοι μηχανικής μάθησης παίζουν έναν καθοριστικό ρόλο στην επεξεργασία και ανάλυση μεγάλων δεδομένων, καθώς επιτρέπουν την εξαγωγή προτύπων, τη δημιουργία προγνωστικών μοντέλων και την αυτοματοποίηση της ανάλυσης δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για να μετατρέψουν τεράστια σύνολα δεδομένων σε χρήσιμες πληροφορίες, υποστηρίζοντας τη λήψη αποφάσεων και την πρόβλεψη μελλοντικών τάσεων [33].

Πιο συγκεκριμένα, η μηχανική μάθηση που έχουν τη δυνατότητα να "μαθαίνουν" από δεδομένα και να βελτιώνουν την απόδοσή τους όσο περισσότερα δεδομένα επεξεργάζονται.. Σε περιβάλλοντα μεγάλων δεδομένων, οι αλγόριθμοι μηχανικής μάθησης πρέπει να είναι σε θέση να επεξεργαστούν τεράστια σύνολα δεδομένων σε κλίμακα που υπερβαίνει κατά πολύ τα παραδοσιακά συστήματα [15]. Αυτό σημαίνει ότι οι αλγόριθμοι πρέπει να είναι σχεδιασμένοι για να λειτουργούν αποδοτικά σε κατανεμημένα περιβάλλοντα και να χρησιμοποιούν πόρους με βέλτιστο τρόπο για να επιτυγχάνουν αποτελεσματική ανάλυση.

Οι πιο κοινές κατηγορίες αλγορίθμων μηχανικής μάθησης για μεγάλα δεδομένα περιλαμβάνουν αλγόριθμους ταξινόμησης, ομαδοποίησης, καθώς και πιο εξειδικευμένους αλγόριθμους όπως τα SVM (Support Vector Machines) και τα νευρωνικά δίκτυα [34]. Στον ακόλουθο πίνακα (Πίνακας 1) συνοψίζονται οι βασικές κατηγορίες αλγορίθμων, τα κύρια χαρακτηριστικά τους, τα πλεονεκτήματα και οι βασικές εφαρμογές τους, και στη συνέχεια, αναλύονται πιο διεξοδικά.

Πίνακας 1. Βασικοί αλγόριθμοι μηχανικής μάθησης για μεγάλα δεδομένα

Κατηγορία αλγορίθμου	Παραδείγματα	Κύρια χαρακτηριστικά	Πλεονεκτήματα	Ενδείκνυται για Big Data	Τυπικές εφαρμογές
Ταξινόμηση (Classification)	Decision Trees, Random Forest, Logistic Regression, SVM	Εποπτευόμενη μάθηση, προβλέπει κατηγορίες	Γρήγορη πρόβλεψη, ερμηνευσιμότητα, υψηλή ακρίβεια (SVM)	Ενδείκνυται, με κατάλληλη παραλληλοποίηση	Ανίχνευση απάτης, φίλτρα spam, ανάλυση συναισθήματος
Ομαδοποίηση (Clustering)	K-Means, Hierarchical Clustering, DBSCAN	Μη εποπτευόμενη, ανακαλύπτει ομάδες	Καλή κατανόηση δομών, μείωση πολυπλοκότητας	Ενδείκνυται, ειδικά K-Means & DBSCAN)	Τμηματοποίηση πελατών, ανάλυση κοινωνικών δικτύων
SVM	SVM με linear ή kernel	Δυαδική ταξινόμηση, υπερεπίπεδο διαχωρισμού	Υψηλή ακρίβεια, διαχείριση υψηλών διαστάσεων, ανθεκτικότητα	Περιορισμένη κλιμάκωση (απαιτεί ισχυρή υπολογιστική ισχύ)	Ανάλυση εικόνας/κειμένου, βιοϊατρικά δεδομένα
Νευρωνικά Δίκτυα (NNs)	MLP, CNN, RNN, DNN	Βαθιά εκμάθηση, αυτόματη εξαγωγή χαρακτηριστικών	Μάθηση σύνθετων προτύπων, scalability, διαχείριση πολυμέσων	Ενδείκνυται, σε κατανεμημένα περιβάλλοντα με GPU/TPU)	Ανάλυση εικόνων, NLP, προτάσεις προϊόντων, αυτόνομα συστήματα

3.1.1 Αλγόριθμοι ταξινόμησης (Classification Algorithms)

Οι αλγόριθμοι ταξινόμησης είναι από τους πιο κοινούς αλγορίθμους στη μηχανική μάθηση. Είναι αλγόριθμοι εποπτευόμενης μάθησης που εκπαιδεύονται σε σύνολα δεδομένων με ετικέτες (δηλαδή, δεδομένα με προκαθορισμένες κατηγορίες) και στη συνέχεια χρησιμοποιούνται για την κατηγοριοποίηση νέων, αόρατων δεδομένων. Ο στόχος είναι να ανακαλύψουν τα πρότυπα και τις σχέσεις στα δεδομένα εισόδου που σχετίζονται με τις κατηγορίες εξόδου, ώστε να μπορούν να κάνουν ακριβείς προβλέψεις για νέα δεδομένα [15].

Οι κύριοι αλγόριθμοι ταξινόμησης είναι:

- Δέντρα απόφασης (Decision trees): Δημιουργούν ένα δέντρο απόφασης με κόμβους που βασίζονται σε χαρακτηριστικά των δεδομένων. Είναι γρήγοροι, εύκολοι στην ερμηνεία, αλλά ενδέχεται να υπερεφαρμόζουν στα δεδομένα εκπαίδευσης.
- Random forest: Συνδυάζει πολλά δέντρα απόφασης για να βελτιώσει την ακρίβεια και τη γενίκευση του μοντέλου. Χρησιμοποιείται ευρέως σε μεγάλα δεδομένα λόγω της ανθεκτικότητάς του στον υπερπροσανατολισμό (overfitting).

- Logistic regression: Χρησιμοποιείται για δυαδική ταξινόμηση και προβλέπει την πιθανότητα ένα δείγμα να ανήκει σε μια κατηγορία βάσει γραμμικής συνάρτησης [34].

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούνται σε πολλές εφαρμογές, όπως η ανίχνευση απάτης, η ταξινόμηση ηλεκτρονικού ταχυδρομείου (spam filtering) και η ανάλυση συναισθημάτων σε κοινωνικά δίκτυα [15]. Μια από τις κύριες προκλήσεις αυτών των αλγορίθμων είναι ότι πρέπει να κλιμακώνονται αποδοτικά για να χειρίζονται μεγάλα σύνολα δεδομένων, και συχνά εφαρμόζονται σε κατανεμημένα περιβάλλοντα για να βελτιστοποιηθεί η απόδοση.

3.1.2 Αλγόριθμοι ομαδοποίησης (Clustering Algorithms)

Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για τη διαίρεση δεδομένων σε ομάδες (clusters) με βάση τις ομοιότητές τους. Η ομαδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες αλλά ανακαλύπτει τις ομάδες μέσα στα δεδομένα. Στόχος των αλγορίθμων ομαδοποίησης είναι να χωρίσουν τα δεδομένα σε ομάδες έτσι ώστε τα δεδομένα μέσα σε κάθε ομάδα να είναι όσο το δυνατόν πιο ομοιογενή, ενώ οι διαφορετικές ομάδες να είναι όσο το δυνατόν πιο διακριτές μεταξύ τους [35]. Στο πλαίσιο των μεγάλων δεδομένων, η ομαδοποίηση αποτελεί κρίσιμο εργαλείο για την εξαγωγή προτύπων και τη μείωση της πολυπλοκότητας των δεδομένων, προσφέροντας βαθύτερη κατανόηση των υποκείμενων δομών τους.

Οι βασικοί αλγόριθμοι ομαδοποίησης είναι:

- K-Means: Διαχωρίζει τα δεδομένα σε K ομάδες με βάση τις αποστάσεις τους από τα κεντροειδή (centroid) των ομάδων. Είναι γρήγορος και αποδοτικός, αλλά η απόδοσή του εξαρτάται από την αρχική επιλογή των κεντροειδών και τον αριθμό των ομάδων (K) [35].
- Hierarchical Clustering: Δημιουργεί μια ιεραρχία ομάδων με βάση τις ομοιότητες μεταξύ των δεδομένων, όπου οι μικρές ομάδες συγχωνεύονται σε μεγαλύτερες. Το αποτέλεσμα είναι ένα δενδροδιάγραμμα (dendrogram) που δείχνει τις σχέσεις μεταξύ των δεδομένων σε διαφορετικά επίπεδα ομαδοποίησης.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Χρησιμοποιείται για την ανίχνευση ομάδων με βάση την πυκνότητα των δεδομένων και είναι αποτελεσματικός για την αναγνώριση ανωμαλιών. Αναγνωρίζει ομάδες δεδομένων που βρίσκονται κοντά μεταξύ τους σε περιοχές υψηλής πυκνότητας

και αγνοεί περιοχές χαμηλής πυκνότητας, που μπορεί να θεωρηθούν ως θόρυβος (outliers) [36].

Η ομαδοποίηση είναι χρήσιμη σε εφαρμογές όπως η τμηματοποίηση πελατών, η ανάλυση κοινωνικών δικτύων, η αναγνώριση προτύπων και η βιολογική ανάλυση. Σε μεγάλα δεδομένα, η ομαδοποίηση μπορεί να είναι απαιτητική, ειδικά όταν τα δεδομένα είναι υψηλής διάστασης ή περιέχουν θόρυβο τους [35].

3.1.3 Support Vector Machines (SVM)

Οι Support Vector Machines (SVM) είναι ένας από τους πιο δημοφιλείς αλγόριθμους ταξινόμησης και έχουν αποδειχθεί εξαιρετικά αποτελεσματικοί στην ανάλυση μεγάλων δεδομένων. Ο SVM είναι ένας εποπτευόμενος αλγόριθμος μάθησης που στοχεύει να βρει το καλύτερο υπερεπίπεδο (hyperplane) που διαχωρίζει τα δεδομένα σε διαφορετικές κατηγορίες με το μέγιστο δυνατό περιθώριο (margin) [37]. Αυτό σημαίνει ότι προσπαθεί να βρει το υπερεπίπεδο που χωρίζει τα δεδομένα με τέτοιο τρόπο ώστε τα δεδομένα των δύο κατηγοριών να είναι όσο το δυνατόν πιο απομακρυσμένα από το υπερεπίπεδο, δημιουργώντας τη μέγιστη απόσταση μεταξύ των σημείων και του επιπέδου.

Τα βασικά στοιχεία του SVM είναι:

- Δυαδική ταξινόμηση: Οι SVM χρησιμοποιούνται κυρίως για δυαδική ταξινόμηση, όπου τα δεδομένα διαχωρίζονται σε δύο κατηγορίες.
- Margin (Περιθώριο): Το περιθώριο είναι η απόσταση μεταξύ του υπερεπιπέδου και των πιο κοντινών σημείων δεδομένων από κάθε κατηγορία, που ονομάζονται support vectors. Ο αλγόριθμος προσπαθεί να μεγιστοποιήσει αυτό το περιθώριο [37].
- Support vectors: Αυτά είναι τα σημεία δεδομένων που βρίσκονται κοντά στο υπερεπίπεδο διαχωρισμού και είναι καθοριστικά για τη δημιουργία του μοντέλου SVM.
- Πυρήνας (Kernel): Για δεδομένα που δεν είναι γραμμικά διαχωρίσιμα, ο SVM χρησιμοποιεί μια τεχνική που ονομάζεται πυρηνική συνάρτηση (kernel function), η οποία μετασχηματίζει τα δεδομένα σε υψηλότερες διαστάσεις, όπου γίνεται ευκολότερο να βρεθεί το διαχωριστικό υπερεπίπεδο [38].

Τα κύρια πλεονεκτήματα του αλγορίθμου SVM είναι:

- Υψηλή ακρίβεια: Οι SVM είναι γνωστοί για την υψηλή τους ακρίβεια σε προβλήματα ταξινόμησης, ειδικά όταν τα δεδομένα είναι καλά διαχωρίσιμα.
- Ανθεκτικότητα σε υπερεκπαίδευση (Overfitting): Οι SVM προσπαθούν να μεγιστοποιήσουν το περιθώριο, κάτι που συνήθως οδηγεί σε καλύτερη γενίκευση και μείωση του κινδύνου υπερεκπαίδευσης.
- Διαχείριση υψηλών διαστάσεων: Οι SVM μπορούν να διαχειριστούν δεδομένα υψηλών διαστάσεων πολύ αποτελεσματικά, ειδικά με τη χρήση πυρηνικών συναρτήσεων.
- Ανθεκτικότητα στα outliers: Εφόσον ο SVM προσπαθεί να μεγιστοποιήσει το περιθώριο μεταξύ των κατηγοριών, μπορεί να είναι πιο ανθεκτικός σε ορισμένα είδη θορύβου και ανωμαλιών [39].

Ο SVM χρησιμοποιείται ευρέως στην ανάλυση κειμένων, στην αναγνώριση εικόνων, καθώς και σε άλλες εφαρμογές όπου η ταξινόμηση είναι απαραίτητη. Σε μεγάλα δεδομένα, ο αλγόριθμος αυτός μπορεί να αντιμετωπίσει προβλήματα κλιμάκωσης, καθώς απαιτεί υψηλή υπολογιστική ισχύ για την εκπαίδευση του μοντέλου, ιδιαίτερα όταν τα δεδομένα περιέχουν πολλές διαστάσεις [37].

3.1.4 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης εμπνευσμένη από τη λειτουργία του ανθρώπινου εγκεφάλου. Χρησιμοποιούνται ευρέως σε προβλήματα εποπτευόμενης, μη εποπτευόμενης και ενισχυτικής μάθησης, και είναι ιδιαίτερα ισχυρά σε εφαρμογές που σχετίζονται με μεγάλα δεδομένα, λόγω της ικανότητάς τους να μαθαίνουν σύνθετα πρότυπα και να διαχειρίζονται μεγάλα και πολύπλοκα σύνολα δεδομένων [40].

Ένα νευρωνικό δίκτυο αποτελείται από πολλά στρώματα νευρώνων (τεχνητές μονάδες που προσομοιώνουν βιολογικούς νευρώνες), που είναι οργανωμένα σε:

- Επίπεδο εισόδου (Input Layer): Λαμβάνει τα χαρακτηριστικά του συνόλου δεδομένων ως εισόδους.
- Κρυφά επίπεδα (Hidden Layers): Αποτελούνται από νευρώνες που επεξεργάζονται τις πληροφορίες από τα επίπεδα εισόδου μέσω συναρτήσεων ενεργοποίησης (activation functions) και μεταδίδουν το αποτέλεσμα στα επόμενα επίπεδα.
- Επίπεδο εξόδου (Output Layer): Παράγει τις τελικές προβλέψεις ή ταξινομήσεις του μοντέλου [41].

Η εκπαίδευση ενός νευρωνικού δικτύου βασίζεται σε μια διαδικασία εμπρόσθιας διάδοσης (forward propagation) και οπισθοδιάδοσης (backpropagation), όπου τα σφάλματα που προκύπτουν από την απόκλιση της πρόβλεψης από την πραγματική τιμή μεταδίδονται προς τα πίσω στο δίκτυο για να προσαρμοστούν τα βάρη των νευρώνων και να βελτιωθεί η ακρίβεια του μοντέλου [40].

Οι τύποι νευρωνικών δικτύων είναι:

1. Πολυεπίπεδα νευρωνικά δίκτυα (Multilayer Perceptrons - MLP):

Είναι τα πιο απλά νευρωνικά δίκτυα, όπου κάθε νευρώνας συνδέεται με κάθε νευρώνα των επόμενων επιπέδων. Χρησιμοποιούνται σε προβλήματα εποπτευόμενης μάθησης, όπως η ταξινόμηση και η παλινδρόμηση.

2. Συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN):

Χρησιμοποιούνται κυρίως για προβλήματα επεξεργασίας εικόνας και βίντεο. Αποτελούνται από επίπεδα συνελίξεων που εξάγουν χαρακτηριστικά από εικόνες μέσω φίλτρων, και επίπεδα pooling που μειώνουν τη διάσταση των χαρακτηριστικών [42].

3. Αναδραστικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNN):

Είναι σχεδιασμένα για προβλήματα με δεδομένα ακολουθιακής φύσης, όπως επεξεργασία φυσικής γλώσσας (NLP) και ανάλυση χρονοσειρών. Διαθέτουν βρόχους που επιτρέπουν την ανάκληση προηγούμενων καταστάσεων του δικτύου, δίνοντάς τους τη δυνατότητα να διαχειρίζονται χρονικά εξαρτημένες πληροφορίες [41].

4. Βαθιά νευρωνικά δίκτυα (Deep Neural Networks - DNN):

Πρόκειται για νευρωνικά δίκτυα με πολλά κρυφά επίπεδα, που τους επιτρέπουν να μαθαίνουν πιο σύνθετα πρότυπα από τα δεδομένα. Χρησιμοποιούνται σε περίπλοκα προβλήματα όπως αναγνώριση φωνής, αυτόνομα οχήματα και προτάσεις προϊόντων [42].

Τα κύρια πλεονεκτήματα των νευρωνικών δικτύων είναι:

- Ικανότητα μάθησης σύνθετων προτύπων: Τα νευρωνικά δίκτυα μπορούν να μάθουν σύνθετα και μη γραμμικά πρότυπα, κάτι που τα καθιστά εξαιρετικά αποτελεσματικά σε εφαρμογές όπως η αναγνώριση εικόνων και η επεξεργασία φυσικής γλώσσας.

- Αυτοματοποίηση εξαγωγής χαρακτηριστικών: Σε αντίθεση με πολλούς παραδοσιακούς αλγορίθμους, τα νευρωνικά δίκτυα μπορούν να αυτοματοποιήσουν τη διαδικασία εξαγωγής χαρακτηριστικών, αποφεύγοντας την ανάγκη για χειροκίνητη μηχανική χαρακτηριστικών (feature engineering) [42].
- Κλιμάκωση με μεγάλα δεδομένα: Με την κατάλληλη υποδομή, τα νευρωνικά δίκτυα μπορούν να κλιμακωθούν για να διαχειριστούν τεράστια σύνολα δεδομένων και να βελτιώσουν την απόδοσή τους όσο αυξάνονται τα δεδομένα.
- Διαχείριση πολυδιάστατων δεδομένων: Τα νευρωνικά δίκτυα μπορούν να διαχειριστούν δεδομένα με πολλές διαστάσεις, όπως εικόνες, βίντεο, ήχους και άλλους τύπους πολυμέσων [40].

3.2 Αλγόριθμοι εξόρυξης δεδομένων

Οι αλγόριθμοι εξόρυξης δεδομένων (data mining) αποτελούν μια από τις πιο ισχυρές μεθόδους για την ανάλυση μεγάλων δεδομένων, καθώς επιτρέπουν την ανίχνευση κρυμμένων προτύπων, τη δημιουργία μοντέλων και τη λήψη χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων. Αυτοί οι αλγόριθμοι χρησιμοποιούνται σε διάφορους τομείς, όπως οι επιχειρήσεις, η υγειονομική περίθαλψη, τα κοινωνικά δίκτυα και οι επιστήμες, για να βοηθήσουν στην κατανόηση πολύπλοκων δεδομένων και στη λήψη τεκμηριωμένων αποφάσεων [43].

Η εξόρυξη δεδομένων περιλαμβάνει μια σειρά τεχνικών και αλγορίθμων που επιτρέπουν στους οργανισμούς να ανακαλύπτουν πρότυπα, τάσεις και συνδέσεις που δεν είναι εύκολα ορατά μέσω παραδοσιακών μεθόδων ανάλυσης. Οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να εφαρμοστούν τόσο σε δομημένα όσο και σε αδόμητα δεδομένα και συχνά περιλαμβάνουν μεθόδους μηχανικής μάθησης, στατιστικής και αναγνώρισης προτύπων [43].

Οι κύριοι στόχοι των αλγορίθμων εξόρυξης δεδομένων είναι:

- Η ταξινόμηση δεδομένων σε διαφορετικές κατηγορίες.
- Η ομαδοποίηση δεδομένων με βάση ομοιότητες.
- Η συσχέτιση διαφορετικών στοιχείων δεδομένων μεταξύ τους.
- Η αναγνώριση ανωμαλιών ή ακραίων τιμών στα δεδομένα.
- Η δημιουργία προβλέψεων για μελλοντικές τάσεις ή συμπεριφορές με βάση τα υφιστάμενα δεδομένα [44].

Οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να κατηγοριοποιηθούν σε διάφορες κατηγορίες ανάλογα με τον τύπο ανάλυσης που εκτελούν:

1. Αλγόριθμοι ταξινόμησης
2. Αλγόριθμοι ομαδοποίησης
3. Αλγόριθμοι συσχέτισης
4. Αλγόριθμοι αναγνώρισης ανωμαλιών
5. Αλγόριθμοι αναγνώρισης προτύπων και πρόβλεψης [43].

Στη συνέχεια παρουσιάζονται, αρχικά συνοπτικά στον Πίνακα 2, και έπειτα πιο αναλυτικά, οι βασικοί αλγόριθμοι εξόρυξης δεδομένων.

Πίνακας 2. Βασικοί αλγόριθμοι εξόρυξης δεδομένων

Κατηγορία αλγορίθμου	Παραδείγματα	Κύρια χαρακτηριστικά	Πλεονεκτήματα	Προκλήσεις / Περιορισμοί	Ενδεικτικές εφαρμογές
Συσχέτισης (Association)	Apriori, FP-Growth	Εξαγωγή κανόνων συσχέτισης (support, confidence, lift), συχνή συν-εμφάνιση αντικειμένων	Κατάλληλο για market basket analysis, εξατομικευμένες συστάσεις	Υψηλό υπολογιστικό κόστος για μεγάλα δεδομένα (Apriori)	Ανάλυση καλαθιού αγορών, σύσταση προϊόντων
Αναγνώρισης ανωμαλιών (Anomaly detection)	Isolation Forest, Local Outlier Factor (LOF)	Εντοπισμός σπάνιων ή μη φυσιολογικών τιμών, βασισμένο σε απομόνωση ή τοπική πυκνότητα	Κλιμακώσιμο (Isolation Forest), εντοπίζει τοπικές ανωμαλίες (LOF)	Αστάθεια ή υπολογιστική βαρύτητα σε μεγάλα σύνολα (LOF)	Ανίχνευση απάτης, κυβερνοασφάλεια, ανίχνευση ελαττωμάτων
Αναγνώρισης προτύπων και πρόβλεψης (Pattern recognition and prediction)	Regression Models, Neural Networks, Time Series Models	Ανάλυση ιστορικών δεδομένων για επανάληψη προτύπων και μελλοντική πρόβλεψη	Προβλέπει τάσεις, εφαρμόζεται σε δυναμικά δεδομένα, υψηλή χρηστικότητα	Απαιτεί ποιότητα δεδομένων, μπορεί να εμφανίσει overfitting	Πρόβλεψη αγορών, διάγνωση ασθενειών, χρηματοοικονομική ανάλυση

3.2.1 Αλγόριθμοι συσχέτισης

Οι αλγόριθμοι συσχέτισης είναι μια σημαντική κατηγορία αλγορίθμων εξόρυξης δεδομένων που χρησιμοποιούνται για την εύρεση σχέσεων μεταξύ μεταβλητών ή αντικειμένων μέσα σε μεγάλα σύνολα δεδομένων. Αυτοί οι αλγόριθμοι επιδιώκουν να εντοπίσουν μοτίβα ή κανόνες συσχέτισης που αναδεικνύουν πώς συνδέονται διαφορετικά αντικείμενα ή χαρακτηριστικά μεταξύ τους, και χρησιμοποιούνται ευρέως σε εφαρμογές όπως

η ανάλυση αγοραστικών συμπεριφορών, οι συστάσεις προϊόντων και η βελτιστοποίηση διαδικασιών [45]. Ένα κλασικό παράδειγμα είναι η ανάλυση καλαθιού αγορών (market basket analysis), όπου αναλύονται τα προϊόντα που αγοράζουν μαζί οι πελάτες. Στόχος είναι να εντοπιστούν συνδυασμοί αντικειμένων που εμφανίζονται συχνά μαζί, ώστε να ανιχνευτούν κανόνες του τύπου "εάν ο πελάτης αγοράσει το προϊόν Α, τότε πιθανότατα θα αγοράσει και το προϊόν Β" [3].

Από τους δημοφιλέστερους αλγορίθμους συσχέτισης είναι:

- Αλγόριθμος Apriori: Είναι ένας από τους πιο διαδεδομένους αλγορίθμους για την εξαγωγή κανόνων συσχέτισης. Η βασική ιδέα πίσω από τον Apriori είναι να εντοπίσει υποσύνολα αντικειμένων που εμφανίζονται συχνά μαζί στα δεδομένα, ξεκινώντας από τα μεμονωμένα αντικείμενα και επεκτείνοντας τα σε μεγαλύτερους συνδυασμούς. Ο αλγόριθμος δημιουργεί σύνολα αντικειμένων (itemsets) που εμφανίζονται συχνά στα δεδομένα. Έπειτα, υπολογίζει πόσο συχνά κάθε συνδυασμός αντικειμένων εμφανίζεται μαζί στα δεδομένα (support) και εξάγει κανόνες συσχέτισης βάσει της συν-εμφάνισης αυτών των συνδυασμών, χρησιμοποιώντας μέτρα όπως η υποστήριξη (support), η εμπιστοσύνη (confidence) και η ανύψωση (lift) [45]. Ο αλγόριθμος Apriori είναι εύκολος στην κατανόηση και μπορεί να προσαρμοστεί σε διαφορετικές εφαρμογές εξόρυξης δεδομένων [46]. Ωστόσο, μπορεί να είναι αργός και υπολογιστικά δαπανηρός για μεγάλα σύνολα δεδομένων, καθώς πρέπει να υπολογίζει επαναληπτικά τους συνδυασμούς αντικειμένων [3].
- FP-Growth (Frequent Pattern Growth): Είναι ένας αλγόριθμος που επιλύει το πρόβλημα της συχνής εμφάνισης συνόλων αντικειμένων (frequent itemset mining) με τρόπο πιο αποδοτικό από τον Apriori. Ο FP-Growth χρησιμοποιεί μια δομή δεδομένων που ονομάζεται FP-tree (Frequent Pattern tree) για την αποθήκευση των συχνών προτύπων. Ο FP-Growth είναι ταχύτερος από τον Apriori επειδή αποφεύγει την επαναληπτική δημιουργία υποψηφίων συνδυασμών και αποθηκεύει πληροφορίες σε μια συμπιεσμένη μορφή μέσω του FP-tree. Όμως, η κατασκευή και αποθήκευση του FP-tree μπορεί να είναι απαιτητική σε μνήμη, ειδικά όταν τα δεδομένα είναι πολύ μεγάλα [47].

3.2.2 Αλγόριθμοι αναγνώρισης ανωμαλιών

Οι αλγόριθμοι αναγνώρισης ανωμαλιών είναι μια κατηγορία αλγορίθμων εξόρυξης δεδομένων που έχουν σχεδιαστεί για να εντοπίζουν μη φυσιολογικά ή ακραία σημεία δεδομένων που αποκλίνουν από το αναμενόμενο μοτίβο ή τη συμπεριφορά. Οι ανωμαλίες αυτές μπορεί να αντιπροσωπεύουν σφάλματα δεδομένων, σπάνια γεγονότα ή ακόμη και σημαντικές πληροφορίες, όπως η ανίχνευση απάτης, η ανίχνευση ελαττωμάτων σε βιομηχανικές διαδικασίες ή η ανίχνευση κυβερνοεπιθέσεων [48].

Υπάρχουν διάφορες προσεγγίσεις που χρησιμοποιούνται για την ανίχνευση ανωμαλιών:

- Μοντέλα μάθησης: Εποπτευόμενα ή μη εποπτευόμενα μοντέλα που εκπαιδεύονται για να διακρίνουν φυσιολογικά από μη φυσιολογικά δεδομένα.
- Μοντέλα πυκνότητας: Αλγόριθμοι που βασίζονται στην ανάλυση της πυκνότητας των δεδομένων και εντοπίζουν ανωμαλίες ως αραιές περιοχές.
- Αλγόριθμοι κανόνων και στατιστικές προσεγγίσεις: Αυτοί εντοπίζουν ανωμαλίες με βάση την απόκλιση από προκαθορισμένους κανόνες ή στατιστικά πρότυπα [49].

Από τους πιο δημοφιλείς αλγόριθμους αναγνώρισης ανωμαλιών είναι:

- Isolation Forest: Είναι ένας αλγόριθμος βασισμένος σε δέντρα απόφασης που απομονώνουν τα σημεία δεδομένων. Η βασική ιδέα πίσω από τον αλγόριθμο είναι ότι οι ανωμαλίες είναι σπάνιες και διαφορετικές από τα υπόλοιπα δεδομένα, οπότε μπορούν να απομονωθούν πιο γρήγορα από τα κανονικά δεδομένα. Ο αλγόριθμος χτίζει πολλά δέντρα απομόνωσης (isolation trees) και χρησιμοποιεί τον μέσο αριθμό διασπάσεων που απαιτούνται για την απομόνωση ενός σημείου για να το χαρακτηρίσει ως ανωμαλία. Το κύριο πλεονέκτημα αυτού του αλγόριθμου είναι ότι είναι γρήγορος και κλιμακώσιμος, κατάλληλος για μεγάλα σύνολα δεδομένων. Ωστόσο, εξαρτάται από την τυχαιότητα των διασπάσεων, κάτι που μπορεί να οδηγήσει σε αστάθεια όταν τα δεδομένα δεν είναι αρκετά διακριτά [50].
- Local Outlier Factor (LOF): Είναι ένας αλγόριθμος βασισμένος στην ανάλυση πυκνότητας που συγκρίνει την πυκνότητα ενός σημείου δεδομένων με την πυκνότητα των γειτόνων του. Εάν η πυκνότητα του σημείου είναι σημαντικά χαμηλότερη από αυτήν των γειτόνων του, τότε το σημείο θεωρείται ανωμαλία. Το κύριο πλεονέκτημά του είναι πως είναι κατάλληλος για δεδομένα όπου οι ανωμαλίες έχουν τοπικά

χαρακτηριστικά και όχι παγκόσμια. Όμως, είναι αργός για πολύ μεγάλα δεδομένα λόγω του κόστους υπολογισμού των γειτόνων και της τοπικής πυκνότητας [50].

3.2.3 Αλγόριθμοι αναγνώρισης προτύπων και πρόβλεψης

Η αναγνώριση προτύπων και η πρόβλεψη είναι κεντρικές για την εξόρυξη δεδομένων. Αυτοί οι αλγόριθμοι αξιοποιούν τα υπάρχοντα δεδομένα για να εντοπίσουν τάσεις, συσχετίσεις και σχέσεις, ώστε να μπορούν να κάνουν προβλέψεις για νέα δεδομένα. Συγκεκριμένα, βασίζονται στην ανάλυση ιστορικών δεδομένων για να αναγνωρίσουν μοτίβα που επαναλαμβάνονται και να μάθουν από αυτά. Μόλις αναγνωριστούν τα πρότυπα, οι αλγόριθμοι χρησιμοποιούν αυτή τη γνώση για να κάνουν προβλέψεις σχετικά με νέα ή μελλοντικά δεδομένα [51]. Οι αλγόριθμοι αναγνώρισης προτύπων και πρόβλεψης συνήθως εντάσσονται είτε στην εποπτευόμενη είτε στη μη εποπτευόμενη μάθηση, ανάλογα με το αν τα δεδομένα περιέχουν ετικέτες ή όχι. Χρησιμοποιούνται σε ποικίλες εφαρμογές, όπως η πρόβλεψη αγοραστικών συμπεριφορών, η ανάλυση χρηματοοικονομικών δεδομένων, η διάγνωση σε ιατρικά δεδομένα, και πολλά άλλα [51].

3.3 Αλγόριθμοι ανάλυσης κειμένου

Οι αλγόριθμοι ανάλυσης κειμένου και οι τεχνικές ανάλυσης φυσικής γλώσσας (Natural Language Processing - NLP) έχουν αποκτήσει κεντρικό ρόλο στην ανάλυση μεγάλων συλλογών κειμένων, ιδιαίτερα σε περιβάλλοντα με μεγάλα δεδομένα. Το NLP επιτρέπει την κατανόηση και επεξεργασία της ανθρώπινης γλώσσας από υπολογιστές, προσφέροντας λύσεις για την εξαγωγή πληροφοριών από κείμενα, την κατηγοριοποίηση, την ανάλυση συναισθημάτων και πολλά άλλα [52].

Το NLP συνδυάζει την υπολογιστική γλωσσολογία με τη μηχανική μάθηση και την τεχνητή νοημοσύνη για να αναλύσει και να επεξεργαστεί δεδομένα φυσικής γλώσσας. Οι τεχνικές NLP μπορούν να εφαρμοστούν σε διάφορα είδη κειμένων, από άρθρα ειδήσεων και μηνύματα ηλεκτρονικού ταχυδρομείου έως δεδομένα κοινωνικών δικτύων [53]. Οι κυριότερες από αυτές τις τεχνικές αναφέρονται παρακάτω (Πίνακας 3).

Πίνακας 3. Βασικοί αλγόριθμοι ανάλυσης κειμένου (NLP)

Τεχνική NLP / Ανάλυση	Κύριες μέθοδοι / Αλγόριθμοι	Περιγραφή	Εφαρμογές
Προεπεξεργασία κειμένου	Tokenization, Stop Words Removal, Stemming, Lemmatization	Καθαρισμός και μετατροπή κειμένου σε μορφή κατάλληλη για ανάλυση	Όλα τα έργα NLP (πρώτο βήμα), ανάλυση κοινωνικών μέσων
Ανάλυση Συναισθημάτων	SVM, Lexicons, Νευρωνικά Δίκτυα	Εντοπισμός θετικών, αρνητικών ή ουδέτερων συναισθημάτων σε κείμενα	Αξιολογήσεις προϊόντων, πολιτικά σχόλια, κοινωνικά δίκτυα
Ανάλυση θεμάτων	LDA, NMF	Αυτόματη ανάδυση θεματικών ενοτήτων σε μεγάλες συλλογές κειμένων	Αναφορές, άρθρα ειδήσεων, θεματική ανάλυση εγγράφων
Αναγνώριση οντοτήτων	Rule-Based Systems, CRF, Νευρωνικά Δίκτυα	Αναγνώριση ονομάτων, τοποθεσιών, ημερομηνιών και άλλων οντοτήτων	Ψηφιακές βιβλιοθήκες, Ειδησεογραφικές υπηρεσίες, νομικά έγγραφα
Ανάλυση χρονοσειρών κειμένου	Time Series NLP, Temporal Topic Models	Μελέτη της εξέλιξης θεμάτων και συναισθημάτων με την πάροδο του χρόνου	Blog posts, άρθρα, κοινωνικά δίκτυα, παρακολούθηση τάσεων

3.3.1 Προεπεξεργασία κειμένου

Η προεπεξεργασία κειμένου είναι το πρώτο βήμα σε κάθε έργο NLP και περιλαμβάνει τη μετατροπή του ακατέργαστου κειμένου σε μορφή κατάλληλη για ανάλυση. Οι κύριες τεχνικές προεπεξεργασίας περιλαμβάνουν:

- Αφαίρεση σημείων στίξης και αριθμών: Καθαρισμός του κειμένου από σημεία στίξης, αριθμούς και ειδικούς χαρακτήρες που δεν συμβάλλουν στην ανάλυση.
- Κανονικοποίηση: Μετατροπή του κειμένου σε μια κοινή μορφή, π.χ. μετατροπή όλων των γραμμάτων σε πεζά ή τη μείωση των λέξεων στις ρίζες τους (stemming/lemmatization) [54].
- Διακοπή λέξεων (Stop Words Removal): Αφαίρεση κοινών λέξεων όπως "και", "ο", "η", που δεν περιέχουν χρήσιμη πληροφορία για την ανάλυση.
- Tokenization: Η διάσπαση του κειμένου σε μικρότερα κομμάτια, που ονομάζονται tokens (λέξεις, φράσεις, σύμβολα ή ακόμα και ολόκληρες προτάσεις) [54].

Αυτές οι τεχνικές είναι απαραίτητες για τη μείωση του όγκου των δεδομένων και για τη βελτίωση της απόδοσης των αλγορίθμων ανάλυσης.

3.3.2 Ανάλυση συναισθημάτων

Η ανάλυση συναισθημάτων είναι μια κοινή εφαρμογή του NLP που χρησιμοποιείται για να αναλύσει τη συναισθηματική φόρτιση των κειμένων. Ο στόχος είναι να ταξινομηθούν τα κείμενα με βάση το συναισθηματικό τους περιεχόμενο, όπως θετικά, αρνητικά ή ουδέτερα συναισθήματα [55].

Οι τεχνικές που χρησιμοποιούνται στην ανάλυση συναισθημάτων περιλαμβάνουν:

- Λεξικά συναισθημάτων: Τα κείμενα αναλύονται βάσει προκαθορισμένων λεξικών συναισθηματικών λέξεων για να εκτιμηθεί η συναισθηματική φόρτιση.
- Μηχανική μάθηση: Χρησιμοποιούνται μοντέλα ταξινόμησης που έχουν εκπαιδευτεί σε δεδομένα με ετικέτες συναισθημάτων, όπως SVM, δέντρα απόφασης ή νευρωνικά δίκτυα [55].

Η ανάλυση συναισθημάτων εφαρμόζεται σε κριτικές προϊόντων, μέσα κοινωνικής δικτύωσης, πολιτικές αναλύσεις και άλλα.

3.3.3 Ανάλυση θεμάτων

Η ανάλυση θεμάτων χρησιμοποιείται για να εντοπιστούν οι κυρίαρχες θεματικές ενότητες μέσα σε μεγάλα σύνολα κειμένων χωρίς την ύπαρξη προκαθορισμένων κατηγοριών. Είναι μια τεχνική μη εποπτευόμενης μάθησης που επιτρέπει την κατανόηση της δομής και των θεμάτων που αναδύονται μέσα από τα δεδομένα [56].

Κύριοι αλγόριθμοι ανάλυσης θεμάτων είναι:

- Latent Dirichlet Allocation (LDA): Ένας από τους πιο διαδεδομένους αλγορίθμους, που διαχωρίζει το κείμενο σε ένα σύνολο θεμάτων με βάση τις συσχετίσεις των λέξεων [56].
- Non-Negative Matrix Factorization (NMF): Χρησιμοποιείται για να διαχωρίσει τις λέξεις σε θεματικές ενότητες μέσω ανάλυσης πινάκων, όπου τα δεδομένα είναι μη αρνητικά [57].

Η ανάλυση θεμάτων είναι ιδιαίτερα χρήσιμη για την ανάλυση μεγάλων συλλογών κειμένων, όπως άρθρα, αναφορές, κριτικές και αρχεία κοινωνικών δικτύων [56].

3.3.4 Αναγνώριση οντοτήτων

Η αναγνώριση οντοτήτων είναι μια τεχνική NLP που χρησιμοποιείται για να εντοπίσει και να κατηγοριοποιήσει οντότητες σε ένα κείμενο, όπως ονόματα, τοποθεσίες, ημερομηνίες και οργανισμούς. Η αναγνώριση οντοτήτων χρησιμοποιείται σε εφαρμογές όπως οι ψηφιακές βιβλιοθήκες, τα αρχεία ειδήσεων και οι βιομηχανίες που απαιτούν αναγνώριση ονομάτων [53].

Κύριες τεχνικές για την αναγνώριση οντοτήτων είναι:

- Κανόνας βασισμένος σε συστήματα (Rule-Based Systems): Χρήση συνόλου κανόνων και μοτίβων για την αναγνώριση συγκεκριμένων οντοτήτων σε ένα κείμενο.
- Μηχανική μάθηση: Εκπαίδευση μοντέλων που μπορούν να εντοπίσουν αυτόματα οντότητες, όπως τα νευρωνικά δίκτυα ή τα CRF (Conditional Random Fields) [53].

3.3.5 Ανάλυση χρονοσειρών κειμένου

Η ανάλυση χρονοσειρών κειμένου επικεντρώνεται στη μελέτη του τρόπου με τον οποίο τα μοτίβα και τα θέματα στα κείμενα αλλάζουν με την πάροδο του χρόνου. Αυτή η τεχνική είναι χρήσιμη για την κατανόηση της εξέλιξης των θεμάτων, των απόψεων ή των συναισθημάτων σε μια μεγάλη συλλογή δεδομένων, όπως οι δημοσιεύσεις σε ιστολόγια, τα νέα άρθρα ή οι αναρτήσεις στα κοινωνικά δίκτυα. Οι τεχνικές αυτές χρησιμοποιούνται για την ανάλυση τάσεων ή για την παρακολούθηση της αλλαγής στις αντιλήψεις και τα συναισθήματα σε βάθος χρόνου [58].

Η ανάλυση κειμένου με χρήση τεχνικών NLP εφαρμόζεται σε πολλές περιοχές, όπως:

- Κοινωνικά δίκτυα: Ανάλυση μηνυμάτων από πλατφόρμες κοινωνικών μέσων για να εντοπιστούν τάσεις, συναισθήματα και θέματα ενδιαφέροντος.
- Ψηφιακό μάρκετινγκ: Ανάλυση σχολίων πελατών και αξιολογήσεων για την κατανόηση των αντιλήψεων του κοινού και την προσαρμογή στρατηγικών μάρκετινγκ [52].
- Υγειονομική περίθαλψη: Ανάλυση ιατρικών κειμένων και αρχείων για την εξαγωγή χρήσιμων πληροφοριών σχετικά με την υγεία των ασθενών και την εξέλιξη ασθενειών.

- Νομική ανάλυση: Ανάλυση νομικών εγγράφων για να εξαχθούν πρότυπα ή να αναγνωριστούν σημαντικές πληροφορίες που βοηθούν στη νομική στρατηγική [53].

3.4 Αλγόριθμοι ανάλυσης δικτύων

Οι αλγόριθμοι ανάλυσης δικτύων χρησιμοποιούνται για την ανάλυση των σχέσεων και των συνδέσεων μεταξύ των οντοτήτων σε διάφορα είδη δικτύων, όπως τα κοινωνικά δίκτυα, τα δίκτυα επικοινωνιών, τα βιολογικά δίκτυα και τα δίκτυα μεταφορών. Αυτοί οι αλγόριθμοι αξιοποιούνται για να κατανοήσουν τις δομές των δικτύων, να εντοπίσουν σημαντικούς κόμβους, να μελετήσουν τις ροές πληροφορίας και να αναλύσουν τις συνδέσεις που σχηματίζουν τα δίκτυα. Η ανάλυση δικτύων έχει αποκτήσει ιδιαίτερη σημασία τα τελευταία χρόνια λόγω της ανάπτυξης των κοινωνικών μέσων και των τεχνολογιών που βασίζονται σε δίκτυα [59].

Ένα δίκτυο ανάλυσης δικτύων αποτελείται από:

- Κόμβους (Nodes): Τα στοιχεία ή οι οντότητες που συνδέονται μεταξύ τους. Σε ένα κοινωνικό δίκτυο, για παράδειγμα, οι κόμβοι αντιπροσωπεύουν άτομα ή οργανισμούς.
- Ακμές (Edges): Οι σύνδεσμοι ή οι σχέσεις που συνδέουν τους κόμβους. Στο πλαίσιο ενός κοινωνικού δικτύου, οι ακμές αντιπροσωπεύουν φιλίες, επαφές ή αλληλεπιδράσεις [60].

Τα δίκτυα μπορούν να είναι κατευθυνόμενα ή μη κατευθυνόμενα, δηλαδή οι συνδέσεις μεταξύ των κόμβων μπορεί να έχουν φορά ή να είναι συμμετρικές, ενώ μπορούν επίσης να είναι σταθμισμένα (weighted) ή μη σταθμισμένα [60].

Μια θεμελιώδης έννοια στην ανάλυση δικτύων είναι η κεντρικότητα και χρησιμοποιείται για να μετρήσει τη σημασία ή την επιρροή ενός κόμβου μέσα στο δίκτυο [61].

- Βαθμός κεντρικότητας (Degree Centrality): Υπολογίζει τον αριθμό των άμεσων συνδέσεων ενός κόμβου με άλλους κόμβους. Ο κόμβος με τον μεγαλύτερο βαθμό κεντρικότητας θεωρείται πιο σημαντικός επειδή έχει περισσότερες συνδέσεις.
- Κεντρικότητα ενδιαμεσότητας (Betweenness Centrality): Μετρά πόσες φορές ένας κόμβος βρίσκεται στη "μέση" του συντομότερου μονοπατιού μεταξύ

άλλων κόμβων. Αυτό δείχνει τη σημασία του κόμβου για τη ροή της πληροφορίας.

- Κεντρικότητα εγγύτητας (Closeness Centrality): Υπολογίζει την απόσταση ενός κόμβου από όλους τους άλλους κόμβους του δικτύου. Κόμβοι με υψηλή κεντρικότητα εγγύτητας θεωρούνται πιο κοντά στους υπόλοιπους κόμβους του δικτύου και μπορούν να διαδώσουν πληροφορίες γρήγορα.
- Κεντρικότητα διάνυσης (Eigenvector Centrality): Υπολογίζει τη σημασία ενός κόμβου όχι μόνο με βάση τις συνδέσεις του, αλλά και με βάση τη σημασία των κόμβων με τους οποίους συνδέεται. Χρησιμοποιείται σε εφαρμογές όπως ο αλγόριθμος PageRank της Google [61].

Η ανάλυση δικτύων χρησιμοποιεί διάφορους αλγόριθμους και τεχνικές για να αναλύσει τις δομές και τις συνδέσεις μέσα σε ένα δίκτυο. Παρακάτω παρατίθενται μερικοί από τους πιο σημαντικούς αλγόριθμους ανάλυσης δικτύων (Πίνακας 4).

Πίνακας 4. Βασικοί αλγόριθμοι ανάλυσης δικτύων

Κατηγορία αλγόριθμου	Ενδεικτικοί αλγόριθμοι	Περιγραφή	Εφαρμογές
Ανάλυση κεντρικότητας	Degree, Betweenness, Closeness, Eigenvector	Μέτρηση σημασίας ενός κόμβου σε δίκτυο βάσει συνδέσεων ή επιρροής	Κοινωνικά & βιολογικά δίκτυα, επιρροή χρηστών, διάδοση πληροφορίας
Εντοπισμός κοινοτήτων	Girvan-Newman, Louvain, Label Propagation	Διαχωρισμός κόμβων σε κοινότητες με υψηλή εσωτερική συνδεσιμότητα	Ανάλυση κοινοτήτων σε κοινωνικά δίκτυα, ανίχνευση ομάδων
Συντομότερες διαδρομές	Dijkstra, Bellman-Ford, Floyd-Warshall	Εύρεση της ταχύτερης ή πιο αποδοτικής διαδρομής μεταξύ κόμβων	Βελτιστοποίηση διαδρομών, μελέτη κυκλοφορίας, ροή πληροφορίας
Κατάταξη κόμβων (PageRank)	PageRank	Κατάταξη κόμβων βάσει της συνδεσιμότητάς τους με σημαντικούς κόμβους	Μηχανές αναζήτησης, web ranking, ανάλυση επιρροής
Αλγόριθμοι ροής	Edmonds-Karp, Dinic	Ανάλυση της μέγιστης ροής μέσα σε ένα δίκτυο	Δίκτυα μεταφορών, παροχής υπηρεσιών, logistics

3.4.1 Αλγόριθμοι εντοπισμού κοινοτήτων

Ο εντοπισμός κοινοτήτων στοχεύει στον διαχωρισμό ενός δικτύου σε υποδίκτυα ή κοινότητες, όπου οι κόμβοι εντός κάθε κοινότητας είναι πιο στενά συνδεδεμένοι μεταξύ τους από ό,τι με τους κόμβους άλλων κοινοτήτων [62]. Τέτοιου τύπου αλγόριθμοι είναι:

- Αλγόριθμος Girvan-Newman: Βασίζεται στην αφαίρεση ακμών με την υψηλότερη ενδιαμεσότητα (betweenness) και χρησιμοποιείται για τη διάσπαση του δικτύου σε κοινότητες.
- Αλγόριθμος Louvain: Χρησιμοποιείται για τον εντοπισμό κοινοτήτων μεγιστοποιώντας τη διαμόρφωση (modularity) του δικτύου. Είναι ένας από τους πιο δημοφιλείς αλγόριθμους για την ανάλυση κοινοτήτων σε μεγάλα δίκτυα λόγω της απόδοσής του [62].
- Αλγόριθμος Label Propagation: Ένας επαναληπτικός αλγόριθμος όπου οι κόμβοι παίρνουν την "ταυτότητα" των γειτόνων τους, οδηγώντας τελικά στον σχηματισμό κοινοτήτων [60].

3.4.2 Αλγόριθμοι εύρεσης συντομότερων διαδρομών

Η εύρεση της συντομότερης διαδρομής μεταξύ δύο κόμβων σε ένα δίκτυο είναι μια κοινή ανάλυση που χρησιμοποιείται για τη μελέτη της ροής πληροφορίας, της ταχύτητας διάδοσης και άλλων φαινομένων σε δίκτυα.

- Αλγόριθμος Dijkstra: Χρησιμοποιείται για να βρει τη συντομότερη διαδρομή από έναν κόμβο προς όλους τους άλλους κόμβους σε ένα σταθμισμένο δίκτυο.
- Αλγόριθμος Bellman-Ford: Μπορεί να χειριστεί δίκτυα με αρνητικά βάρη στις ακμές και χρησιμοποιείται επίσης για την εύρεση των συντομότερων διαδρομών [63].
- Αλγόριθμος Floyd-Warshall: Υπολογίζει τη συντομότερη διαδρομή μεταξύ όλων των ζευγών κόμβων σε ένα δίκτυο [64].

3.4.3 Αλγόριθμος PageRank

Ο PageRank είναι ένας αλγόριθμος που αναπτύχθηκε από την Google για την κατάταξη ιστοσελίδων στο διαδίκτυο. Μετρά τη σημασία μιας σελίδας βάσει του αριθμού και της ποιότητας των συνδέσεων που δείχνουν προς αυτήν. Ο PageRank είναι ουσιαστικά μια μορφή κεντρικότητας διάνυσης, όπου οι κόμβοι που συνδέονται με σημαντικούς άλλους κόμβους λαμβάνουν μεγαλύτερη βαθμολογία [65].

3.4.4 Αλγόριθμοι ροής

Οι αλγόριθμοι ροής εξετάζουν τη ροή της πληροφορίας, των αγαθών ή των δεδομένων σε ένα δίκτυο και χρησιμοποιούνται σε εφαρμογές όπως τα δίκτυα μεταφορών και τα δίκτυα παροχής υπηρεσιών.

- Αλγόριθμος Edmonds-Karp: Χρησιμοποιείται για να βρει τη μέγιστη ροή σε ένα δίκτυο, βελτιώνοντας την κλασική μέθοδο Ford-Fulkerson.
- Αλγόριθμος Dinic: Βελτιστοποιημένος αλγόριθμος για την εύρεση της μέγιστης ροής σε ένα δίκτυο με πολυωνυμικό χρόνο [66].

Συνολικά, οι αλγόριθμοι ανάλυσης δικτύων προσφέρουν μια ισχυρή εργαλειοθήκη για την κατανόηση των πολύπλοκων δομών και αλληλεπιδράσεων σε διάφορα είδη δικτύων. Η ανάλυση δικτύων εφαρμόζεται σε πολλούς διαφορετικούς τομείς, όπως τα κοινωνικά δίκτυα (ανάλυση σχέσεων και επιρροών μεταξύ ατόμων ή ομάδων, μελέτη της διάδοσης πληροφοριών ή φημών κ.α.), τα βιολογικά δίκτυα (μελέτη αλληλεπιδράσεων μεταξύ γονιδίων, πρωτεϊνών ή άλλων βιολογικών στοιχείων για την κατανόηση των μοριακών μηχανισμών σε ζωντανούς οργανισμούς), τα δίκτυα μεταφορών (βελτίωση της διαχείρισης ροών κυκλοφορίας, σχεδιασμός βέλτιστων διαδρομών και ανάλυση συμφόρησης σε δίκτυα μεταφορών) [59].

3.5 Προκλήσεις αλγορίθμων

3.5.1 Προκλήσεις αλγορίθμων μηχανικής μάθησης για μεγάλα δεδομένα

Η χρήση αλγορίθμων μηχανικής μάθησης για την επεξεργασία και ανάλυση μεγάλων δεδομένων συνοδεύεται από μια σειρά προκλήσεων. Οι κυριότερες από αυτές είναι:

- Όγκος και ποικιλία δεδομένων: Ο τεράστιος όγκος δεδομένων και η ποικιλία τους (δομημένα, ημιδομημένα και αδόμητα δεδομένα) κάνουν δύσκολη την αποθήκευση, επεξεργασία και ανάλυση. Οι αλγόριθμοι πρέπει να μπορούν να κλιμακώνονται και να χειρίζονται μεγάλα και ανομοιογενή σύνολα δεδομένων.
- Υπολογιστική πολυπλοκότητα: Οι αλγόριθμοι μηχανικής μάθησης, ειδικά τα βαθιά νευρωσικά δίκτυα (Deep Learning), απαιτούν τεράστιους υπολογιστικούς πόρους και χρόνο για την εκπαίδευση, λόγω των επαναληπτικών διαδικασιών και της ανάγκης για μεγάλες ποσότητες δεδομένων [68].

- Υπερεκπαίδευση (Overfitting): Σε μεγάλους όγκους δεδομένων, υπάρχει κίνδυνος οι αλγόριθμοι να απομνημονεύουν τα δεδομένα εκπαίδευσης αντί να γενικεύουν τα μοτίβα τους, κάτι που οδηγεί σε χαμηλή απόδοση σε νέα δεδομένα [67].

3.5.2 Προκλήσεις στην εξόρυξη δεδομένων

Η εφαρμογή των αλγορίθμων εξόρυξης δεδομένων σε μεγάλα δεδομένα έρχεται αντιμέτωπη με αρκετές προκλήσεις:

- Αντιμετώπιση πολυδιάστατων δεδομένων: Η εξόρυξη δεδομένων από πολύ μεγάλα και πολυδιάστατα σύνολα δεδομένων αυξάνει την πολυπλοκότητα των υπολογισμών και τη δυσκολία στην εξαγωγή χρήσιμων μοτίβων.
- Ανάλυση σε πραγματικό χρόνο: Η εξόρυξη δεδομένων σε πραγματικό χρόνο είναι δύσκολη, καθώς οι αλγόριθμοι πρέπει να είναι αρκετά γρήγοροι για να αναλύουν δεδομένα που ρέουν συνεχώς, π.χ. σε χρηματοοικονομικές συναλλαγές ή αισθητήρες IoT [51].
- Καθαρότητα και ποιότητα δεδομένων: Τα δεδομένα μπορεί να περιέχουν θόρυβο, ελλιπή ή ανακριβή δεδομένα, γεγονός που μπορεί να υποβαθμίσει την απόδοση των αλγορίθμων εξόρυξης δεδομένων [68].

3.5.3 Προκλήσεις στην ανάλυση κειμένου

Οι κύριες προκλήσεις που σχετίζονται με τους αλγόριθμους ανάλυσης κειμένου είναι:

- Πολυπλοκότητα της φυσικής γλώσσας: Η ανθρώπινη γλώσσα είναι εξαιρετικά περίπλοκη και πολυδιάστατη, με πολλές ασάφειες, συμφραζόμενα και εκφράσεις που μπορεί να είναι δύσκολο να αναγνωριστούν από τους αλγορίθμους ανάλυσης κειμένου. Οι γλωσσικές ιδιαιτερότητες, όπως η συνωνυμία, η πολυσημία και τα ιδιωματικά εκφραστικά μέσα, προκαλούν επιπλοκές στην ακριβή κατανόηση.
- Υπολογιστικό κόστος: Η επεξεργασία μεγάλων ποσοτήτων αδόμητων δεδομένων κειμένου είναι απαιτητική σε υπολογιστικούς πόρους, ειδικά κατά την εκπαίδευση πολύπλοκων μοντέλων NLP [53].
- Ερμηνευσιμότητα και εμπιστοσύνη: Τα μοντέλα ανάλυσης κειμένου συχνά λειτουργούν ως "μαύρα κουτιά" και μπορεί να είναι δύσκολο να ερμηνευτούν οι προβλέψεις τους. Αυτό καθιστά δύσκολη την αξιολόγηση της αξιοπιστίας τους, ειδικά σε κρίσιμους τομείς, όπως η ιατρική ή η νομική [69].

3.5.4 Προκλήσεις στην ανάλυση δικτύων

Οι αλγόριθμοι ανάλυσης δικτύων κατά την εφαρμογή τους στα μεγάλα δεδομένα έχουν να αντιμετωπίσουν τις εξής προκλήσεις:

- Κλίμακα και πολυπλοκότητα δικτύων: Τα δίκτυα σε μεγάλες κλίμακες, όπως τα κοινωνικά δίκτυα ή τα δίκτυα υπολογιστών, περιέχουν τεράστιους αριθμούς κόμβων και ακμών, καθιστώντας δύσκολη την ανάλυση και την εξαγωγή συμπερασμάτων σε πραγματικό χρόνο.
- Αναγνώριση κοινοτήτων και ιεραρχιών: Η ανίχνευση κοινοτήτων ή η ιεραρχική ανάλυση σε μεγάλα δίκτυα μπορεί να είναι υπολογιστικά έντονη, ειδικά όταν τα δίκτυα είναι πολύπλοκα και δυναμικά (δηλαδή, αλλάζουν συνεχώς με την πάροδο του χρόνου) [62].
- Ανάλυση σε δυναμικά δίκτυα: Σε δυναμικά δίκτυα, όπου οι συνδέσεις μεταξύ κόμβων αλλάζουν με την πάροδο του χρόνου, η στατική ανάλυση δεν επαρκεί. Οι αλγόριθμοι πρέπει να μπορούν να παρακολουθούν τις αλλαγές σε πραγματικό χρόνο και να προσαρμόζονται στις νέες συνθήκες [60].

Οι προκλήσεις που αντιμετωπίζουν οι αλγόριθμοι στην επεξεργασία και ανάλυση μεγάλων δεδομένων καθιστούν επιτακτική την ανάγκη για συνεχή βελτίωση και βελτιστοποίηση αυτών των αλγορίθμων, ώστε να μπορούν να ανταποκριθούν αποτελεσματικά στις απαιτήσεις των σύγχρονων εφαρμογών και να αξιοποιήσουν πλήρως τις δυνατότητες των μεγάλων δεδομένων [69].

4 Τεχνικές και Μέθοδοι Βελτιστοποίησης Αλγορίθμων

4.1 Ανάγκη για βελτιστοποίηση αλγορίθμων

Η ανάγκη για βελτιστοποίηση αλγορίθμων γίνεται όλο και πιο επιτακτική καθώς τα δεδομένα αυξάνονται ραγδαία και οι απαιτήσεις για γρήγορη και αποτελεσματική ανάλυση αυξάνονται αντίστοιχα. Οι αλγόριθμοι που χρησιμοποιούνται για την επεξεργασία μεγάλων δεδομένων και την εξαγωγή χρήσιμων πληροφοριών πρέπει να είναι αποδοτικοί τόσο σε υπολογιστικό επίπεδο όσο και σε επίπεδο μνήμης, καθώς οποιαδήποτε αναποτελεσματικότητα μπορεί να οδηγήσει σε τεράστια καθυστέρηση ή ακόμα και σε αδυναμία ολοκλήρωσης της ανάλυσης [68]. Αυτή η ανάγκη για βελτιστοποίηση προκύπτει από διάφορους παράγοντες και έχει συγκεκριμένους στόχους.

Αναλυτικότερα, οι παράγοντες που επηρεάζουν την ανάγκη για βελτιστοποίηση των αλγορίθμων είναι:

- Αύξηση όγκου και ταχύτητας δεδομένων:

  Καθώς τα δεδομένα αυξάνονται εκθετικά σε όγκο και ταχύτητα παραγωγής (π.χ. μέσω εφαρμογών IoT, κοινωνικών δικτύων και άλλων πηγών), οι υπάρχοντες αλγόριθμοι πρέπει να προσαρμόζονται και να βελτιστοποιούνται για να μπορούν να τα επεξεργάζονται αποτελεσματικά σε πραγματικό χρόνο ή σχεδόν πραγματικό χρόνο [70].

- Ανάγκη για αναλυτικά αποτελέσματα σε πραγματικό χρόνο:

  Πολλές σύγχρονες εφαρμογές απαιτούν ανάλυση σε πραγματικό χρόνο, όπως στη χρηματοοικονομική ανάλυση, στις διαδικασίες παραγωγής και στη διαχείριση κυκλοφορίας. Οι αλγόριθμοι πρέπει να βελτιστοποιηθούν ώστε να παρέχουν ακριβή αποτελέσματα χωρίς καθυστερήσεις [1].

- Ενεργειακές απαιτήσεις:

  Οι αλγόριθμοι που εφαρμόζονται σε μεγάλα δεδομένα συχνά εκτελούνται σε μεγάλα κέντρα δεδομένων με υψηλές απαιτήσεις ενέργειας. Η βελτιστοποίηση των αλγορίθμων για μείωση της κατανάλωσης ενέργειας έχει οικονομικά και

περιβαλλοντικά οφέλη και αποτελεί σημαντική πρόκληση για την αποδοτικότητα των σύγχρονων υπολογιστικών συστημάτων.

- Κατανεμημένα συστήματα και Cloud Computing:

  Η στροφή προς τα κατανεμημένα συστήματα και το cloud computing απαιτεί αλγορίθμους που είναι ειδικά σχεδιασμένοι να εκτελούνται σε πολλούς κόμβους και να αξιοποιούν τους πόρους του δικτύου με βέλτιστο τρόπο. Η βελτιστοποίηση των αλγορίθμων για αυτά τα περιβάλλοντα είναι κρίσιμη για την επίτευξη υψηλής απόδοσης και αξιοπιστίας [70].

Οι στόχοι της βελτιστοποίησης των αλγορίθμων είναι:

- Αύξηση της αποδοτικότητας:

  Οι βελτιστοποιημένοι αλγόριθμοι πρέπει να εκτελούνται ταχύτερα και να χρησιμοποιούν λιγότερους υπολογιστικούς πόρους. Σε μεγάλα δεδομένα, αυτό σημαίνει ότι οι αλγόριθμοι πρέπει να είναι ικανοί να κλιμακώνονται και να επεξεργάζονται δεδομένα με μεγάλη ταχύτητα, ακόμα και όταν αυτά βρίσκονται σε κατανεμημένα περιβάλλοντα. Η αποδοτικότητα μπορεί να βελτιωθεί μέσω του παραλληλισμού, της βελτιστοποίησης της μνήμης και της αποφυγής περιττών υπολογισμών [67].

  - Μείωση πολυπλοκότητας:

  Οι αλγόριθμοι που λειτουργούν με μεγάλο όγκο δεδομένων συχνά έχουν υψηλή πολυπλοκότητα, γεγονός που καθιστά τη λειτουργία τους αργή και απαιτητική. Η βελτιστοποίηση των αλγορίθμων συχνά επικεντρώνεται στη μείωση της υπολογιστικής πολυπλοκότητας, χρησιμοποιώντας πιο αποδοτικούς αλγόριθμους ή μαθηματικά μοντέλα που επιταχύνουν την επίλυση προβλημάτων [13].

  - Εξοικονόμηση πόρων:

  Η βελτιστοποίηση δεν αφορά μόνο την ταχύτητα, αλλά και την κατανάλωση πόρων, όπως η μνήμη και η ενέργεια. Οι αλγόριθμοι πρέπει να χρησιμοποιούν λιγότερη μνήμη, ειδικά όταν επεξεργάζονται δεδομένα υψηλής διάστασης ή μεγάλης κλίμακας, και να λειτουργούν αποδοτικά σε περιβάλλοντα με περιορισμένους πόρους [69].

  - Ανθεκτικότητα και κλιμάκωση:

Οι αλγόριθμοι πρέπει να είναι ανθεκτικοί και να μπορούν να κλιμακώνονται αποδοτικά καθώς τα δεδομένα αυξάνονται. Αυτό σημαίνει ότι οι βελτιστοποιημένοι αλγόριθμοι πρέπει να είναι σε θέση να διαχειρίζονται μεγάλες αλλαγές στον όγκο των δεδομένων χωρίς να υποβαθμίζεται η απόδοσή τους [13].

- Προσαρμοστικότητα σε ποικίλα δεδομένα:

Οι βελτιστοποιημένοι αλγόριθμοι πρέπει να μπορούν να προσαρμόζονται σε δεδομένα που ποικίλουν σε μορφή, διάσταση και ποιότητα. Οι τεχνικές βελτιστοποίησης στοχεύουν να αυξήσουν την ευελιξία και την ευρωστία των αλγορίθμων, ώστε να αποδίδουν καλά σε διαφορετικά είδη δεδομένων, όπως δομημένα, ημιδομημένα και αδόμητα δεδομένα [1].

4.2 Τεχνικές βελτιστοποίησης

Οι τεχνικές βελτιστοποίησης είναι κρίσιμες για την αποδοτική επεξεργασία και ανάλυση μεγάλων δεδομένων. Αυτές οι τεχνικές στοχεύουν στην αύξηση της απόδοσης, τη μείωση των υπολογιστικών απαιτήσεων και την καλύτερη χρήση των διαθέσιμων πόρων [71].

Στη συνέχεια, αναλύονται μερικές από τις πιο σημαντικές τεχνικές βελτιστοποίησης, όπως ο παραλληλισμός, η κατανομή φόρτου, η μείωση της πολυπλοκότητας, καθώς και η χρήση κατανεμημένων υπολογιστικών μοντέλων, όπως παρουσιάζονται συνοπτικά στον ακόλουθο πίνακα.

Πίνακας 5. Σύγκριση τεχνικών βελτιστοποίησης για μεγάλα δεδομένα

Τεχνική	Πλεονεκτήματα	Μειονεκτήματα	Παραδείγματα εργαλείων/εφαρμογών
Παραλληλισμός	Ταχύτητα, κατανομή πόρων	Δυσκολία συγχρονισμού, bottlenecks	Apache Spark, CUDA, MapReduce
Κατανομή φορτίου	Βέλτιστη αξιοποίηση κόμβων	Πολυπλοκότητα υλοποίησης	YARN, Kubernetes, Spark scheduler
Μείωση πολυπλοκότητας	Ταχύτερη ανάλυση, εξοικονόμηση πόρων	Πιθανή απώλεια ακρίβειας	PCA, Approx. K-means, Sampling
Κατανεμημένα μοντέλα	Κλιμακωσιμότητα, ανθεκτικότητα	Απαιτεί υποδομές cloud / δικτύου	Hadoop, Flink, Hive

4.2.1 Παραλληλισμός

Ο παραλληλισμός αποτελεί μία από τις πιο βασικές και αποτελεσματικές τεχνικές βελτιστοποίησης αλγορίθμων για την επεξεργασία μεγάλων δεδομένων. Η ανάγκη για παραλληλισμό προκύπτει από τον τεράστιο όγκο δεδομένων που απαιτείται να επεξεργαστούν οι σύγχρονοι αλγόριθμοι, καθώς και από τις αυξημένες απαιτήσεις σε υπολογιστικούς πόρους που σχετίζονται με αυτές τις διεργασίες. Ο παραλληλισμός επιτρέπει τη διάσπαση μιας πολύπλοκης διεργασίας σε μικρότερα, ανεξάρτητα τμήματα, τα οποία μπορούν να εκτελεστούν ταυτόχρονα σε πολλαπλούς επεξεργαστές ή υπολογιστικούς κόμβους, με αποτέλεσμα τη σημαντική επιτάχυνση της συνολικής επεξεργασίας [71].

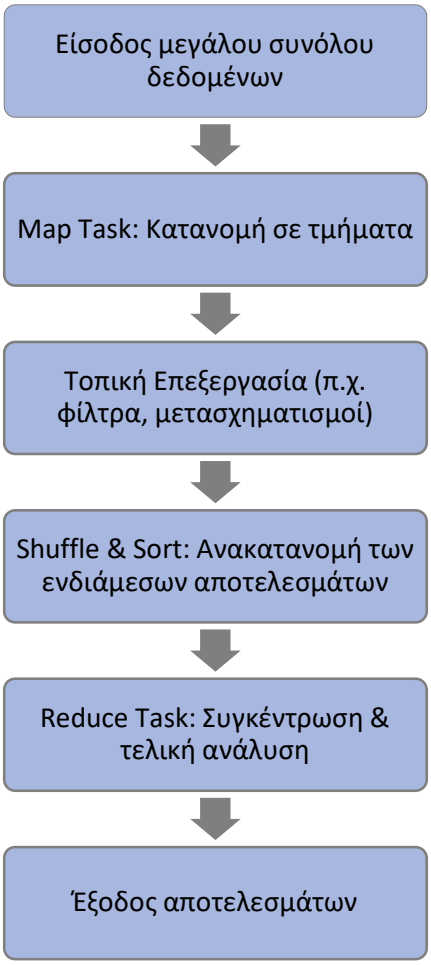
Ειδικότερα, ο παραλληλισμός εκμεταλλεύεται την ικανότητα ενός υπολογιστικού συστήματος να εκτελεί πολλές εργασίες ταυτόχρονα, είτε σε πολλούς επεξεργαστικούς πυρήνες ενός υπολογιστή, είτε σε ένα δίκτυο διανεμημένων κόμβων. Σε ένα τυπικό σενάριο παραλληλισμού για μεγάλα δεδομένα, τα δεδομένα χωρίζονται σε μικρότερα υποσύνολα, και κάθε υποσύνολο επεξεργάζεται ανεξάρτητα. Οι αλγόριθμοι παραλληλισμού χρησιμοποιούνται για τη διαχείριση αυτών των διεργασιών, εξασφαλίζοντας ότι τα αποτελέσματα από κάθε κόμβο ή επεξεργαστή συγχωνεύονται με σωστό τρόπο για να παραχθεί το τελικό αποτέλεσμα [72].

Υπάρχουν δύο κύρια μοντέλα παραλληλισμού που χρησιμοποιούνται στην επεξεργασία μεγάλων δεδομένων:

1. Παραλληλισμός δεδομένων (Data Parallelism): Στο μοντέλο αυτό, τα δεδομένα διαχωρίζονται σε ανεξάρτητα τμήματα και επεξεργάζονται ταυτόχρονα από διαφορετικούς επεξεργαστές. Για παράδειγμα, σε ένα σύστημα με πολλές μηχανές, κάθε μηχανή μπορεί να αναλάβει την επεξεργασία ενός μέρους των δεδομένων, ενώ οι υπολογισμοί εκτελούνται ταυτόχρονα.
2. Παραλληλισμός εργασιών (Task Parallelism): Εδώ, αντί να διαχωρίζονται τα δεδομένα, οι εργασίες του αλγορίθμου χωρίζονται σε μικρότερα, ανεξάρτητα βήματα που εκτελούνται παράλληλα. Κάθε εργασία μπορεί να απαιτεί διαφορετικά δεδομένα ή να εκτελεί διαφορετικές λειτουργίες σε αυτά τα δεδομένα [71].

Ένα κλασικό παράδειγμα παραλληλισμού στην ανάλυση μεγάλων δεδομένων είναι η χρήση του MapReduce, ενός κατανεμημένου μοντέλου επεξεργασίας που επιτρέπει την επεξεργασία τεράστιων όγκων δεδομένων σε παράλληλα συστήματα. Στο μοντέλο αυτό, το Map χωρίζει τα δεδομένα σε μικρότερα κομμάτια και τα κατανέμει σε διάφορους

κόμβους για παράλληλη επεξεργασία, ενώ το Reduce συνδυάζει τα αποτελέσματα από τους διαφορετικούς κόμβους σε ένα τελικό αποτέλεσμα (Εικόνα 2). Το MapReduce είναι ιδιαίτερα χρήσιμο για εφαρμογές που επεξεργάζονται δεδομένα σε μεγάλο όγκο, όπως ανάλυση log files, αναζήτηση και ανάλυση κειμένου, και επεξεργασία μεγάλων βάσεων δεδομένων [73].



Εικόνα 2. Διάγραμμα Ροής. Διαδικασία Παραλληλισμού Αλγορίθμου με MapReduce

Ένα άλλο παράδειγμα είναι η χρήση παραλληλισμού στη μηχανική μάθηση, όπου τα νευρωνικά δίκτυα εκπαιδεύονται σε παράλληλους επεξεργαστές ή GPU (Graphics Processing Units). Η δυνατότητα να διαχωρίζεται η εκπαίδευση σε τμήματα και να κατανέμεται σε πολλές GPU μειώνει δραστικά τον χρόνο εκπαίδευσης, επιτρέποντας τη χρήση πιο σύνθετων και βαθιών μοντέλων σε μεγάλα σύνολα δεδομένων [74].

Ο παραλληλισμός προσφέρει σημαντικά πλεονεκτήματα για την επεξεργασία μεγάλων δεδομένων, καθώς βελτιώνει την απόδοση και μειώνει τον χρόνο που απαιτείται για την εκτέλεση μεγάλων υπολογιστικών εργασιών. Επιπλέον, επιτρέπει την καλύτερη εκμετάλλευση των διαθέσιμων υπολογιστικών πόρων, ιδιαίτερα σε κατανεμημένα περιβάλλοντα, όπου μπορούν να χρησιμοποιηθούν πολλοί υπολογιστικοί κόμβοι ταυτόχρονα [70].

Ωστόσο, ο παραλληλισμός συνοδεύεται και από προκλήσεις. Ο συγχρονισμός μεταξύ των παράλληλων διαδικασιών, η κατανομή των δεδομένων με τρόπο που να αποφεύγονται οι αλληλοεπικαλύψεις και η σωστή διαχείριση των πόρων είναι περίπλοκες διαδικασίες που απαιτούν προσεκτική σχεδίαση. Επιπλέον, σε κάποιες περιπτώσεις, η απόδοση μπορεί να περιοριστεί από το "πρόβλημα του λαιμού του μπουκαλιού", όπου η ταχύτητα ολόκληρου του συστήματος περιορίζεται από έναν αργό κόμβο ή διαδικασία [74]. Παρά τις προκλήσεις του όμως, όταν σχεδιαστεί και εφαρμοστεί σωστά, ο παραλληλισμός μπορεί να μειώσει δραματικά τον χρόνο επεξεργασίας και να αυξήσει την αποδοτικότητα των συστημάτων, κάνοντάς τον αναπόσπαστο εργαλείο στην αντιμετώπιση των απαιτήσεων των μεγάλων δεδομένων [72].

4.2.2 Κατανομή φόρτου

Η κατανομή φορτίου (load balancing) αποτελεί μια κρίσιμη τεχνική βελτιστοποίησης των αλγορίθμων στα μεγάλα δεδομένα, που αποσκοπεί στη διανομή των υπολογιστικών εργασιών και των δεδομένων μεταξύ των διαθέσιμων πόρων με ισόρροπο τρόπο [75]. Η τεχνική αυτή εξασφαλίζει ότι όλοι οι υπολογιστικοί κόμβοι ή οι επεξεργαστές χρησιμοποιούνται αποδοτικά, αποφεύγοντας την υπερφόρτωση κάποιων εξ αυτών και την υποχρησιμοποίηση άλλων. Η σωστή κατανομή του φορτίου είναι ζωτικής σημασίας για την αποδοτικότητα των συστημάτων που επεξεργάζονται μεγάλους όγκους δεδομένων, καθώς συμβάλλει στη μείωση του χρόνου επεξεργασίας, στην καλύτερη εκμετάλλευση των υπολογιστικών πόρων και στην αποφυγή των σημείων συμφόρησης (bottlenecks) [76].

Σε ένα περιβάλλον μεγάλων δεδομένων, η κατανομή φορτίου αποσκοπεί στο να διανέμει τις υπολογιστικές εργασίες και τα δεδομένα μεταξύ πολλαπλών επεξεργαστών ή υπολογιστικών κόμβων. Αυτό γίνεται με στόχο να επιτευχθεί ισορροπία στην εκτέλεση των εργασιών, ώστε να μην υπερφορτωθεί κάποιος κόμβος ενώ άλλοι παραμένουν αδρανείς. Η κακή κατανομή φορτίου μπορεί να οδηγήσει σε αναποτελεσματική εκμετάλλευση των υπολογιστικών πόρων, με ορισμένους κόμβους να λειτουργούν υπό πλήρες φορτίο

και άλλους να παραμένουν ανεκμετάλλευτοι, αυξάνοντας έτσι τον συνολικό χρόνο εκτέλεσης [1].

Υπάρχουν διάφορες στρατηγικές κατανομής φορτίου που μπορούν να χρησιμοποιηθούν, ανάλογα με το είδος του συστήματος και τα χαρακτηριστικά των δεδομένων:

1. Στατική κατανομή φορτίου (Static Load Balancing): Σε αυτή τη στρατηγική, η κατανομή των εργασιών γίνεται πριν από την έναρξη της εκτέλεσης και παραμένει σταθερή καθ' όλη τη διάρκεια της επεξεργασίας. Η στατική κατανομή φορτίου είναι κατάλληλη όταν η φύση των εργασιών και των δεδομένων είναι γνωστή εκ των προτέρων και δεν αλλάζει κατά τη διάρκεια της επεξεργασίας [76]. Για παράδειγμα, μπορεί να εφαρμοστεί σε σενάρια όπου τα δεδομένα είναι ισοκατανεμημένα και οι κόμβοι έχουν παρόμοιες υπολογιστικές ικανότητες.
2. Δυναμική κατανομή φορτίου (Dynamic Load Balancing): Στη δυναμική κατανομή φορτίου, η κατανομή των εργασιών προσαρμόζεται κατά τη διάρκεια της επεξεργασίας, ανάλογα με το τρέχον φορτίο κάθε κόμβου. Αυτή η στρατηγική είναι ιδιαίτερα χρήσιμη όταν η κατανομή των δεδομένων ή των εργασιών είναι άνιση ή όταν οι υπολογιστικοί πόροι μεταβάλλονται δυναμικά. Η δυναμική κατανομή εξασφαλίζει ότι οι κόμβοι που αντιμετωπίζουν υψηλότερο φορτίο μπορούν να ανακουφιστούν, μεταφέροντας κάποιες εργασίες σε άλλους λιγότερο φορτωμένους κόμβους [77].
3. Κατανεμημένη κατανομή φορτίου (Distributed Load Balancing): Σε κατανεμημένα συστήματα, η κατανομή του φορτίου πραγματοποιείται με τη συνεργασία των κόμβων μεταξύ τους. Κάθε κόμβος μπορεί να διαχειρίζεται τοπικά την κατανομή του φορτίου του, ενώ παράλληλα επικοινωνεί με άλλους κόμβους για να διαμοιράσει τις εργασίες όταν παρατηρεί υπερφόρτωση. Αυτό το μοντέλο χρησιμοποιείται συχνά σε συστήματα μεγάλης κλίμακας, όπου οι κόμβοι λειτουργούν ανεξάρτητα και δεν υπάρχει κεντρικός διαχειριστής [76].

Ένα από τα πιο χαρακτηριστικά παραδείγματα κατανομής φορτίου είναι το Hadoop YARN (Yet Another Resource Negotiator), το οποίο είναι υπεύθυνο για τη διαχείριση των πόρων σε κατανεμημένα συστήματα και τη δυναμική κατανομή των εργασιών στους διαθέσιμους κόμβους. Το YARN αναθέτει εργασίες στους υπολογιστικούς κόμβους με βάση τη διαθεσιμότητά τους και προσαρμόζεται στις αλλαγές στο φορτίο κατά τη διάρκεια της εκτέλεσης, εξασφαλίζοντας την ισορροπία του φορτίου σε όλο το σύστημα [77].

Η κατανομή φορτίου προσφέρει σημαντικά πλεονεκτήματα για την επεξεργασία μεγάλων δεδομένων. Βελτιώνει την απόδοση του συστήματος, μειώνοντας το χρόνο εκτέλεσης των εργασιών μέσω της ορθής διανομής του φόρτου εργασίας σε πολλούς υπολογιστικούς πόρους. Επίσης, ενισχύει τη σταθερότητα του συστήματος, καθώς αποφεύγονται σημεία συμφόρησης που θα μπορούσαν να προκαλέσουν καθυστερήσεις ή ακόμα και καταρρεύσεις του συστήματος [75].

Όμως, η κατανομή φορτίου δεν είναι χωρίς προκλήσεις. Η σωστή ισορροπία μπορεί να είναι δύσκολη να επιτευχθεί, ιδιαίτερα σε δυναμικά περιβάλλοντα όπου οι εργασίες και οι πόροι μεταβάλλονται συχνά. Επιπλέον, ο συγχρονισμός μεταξύ των κόμβων και η διαχείριση της επικοινωνίας μπορεί να προσθέσουν πολυπλοκότητα στο σύστημα, καθώς απαιτείται αποτελεσματική οργάνωση για να αποφευχθεί η υπερφόρτωση κάποιων κόμβων. Τέλος, η υλοποίηση της κατανεμημένης κατανομής φορτίου μπορεί να είναι τεχνικά απαιτητική, ιδιαίτερα σε κατανεμημένα συστήματα μεγάλης κλίμακας [69].

4.2.3 Μείωση Πολυπλοκότητας

Η μείωση της πολυπλοκότητας αποτελεί μια από τις πιο σημαντικές τεχνικές βελτιστοποίησης αλγορίθμων στα big data, επιδιώκοντας την ελαχιστοποίηση των υπολογιστικών πόρων και του χρόνου που απαιτούνται για την επεξεργασία των δεδομένων, χωρίς να θυσιάζεται η ακρίβεια των αποτελεσμάτων. Καθώς τα δεδομένα αυξάνονται εκθετικά, η πολυπλοκότητα των αλγορίθμων γίνεται κρίσιμης σημασίας, καθώς επηρεάζει άμεσα την αποδοτικότητα, τον χρόνο εκτέλεσης και τις απαιτήσεις σε μνήμη και υπολογιστική ισχύ [78]. Η μείωση της πολυπλοκότητας περιλαμβάνει στρατηγικές που απλοποιούν τους αλγόριθμους, περιορίζουν τον αριθμό των υπολογισμών και επιτρέπουν την ταχύτερη επεξεργασία των δεδομένων, ακόμα και όταν αυτά είναι τεράστια [28].

Οι κυριότερες τεχνικές που μπορούν να χρησιμοποιηθούν για τη μείωση της πολυπλοκότητας των αλγορίθμων, επιτρέποντας την αποδοτικότερη επεξεργασία μεγάλων δεδομένων, είναι:

1. Μείωση διαστάσεων (Dimensionality Reduction): Ένα από τα πιο κοινά προβλήματα στα μεγάλα δεδομένα είναι η υπερβολική διάσταση των χαρακτηριστικών, γεγονός που αυξάνει την πολυπλοκότητα των υπολογισμών. Οι τεχνικές μείωσης διαστάσεων, όπως η Ανάλυση Κύριων Συνιστωσών (PCA) ή η Μερική Λιγότερη Πλατεία Παλινδρόμηση (PLS), επιτρέπουν τη μείωση του αριθμού των χαρακτηριστικών που πρέπει να επεξεργαστεί ένας αλγόριθμος, διατηρώντας ταυτόχρονα

τις βασικές πληροφορίες των δεδομένων. Αυτό μειώνει την πολυπλοκότητα και τον υπολογιστικό χρόνο, βελτιώνοντας την ταχύτητα και αποδοτικότητα των αλγορίθμων [78].

2. Προσεγγιστικοί αλγόριθμοι (Approximation Algorithms): Οι προσεγγιστικοί αλγόριθμοι προσφέρουν λύσεις που είναι "αρκετά καλές", χωρίς να χρειάζεται να επιτευχθεί η απόλυτη βέλτιστη λύση. Στα μεγάλα δεδομένα, αυτό μπορεί να είναι επαρκές, καθώς η ακριβής λύση μπορεί να είναι εξαιρετικά δαπανηρή σε υπολογιστικό χρόνο και πόρους. Για παράδειγμα, οι προσεγγιστικές εκδοχές του αλγορίθμου K-means χρησιμοποιούνται ευρέως για την ομαδοποίηση δεδομένων σε μεγάλα σύνολα δεδομένων, προσφέροντας αποδεκτές λύσεις με σημαντικά χαμηλότερη πολυπλοκότητα [79].
3. Απλοποίηση υπολογισμών (Algorithm Simplification): Μερικές φορές, οι αλγόριθμοι μπορούν να απλοποιηθούν αφαιρώντας περιττές λειτουργίες ή εφαρμόζοντας έξυπνες στρατηγικές για τη μείωση του αριθμού των απαραίτητων υπολογισμών. Ένα παράδειγμα είναι η χρήση ευρετικών μεθόδων για τη βελτίωση της ταχύτητας των αλγορίθμων χωρίς να απαιτείται η επεξεργασία κάθε δυνατού συνδυασμού δεδομένων [79].
4. Μείωση της υπολογιστικής πολυπλοκότητας (Reducing Time Complexity): Η βελτιστοποίηση των αλγορίθμων με στόχο τη μείωση της υπολογιστικής πολυπλοκότητας είναι κεντρική για την επεξεργασία μεγάλων δεδομένων. Για παράδειγμα, η αντικατάσταση αλγορίθμων με υψηλή πολυπλοκότητα, όπως O(n^2), με αποδοτικότερους αλγορίθμους που έχουν μικρότερη πολυπλοκότητα, όπως O(n log n), μπορεί να έχει δραστικά αποτελέσματα στη μείωση του χρόνου εκτέλεσης όταν ο αριθμός των δεδομένων αυξάνεται [80].
5. Δειγματοληψία δεδομένων (Data Sampling): Σε περιπτώσεις όπου η πλήρης ανάλυση ενός συνόλου δεδομένων είναι υπερβολικά αργή ή δαπανηρή, η δειγματοληψία μπορεί να χρησιμοποιηθεί για να επιλέξει ένα υποσύνολο δεδομένων που αντιπροσωπεύει το συνολικό σύνολο. Αυτό μειώνει την πολυπλοκότητα των υπολογισμών και επιταχύνει τους αλγορίθμους χωρίς σημαντική απώλεια ακρίβειας, ιδιαίτερα σε προβλήματα όπου η ανάλυση μεγάλων δειγμάτων είναι επαρκής για τη λήψη αποφάσεων [78].

Η μείωση της πολυπλοκότητας προσφέρει σημαντικά πλεονεκτήματα για την επεξεργασία μεγάλων δεδομένων. Πρώτον, βελτιώνει την απόδοση των συστημάτων,

μειώνοντας τον χρόνο επεξεργασίας και την ανάγκη για υπολογιστικούς πόρους, όπως μνήμη και επεξεργαστική ισχύ. Αυτό είναι ιδιαίτερα κρίσιμο όταν οι αλγόριθμοι πρέπει να επεξεργαστούν τεράστιους όγκους δεδομένων ή να λειτουργούν σε πραγματικό χρόνο [68]. Δεύτερον, η μείωση της πολυπλοκότητας επιτρέπει την καλύτερη κλιμάκωση των αλγορίθμων, ώστε να μπορούν να επεξεργάζονται μεγαλύτερα σύνολα δεδομένων χωρίς να υποβαθμίζεται η αποδοτικότητα του συστήματος [28].

Επιπλέον, η μείωση της πολυπλοκότητας μπορεί να μειώσει τον κίνδυνο εμφάνισης λαθών ή προβλημάτων που σχετίζονται με την πολυπλοκότητα των συστημάτων, όπως η υπερβολική χρήση πόρων ή η εμφάνιση συμφόρησης σε υπολογιστικά συστήματα. Η απλούστευση των αλγορίθμων μπορεί επίσης να διευκολύνει τη συντήρηση και την αναβάθμιση των συστημάτων, ενώ παράλληλα μειώνει τον κίνδυνο αποτυχίας κατά την επεξεργασία μεγάλων δεδομένων [81].

Παρά τα πλεονεκτήματα της μείωσης της πολυπλοκότητας, η διαδικασία αυτή δεν είναι χωρίς προκλήσεις. Η απλοποίηση ενός αλγορίθμου μπορεί να οδηγήσει σε απώλεια ακρίβειας ή πληροφορίας, ιδιαίτερα όταν χρησιμοποιούνται προσεγγιστικοί αλγόριθμοι ή δειγματοληψία δεδομένων. Επιπλέον, η επιλογή της σωστής τεχνικής μείωσης πολυπλοκότητας απαιτεί εξειδικευμένη γνώση και προσεκτική ανάλυση του συγκεκριμένου προβλήματος και των δεδομένων [30]. Σε ορισμένες περιπτώσεις, μπορεί να είναι δύσκολο να επιτευχθεί ισορροπία μεταξύ της μείωσης της πολυπλοκότητας και της διατήρησης της ποιότητας των αποτελεσμάτων.

4.2.4 Κατανεμημένα υπολογιστικά μοντέλα

Τα κατανεμημένα υπολογιστικά μοντέλα, όπως το Hadoop και το Apache Spark, αποτελούν βασικές τεχνικές βελτιστοποίησης των αλγορίθμων για την επεξεργασία μεγάλων δεδομένων [82]. Αυτά τα μοντέλα επιτρέπουν την κατανομή των υπολογιστικών εργασιών σε ένα δίκτυο υπολογιστικών κόμβων, διευκολύνοντας την αποδοτική επεξεργασία τεράστιων όγκων δεδομένων που δεν μπορούν να επεξεργαστούν από έναν μόνο υπολογιστή ή σε κεντρικά συστήματα [80]. Με την αξιοποίηση της κατανεμημένης επεξεργασίας, τα μοντέλα αυτά παρέχουν μια αποτελεσματική λύση για τη βελτιστοποίηση της απόδοσης των αλγορίθμων και τη διαχείριση των μεγάλων δεδομένων με μεγαλύτερη ταχύτητα και αντοχή.

- Hadoop

Το Hadoop είναι ένα κατανεμημένο υπολογιστικό πλαίσιο ανοιχτού κώδικα που επιτρέπει την επεξεργασία μεγάλων δεδομένων με τη χρήση κατανεμημένων υπολογιστικών πόρων. Το βασικό συστατικό του Hadoop είναι το Hadoop Distributed File System (HDFS), το οποίο αποθηκεύει δεδομένα κατανεμημένα σε πολλούς κόμβους, επιτρέποντας την ταυτόχρονη πρόσβαση και επεξεργασία τους. Επιπλέον, το MapReduce, ένα από τα κύρια προγραμματιστικά μοντέλα του Hadoop, επιτρέπει την κατανομή της επεξεργασίας δεδομένων σε πολλά μικρότερα κομμάτια (tasks), τα οποία εκτελούνται σε διαφορετικούς κόμβους [82].

Τα κύρια πλεονεκτήματα του Hadoop είναι:

1. Κατανεμημένη επεξεργασία: Το Hadoop κατανεμεί τα δεδομένα και την επεξεργασία τους σε ένα σύνολο κόμβων, γεγονός που του επιτρέπει να επεξεργάζεται τεράστιους όγκους δεδομένων με τρόπο αποδοτικό και κλιμακούμενο.
2. Ανοχή σε σφάλματα (Fault Tolerance): Το Hadoop παρέχει μηχανισμούς ανοχής σε σφάλματα, καθώς αν κάποιος κόμβος αποτύχει κατά τη διάρκεια της επεξεργασίας, οι εργασίες του κατανέμονται σε άλλους κόμβους, εξασφαλίζοντας τη συνέχιση της επεξεργασίας [83].
3. Χαμηλό κόστος και ευκολία υλοποίησης: Το Hadoop είναι ανοιχτού κώδικα και μπορεί να λειτουργήσει σε χαμηλού κόστους hardware, καθιστώντας το προσιτό για μεγάλες επιχειρήσεις αλλά και μικρότερους οργανισμούς [82].

Ωστόσο, το Hadoop βασίζεται σε ένα batch processing μοντέλο, το οποίο σημαίνει ότι η επεξεργασία των δεδομένων γίνεται σε διακριτά βήματα και δεν υποστηρίζει εύκολα την ανάλυση σε πραγματικό χρόνο. Επιπλέον, το MapReduce μπορεί να είναι υπολογιστικά ακριβό για ορισμένα είδη επεξεργασίας, λόγω της ανάγκης για συνεχή ανάγνωση και εγγραφή δεδομένων στο HDFS [73].

- Apache Spark

Το Apache Spark είναι ένα πιο εξελιγμένο κατανεμημένο υπολογιστικό πλαίσιο που χτίστηκε για να υπερβεί τις αδυναμίες του Hadoop, ειδικά σε ό,τι αφορά την ταχύτητα και την ευελιξία. Το Spark προσφέρει ένα σύστημα in-memory επεξεργασίας, το οποίο διατηρεί τα δεδομένα στη μνήμη RAM αντί να τα γράφει και να τα διαβάζει συνεχώς από το δίσκο. Αυτό επιταχύνει σημαντικά την επεξεργασία δεδομένων, ιδιαίτερα σε επαναληπτικές διεργασίες, όπως η εκπαίδευση μηχανικών μοντέλων μάθησης [84].

Τα βασικά πλεονεκτήματα του Apache Spark είναι:

1. Επεξεργασία στη μνήμη (In-Memory Processing): Το Spark μπορεί να επεξεργάζεται δεδομένα απευθείας στη μνήμη, επιτρέποντας πολύ ταχύτερη ανάλυση, ειδικά σε επαναληπτικά καθήκοντα, όπου η ανάγκη για πρόσβαση στο δίσκο μειώνεται δραματικά.
2. Υποστήριξη ανάλυσης σε πραγματικό χρόνο: Το Spark διαθέτει πλαίσια ανάλυσης ροών δεδομένων, όπως το Spark Streaming, που επιτρέπει την επεξεργασία δεδομένων σε πραγματικό χρόνο [84].
3. Ευελιξία: Το Spark υποστηρίζει πολλαπλά προγραμματιστικά μοντέλα, όπως MapReduce, Graph Processing (GraphX), Machine Learning (MLlib) και SQL μέσω του Spark SQL. Αυτό το καθιστά εξαιρετικά ευέλικτο για την επεξεργασία και ανάλυση μεγάλων δεδομένων σε διάφορες μορφές [85].

Η ενσωμάτωση της επεξεργασίας στη μνήμη επιτρέπει στο Spark να υπερτερεί σε σενάρια που απαιτούν επαναλαμβανόμενη πρόσβαση στα δεδομένα, όπως η εκπαίδευση αλγορίθμων μηχανικής μάθησης και οι προσομοιώσεις. Παρόλα αυτά, αυτή η προσέγγιση απαιτεί περισσότερη μνήμη RAM, κάτι που μπορεί να αυξήσει το κόστος των υποδομών, ιδιαίτερα σε μεγάλα σύνολα δεδομένων [84].

Γενικά, τα κατανεμημένα υπολογιστικά μοντέλα όπως το Hadoop και το Spark επιτρέπουν τη διαχείριση μεγάλων δεδομένων με έναν τρόπο που καθιστά την επεξεργασία ταχύτερη και πιο αποδοτική, χρησιμοποιώντας πολλούς υπολογιστικούς κόμβους ταυτόχρονα [85]. Για παράδειγμα, σε μια ανάλυση δεδομένων πελατών σε έναν μεγάλο οργανισμό, τα δεδομένα μπορούν να αποθηκευτούν και να επεξεργαστούν κατανεμημένα σε πολλούς κόμβους χρησιμοποιώντας το Hadoop. Με το Spark, τα δεδομένα μπορούν να υποστούν ανάλυση σε πραγματικό χρόνο, όπως η παρακολούθηση της συμπεριφοράς των πελατών και η παροχή προσωποποιημένων προτάσεων σε πραγματικό χρόνο [86].

Με βάση όσα προαναφέρθηκαν, τα κατανεμημένα υπολογιστικά μοντέλα προσφέρουν πλεονεκτήματα όπως η αυξημένη ταχύτητα, η κλιμακούμενη επεξεργασία και η αξιοπιστία σε περιβάλλοντα μεγάλων δεδομένων. Μπορούν να διαχειριστούν τεράστιες ποσότητες δεδομένων που δεν θα μπορούσαν να υποστηριχθούν από μεμονωμένους υπολογιστές και να προσαρμόζονται στις ανάγκες των επιχειρήσεων, επιτρέποντας παράλληλα την επεξεργασία των δεδομένων σε πραγματικό χρόνο [80].

Ωστόσο, αυτές οι τεχνολογίες συνοδεύονται και με προκλήσεις. Η εγκατάσταση, συντήρηση και λειτουργία κατανεμημένων υπολογιστικών συστημάτων απαιτεί

εξειδικευμένες γνώσεις και προγραμματιστικές ικανότητες. Η αποδοτική κατανομή των πόρων και η σωστή διαχείριση των σφαλμάτων είναι επίσης κρίσιμες για τη διασφάλιση της σταθερότητας και της αξιοπιστίας αυτών των συστημάτων. Επιπλέον, το κόστος των υποδομών, ειδικά για το Apache Spark με τις αυξημένες ανάγκες σε μνήμη, μπορεί να είναι υψηλό για ορισμένες επιχειρήσεις [84]. Παρά τις προκλήσεις που φέρνουν όμως, προσφέρουν σημαντικές δυνατότητες που μπορούν να μεταμορφώσουν τη διαχείριση των μεγάλων δεδομένων και να ενισχύσουν την απόδοση των αλγορίθμων που χρησιμοποιούνται σε αυτά τα συστήματα.

5 Συμπεράσματα

Με την ολοκλήρωση της παρούσας μελέτης, αναδεικνύεται η σημασία της βελτιστοποίησης αλγορίθμων για την επεξεργασία μεγάλων δεδομένων (Big Data), εστιάζοντας σε τεχνικές που βελτιώνουν την απόδοση, την κλιμάκωση και την ακρίβεια των αλγορίθμων. Μέσα από την ανάλυση των διαφόρων τεχνικών βελτιστοποίησης, αποδείχθηκε ότι η βελτιστοποίηση μπορεί να βελτιώσει δραστικά τις επιδόσεις των αλγορίθμων και να ενισχύσει τη δυνατότητα επεξεργασίας μεγάλων όγκων δεδομένων [4].

Η μελέτη αυτή αποκαλύπτει ότι δεν υπάρχει μία και μοναδική προσέγγιση που να είναι επαρκής για όλα τα είδη προβλημάτων. Η αποτελεσματικότητα κάθε τεχνικής εξαρτάται από τη φύση των δεδομένων, τους στόχους της ανάλυσης, την πολυπλοκότητα του προβλήματος, καθώς και τους διαθέσιμους υπολογιστικούς πόρους [67]. Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης, όπως η λογιστική παλινδρόμηση, τα δέντρα απόφασης και οι υποστηρικτές διανυσμάτων (SVM), εξακολουθούν να είναι ιδιαίτερα χρήσιμοι σε προβλήματα με δομημένα δεδομένα και ερμηνεύσιμα αποτελέσματα. Ωστόσο, υστερούν στην κλιμάκωση και παρουσιάζουν περιορισμούς όταν καλούνται να αντιμετωπίσουν πολυδιάστατα, ετερογενή ή αδόμητα δεδομένα [38].

Αντίθετα, οι αλγόριθμοι βαθιάς μάθησης (Deep Learning), όπως τα συνελικτικά και αναδρομικά νευρωνικά δίκτυα, υπερέχουν σε προβλήματα υψηλής πολυπλοκότητας, όπως η επεξεργασία εικόνας, ήχου και φυσικής γλώσσας [40]. Αυτές οι τεχνικές προσφέρουν δυνατότητες αυτόματης εξαγωγής χαρακτηριστικών και υψηλή ακρίβεια, αλλά απαιτούν μεγάλο υπολογιστικό κόστος και όγκο δεδομένων για την εκπαίδευσή τους, ενώ συχνά υστερούν σε διαφάνεια και ερμηνευσιμότητα [81].

Στον τομέα της μη εποπτευόμενης μάθησης, οι αλγόριθμοι ομαδοποίησης όπως ο K-Means και ο DBSCAN επιτρέπουν την ανακάλυψη υποκείμενων προτύπων χωρίς προγενέστερη γνώση. Ο K-Means είναι ιδιαίτερα γρήγορος και αποδοτικός, αλλά ευαίσθητος στην αρχική επιλογή των κεντροειδών, ενώ ο DBSCAN είναι πιο ευέλικτος, αν και απαιτητικότερος υπολογιστικά [36, 90]. Αντίστοιχα, οι αλγόριθμοι εξόρυξης συσχετίσεων, όπως ο Apriori και ο FP-Growth, είναι χρήσιμοι για την εύρεση κανόνων συσχέτισης, με

τον δεύτερο να υπερέχει σε απόδοση χάρη στη συμπιεσμένη αναπαράσταση FP-tree, αλλά με αυξημένες απαιτήσεις σε μνήμη [46].

Αναπόσπαστο στοιχείο της σύγχρονης επεξεργασίας μεγάλων δεδομένων αποτελούν οι τεχνικές παραλληλισμού και κατανεμημένης επεξεργασίας, όπως αυτές που εφαρμόζονται στις πλατφόρμες Hadoop και Apache Spark [84-86]. Το Spark, ειδικά, υπερέχει λόγω της in-memory επεξεργασίας του, μειώνοντας δραστικά τον χρόνο εκτέλεσης. Οι τεχνικές αυτές δεν επιλύουν τα ίδια τα προβλήματα αλγοριθμικά, αλλά καθιστούν δυνατή την εφαρμογή των τεχνικών σε κλίμακα, αυξάνοντας την αποδοτικότητα και την αξιοπιστία [73]. Τέλος, σημαντικό ρόλο παίζουν οι τεχνικές μείωσης της υπολογιστικής πολυπλοκότητας, όπως η δειγματοληψία δεδομένων, η μείωση διαστάσεων (π.χ. PCA) και η ευρετική απλοποίηση αλγορίθμων. Αυτές οι τεχνικές διευκολύνουν την ανάλυση σε περιβάλλοντα με περιορισμένους πόρους, χωρίς να θυσιάζουν σημαντικά την ακρίβεια των αποτελεσμάτων [88].

Γενικά, διαπιστώθηκε ότι οι πιο επιτυχημένες προσεγγίσεις είναι εκείνες που συνδυάζουν διαφορετικές μεθόδους σε υβριδικά μοντέλα, αξιοποιώντας παράλληλα τα πλεονεκτήματα της κατανεμημένης επεξεργασίας, της μηχανικής μάθησης και της βελτιστοποίησης πόρων [91]. Σε περιβάλλοντα με αυξανόμενες απαιτήσεις, τέτοιες προσεγγίσεις εξασφαλίζουν μεγαλύτερη αποδοτικότητα, προσαρμοστικότητα και δυνατότητα εφαρμογής σε πραγματικό χρόνο [93].

Τα αποτελέσματα της μελέτης αυτής συμβάλλουν στη βελτίωση της επεξεργασίας μεγάλων δεδομένων, καθώς παρέχουν μια ολοκληρωμένη ανάλυση και σύνθεση των τεχνικών βελτιστοποίησης αλγορίθμων που είναι απαραίτητες για την αποδοτική επεξεργασία μεγάλων δεδομένων. Αποδεικνύεται ότι οι αλγόριθμοι μπορούν να βελτιστοποιηθούν για να αντιμετωπίσουν τις προκλήσεις που προκύπτουν από τον αυξανόμενο όγκο, την ταχύτητα και την ποικιλία των δεδομένων [31]. Οι βελτιστοποιημένοι αλγόριθμοι επιτρέπουν την αποτελεσματική και γρήγορη ανάλυση δεδομένων, τη μείωση του χρόνου και του κόστους επεξεργασίας και τη δημιουργία πιο ακριβών και χρήσιμων μοντέλων [67]. Επιπλέον, η εργασία προωθεί την ανάπτυξη νέων πρακτικών και τεχνικών που μπορούν να συμβάλουν στην περαιτέρω εξέλιξη του πεδίου των μεγάλων δεδομένων, προσφέροντας κατευθυντήριες γραμμές για τη βελτιστοποίηση των συστημάτων αυτών.

Σε ανάλογα συμπεράσματα έχουν καταλήξει και άλλες πρόσφατες έρευνες. Για παράδειγμα, η μελέτη των Naeem [87] παρέχει μια ολοκληρωμένη επισκόπηση των τάσεων

και των μελλοντικών προκλήσεων που σχετίζονται με τα μεγάλα δεδομένα. Οι συγγραφείς αναγνωρίζουν μια σειρά από προκλήσεις που αντιμετωπίζει ο τομέας των μεγάλων δεδομένων, όπως το ζήτημα της αποθήκευσης και της υπολογιστικής ισχύος, της διαχείρισης ετερογενών δεδομένων, της ιδιωτικότητας και ασφάλειας. Η μελέτη τονίζει τη σημασία της συνεχιζόμενης έρευνας σε τεχνολογίες αιχμής όπως τα κατανεμημένα υπολογιστικά συστήματα, το Internet of Things (IoT), το 5G και η τεχνητή νοημοσύνη, που μπορούν να βελτιώσουν την αποτελεσματικότητα της ανάλυσης μεγάλων δεδομένων. Αναγνωρίζεται επίσης η ανάγκη για ανάπτυξη πιο αποτελεσματικών και κλιμακούμενων αλγορίθμων που μπορούν να επεξεργάζονται σε πραγματικό χρόνο. Επιπλέον, οι ερευνητές τονίζουν τη σημασία της διεπιστημονικής προσέγγισης στην αντιμετώπιση των μελλοντικών προκλήσεων των μεγάλων δεδομένων. Η συνεργασία μεταξύ διαφόρων τομέων, όπως η πληροφορική, η στατιστική, η μηχανική μάθηση και η ανάλυση δεδομένων, θεωρείται κρίσιμη για την επιτυχία της έρευνας και των εφαρμογών σε big data.

Η μελέτη του Adadi [88] επικεντρώνεται στην αποδοτική χρήση δεδομένων κατά την ανάπτυξη και την εφαρμογή αλγορίθμων στην εποχή των μεγάλων δεδομένων. Αναγνωρίζει ότι ενώ τα big data προσφέρουν πολλές ευκαιρίες, η τεράστια κλίμακα τους μπορεί να οδηγήσει σε προκλήσεις όσον αφορά την επεξεργασία, αποθήκευση και ανάλυση τους. Το άρθρο τονίζει την ανάγκη για αλγόριθμους που μπορούν να επιτύχουν υψηλή απόδοση με λιγότερα δεδομένα, εστιάζοντας στην αποδοτική μάθηση από περιορισμένα δεδομένα. Περιλαμβάνει μεθόδους όπως: Few-shot Learning (εκπαίδευση μοντέλων με ελάχιστα δείγματα δεδομένων), Transfer Learning (αξιοποίηση γνώσης από ένα πεδίο για να βελτιωθεί η απόδοση σε ένα άλλο πεδίο, με τη χρήση λιγότερων δεδομένων), Meta-Learning (εκπαίδευση μοντέλων να μαθαίνουν πιο αποτελεσματικά από μικρές ποσότητες δεδομένων), Self-supervised Learning (ανάπτυξη μοντέλων που χρησιμοποιούν μη επισημασμένα δεδομένα για να βελτιώσουν τις προβλέψεις με ελάχιστη επιβλεπόμενη μάθηση). Ο Adadi [88] καταλήγει ότι, καθώς οι τεχνολογίες μεγάλων δεδομένων συνεχίζουν να εξελίσσονται, η ανάπτυξη καινοτόμων αλγορίθμων που μπορούν να εκμεταλλευτούν αποτελεσματικά μικρότερα σύνολα δεδομένων θα παραμείνει κρίσιμη.

Οι Gaye, Zhang και Wulamu [89] ασχολήθηκαν με τη βελτίωση του αλγορίθμου Support Vector Machine (SVM) στο πλαίσιο των μεγάλων δεδομένων. Αναγνωρίζοντας τις προκλήσεις που αντιμετωπίζει ο συγκεκριμένος αλγόριθμος σε περιβάλλοντα μεγάλων δεδομένων, οι συγγραφείς προτείνουν διάφορες βελτιώσεις και τροποποιήσεις για να τον καταστήσουν πιο αποδοτικό. Αυτές οι βελτιώσεις περιλαμβάνουν: χρήση τεχνικών για

την ταχύτερη εκπαίδευση του SVM με παράλληλη μείωση της πολυπλοκότητας, χρήση δειγματοληψίας για να περιοριστεί το μέγεθος των δεδομένων που χρησιμοποιούνται στον αλγόριθμο, διατηρώντας παράλληλα την απόδοση του SVM, εφαρμογή παραλληλισμού και κατανεμημένων συστημάτων για την κατανομή του φόρτου εργασίας, κάτι που επιτρέπει την ταχύτερη εκπαίδευση και ταξινόμηση σε μεγάλους όγκους δεδομένων, προσαρμογή και βελτίωση του kernel trick, που είναι σημαντικό για την επεξεργασία μη γραμμικών προβλημάτων, με στόχο την καλύτερη απόδοση σε μεγάλα δεδομένα. Τα αποτελέσματα της μελέτης δείχνουν ότι οι βελτιωμένες εκδόσεις του SVM επιτυγχάνουν καλύτερη ταχύτητα εκπαίδευσης και ταξινόμησης, μειώνοντας ταυτόχρονα την κατανάλωση πόρων, χωρίς να θυσιάζεται η ακρίβεια του αλγορίθμου. Οι ερευνητές καταλήγουν πως οι βελτιωμένες τεχνικές μπορούν να επιταχύνουν την ανάλυση και να επιτρέψουν την εφαρμογή του SVM σε μεγάλης κλίμακας συστήματα.

Η μελέτη των Ikotun [90] εστιάζει στον αλγόριθμο K-means, εξετάζοντας τις παραλλαγές του και τις τελευταίες εξελίξεις στην εποχή των μεγάλων δεδομένων. Το άρθρο διερευνά διάφορες παραλλαγές και βελτιώσεις του K-means που έχουν αναπτυχθεί με στόχο να ξεπεραστούν οι περιορισμοί του κλασικού αλγορίθμου. Ορισμένες από τις παραλλαγές αυτές είναι: K-means++ (μια μέθοδος βελτίωσης της αρχικής επιλογής των κεντροειδών για να επιτευχθεί καλύτερη σύγκλιση), Fuzzy K-means (μια προσέγγιση που επιτρέπει στα δεδομένα να ανήκουν σε πολλαπλά clusters με διαφορετικούς βαθμούς συμμετοχής), Mini-batch K-means (μια παραλλαγή που χρησιμοποιεί μικρότερα τμήματα δεδομένων (mini-batches) για την εκπαίδευση του μοντέλου, μειώνοντας έτσι το υπολογιστικό κόστος και επιτρέποντας την κλιμάκωση του αλγορίθμου σε μεγάλα δεδομένα). Οι συγγραφείς εξετάζουν πρακτικές εφαρμογές των παραλλαγών του K-means σε διάφορους τομείς και επισημαίνουν την ανάγκη μελλοντικών ερευνών για την περαιτέρω βελτίωση και την ανάπτυξη νέων παραλλαγών του K-means που θα μπορούν να αντιμετωπίσουν τις προκλήσεις των μεγάλων δεδομένων.

Ακόμα μια μελέτη, αυτή του Potla [91], εστιάζει στις προκλήσεις και τις ευκαιρίες που προκύπτουν από την ανάπτυξη κλιμακούμενων αλγορίθμων μηχανικής μάθησης για την ανάλυση μεγάλων δεδομένων. Το άρθρο ξεκινά με την παρουσίαση του πλαισίου των μεγάλων δεδομένων και τον καθοριστικό ρόλο που διαδραματίζουν οι αλγόριθμοι μηχανικής μάθησης στην εξαγωγή γνώσης από τεράστιες ποσότητες δεδομένων. Στη συνέχεια, ο συγγραφέας εξετάζει τις κύριες προκλήσεις που προκύπτουν από την εφαρμογή αλγορίθμων μηχανικής μάθησης σε περιβάλλοντα μεγάλων δεδομένων και παρουσιάζει

διάφορους αλγορίθμους μηχανικής μάθησης που είναι κλιμακούμενοι και κατάλληλοι για ανάλυση μεγάλων δεδομένων. Αυτοί οι αλγόριθμοι περιλαμβάνουν: παραλλαγές του Linear Models (όπως ο γραμμικός ταξινομητής και οι παραλλαγές του για γρήγορη εκπαίδευση σε big data), Random Forests και Decision Trees (προσαρμογές που επιτρέπουν την κλιμάκωση αυτών των αλγορίθμων σε κατανεμημένα συστήματα), Stochastic Gradient Descent (SGD) (χρήση της τεχνικής αυτής για τη γρήγορη εκπαίδευση μοντέλων σε μεγάλα δεδομένα μέσω βελτιστοποίησης). Οι Potla [91] αναλύει επίσης τεχνολογίες και εργαλεία που επιτρέπουν την ανάπτυξη και εφαρμογή κλιμακούμενων αλγορίθμων μηχανικής μάθησης σε περιβάλλοντα μεγάλων δεδομένων. Αυτές περιλαμβάνουν πλατφόρμες όπως το Apache Spark και το Hadoop, τα οποία επιτρέπουν την κατανεμημένη επεξεργασία μεγάλων δεδομένων με αποδοτικό τρόπο. Η μελέτη υπογραμμίζει ότι η συνεχιζόμενη ανάπτυξη νέων τεχνικών για την κλιμάκωση της μηχανικής μάθησης θα επιτρέψει την ανάλυση όλο και μεγαλύτερων συνόλων δεδομένων, προάγοντας εφαρμογές όπως η εξατομικευμένη ιατρική, οι έξυπνες πόλεις, και η πρόβλεψη καταναλωτικών τάσεων.

Το άρθρο των Abualigah et al. [92] πραγματεύεται τις τελευταίες εξελίξεις στους μεταευρετικούς αλγορίθμους βελτιστοποίησης (meta-heuristic optimization algorithms) και την εφαρμογή τους στην ανάλυση και ομαδοποίηση μεγάλων δεδομένων κειμένου. Οι μεταευρετικοί αλγόριθμοι αποτελούν έναν ευέλικτο τρόπο για την εύρεση κατάλληλων λύσεων σε προβλήματα που είναι δύσκολο να λυθούν με κλασικούς αλγόριθμους βελτιστοποίησης. Το άρθρο παρουσιάζει και αναλύει την εφαρμογή μεταευρετικών αλγορίθμων βελτιστοποίησης, όπως είναι οι αλγόριθμοι εμπνευσμένοι από τη φύση, για την αντιμετώπιση των προκλήσεων που αφορούν στη βελτιστοποίηση της διαδικασίας ομαδοποίησης δεδομένων. Το άρθρο παρέχει μια συγκριτική ανάλυση των διαφόρων μεταευρετικών αλγορίθμων, όπως οι Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) κ.α. Εξετάζει επίσης νέες εξελίξεις και προσαρμογές αυτών των αλγορίθμων σε περιπτώσεις μεγάλης κλίμακας κειμενικών δεδομένων, όπως οι αναζητήσεις στο διαδίκτυο, τα κοινωνικά δίκτυα και άλλες πηγές πληροφορίας που παράγουν τεράστιους όγκους δεδομένων κειμένου. Οι ερευνητές καταλήγουν ότι οι μεταευρετικοί αλγόριθμοι έχουν μεγάλες δυνατότητες στην ομαδοποίηση δεδομένων κειμένου, κυρίως όταν πρόκειται για πολύπλοκα και μεγάλης κλίμακας προβλήματα. Οι προτεινόμενες βελτιώσεις στους αλγορίθμους προσφέρουν καλύτερη ακρίβεια, ταχύτητα και αποδοτικότητα στη διαδικασία ομαδοποίησης.

Σύμφωνα με τα παραπάνω, παρέχονται συγκεκριμένες κατευθύνσεις για τις μελλοντικές έρευνες στο συγκεκριμένο πεδίο, οι οποίες θα πρέπει να επικεντρώνονται στην ανάπτυξη νέων, ακόμη πιο αποδοτικών τεχνικών και αλγορίθμων που θα αντιμετωπίσουν τις συνεχώς αυξανόμενες απαιτήσεις και την πολυπλοκότητα των δεδομένων. Πιο συγκεκριμένα, η μελλοντική έρευνα θα πρέπει να επικεντρωθεί στην ανάπτυξη υβριδικών αλγορίθμων που να συνδυάζουν τεχνικές παραλληλισμού, κατανεμημένης επεξεργασίας και μηχανικής μάθησης. Στόχος είναι η ενίσχυση της αποδοτικότητας και η ευελιξία στην επεξεργασία πολυδιάστατων και δυναμικών δεδομένων [91]. Οι υβριδικοί αυτοί αλγόριθμοι μπορούν να συνδυάζουν παραδοσιακές μεθόδους με βαθιά μάθηση και ευφυείς ευρετικές τεχνικές για βέλτιστη επίδοση σε ετερογενή περιβάλλοντα [91].

Επιπλέον, οι μελλοντικές μελέτες μπορούν να εστιάσουν στη δημιουργία αλγορίθμων που να προσαρμόζουν δυναμικά τις παραμέτρους και τη δομή τους κατά την εκτέλεση. Τέτοια συστήματα μπορούν να αξιοποιούν τεχνητή νοημοσύνη (AI) για την αυτόματη παρακολούθηση της απόδοσης και την επιλογή των βέλτιστων ρυθμίσεων, βελτιώνοντας την αποτελεσματικότητα χωρίς ανθρώπινη παρέμβαση [93]. Ακόμη, η ανάπτυξη κβαντικών αλγορίθμων για προβλήματα βελτιστοποίησης σε Big Data αποτελεί μία ιδιαίτερα ελπιδοφόρα περιοχή έρευνας. Οι μελέτες αυτές μπορούν να εξετάσουν την αποδοτικότητα κβαντικών μοντέλων στην επίλυση προβλημάτων που σήμερα θεωρούνται υπολογιστικά απαγορευτικά [94].

Καθώς η ανάγκη για real-time ανάλυση αυξάνεται (π.χ. σε συστήματα IoT, μεταφορές, υγεία), η μελλοντική έρευνα μπορεί να στραφεί στην ανάπτυξη αλγορίθμων χαμηλού latency που ενσωματώνουν τεχνικές ροής δεδομένων και δυναμικής επεξεργασίας [67]. Ένα κρίσιμο ζήτημα για την επεξεργασία σε κινητές και edge πλατφόρμες είναι η κατανάλωση πόρων. Συνεπώς, η έρευνα θα μπορούσε να εστιάσει στην ανάπτυξη αλγορίθμων με μειωμένες απαιτήσεις σε μνήμη και ενέργεια, κατάλληλων για εφαρμογές σε αισθητήρες, μικροεπεξεργαστές και κινητές συσκευές [78].

Μια ακόμα κατεύθυνση για μελλοντικές έρευνες αφορά την προώθηση της ανάπτυξης αλγορίθμων που ενσωματώνουν εξαρχής μεθόδους προστασίας ιδιωτικότητας, όπως differential privacy και κρυπτογραφημένη ανάλυση [21]. Αυτό θα καταστήσει την ανάλυση μεγάλων ευαίσθητων δεδομένων ασφαλέστερη και συμμορφωμένη με τις νομικές απαιτήσεις [24]. Τέλος, μια σημαντική κατεύθυνση είναι η δημιουργία γενικεύσιμων πλαισίων που θα μπορούν να εφαρμοστούν σε πολλούς τομείς (υγεία, ενέργεια, εμπόριο, κυβερνοασφάλεια) με ελάχιστες προσαρμογές.

Ενώ η παρούσα εργασία επικεντρώθηκε σε τεκμηριωμένες τεχνικές βελτιστοποίησης, το πεδίο αυτό είναι σε διαρκή κίνηση, με την έρευνα να εξελίσσεται προς νέες κατευθύνσεις, όπως η ενσωμάτωση τεχνητής νοημοσύνης και μηχανικής μάθησης για αυτοβελτιστοποιούμενους αλγορίθμους. Η αυξανόμενη ανάγκη για διαχείριση δεδομένων σε πραγματικό χρόνο, καθώς και η προσαρμογή σε κατανεμημένα, παγκόσμια δίκτυα απαιτούν συνεχή καινοτομία [17]. Η επόμενη γενιά αλγορίθμων και υπολογιστικών μοντέλων αναμένεται να προσφέρει πρωτοποριακές λύσεις, που όχι μόνο θα διαχειρίζονται αποτελεσματικά τα δεδομένα, αλλά θα ενισχύουν και τη δημιουργικότητα και τη λήψη στρατηγικών αποφάσεων σε έναν συνεχώς μεταβαλλόμενο κόσμο.

Βιβλιογραφία

[1] Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231-1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009

[2] Bhadani, A. K. & Jothimani, D. (2016). Big Data: Challenges, Opportunities, and Realities. In M. Singh & D. G. (Eds.), *Effective Big Data Management and Opportunities for Implementation* (pp. 1-24). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-5225-0182-4.ch001

[3] Zhou, X., Liang, W., Kevin, I., Wang, K., & Yang, L. T. (2020). Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. *IEEE Transactions on Computational Social Systems*, *8*(1), 171-178. https://doi.org/10.1109/TCSS.2020.2987846

[4] Roy, C., Rautaray, S. S., & Pandey, M. (2018). Big Data Optimization Techniques: A Survey. *International Journal of Information Engineering & Electronic Business*, *10*(4). https://doi.org/10.5815/ijieeb.2018.04.06

[5] Chebbi, I., Boulila, W., & Farah, I. R. (2015). Big data: Concepts, challenges and applications. In Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part II (pp. 638-647). Springer International Publishing. https://doi.org/10.1007/978-3-319-24306-1_62

[6] Mishra, D., Luo, Z., Jiang, S., Papadopoulos, T., & Dubey, R. (2017). A bibliographic study on big data: concepts, trends and challenges. *Business Process Management* Journal, 23(3), 555-573. https://doi.org/10.1108/BPMJ-10-2015-0149

[7] Iqbal, M., Kazmi, S. H. A., Manzoor, A., Soomrani, A. R., Butt, S. H., & Shaikh, K. A. (2018). A study of big data for business growth in SMEs: Opportunities & challenges. In 2018 International conference on computing, mathematics and engineering *technologies (iCoMET)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICOMET.2018.8346368

[8] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data*, *6*(1), 1-25. https://doi.org/10.1186/s40537-019-0217-0

[9] Faaique, M. (2024). Overview of big data analytics in modern astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 96-113. https://doi.org/10.59543/ijmscs.v2i.8561

[10] Kumar, D., Kamesh, D. B. K., & Syed, U. (2014). A study on Big Data and its importance. *International Journal of Applied Engineering Research*, *9*(20), 7469-7479.

[11] Kapil, G., Agrawal, A., & Khan, R. A. (2016). A study of big data characteristics. In *2016 international conference on communication and electronics systems (ICCES)* (pp. 1-4). IEEE. https://doi.org/10.1109/CESYS.2016.7889917

[12] Kaur, N., & Sood, S. K. (2017). Dynamic resource allocation for big data streams based on data characteristics (5 Vs). *International Journal of Network Management*, *27*(4), e1978. https://doi.org/10.1002/nem.1978

[13] Rawat, R., & Yadav, R. (2021). Big data: Big data analysis, issues and challenges and technologies. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012014). IOP Publishing. https://doi.org/10.1088/1757-899X/1022/1/012014

[14] Bagga, S., & Sharma, A. (2018). Big data and its challenges: a review. In *2018 4th International Conference on Computing Sciences (ICCS)* (pp. 183-187). IEEE.

[15] Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst*, *36*, 1-12. https://doi.org/10.1109/ICCS.2018.00037

[16] Patgiri, R. (2018). Issues and challenges in big data: a survey. In *Distributed Computing and Internet Technology: 14th International Conference, ICDCIT 2018, Bhubaneswar, India, January 11–13, 2018, Proceedings 14* (pp. 295-300). Springer International Publishing. https://doi.org/10.1007/978-3-319-72344-0_25

[17] Prakash, A., Navya, N., & Natarajan, J. (2019). Big data preprocessing for modern world: opportunities and challenges. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (pp. 335-343). Springer International Publishing. https://doi.org/10.1007/978-3-030-03146-6_37

[18] Idrees, S. M., Alam, M. A., & Agarwal, P. (2019). A study of big data and its challenges. *International Journal of Information Technology*, *11*, 841-846. https://doi.org/10.1007/s41870-018-0185-1

[19] Mishra, S., & Misra, A. (2017). Structured and unstructured big data analytics. In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 740-746). IEEE. https://doi.org/10.1109/CTCEEC.2017.8454999

[20] Lawal, Z. K., Zakari, R. Y., Shuaibu, M. Z., & Bala, A. (2016). A review: Issues and Challenges in Big Data from Analytic and Storage perspectives. *International Journal of Engineering and Computer Science*, *5*(3), 15947-15961.

[21] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big data*, *6*(1), 1-16. https://doi.org/10.1186/s40537-019-0206-3

[22] Komalavalli, C., & Laroiya, C. (2019). Challenges in big data analytics techniques: a survey. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 223-228). IEEE. https://doi.org/10.1109/CONFLUENCE.2019.8776932

[23] Bertino, E., & Ferrari, E. (2017). Big data security and privacy. In *A comprehensive guide through the Italian database research over the last 25 years* (pp. 425-439). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61893-7_25

[24] Joshi, N., & Kadhiwala, B. (2017). Big data security and privacy issues—A survey. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-5). IEEE. https://doi.org/10.1109/IPACT.2017.8245064

[25] Martinez, D., & Herrera, S. (2023). Examining the Ethical and Legal Challenges of Anonymized Data Sharing in the Era of Big Data Analytics. *Journal of Sustainable Technologies and Infrastructure Planning*, *7*(5), 59-77.

[26] Venkatraman, R., & Venkatraman, S. (2019). Big Data Infrastructure, Data Visualisation and Challenges. In *Proceedings of the 3rd International Conference on Big Data and Internet of Things* (pp. 13-17). https://doi.org/10.1145/3361758.3361768

[27] Arfat, Y., Usman, S., Mehmood, R., & Katib, I. (2020). Big Data for Smart Infrastructure Design: Opportunities and Challenges. *Smart Infrastructure and Applications:*

Foundations for Smarter Cities and Societies, 491-518. https://doi.org/10.1007/978-3-030-13705-2_20

[28] Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big Data Analytics: Applications, Prospects and Challenges. *Mobile big data: A roadmap from models to technologies*, 3-20. https://doi.org/10.1007/978-3-319-67925-9_1

[29] Mondal, K. (2016). Design issues of big data parallelisms. In *Information Systems Design and Intelligent Applications: Proceedings of Third International Conference INDIA 2016, Volume 2* (pp. 209-217). Springer India. https://doi.org/10.1007/978-81-322-2752-6_20

[30] Vargas-Solar, G., Zechinelli-Martini, J. L., & Espinosa-Oviedo, J. A. (2017). Big data management: what to keep from the past to face future challenges?. *Data Science and Engineering*, *2*(4), 328-345. https://doi.org/10.1007/s41019-017-0043-3

[31] Wu, Z., Sun, J., Zhang, Y., Wei, Z., & Chanussot, J. (2021). Recent developments in parallel and distributed computing for remotely sensed big data processing. *Proceedings of the IEEE*, *109*(8), 1282-1305. https://doi.org/10.1109/JPROC.2021.3087029

[32] Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W. (2016). Data wrangling for big data: Challenges and opportunities. In 19th International Conference on Extending DatabaseTechnology (pp. 473-478). https://doi.org/10.5441/002/edbt.2016.44

[33] Dulhare, U. N., Ahmad, K., & Ahmad, K. A. B. (Eds.). (2020). *Machine learning and big data: concepts, algorithms, tools and applications*. John Wiley & sons.

[34] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138. https://doi.org/10.14445/22312803/IJCTT-V48P126

[35] Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of data science*, *2*, 165-193. https://doi.org/10.1007/s40745-015-0040-1

[36] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236. https://doi.org/10.48550/arXiv.1612.08388

[37] Mahmood, S. A., & Qasim, Q. Q. (2020). Big data sentimental analysis using document to vector and optimized Support vector machine. *UHD Journal of Science and Technology*, *4*(1), 18-28. https://doi.org/10.21928/uhdjst.v4n1y2020.pp18-28

[38] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[39] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[40] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of big data*, *7*(1), 28. https://doi.org/10.1186/s40537-020-00305-w

[41] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, *33*(12), 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

[42] Sharkawy, A. N. (2020). Principle of neural network and its main types. *Journal of Advances in Applied & Computational Mathematics*, *7*, 8-19. https://doi.org/-10.15377/2409-5761.2020.07.2

[43] Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology*, *8*(1), 13-19.

[44] Vuppula, A. (2019). Efficiency And Scalability Of Data Mining Algorithms. *International Journal Of Scientific Development And Research*, *4*(9), 322-328.

[45] Chang, X., & Yang, Y. (2016). Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE transactions on neural networks and learning systems*, *28*(10), 2294-2305. https://doi.org/10.1109/TNNLS.2016.2582746

[46] Harikumar, S., & Dilipkumar, D. U. (2016). Apriori algorithm for association rule mining in high dimensional data. In *2016 International Conference on Data Science and Engineering (ICDSE)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICDSE.2016.7823952

[47] Ye, Q., Lu, W., & Ning, S. (2023). Improved Algorithm of FP-Growth Based on Strong Data Correlation. In *International Conference on Business Intelligence and Information Technology* (pp. 27-36). Singapore: Springer Nature Singapore.

[48] Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, *60*, 708-713. https://doi.org/10.1016/j.procs.2015.08.220

[49] Xu, X., Liu, H., & Yao, M. (2019). Recent progress of anomaly detection. *Complexity*, *2019*(1), 2686378. https://doi.org/10.1155/2019/2686378

[50] Fan, L., Ma, J., Tian, J., Li, T., & Wang, H. (2021). Comparative Study of Isolation Forest and LOF algorithm in anomaly detection of data mining. In *2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICBAR55169.2021.00008

[51] Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., & Li, J. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, *1*(2), 149-164. https://doi.org/10.1016/j.enbenv.2019.11.003

[52] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, *51*(5), 1-36. https://doi.org/10.1145/3234150

[53] Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: methodological study. Journal of medical Internet research, 20(6), e231. https://doi.org/10.2196/jmir.9702

[54] Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509-553. https://doi.org/10.1017/S1351324922000213

[55] Bhati, R. G. (2019). A survey on sentiment analysis algorithms and datasets. *Review of Computer Engineering Research*, *6*(2), 84-91. https://doi.org/10.18488/journal.76.2019.62.84.91

[56] Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched latent dirichlet allocation for sentiment analysis. *Expert Systems*, *37*(4), e12527. https://doi.org/10.1111/exsy.12527

[57] Shahbazi, Z., & Byun, Y. C. (2020). Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. *Journal of Intelligent & Fuzzy Systems*, *39*(1), 753-770. https://doi.org/10.3233/JIFS-191690

[58] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, *33*(4), 917-963. https://doi.org/10.1007/s10618-019-00619-1

[59] Hevey, D. (2018). Network analysis: a brief overview and tutorial. Health psychology and behavioral medicine, 6(1), 301-328. https://doi.org/10.1080/21642850.2018.1521283

[60] Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, *6*(1), 30750. https://doi.org/10.1038/srep30750

[61] Bloch, F., Jackson, M. O., & Tebaldi, P. (2023). Centrality measures in networks. *Social Choice and Welfare*, *61*(2), 413-453. https://doi.org/10.1007/s00355-023-01456-4

[62] Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, *108*, 87-111. https://doi.org/10.1016/j.jnca.2018.02.011

[63] AbuSalim, S. W., Ibrahim, R., Saringat, M. Z., Jamel, S., & Wahab, J. A. (2020). Comparative analysis between dijkstra and bellman-ford algorithms in shortest path optimization. In *IOP Conference Series: Materials Science and Engineering* (Vol. 917, No. 1, p. 012077). IOP Publishing. https://doi.org/10.1088/1757899X-/917/1/012077

[64] Ramadiani et al (2018). Floyd-warshall algorithm to determine the shortest path based on android. In *IOP Conference Series: Earth and Environmental Science* (Vol. 144, No. 1, p. 012019). IOP Publishing. https://doi.org/10.1088/1755-1315/144/1/012019

[65] Dode, A., & Hasani, S. (2017). PageRank algorithm. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *19*(1), 01-07. https://doi.org/10.9790/0661-1901030107

[66] De Loera, J. A., Hemmecke, R., & Lee, J. (2015). On augmentation algorithms for linear and integer-linear programming: From Edmonds--Karp to Bland and beyond. *SIAM Journal on Optimization*, *25*(4), 2494-2511. https://doi.org/10.1137-/151002915

[67] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350-361. https://doi.org/10.-1016/j.neucom.2017.01.026

[68] L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *Ieee Access*, *5*, 7776-7797. https://doi.org/10.1109/ACCESS.2017.2696365

[69] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of business research*, *70*, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[70] Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth, 10(1), 13-53. https://doi.org/10.1080/17538947.2016.1239771

[71] Richtárik, P., & Takáč, M. (2016). Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, *156*, 433-484. https://doi.org/10.1007/s-10107-015-0901-6

[72] Zhang, Y., Cao, T., Li, S., Tian, X., Yuan, L., Jia, H., & Vasilakos, A. V. (2016). Parallel processing systems for big data: a survey. *Proceedings of the IEEE*, *104*(11), 2114-2136. https://doi.org/10.1109/JPROC.2016.2591592

[73] Tsai, C. F., Lin, W. C., & Ke, S. W. (2016). Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies. *Journal of Systems and Software*, *122*, 83-92. https://doi.org/10.1016/j.jss.2016.09.007

[74] Hernández, Á. B., Perez, M. S., Gupta, S., & Muntés-Mulero, V. (2018). Using machine learning to optimize parallelism in big data applications. *Future Generation Computer Systems*, *86*, 1076-1092. https://doi.org/10.1016/j.future.2017.07.003

[75] Ghomi, E. J., Rahmani, A. M., & Qader, N. N. (2017). Load-balancing algorithms in cloud computing: A survey. *Journal of Network and Computer Applications*, *88*, 50-71. https://doi.org/10.1016/j.jnca.2017.04.007

[76] Mishra, S. K., Sahoo, B., & Parida, P. P. (2020). Load balancing in cloud computing: a big picture. *Journal of King Saud University-Computer and Information Sciences*, *32*(2), 149-158. https://doi.org/10.1016/j.jksuci.2018.01.003

[77] Perwej, Y., Kerim, B., Adrees, M. S., & Sheta, O. E. (2017). An empirical exploration of the yarn in big data. *International Journal of Applied Information Systems (IJAIS)*, *12*(9), 19-29. https://doi.org/10.5120/ijais2017451730

[78] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, *8*, 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

[79] Capó, M., Pérez, A., & Lozano, J. A. (2017). An efficient approximation to the K-means clustering for massive data. *Knowledge-Based Systems*, *117*, 56-69. https://doi.org/10.1016/j.knosys.2016.06.031

[80] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448. https://doi.org/10.1016/j.jksuci.2017.06.001

[81] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data*, *2*, 1-21. https://doi.org/10.1186/s40537-014-0007-7

[82] Sethy, R., & Panda, M. (2015). Big data analysis using Hadoop: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(7).

[83] Hannan, S. A. (2016). An overview on big data and hadoop. *International Journal of Computer Applications*, *154*(10). https://doi.org/10.5120/ijca2016912241

[84] Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on Apache Spark. *International Journal of Data Science and Analytics*, *1*, 145-164. https://doi.org/10.1007/s41060-016-0027-9

[85] Xu, Y., Liu, H., & Long, Z. (2020). A distributed computing framework for wind speed big data forecasting on Apache Spark. *Sustainable Energy Technologies and Assessments*, *37*, 100582. https://doi.org/10.1016/j.seta.2019.100582

[86] Ketu, S., Mishra, P. K., & Agarwal, S. (2020). Performance analysis of distributed computing frameworks for big data analytics: hadoop vs spark. *Computación y Sistemas*, *24*(2), 669-686. https://doi.org/10.13053/CyS-24-2-3401

[87] Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., ... & De-La-Hoz-Valdiris, E. (2022). Trends and future perspective challenges in big data. In *Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania* (pp. 309-325). Springer Singapore. https://doi.org/-10.1007/978-981-16-5036-9_30

[88] Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data*, *8*(1), 24. https://doi.org/10.1186/s40537-021-00419-9

[89] Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering*, *2021*(1), 5594899. https://doi.org/10.1155/2021/5594899

[90] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178-210. https://doi.org/10.1016/j.ins.2022.11.139

[91] Potla, R. T. (2022). Scalable Machine Learning Algorithms for Big Data Analytics: Challenges and Opportunities. *Journal of Artificial Intelligence Research*, *2*(2), 124-141.

[92] Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hamad, H. A., Omari, M., Alshinwan, M., & Khasawneh, A. M. (2021). Advances in meta-heuristic optimization algorithms in big data text clustering. *Electronics*, *10*(2), 101. https://doi.org/10.3390/electronics10020101

[93] Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International journal of information management*, *48*, 63-71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[94] Gill, S. S., Kumar, A., Singh, H., Singh, M., Kaur, K., Usman, M., & Buyya, R. (2022). Quantum computing: A taxonomy, systematic review and future directions. *Software: Practice and Experience*, *52*(1), 66-114. https://doi.org/10.1002/SPE.3039