ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών
Δίκτυα Επικοινωνιών Νέας Γενιάς και Κατανεμημένα Περιβάλλοντα Εφαρμογών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σχεδίαση και Εφαρμογή Κατανεμημένου Συστήματος Επεξεργασίας Δεδομένων Κρυπτονομισμάτων σε πραγματικό χρόνο

Ιωάννης Πασπάτης
ΑΜ: 21008
Ανδριανή Κ. Ντουράκη
ΑΜ: 21015

Εισηγητής: Βασίλειος Μάμαλης, Καθηγητής

(Κενό φύλλο)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σχεδίαση και Εφαρμογή Κατανεμημένου Συστήματος Επεξεργασίας Δεδομένων Κρυπτονομισμάτων σε πραγματικό χρόνο

Ιωάννης Πασπάτης
ΑΜ: 21008
Ανδριανή Κ. Ντουράκη
ΑΜ: 21015

Εισηγητής:

Βασίλειος Μάμαλης, Καθηγητής

Vasileios Mamalis
Digitally signed by Vasileios Mamalis
Date: 2024.05.28 20:50:47 +03'00'

Εξεταστική Επιτροπή:

Γραμμάτη Πάντζιου, Καθηγήτρια
Αντώνιος Μπόγρης, Καθηγητής

Grammati Pantziou
Digitally signed by Grammati Pantziou
Date: 2024.05.29 15:36:38 +03'00'

Antonios Bogris
Digitally signed by Antonios Bogris
Date: 2024.05.29 09:30:22 +03'00'

Ημερομηνία Εξέτασης 22/05/2024

(Κενό φύλλο)

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Οι κάτωθι υπογεγραμμένοι: Ο Ιωάννης Πασπάτης του Δημητρίου, με αριθμό μητρώου 21008 και η Ανδριανή Ντουράκη του Κωνσταντίνου, με αριθμό μητρώου 21015, φοιτητές του Προγράμματος Μεταπτυχιακών Σπουδών Δίκτυα Επικοινωνιών Νέας Γενιάς και Κατανεμημένα Περιβάλλοντα Εφαρμογών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνουμε ότι:

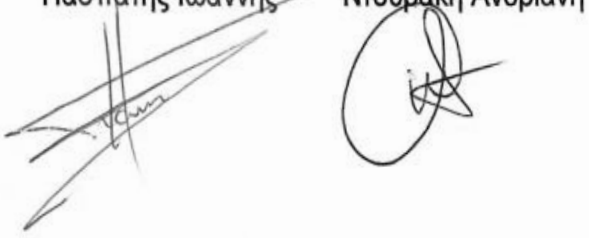
«Είμαστε συγγραφείς αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχαμε για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες κάναμε χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνουμε ότι αυτή η εργασία έχει συγγραφεί από εμάς αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μας, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μας ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μας».

Επιθυμούμε την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μας μέχρι ……και έπειτα από αίτηση μας στη Βιβλιοθήκη και έγκριση του επιβλέποντα καθηγητή.

Οι Δηλώντες

Πασπάτης Ιωάννης Ντουράκη Ανδριανή



(Κενό φύλλο)

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία ολοκληρώθηκε μετά από επίμονες προσπάθειες. Την προσπάθειά μας αυτή υποστήριξε ο επιβλέπων καθηγητής μας, τον οποίο θα θέλαμε να ευχαριστήσουμε.

(Κενό φύλλο)

ΠΕΡΙΛΗΨΗ

Αυτή η διπλωματική εργασία διερευνά τη διασταύρωση των κρυπτονομισμάτων, την τεχνική ανάλυση, τα μεγάλα δεδομένα, τη μηχανική μάθηση και τα κατανεμημένα συστήματα. Ξεκινά με μια επισκόπηση των κρυπτονομισμάτων και την τεχνική ανάλυση, εμβαθύνοντας σε διάφορους δείκτες που χρησιμοποιούνται στην ανάλυση αγοράς.

Στη συνέχεια, εμβαθύνει σε έννοιες μεγάλων δεδομένων, συμπεριλαμβανομένων των 3Vs (Volume, Velocity, Variety) και της επέκτασής τους σε 5Vs, μαζί με την αρχιτεκτονική και τα εργαλεία που χρησιμοποιούνται στην ανάλυση μεγάλων δεδομένων, εστιάζοντας ιδιαίτερα στην γλώσσα προγραμματισμού python και σε τεχνικές μηχανικής μάθησης με την χρήση της.

Επιπλέον, αναλύει τα κατανεμημένα συστήματα και εργαλεία τους όπως τα Apache Hadoop, Apache Spark, Apache Kafka και MongoDB, δίνοντας έμφαση στον ρόλο τους στη διαχείριση και την αποτελεσματική επεξεργασία μεγάλου όγκου δεδομένων.

Περιγράφεται η διαδικασία συλλογής και προεπεξεργασίας δεδομένων, συμπεριλαμβανομένης της χρήσης **Yfinance** και Kafka για τη συλλογή δεδομένων και τεχνικών για τον καθαρισμό δεδομένων και την εφαρμογή τεχνικών δεικτών.

Διερευνώνται μηχανισμοί αποθήκευσης και διαχείρισης για μεγάλα δεδομένα, επισημαίνοντας το Hadoop Distributed File System (HDFS) και το MongoDB.

Η διατριβή ολοκληρώνεται με το σχεδιασμό και την υλοποίηση ενός μοντέλου μηχανικής μάθησης χρησιμοποιώντας το Apache Spark, συμπεριλαμβανομένων στρατηγικών ανάπτυξης μοντέλων, αγωγών πρόβλεψης σε πραγματικό χρόνο και ενσωμάτωσης με το MongoDB.

Συμπερασματικά, η διατριβή συνοψίζει βασικά ευρήματα, εντοπίζει περιορισμούς και προτείνει πιθανούς τομείς για μελλοντική έρευνα στον τομέα της ανάλυσης κρυπτονομισμάτων, των μεγάλων δεδομένων και της μηχανικής μάθησης.

ABSTRACT

This thesis explores the intersection of cryptocurrencies, technical analysis, big data, machine learning and distributed systems. It starts with an overview of cryptocurrencies and technical analysis, delving into various indicators referred to in technical analysis.

It then delves into big data concepts, including the 3Vs (Volume, Velocity, Variety) and their extension to 5Vs, along with architecture and tools related to big data analysis, with a particular focus on the python programming language and engineering.

In addition, it analyzes distributed systems and their tools such as Apache Hadoop, Apache Spark, Apache Kafka, and MongoDB, emphasizing their role in managing and efficiently processing large volumes of data.

The data collection and preprocessing process is described, including the use of **Yfinance** and Kafka for data collection and techniques for data cleaning and the application of technical indicators.

Furthermore, it explores storage and management mechanisms for big data, highlighting Hadoop Distributed File System (HDFS) and MongoDB.

The thesis concludes with the design and implementation of a machine learning model using Apache Spark, including model development strategies, real-time prediction pipelines, and integration with MongoDB.

In conclusion, the thesis summarizes key findings, identifies limitations, and suggests potential areas for future research in the field of cryptocurrency analytics, big data, and machine learning.

(Κενό φύλλο)

ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ: Κατανεμημένα Περιβάλλοντα
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μεγάλα Δεδομένα, κρυπτονομίσματα, κατανεμημένα συστήματα, μηχανική μάθηση

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Οι αγορές κρυπτονομισμάτων, που χαρακτηρίζονται από την ασταθή και περίπλοκη φύση τους, παρουσιάζουν στους εμπόρους και τους επενδυτές μια διπλή πρόκληση και ευκαιρία. Η ακριβής πρόβλεψη της τιμής αποτελεί τον ακρογωνιαίο λίθο της επιτυχίας, μια προσπάθεια να ξεκλειδώσετε τις δυνατότητες για κέρδος σε αυτά τα δυναμικά νερά. Αυτό το έργο στοχεύει στο σχεδιασμό και την εκτέλεση ενός κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο, εστιάζοντας τον πυρήνα του ζωτικού καθήκοντος της πρόβλεψης τιμών. Τα αρχιτεκτονικά θεμέλια αυτού του συστήματος στηρίζονται στα εργαλεία Hadoop 3.2.1, Spark 3.0.0, Kafka 2.6 και MongoDB 4.4, τα οποία λειτουργούν συλλογικά ως η ισχυρή ραχοκοκαλιά για την απορρόφηση, την επεξεργασία, την αποθήκευση και τον έλεγχο των προβλεπόμενων τιμών.

Το τοπίο των αγορών κρυπτονομισμάτων είναι γεμάτο με έναν χείμαρρο δεδομένων συναλλαγών και αγοράς, μια αδιάκοπη ροή που απαιτεί ευέλικτη επεξεργασία και ανάλυση σε πραγματικό χρόνο. Για να αντιμετωπίσουμε αυτόν τον κατακλυσμό δεδομένων, απευθυνόμαστε στο Hadoop HDFS, ένα κατανεμημένο σύστημα αρχείων που φημίζεται για την ικανότητα του στο χειρισμό ογκωδών δεδομένων κρυπτονομισμάτων. Πέρα από την απόλυτη χωρητικότητα αποθήκευσης, το Hadoop προσθέτει ένα επίπεδο ανθεκτικότητας και υψηλής απόδοσης, ενισχύοντας την επεκτασιμότητα που απαιτείται για την αποτελεσματική επεξεργασία και διατήρηση δεδομένων.

Επιδιώκοντας την πρόβλεψη τιμών σε πραγματικό χρόνο, υιοθετούμε το Spark 3.0.0 ως την βέλτιστη επιλογή. Η ικανότητα του Spark για επεξεργασία και παραλληλοποίηση στη μνήμη μας φέρνει στη σφαίρα της ανάλυσης σε πραγματικό χρόνο, ακόμη και όταν έχουμε να κάνουμε με σημαντικά σύνολα δεδομένων. Η αξιοποίηση των βιβλιοθηκών μηχανικής μάθησης του Spark εξουσιοδοτεί το σύστημά μας να δημιουργεί προγνωστικά μοντέλα χρησιμοποιώντας ιστορικά δεδομένα

κρυπτονομισμάτων και να δημιουργεί προβλέψεις τιμών σε πραγματικό χρόνο. Η εγγενής κατανεμημένη φύση του Spark διευκολύνει τη διαδρομή προς την επεκτασιμότητα, διασφαλίζοντας ότι το σύστημα μπορεί να πλοηγηθεί επιδέξια στις ροές δεδομένων κρυπτονομισμάτων υψηλής ταχύτητας.

Για τη ζωτική εργασία της απορρόφησης δεδομένων σε πραγματικό χρόνο, το Kafka 2.6 χρησιμεύει ως η κατανεμημένη πλατφόρμα ροής μας. Ο Kafka απλοποιεί τη συλλογή και ροή δεδομένων κρυπτονομισμάτων από διαφορετικές πηγές, διασφαλίζοντας ταυτόχρονα ανοχή σφαλμάτων και υψηλή απόδοση. Η κατανεμημένη αρχιτεκτονική του παρέχει στο σύστημα την επεκτασιμότητα και την αξιοπιστία που απαιτούνται για εφαρμογές ροής, επιτρέποντάς του να απορροφά και να επεξεργάζεται συνεχώς τα πιο πρόσφατα δεδομένα της αγοράς για ακριβή πρόβλεψη τιμών.

Η καρδιά των αναλυτικών μας γνώσεων βρίσκεται στο MongoDB 4.4, μια βάση δεδομένων NoSQL που είναι γνωστή για τις δομημένες δυνατότητες αποθήκευσης και ανάκτησης δεδομένων υψηλής απόδοσης. Προσφέρει ένα ιδανικό καταφύγιο για τη διαφύλαξη των προβλεπόμενων αξιών κρυπτονομισμάτων. Το ευέλικτο μοντέλο δεδομένων της MongoDB και οι ισχυρές δυνατότητες ερωτημάτων ενισχύουν την αποτελεσματική ανάλυση και οπτικοποίηση αυτών των προβλέψεων, παρέχοντας στους εμπόρους και στους επενδυτές τις γνώσεις βάσεων δεδομένων για τη λήψη τεκμηριωμένων αποφάσεων.

Αυτή η ένωση των Hadoop, Spark, Kafka και MongoDB συγκεντρώνει ένα κατανεμημένο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο με την υπόσχεση επεκτασιμότητας, ανοχής σφαλμάτων και αποτελεσματικότητας στην ακριβή πρόβλεψη τιμών. Το σύστημα αποκαλύπτει πληροφορίες σε πραγματικό χρόνο σχετικά με τη συνεχώς εξελισσόμενη δυναμική των κρυπτονομισμάτων, παρέχοντας στους χρήστες του τη δύναμη να λαμβάνουν ενημερωμένες αποφάσεις συναλλαγών που βασίζονται σε προβλέψεις που είναι χαραγμένες στο MongoDB.

Συμπερασματικά, αυτό το έργο προσπαθεί να γεννήσει ένα προηγμένο κατανεμημένο σύστημα, αξιοποιώντας τη συλλογική δύναμη των Hadoop, Spark, Kafka και MongoDB για να πλοηγηθεί στον περίπλοκο κόσμο της πρόβλεψης τιμών κρυπτονομισμάτων σε πραγματικό χρόνο. Οπλισμένο με εργαλεία αιχμής, το σύστημα είναι έτοιμο να αντιμετωπίσει τις προκλήσεις που είναι εγγενείς στην πρόβλεψη τιμών

κρυπτονομισμάτων, εγκαινιάζοντας μια νέα εποχή δυνατοτήτων τόσο για τους εμπόρους όσο και για τους επενδυτές.

1.1 Περιγραφή του αντικειμένου της διπλωματικής εργασίας

Η αγορά κρυπτονομισμάτων, μια σφαίρα άνευ προηγουμένου ανάπτυξης και αστάθειας, έχει γοητεύσει τους εμπόρους, τους επενδυτές και τους ερευνητές σε όλο τον κόσμο. Αποτελεί μια δυναμική αρένα όπου τα ψηφιακά περιουσιακά στοιχεία κυμαίνονται με εκπληκτική ταχύτητα, καθιστώντας την πρακτική της πρόβλεψης τιμών όχι μόνο κρίσιμη αλλά, κατά καιρούς, τον βασικό καθοριστικό παράγοντα της επιτυχίας. Ωστόσο, μέσα σε αυτό το ζωντανό οικοσύστημα, οι συμβατικές λύσεις συχνά παραπαίουν, αδυνατώντας να αντιμετωπίσουν τον κολοσσιαίο και συνεχώς αυξανόμενο όγκο δεδομένων κρυπτονομισμάτων, αποτυγχάνοντας να προσφέρουν την επεκτασιμότητα και τις δυνατότητες σε πραγματικό χρόνο που απαιτούνται από αυτό το γρήγορο βασίλειο.

Το αναμφισβήτητο κίνητρο πίσω από αυτό το εγχείρημα πηγάζει από την αναγνώριση ότι η αγορά κρυπτονομισμάτων λειτουργεί σε διαφορετικό ρολόι - όπου οι ώρες μπορεί να φαίνονται σαν απλά δευτερόλεπτα και ένα κλάσμα του δευτερολέπτου μπορεί να διαμορφώσει περιουσίες. Είναι μια σφαίρα όπου τα καλά τεκμηριωμένα ιστορικά δεδομένα και οι έγκαιρες προβλέψεις δεν είναι απλώς επιθυμητές, αλλά επιβεβλημένες για όσους πλοηγούνται στα δυναμικά της ρεύματα. Σε απάντηση, αυτό το έργο αναδύεται με έναν σαφή και αποφασιστικό στόχο: να αντιμετωπίσει αυτές τις προκλήσεις κατά μέτωπο και να γεφυρώσει τα υπάρχοντα κενά στο τοπίο δεδομένων κρυπτονομισμάτων. Επιδιώκει να το πράξει δημιουργώντας και εκτελώντας ένα κατανεμημένο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο, εστιασμένο με λέιζερ στο βασικό καθήκον της πρόβλεψης τιμών.

Μέσα στο χαοτικό και συνεχώς εξελισσόμενο τοπίο των κρυπτονομισμάτων, όπου η ακρίβεια και η ευελιξία είναι τα νομίσματα της επιβίωσης, η ανάγκη για ολοκληρωμένη τεκμηρίωση και έγκαιρες γνώσεις είναι εμφανής. Η αποστολή αυτού του έργου δεν είναι παρά να ενδυναμώσει τους εμπόρους, τους επενδυτές και τους ερευνητές με τα εξελιγμένα εργαλεία και τις γνώσεις που βασίζονται σε δεδομένα που είναι απαραίτητες σε αυτήν την ψηφιακή αρένα. Με τη συνέργεια των τεχνολογιών αιχμής και την ακλόνητη αφοσίωση, αυτό το έργο φιλοδοξεί όχι μόνο να ανταποκριθεί

στις προκλήσεις της πρόβλεψης τιμών κρυπτονομισμάτων αλλά να θέσει ένα νέο πρότυπο για τη λήψη αποφάσεων βάσει δεδομένων σε αυτή τη δυναμική και συναρπαστική αγορά.

1.2 Δήλωση Προβλήματος και Ερευνητική Ερώτηση

Στην καρδιά αυτού του έργου βρίσκεται μια θεμελιώδης πρόκληση - η σπανιότητα συστημάτων πρόβλεψης τιμών κρυπτονομισμάτων που ενσωματώνουν απρόσκοπτα τα βασικά χαρακτηριστικά της επεκτασιμότητας και της δυνατότητας σε πραγματικό χρόνο. Οι υπάρχουσες λύσεις παλεύουν με τον αδιάκοπο κατακλυσμό δεδομένων κρυπτονομισμάτων, που συχνά βρίσκονται παγιδευμένοι στον ιστό των ροών δεδομένων υψηλής ταχύτητας, προσπαθώντας να παρέχουν ακριβείς προβλέψεις με τον ιλιγγιώδη ρυθμό που απαιτεί η δυναμική αγορά.

Για να αντιμετωπίσει αυτήν την πιεστική πρόκληση, αυτό το έργο ξεκινά ένα μετασχηματιστικό ταξίδι που καθοδηγείται από ένα κεντρικό ερευνητικό ερώτημα:

Ερευνητικό ερώτημα: Πώς μπορεί ένα κατανεμημένο σύστημα να αξιοποιήσει την ισχυρή συνέργεια των Hadoop HDFS, Spark, Kafka και MongoDB για να προβλέψει την τιμή κρυπτονομισμάτων σε πραγματικό χρόνο;

Σε αυτήν την αναζήτηση, στοχεύουμε να πρωτοστατήσουμε σε μια λύση που όχι μόνο πλοηγείται στην πολυπλοκότητα της αγοράς κρυπτονομισμάτων, αλλά περιλαμβάνει επίσης τις δίδυμες επιταγές της επεκτασιμότητας και της επεξεργασίας σε πραγματικό χρόνο. Το ταξίδι μας ξεκινά συνδυάζοντας τα δυνατά σημεία του Hadoop HDFS για αποθήκευση δεδομένων, του Spark για αστραπιαίους υπολογισμούς, του Kafka για απρόσκοπτη απορρόφηση δεδομένων και του MongoDB για αποτελεσματική αποθήκευση και ανάλυση δεδομένων.

Η εξερεύνηση μας καθοδηγείται από τη δέσμευσή μας να δημιουργήσουμε ένα σύστημα που εξουσιοδοτεί τους συμμετέχοντες στην αγορά και τους αναλυτές με έγκαιρες και ακριβείς πληροφορίες, επιτρέποντάς τους να περιηγηθούν στον συνεχώς εξελισσόμενο κόσμο των κρυπτονομισμάτων με σιγουριά.

Καθώς εμβαθύνουμε στις περιπλοκές αυτού του ερευνητικού ερωτήματος, θα αποκαλύψουμε ένα κατανεμημένο σύστημα που όχι μόνο προβλέπει τις τιμές κρυπτονομισμάτων αλλά εισάγει επίσης μια νέα εποχή επεκτασιμότητας σε πραγματικό χρόνο, μεταμορφώνοντας τον τρόπο που αλληλεπιδρούμε με το δυναμικό τοπίο των κρυπτονομισμάτων.

1.3 Επισκόπηση της Διαδικασίας Σχεδιασμού και Εφαρμογής

Η διαδικασία σχεδιασμού και υλοποίησης του κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο περιλαμβάνει αρκετά κρίσιμα στοιχεία και τεχνολογίες. Αυτό το έργο αξιοποιεί το Hadoop HDFS ως το κατανεμημένο σύστημα αρχείων για αποτελεσματική αποθήκευση και ανάκτηση εκτεταμένων όγκων δεδομένων κρυπτονομισμάτων. Χρησιμοποιεί το Spark, ένα ισχυρό κατανεμημένο υπολογιστικό πλαίσιο, το οποίο χρησιμοποιεί τις δυνατότητες επεξεργασίας στη μνήμη και τις βιβλιοθήκες μηχανικής εκμάθησης για ανάλυση δεδομένων σε πραγματικό χρόνο και πρόβλεψη τιμών. Το Kafka χρησιμεύει ως η κατανεμημένη πλατφόρμα ροής, επιτρέποντας τη συνεχή απορρόφηση δεδομένων κρυπτονομισμάτων από διάφορες πηγές. Τέλος, το MongoDB λειτουργεί ως πλατφόρμα αποθήκευσης και ανάλυσης για τις προβλεπόμενες τιμές.

Ο σχεδιασμός του συστήματος είναι οργανωμένος σε διάφορες ενότητες, καθεμία προσαρμοσμένη για να χειρίζεται συγκεκριμένες πτυχές της επεξεργασίας δεδομένων κρυπτονομισμάτων και τις σχετικές προκλήσεις. Αυτές οι ενότητες περιλαμβάνουν:

- Αφομοίωση Δεδομένων: Αυτή η ενότητα είναι υπεύθυνη για τη συλλογή δεδομένων κρυπτονομισμάτων από πολλαπλές πηγές. Μπορεί να περιλαμβάνει χρήση API, απόξεση ιστού ή άλλες μεθόδους συλλογής δεδομένων.

- Προεπεξεργασία: Η προεπεξεργασία δεδομένων είναι ζωτικής σημασίας για τον καθαρισμό και τη μορφοποίηση των εισερχόμενων δεδομένων, διασφαλίζοντας ότι είναι έτοιμα για ανάλυση. Για το σκοπό αυτό χρησιμοποιούνται συχνά βιβλιοθήκες Python όπως η Pandas και η NumPy.

- Εξαγωγή χαρακτηριστικών: Η μηχανική χαρακτηριστικών είναι ένα κρίσιμο βήμα για τη δημιουργία σημαντικών μεταβλητών εισόδου για μοντέλα μηχανικής μάθησης. Εδώ, ενδέχεται να εξαγάγουμε τεχνικούς δείκτες ή άλλα σχετικά χαρακτηριστικά από τα δεδομένα.

- Εκπαίδευση μοντέλου: Αυτή η ενότητα εστιάζει στην εκπαίδευση μοντέλων μηχανικής μάθησης, όπως μοντέλα γραμμικής παλινδρόμησης ή βαθιάς μάθησης, χρησιμοποιώντας ιστορικά δεδομένα. Το PySpark και το Spark MLlib είναι ανεκτίμητα εργαλεία για αυτήν την εργασία.

- Συμπεράσματα σε πραγματικό χρόνο: Η ενότητα προβληματισμού σε πραγματικό χρόνο εφαρμόζει εκπαιδευμένα μοντέλα σε εισερχόμενα δεδομένα ροής για να κάνει προβλέψεις για τις τιμές κρυπτονομισμάτων.

Η διαδικασία υλοποίησης περιλαμβάνει τις ακόλουθες βασικές δραστηριότητες:

- Ενσωμάτωση επιλεγμένων εργαλείων: Τα επιλεγμένα εργαλεία, συμπεριλαμβανομένων των Hadoop HDFS, Spark, Kafka και MongoDB, είναι ενσωματωμένα στην αρχιτεκτονική του συστήματος.

- Διαμόρφωση για κατανεμημένη επεξεργασία: Οι ρυθμίσεις παραμέτρων έχουν ρυθμιστεί για να διασφαλίζεται ότι το σύστημα μπορεί να χειριστεί αποτελεσματικά την κατανεμημένη επεξεργασία. Για παράδειγμα, η διαμόρφωση Spark μπορεί να περιλαμβάνει τον καθορισμό των ρυθμίσεων και των πόρων του συμπλέγματος.

- Ανάπτυξη αλγορίθμων και αγωγών: Αναπτύσσονται οι απαραίτητοι αλγόριθμοι για την προεπεξεργασία δεδομένων, την εξαγωγή χαρακτηριστικών και την εκπαίδευση μοντέλων. Οι αγωγοί Spark κατασκευάζονται για να εξορθολογίσουν την επεξεργασία δεδομένων.

Συνδυάζοντας αυτές τις τεχνολογίες και ενότητες, το σύστημα μπορεί να διαχειριστεί και να αναλύσει αποτελεσματικά δεδομένα κρυπτονομισμάτων σε πραγματικό χρόνο, διευκολύνοντας την πρόβλεψη τιμών και άλλες πολύτιμες πληροφορίες.

Αυτή η ολοκληρωμένη διαδικασία σχεδιασμού και εφαρμογής αποτελεί τη βάση για την επιτυχή εφαρμογή ενός κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Επιτρέπει στο σύστημα να απορροφά, να προεπεξεργάζεται, να αναλύει και να αποθηκεύει δεδομένα κρυπτονομισμάτων

απρόσκοπτα, παρέχοντας πολύτιμες πληροφορίες για συμμετέχοντες και αναλυτές στην αγορά κρυπτονομισμάτων.

ΚΕΦΑΛΑΙΟ 2

ΚΡΥΠΤΟΝΟΜΙΣΜΑΤΑ ΚΑΙ ΤΕΧΝΙΚΗ ΑΝΑΛΥΣΗ

Η κατανόηση των κινήσεων και των τάσεων στις τιμές των κρυπτονομισμάτων είναι μια κρίσιμη πτυχή για τους επενδυτές και τους αναλυτές και η τεχνική ανάλυση παρέχει μια δομημένη προσέγγιση για την επίτευξη αυτής της εικόνας. Αυτή η ενότητα εμβαθύνει σε διάφορους τεχνικούς δείκτες που παίζουν καθοριστικό ρόλο στην κατανόηση της δυναμικής της αγοράς και στην πρόβλεψη μελλοντικών κινήσεων των τιμών στη σφαίρα των κρυπτονομισμάτων. Οι ακόλουθοι δείκτες αποτελούν ένα θεμελιώδες μέρος του κατανεμημένου συστήματος που χρησιμοποιείται για την πρόβλεψη των τιμών κρυπτονομισμάτων, καθένας από τους οποίους συζητείται λεπτομερώς:

- *Κινητός μέσος όρος (MA)*
- *Εκθετικός κινητός μέσος όρος (EMA)*
- *Κινητός μέσος όρος σύγκλισης απόκλισης (MACD)*
- *Δείκτης σχετικής ισχύος (RSI)*
- *Δείκτης Vortex*
- *Average True Range (ATR)*
- *Awesome Oscillator (AO)*
- *Δείκτης καναλιών εμπορευμάτων (CCI)*
- *Μέσος Δείκτης Κατεύθυνσης (ADX)*
- *Williams %R*
- *Στοχαστικός Ταλαντωτής (%K και %D)*
- *On Balance Volume (OBV)*

Αυτό το ολοκληρωμένο σύνολο δεικτών εξοπλίζει τους επενδυτές και τους αναλυτές με πολύτιμα εργαλεία για την πλοήγηση στον περίπλοκο και δυναμικό κόσμο των

αγορών κρυπτονομισμάτων. Με την κατανόηση και τη χρήση αυτών των δεικτών, οι συμμετέχοντες στην αγορά μπορούν να ενισχύσουν την ικανότητά τους να λαμβάνουν τεκμηριωμένες αποφάσεις και να προβλέπουν μελλοντικές κινήσεις τιμών στο συνεχώς εξελισσόμενο τοπίο των κρυπτονομισμάτων.

2.1 Κρυπτονομίσματα

Σύμφωνα με την Ευρωπαϊκή Επιτροπή, ο όρος «κρυπτονομίσματα» μπορεί να είναι παραπλανητικός, καθώς υπονοεί μια μορφή νομίσματος που στερείται βασικών χαρακτηριστικών του παραδοσιακού χρήματος, όπως η κεντρική ρύθμιση και η σταθερότητα. Τα ψηφιακά στοιχεία, που συχνά αναφέρονται ως «κρυπτογραφικά στοιχεία» ή «κρυπτογραφικές μάρκες», είναι ουσιαστικά μοναδικές κρυπτογραφικές αλφαριθμητικές συμβολοσειρές που παρέχουν έλεγχο σε ένα κομμάτι κώδικα υπολογιστή. Παρά την ασταθή φύση τους, αυτά τα περιουσιακά στοιχεία έχουν προσελκύσει σημαντικές παγκόσμιες επενδύσεις, όπως αποδεικνύεται από την αύξηση της αξίας του Bitcoin σε πάνω από 15.000 ευρώ στις αρχές του 2018 [1].

Συμφώνα με την Βιβλιοθήκη του Κογκρέσου, το κρυπτονόμισμα χρησιμεύει ως μέσο ανταλλαγής, αποθήκευση αξίας και μονάδα μέτρησης. Ενώ τα κρυπτονομίσματα έχουν μικρή εγγενή αξία, χρησιμοποιούνται για την τιμολόγηση της αξίας άλλων περιουσιακών στοιχείων. Το Bitcoin, που κυκλοφόρησε το 2009 και θεωρείται ευρέως το πρώτο ψηφιακό περιουσιακό στοιχείο, χρησιμεύει τόσο ως μέσο πληρωμής όσο και ως κερδοσκοπικό εμπόρευμα. Τα ψηφιακά περιουσιακά στοιχεία, γνωστά και ως περιουσιακά στοιχεία κρυπτογράφησης, είναι κρυπτογραφικές αναπαραστάσεις αξίας με δυνατότητα blockchain. Αρχικά προορίζονταν να διευκολύνουν τη μεταφορά αξίας χωρίς την ανάγκη για αξιόπιστη οντότητα τρίτου μέρους. Τα περιουσιακά στοιχεία κρυπτογράφησης κατηγοριοποιούνται σε τρεις κύριους τύπους: κρυπτονομίσματα, εμπορεύματα κρυπτογράφησης και μάρκες κρυπτογράφησης [2].

Ενώ τα κρυπτονομίσματα μπορεί να μην εξελιχθούν σε μια σημαντική κατηγορία χρηματοοικονομικών περιουσιακών στοιχείων, η υποκείμενη τεχνολογία blockchain δείχνει υπόσχεση για τον μετασχηματισμό των επιχειρηματικών πρακτικών. Η αγορά περιλαμβάνει περίπου 1.500 διαφορετικά μάρκες κρυπτογράφησης, με σημαντικούς παίκτες όπως το Bitcoin, το Ethereum και άλλους να κυριαρχούν πάνω από το 80% της αγοράς [1].

Τα κρυπτονομίσματα διαπραγματεύονται ενεργά σε χρηματιστήρια κρυπτογράφησης, με την Ασία να φιλοξενεί τις περισσότερες μεγάλες πλατφόρμες. Εκτός από τις συναλλαγές, τα crypto-tokens έχουν εισαγάγει νέες μεθόδους συγκέντρωσης κεφαλαίων μέσω Αρχικών Προσφορών Νομισμάτων (ICO). Το 2017, οι ICO συγκέντρωσαν περίπου 3,9 δισεκατομμύρια δολάρια, με το ποσό να φτάνει τα 6 δισεκατομμύρια δολάρια το πρώτο τρίμηνο του 2018. Η Ευρώπη αντιπροσωπεύει επί του παρόντος ένα σχετικά μικρό μερίδιο των παγκόσμιων συναλλαγών κρυπτονομισμάτων [1].

Ένα αναδυόμενο θέμα είναι η έννοια των stablecoin, των κρυπτονομισμάτων που είναι συνδεδεμένα με ένα σταθερό περιουσιακό στοιχείο όπως το δολάριο των Ηνωμένων Πολιτειών της Αμερικής, το οποίο μπορεί να διαδραματίσει κρίσιμο ρόλο στην αποκεντρωμένη χρηματοδότηση [2].

Η Ελλάδα ανταποκρινόμενη στο αυξανόμενο ενδιαφέρον για τα κρυπτονομίσματα, το 2014, μέσω της Τράπεζας της Ελλάδος εξέδωσε μια προειδοποίηση σχετικά με τους κινδύνους που συνδέονται με τη χρήση εικονικών νομισμάτων όπως το Bitcoin, συμπεριλαμβανομένης της πιθανής απώλειας χρημάτων. Αυτό συνάδει με παρόμοια ανακοίνωση της Ευρωπαϊκής Αρχής Τραπεζών (EBA) [3]. Τέσσερα χρονιά αργότερα, το 2018, η Ελλάδα εντάχθηκε στην European Blockchain Partnership (EBP). Το EBP στοχεύει στη δημιουργία μιας Ευρωπαϊκής Υποδομής Υπηρεσιών Blockchain (EBSI) για την υποστήριξη της παροχής διασυνοριακών ψηφιακών δημόσιων υπηρεσιών, με ταυτόχρονη τήρηση των καθιερωμένων προτύπων για την ασφάλεια και το απόρρητο [4].

2.2 Τεχνική Ανάλυση

Η τεχνική ανάλυση είναι μια μεθοδολογία που χρησιμοποιείται στις χρηματοπιστωτικές αγορές για την αξιολόγηση και την πρόβλεψη μελλοντικών κινήσεων των τιμών μέσω της εξέτασης ιστορικών δεδομένων τιμών, όγκου, τάσεων και ανοιχτού ενδιαφέροντος [5] [6]. Αυτή η προσέγγιση βασίζεται σε πολλές θεμελιώδεις αρχές που καθοδηγούν την εφαρμογή της.

- *Η τιμή καθορίζει τα πάντα:* Στον πυρήνα της τεχνικής ανάλυσης είναι η προϋπόθεση ότι η δράση της αγοράς ενσωματώνει πλήρως όλες τις σχετικές πληροφορίες, που καλύπτουν οικονομικούς, πολιτικούς, ψυχολογικούς και

άλλους παράγοντες [5]. Αυτή η θεμελιώδης ιδέα, που διευκρινίστηκε από τον John J. Murphy στην εργασία του για την τεχνική ανάλυση, θέτει την αγορά ως ένα αποτελεσματικό σύστημα επεξεργασίας πληροφοριών.

- *Οι τιμές κινούνται σε τάσεις:* Η δεύτερη κατευθυντήρια αρχή υποστηρίζει ότι οι τιμές παρουσιάζουν τάσεις και όχι τυχαίες κινήσεις [5]. Μόλις καθιερωθεί, μια τάση αναμένεται να συνεχιστεί έως ότου μια εξωτερική δύναμη πυροδοτήσει μια αντιστροφή. Οι τεχνικοί αναλυτές επικεντρώνονται στον εντοπισμό, την κατανόηση και τη μόχλευση αυτών των τάσεων για τη λήψη τεκμηριωμένων αποφάσεων συναλλαγών.

    - *Πρωτογενής τάση:* Αυτή η μακροχρόνια τάση, που εκτείνεται από 9 μήνες έως 2 χρόνια, αντανακλά τα συναισθήματα των επενδυτών προς τα εκτυλισσόμενα θεμελιώδη στοιχεία στον επιχειρηματικό κύκλο, επηρεάζοντας το μέγεθος και τη διάρκεια των αγορών bull και bear [6].

    - *Ενδιάμεση τάση:* Διακόπτοντας τις πρωτογενείς ανόδους, αυτή η τάση διαρκεί από 6 εβδομάδες έως 9 μήνες, συχνά είναι παραπλανητική και βασίζεται σε ψευδείς υποθέσεις, επηρεάζοντας τα ποσοστά επιτυχίας των συναλλαγών και απαιτεί προσεκτική ανάλυση [6].

    - *Βραχυπρόθεσμη τάση:* Διαρκούν 3 έως 6 εβδομάδες, αυτές οι τάσεις διακόπτουν τον ενδιάμεσο κύκλο και επηρεάζονται από τυχαία γεγονότα ειδήσεων, θέτοντας μεγαλύτερη πρόκληση για αναγνώριση σε σύγκριση με τις ενδιάμεσες ή τις πρωταρχικές τάσεις [6].

- *Η ιστορία επαναλαμβάνεται:* Χτισμένη στην πίστη στην επανάληψη ορισμένων ιστορικών προτύπων και συμπεριφορών, η τρίτη αρχή προτείνει ότι η ανθρώπινη ψυχολογία παίζει σημαντικό ρόλο στη διαμόρφωση της δυναμικής της αγοράς [5]. Όπως περιγράφεται από τον Murphy, η κατανόηση των ιστορικών προτύπων βοηθά στην πρόβλεψη της μελλοντικής συμπεριφοράς της αγοράς.

Στην τεχνική ανάλυση, η βασική υπόθεση υποστηρίζει ότι οι μακροπρόθεσμες κινήσεις των τιμών αντικατοπτρίζουν την προσφορά και τη ζήτηση, αλλά για τη βραχυπρόθεσμη ανάλυση, η χαοτική και παράλογη φύση της δυναμικής της αγοράς επηρεάζεται από αντιδράσεις σε πληροφορίες, τόσο πραγματικές όσο και φημολογούμενές, από ένα μείγμα θεμελιωδών και τεχνικών στοιχείων και την παράλογη συμπεριφορά των επενδυτών. Γεγονός που υποδηλώνει ότι το να βασίζεσαι αποκλειστικά στην προσφορά και τη ζήτηση μπορεί να μην είναι αξιόπιστο [7].

Στην τεχνική ανάλυση, η αιτία πίσω από την άνοδο ή την πτώση της τιμής μιας μετοχής θεωρείται λιγότερο σημαντική από τη διάρκεια αυτής της αλλαγής, είτε είναι ανοδική είτε καθοδική. Μια άλλη θεμελιώδης υπόθεση στην τεχνική ανάλυση είναι ότι η τιμή μιας μετοχής αντανακλά τα οικονομικά δεδομένα, τις πληροφορίες και τις ειδήσεις της, αλλά επηρεάζεται σημαντικά από τα κυρίαρχα ανθρώπινα συναισθήματα και την ψυχολογία. Η τεχνική ανάλυση εστιάζει αποκλειστικά σε γραφήματα, που προέρχονται από την Αναλυτική Γεωμετρία και τη Στατιστική, αντί να εμβαθύνουμε απαραίτητα σε θεμελιώδεις πτυχές μιας εταιρείας, όπως ισολογισμούς, αναλογία P/E ή κέρδη [8].

Οι αναλυτές και οι επενδυτές, αναγνωρίζοντας την αβεβαιότητα του μέλλοντος, επιδιώκουν να αποκαλύψουν πρότυπα, απο συμπεριφορές του παρελθόντος για να διαμορφώσουν μια βέλτιστη επενδυτική στρατηγική. Οι προηγούμενες κινήσεις των τιμών είναι ζωτικής σημασίας για τον προσδιορισμό της εγκυρότητας των τρεχουσών μετατοπίσεων τιμών, καθώς η μελέτη ιστορικών δεδομένων βοηθά στον εντοπισμό πιθανών μελλοντικών αλλαγών και στον εντοπισμό σημείων αντιστροφής. Αυτή η προσέγγιση πηγάζει από την πεποίθηση ότι οι ανθρώπινες συμπεριφορές ακολουθούν κυκλικά μοτίβα και οι άνθρωποι τείνουν να αντιδρούν παρόμοια σε συγκρίσιμα ερεθίσματα κατά τη διάρκεια των ετών [6] [7] [8].

Ο Charles Dow, ο εκδότης της Wall Street Journal, πιστώνεται ως ο «πατέρας» της τεχνικής ανάλυσης για την σύλληψη του βιομηχανικού μέσου όρου Dow Jones ως δείκτη που αντιπροσωπεύει τη συλλογική κίνηση μιας ομάδας μετοχών, που χρησιμεύει ως βαρόμετρο της αγοράς. Αυτό παραλληλίζει τις παρατηρήσεις στον οικονομικό τομέα, όπου οι επιστήμονες έχουν εντοπίσει φάσεις του οικονομικού κύκλου: άνοδος ή έκρηξη, κρίση, κάθοδος ή ύφεση και ανάκαμψη [5] [6] [7].

Οι προαναφερθείσες πληροφορίες υπογραμμίζουν ότι, σύμφωνα με την Τεχνική Ανάλυση, η δυναμική της αγοράς δεν ωθείται κυρίως από γεγονότα αλλά από τις

προσδοκίες των ανθρώπων. Οι τιμές των μετοχών στους πίνακες της αγοράς θεωρούνται ως προσδοκίες για το μέλλον, αντί να αντανακλούν το παρόν. Τα ανθρώπινα συναισθήματα, που περιλαμβάνουν παράγοντες όπως η απληστία, ο φόβος, η μίμηση και η επιθυμία για ομαδική ένταξη, επηρεάζουν σημαντικά τη συμπεριφορά σε διάφορες αγορές, οδηγώντας συχνά σε αποφάσεις που αργότερα θεωρούνται ανεπιτυχείς. Η Τεχνική Ανάλυση υποστηρίζει ότι αυτές οι συναισθηματικές επιρροές εκδηλώνονται στους σχηματισμούς τιμών, που αναλύονται μέσω διαγραμμάτων για να απομονωθούν οι ψυχολογικές επιπτώσεις στην αγορά.

Σε αυτό το πλαίσιο, οι αναρτήσεις του Έλον Μασκ στο Twitter χρησιμεύουν ως χαρακτηριστικό παράδειγμα, άσκησης επιρροής στη δυναμική της αγοράς. Πολλά άρθρα από τα μεγαλύτερα ειδησεογραφικά πρακτορεία αναφέρουν πώς τα tweets του Μασκ, έχουν επηρεάσει την αγορά. Ο Economist εμβαθύνει στις επιπτώσεις των tweet του Μασκ για την εξαγορά του Twitter [9] [10], ενώ το Forbes εξετάζει τις ξεχωριστές στρατηγικές δέσμευσης στα μέσα κοινωνικής δικτύωσης και τις ευρύτερες επιχειρηματικές τους επιπτώσεις [11] [12]. Η Wall Street Journal και το Bloomberg αναφέρουν τις δηλώσεις του Μασκ σχετικά με τις οικονομικές πτυχές της εξαγοράς και τις επιπτώσεις της στο εργατικό δυναμικό και τα έσοδα του Twitter [13] [14] [15] [16]. Το CNN συγκεντρώνει τα tweets του Μασκ για το Twitter όλα αυτά τα χρόνια, προσφέροντας πληροφορίες για την εξελισσόμενη στάση του [17]. Οι Financial Times συζητούν τις επιπτώσεις της εξαγοράς του Μασκ στην οικονομική απόδοση και τη συνολική κατάσταση του Twitter [18]. Η δραστηριότητα του Έλον Μασκ στο Twitter γίνεται έτσι ένα συναρπαστικό παράδειγμα που απεικονίζει την αλληλεπίδραση μεταξύ των ανθρώπινων συναισθημάτων, των αντιδράσεων της αγοράς και των αρχών της Τεχνικής Ανάλυσης.

2.2.1 Τα Θεμέλια Της Τεχνικής Ανάλυσης

Η βάση της τεχνικής ανάλυσης βρίσκεται στην κατανόηση της τιμής των χρηματοπιστωτικών μέσων, παρόμοια με το DNA που μεταφέρει τη γενετική πληροφορία των ζωντανών οργανισμών. Η τιμή, το βασικό δομικό στοιχείο ενός γραφήματος, ενσωματώνει τη συλλογική γνώση και τα συναισθήματα των επενδυτών σχετικά με ένα συγκεκριμένο περιουσιακό στοιχείο. Κάθε μπάρα τιμής, σχεδιασμένη με τη σειρά, κατασκευάζει την οπτική αναπαράσταση της τάσης ενός γραφήματος. Σε

αυτή την περίπλοκη ταπετσαρία, έξι κρίσιμα στοιχεία ξεδιπλώνονται σε μια μπάρα τιμών [30]:

- *Το Άνοιγμα (Open):* Σηματοδοτεί το σημείο ανοίγματος της συνεδρίασης, μετά την ανασκόπηση των συνθηκών αγοράς μιας νύχτας.

- *Το Υψηλό (High):* Σηματοδοτεί το Υψηλότερο σημείο της συνεδρίασης, δημιουργώντας μια περιοχή αντίστασης για τους αγοραστές.

- *Το χαμηλό (Low):* Σηματοδοτεί το χαμηλότερο σημείο της συνεδρίασης, αντιπροσωπεύοντας μια ζώνη υποστηρικτικής ζήτησης που αποτρέπει περαιτέρω πτώση των τιμών.

- *Το Κλείσιμο (Close):* Σηματοδοτεί το σημείο κλεισίματος της συνεδρίασης.

- *Η Αλλαγή (Change):* Σηματοδοτεί την διαφορά από κλείσιμο σε κλείσιμο της συνεδρίασης, προσφέρει πληροφορίες για την επικρατούσα δυναμική ζήτησης ή προσφοράς.

- *Το Εύρος (Range):* Σηματοδοτεί το διάστημα μεταξύ του υψηλότερου και του χαμηλότερου σημείου σε μια συνεδρία συναλλαγών. Χρησιμεύει ως συνοπτικό μέτρο της αστάθειας των τιμών, μεταφέροντας πληροφορίες σχετικά με την ένταση της προσφοράς και της ζήτησης.

- *Όσον αφορά τον όγκο (Volume):* Ποσοτικοποιεί το ποσό μιας εμπορεύσιμης οντότητας που ανταλλάσσεται μεταξύ αγοραστών και πωλητών. Ο υψηλός όγκος υποδηλώνει ότι περισσότερες μονάδες αλλάζουν ιδιοκτησία ενώ ο χαμηλός το αντίστροφο [30].

2.2.2 Τι είναι οι Δείκτες

Ο όρος "δείκτης" στο πλαίσιο της τεχνικής ανάλυσης αναφέρεται σε έναν υπολογισμό που εφαρμόζεται σε ένα διάγραμμα τιμών για τον προσδιορισμό συγκεκριμένων γεγονότων και χαρακτηριστικών, εστιάζοντας κυρίως στο αν η τιμή είναι σε τάση, στο επίπεδο τάσης και στην πιθανή εμφάνιση αντιστροφής τάσης. Ο

πρωταρχικός στόχος των δεικτών είναι να παρέχουν σαφήνεια και να βελτιώσουν την κατανόηση των κινήσεων των τιμών [19].

Σκοπός των δεικτών είναι εξυπηρετούν στον εντοπισμό γεγονότων γραφημάτων και στην αξιολόγηση της φύσης των τάσεων των τιμών, συμπεριλαμβανομένου του προσδιορισμού της ισχύος της τάσης και της πρόβλεψης των σημείων αντιστροφής της τάσης. Η ταξινόμηση δεικτών μπορεί να γίνει σε δύο κύριες κατηγορίες με βάση τη φύση τους:

- *Δείκτες βάσει κρίσης:* Αυτή η κατηγορία περιλαμβάνει μεθόδους οπτικής αναγνώρισης προτύπων, όπως ανάλυση ράβδων, γραμμών και μοτίβων, μαζί με κηροπήγια. Ενώ είναι αποτελεσματικοί για την ενίσχυση της αντίληψης, αυτοί οι δείκτες μπορεί να είναι χρονοβόροι για να κυριαρχήσουν και δύσκολο να μεταφραστούν σε συνθέσεις λογισμικού για εκ των υστέρων δοκιμές [19]. Τρεις αξιοσημείωτες κατηγορίες τέτοιων εργαλείων περιλαμβάνουν:

    - Μοτίβα τιμών: Τα πρότυπα τιμών είναι σχηματισμοί που παρατηρούνται σε ιστορικά διαγράμματα τιμών που είναι ενδεικτικές πιθανών μελλοντικών κινήσεων της αγοράς. Οι έμποροι συχνά βασίζονται σε αναγνωρισμένα πρότυπα τιμών, όπως κεφάλι και ώμοι, τρίγωνα και σημαίες, για να προβλέψουν τις τάσεις και τις ανατροπές στην αγορά.

    - Μοτίβα κηροπήγια: Τα μοτίβα κηροπήγια είναι ένας συγκεκριμένος τύπος μοτίβου τιμών που περιλαμβάνει την ανάλυση των σχημάτων και των διατάξεων των κηροπήγια σε ένα διάγραμμα τιμών. Αυτά τα μοτίβα παρέχουν πληροφορίες για το κλίμα της αγοράς και μπορούν να σηματοδοτήσουν πιθανές αλλαγές στην κατεύθυνση.

    - Δείκτες εύρους αγοράς: Οι δείκτες εύρους της αγοράς αξιολογούν τη συνολική υγεία και τη συμμετοχή μιας αγοράς αναλύοντας τον αριθμό των περιουσιακών στοιχείων που προχωρούν και μειώνονται. Αυτοί οι δείκτες, όπως η γραμμή προόδου-πτώσης και ο Ταλαντωτής McClellan, προσφέρουν πληροφορίες για την υποκείμενη δύναμη ή αδυναμία μιας τάσης της αγοράς.

Οι έμποροι και οι επενδυτές συχνά ενσωματώνουν έναν συνδυασμό αυτών των εργαλείων στις στρατηγικές τεχνικής ανάλυσης τους για να αποκτήσουν μια ολοκληρωμένη κατανόηση των συνθηκών της αγοράς [20].

- *Δείκτες βασισμένοι σε μαθηματικά:* Αυτή η κατηγορία περιλαμβάνει κινητούς μέσους όρους, παλινδρόμηση, ορμή και άλλους μαθηματικούς υπολογισμούς. Η έκφραση γεγονότων γραφήματος με μαθηματικούς όρους διευκολύνει τον εκ των υστέρων έλεγχο αυτών των δεικτών σε ιστορικά δεδομένα, παρέχοντας πληροφορίες για τις προγνωστικές τους ικανότητες για μελλοντικές ενέργειες τιμών [19]. Αυτοί οι δείκτες μπορούν να κατηγοριοποιηθούν ευρέως σε διαφορετικούς τύπους, καθένας από τους οποίους εξυπηρετεί έναν συγκεκριμένο αναλυτικό σκοπό. Η ταξινόμηση περιλαμβάνει:

    - *Δείκτες ορμής:* Αυτοί οι δείκτες εστιάζουν στην ταχύτητα και την ισχύ των κινήσεων των τιμών, βοηθώντας στον εντοπισμό πιθανών αντιστροφών ή συνέχισης των τάσεων.

    - *Δείκτες τάσης:* Σχεδιασμένοι για να αποκαλύπτουν την επικρατούσα κατεύθυνση της αγοράς, οι δείκτες τάσης βοηθούν τους εμπόρους να αναγνωρίζουν και να ακολουθούν καθιερωμένες τάσεις.

    - *Δείκτες μεταβλητότητας:* Οι δείκτες μεταβλητότητας μετρούν τον βαθμό διακυμάνσεων των τιμών, παρέχοντας πολύτιμες πληροφορίες σχετικά με την αβεβαιότητα της αγοράς και τις πιθανές ευκαιρίες συναλλαγών.

    - *Δείκτες όγκου:* Χρησιμοποιώντας δεδομένα όγκου συναλλαγών, οι δείκτες όγκου προσφέρουν πληροφορίες για τη δύναμη ή την αδυναμία των κινήσεων των τιμών, βοηθώντας τους εμπόρους να επιβεβαιώσουν τις τάσεις.

    - *Δείκτες υποστήριξης και αντίστασης:* Αυτοί οι δείκτες προσδιορίζουν βασικά επίπεδα όπου οι τάσεις των τιμών είναι πιθανό να συναντήσουν εμπόδια ή να αντιμετωπίσουν ανατροπές, βοηθώντας στον προσδιορισμό των σημείων εισόδου και εξόδου.

- *Δείκτες κύκλου:* Εστιασμένοι στον εντοπισμό επαναλαμβανόμενων προτύπων και κύκλων στη συμπεριφορά της αγοράς, οι δείκτες κύκλου βοηθούν στην κατανόηση της περιοδικής φύσης των κινήσεων των τιμών.

- *Δείκτες συναισθήματος:* Οι δείκτες συναισθήματος μετρούν το συνολικό κλίμα της αγοράς και την ψυχολογία των επενδυτών, προσφέροντας πολύτιμες ενδείξεις για πιθανές ανατροπές ή συνέχιση της αγοράς.

Κάθε τύπος δείκτη παίζει μοναδικό ρόλο στην αναλυτική εργαλειοθήκη των εμπόρων και των επενδυτών. Η επιλογή των δεικτών εξαρτάται από τους συγκεκριμένους στόχους ανάλυσης, τα χαρακτηριστικά του τίτλου που διαπραγματεύεται και την προτιμώμενη στρατηγική συναλλαγών. Η πλήρης κατανόηση αυτών των κατηγοριών δεικτών δίνει τη δυνατότητα στους συμμετέχοντες στην αγορά να περιηγούνται στην πολυπλοκότητα των χρηματοπιστωτικών αγορών με μεγαλύτερη ακρίβεια και αποτελεσματικότητα [20].

Η κατανόηση της ταξινόμησης και των χαρακτηριστικών αυτών των δεικτών είναι ζωτικής σημασίας για τους επενδυτές και τους αναλυτές που ασχολούνται με την τεχνική ανάλυση, καθώς συμβάλλει σε μια πιο ενημερωμένη και στρατηγική προσέγγιση στην ερμηνεία της αγοράς.

2.2.3 Κατανοώντας τους Δείκτες

Οι τιμές των τίτλων παρουσιάζουν διάφορα μοτίβα, συμπεριλαμβανομένων των τάσεων, των ανατροπών εντός των τάσεων, των πλάγιων κινήσεων (διαπραγμάτευση εύρους) και της ενδεχόμενης αντιστροφής των τάσεων. Οι δείκτες διαδραματίζουν κρίσιμο ρόλο στην αναγνώριση αυτών των συνθηκών, με διαφορετικούς δείκτες να είναι πιο αποτελεσματικοί σε συγκεκριμένες καταστάσεις. Η παρακάτω λίστα περιγράφει πέντε βασικές συνθήκες που προσδιορίζονται από δείκτες, με προτεινόμενα παραδείγματα για κάθε συνθήκη, αναγνωρίζοντας ότι εναλλακτικοί δείκτες μπορεί επίσης να ισχύουν:

- *Έναρξη μιας τάσης:* Υποδεικνύεται από μια διασταύρωση του κινούμενου μέσου όρου ή ένα ξεκάθαρο μοτίβο.

- *Strength of a Trend:* Καθορίζεται από την κλίση της γραμμικής παλινδρόμησης ή του κινούμενου μέσου όρου.

- *Retracement into a Trend:* Αναγνωρίζεται μέσω δεικτών όπως ο δείκτης σχετικής ισχύος, υποδηλώνοντας μια προσωρινή αναστροφή πριν από την επανέναρξη της τάσης.

- *Τέλος μιας τάσης με πιθανή αντιστροφή:* Αναγνωρίζεται με ορμή, διασταύρωση κινούμενου μέσου όρου ή διάσπαση μοτίβου, σηματοδοτώντας μια πιθανή αλλαγή κατεύθυνσης.

- *Range-Trading:* Υποδεικνύεται από την κλίση της γραμμικής παλινδρόμησης ή του κινούμενου μέσου όρου.

Είναι σημαντικό να σημειωθεί ότι κάθε δείκτης υπερέχει σε συγκεκριμένες καταστάσεις και μπορεί να έχει περιορισμούς σε άλλες. Οι τεχνικοί έμποροι συμμετέχουν σε συνεχείς συζητήσεις σχετικά με τα πλεονεκτήματα και τα μειονεκτήματα διαφορετικών δεικτών για κάθε κατάσταση. Συγκεκριμένα, οι προτιμήσεις για δείκτες ποικίλλουν μεταξύ των εμπόρων, με διαφορετικά άτομα να προτιμούν διαφορετικά εργαλεία για συγκεκριμένες εργασίες. Η επιλογή του δείκτη εξαρτάται όχι μόνο από την ασφάλεια που αναλύεται αλλά και από το αναλυτικό χρονικό πλαίσιο που επιλέγεται για την αξιολόγηση [19].

2.2.4 Κινητός Μέσος Όρος (MA)

Ο κινητός μέσος όρος (MA) ή απλός κινητός μέσος όρος (SMA) είναι μια ευρέως χρησιμοποιούμενη τεχνική εξομάλυνσης στην τεχνική ανάλυση, που χρησιμοποιείται για τη μείωση του θορύβου της αγοράς και τον εντοπισμό των τάσεων των τιμών. Αντιπροσωπεύεται μαθηματικά ως:

$$MA_t = \frac{1}{n}\sum_{i=t-n+1}^{t} p_i \ , n \leq t$$

Υπολογίζει τον μέσο όρο των πιο πρόσφατων n σημείων δεδομένων, παρέχοντας έναν αριθμητικό μέσο που εξομαλύνει τις αλλαγές τιμών. Η επαναληπτική φύση επιτρέπει δυναμικές προσαρμογές με νέες τιμές, επηρεάζοντας τον βαθμό προσαρμογής με βάση τη διαφορά μεταξύ παλαιών και νέων τιμών. Η επιλογή της περιόδου υπολογισμού (n) περιλαμβάνει μια αντιστάθμιση μεταξύ της προγνωστικής ποιότητας και της ανάγκης προσδιορισμού τάσεων σε συγκεκριμένα χρονικά πλαίσια. Η ευελιξία των κινητών μέσων όρων επιτρέπει την προσαρμογή για διάφορες ανάγκες, λαμβάνοντας υπόψη παράγοντες όπως τριμηνιαίες αλλαγές, ετήσια ανάπτυξη ή εμπορικούς κύκλους. Η προσεκτική εξέταση είναι απαραίτητη όταν υπάρχουν κυκλικά ή εποχιακά μοτίβα. Το μήκος του κινητού μέσου όρου μπορεί να ευθυγραμμίζεται με τις εμπορικές απαιτήσεις, όπως φαίνεται από έναν κοσμηματοπώλη που χρησιμοποιεί έναν βραχυπρόθεσμο κινητό μέσο όρο για έγκαιρη λήψη αποφάσεων [21].

2.2.5 Εκθετικός κινητός μέσος όρος (EMA)

Ο εκθετικός κινητός μέσος όρος (EMA), ξεχωρίζει ως μια σταθμισμένη μέση τεχνική που προσφέρει μια διαφοροποιημένη προσέγγιση στην πρόβλεψη τάσεων στη χρηματοοικονομική ανάλυση. Αντιμετωπίζει τα ζητήματα που σχετίζονται με τον απλό κινούμενο μέσο όρο (SMA) αποδίδοντας μεγαλύτερη σημασία στα πρόσφατα δεδομένα, μετριάζοντας έτσι τις απότομες αλλαγές που παρατηρούνται στο SMA. Ο EMA είναι ουσιαστικά μια μορφή ποσοστιαίας εξομάλυνσης, που απαιτεί μόνο την τρέχουσα τιμή, την τελευταία εκθετικά εξομαλυνθείσα τιμή και μια σταθερά εξομάλυνσης για τον υπολογισμό της νέας τιμής [5] [21].

Η τεχνική της εκθετικής εξομάλυνσης αναπτύχθηκε αρχικά κατά τη διάρκεια του Β' Παγκοσμίου Πολέμου για την παρακολούθηση αεροσκαφών, βασιζόμενη στο άμεσο παρελθόν για την πρόβλεψη του άμεσου μέλλοντος. Αυτή η μέθοδος περιλαμβάνει μια γεωμετρική πρόοδο που σχηματίζει τους συντελεστές στάθμισης ενός σταθμισμένου κινητού μέσου όρου με αντίστροφη σειρά, καταδεικνύοντας τη φθίνουσα σημασία κάθε παλαιότερης τιμής.

Παρέχοντας μια ευέλικτη και προσαρμοστική προσέγγιση στην ανάλυση τάσεων, το EMA επιτρέπει στους χρήστες να προσαρμόζουν τη στάθμιση με βάση την επιθυμητή έμφαση στα πρόσφατα δεδομένα. Η επαναληπτική φύση της διαδικασίας της εκθετικής εξομάλυνσης επιτρέπει δυναμικές προσαρμογές, προσφέροντας μια πιο

ανταποκρινόμενη και ακριβή αναπαράσταση των τάσεων της αγοράς [5] [21], Ο τύπος για τον υπολογισμό του EMA έχει ως εξής:

$$EMA_t = (p_t - EMA_{t-1}) \times \frac{2}{n+1} + EMA_{t-1}$$

2.2.6 Κινητός μέσος όρος Σύγκλισης/Απόκλισης (MACD)

Η ανάλυση του MACD είναι ζωτικής σημασίας για την αξιολόγηση των τάσεων, προσφέροντας αξιόπιστες ενδείξεις για τις αλλαγές κατεύθυνσης της αγοράς. Αν και αναγνωρίζεται ευρέως για την αποτελεσματικότητά του, οι χρήστες πρέπει να αναγνωρίσουν ότι, όπως και άλλοι δείκτες τάσης, το MACD λειτουργεί βέλτιστα εντός συγκεκριμένων ορίων. Η σωστή κατανόηση και ερμηνεία των συστατικών του είναι ζωτικής σημασίας για τη μεγιστοποίηση της χρησιμότητάς του στην τεχνική ανάλυση [22].

Η γραμμή MACD (MACD Line), που συχνά αναφέρεται ως η γρήγορη γραμμή (Fast Line), υπολογίζεται ως ο εκθετικός κινούμενος μέσος όρος (EMA) δεκατεσσάρων ημερών μείον το EMA για είκοσι έξη ημέρων, χωρίς αυτό να είναι απόλυτο. Λειτουργεί ως δυναμικός δείκτης της βραχυπρόθεσμης ορμής τιμών, επιτρέποντας στους εμπόρους να μετρούν την άμεση τάση της αγοράς με μεγαλύτερη ανταπόκριση [22], η μαθηματική του αποτύπωσή ορίζεται ως εξής:

$$MACD\ Line = EMA_{14} - EMA_{26}$$

Η γραμμή σήματος (Signal Line), ή αργή γραμμή (Slow Line), προκύπτει μέσω δύο μεθόδων υπολογισμού: ενός απλού κινούμενου μέσου όρου εννέα ημερών (SMA) ή ενός εκθετικά εξομαλυνόμενου κινητού μέσου εννέα ημερών της βασικής γραμμής MACD. Συνήθως, οι χρήστες προτιμούν τη γραμμή σήματος που υπολογίζεται με χρήση EMA για την ακρίβεια και την ευρωστία της [22], ο τύπος για το υπολογισμό της γραμμής με την χρήση EMA ορίζεται ως εξης:

$$Signal\ Line = \frac{2}{10} \times MACD\ Line + \left(1 - \frac{2}{10}\right) \times Previous\ Signal\ Line$$

Το ιστόγραμμα MACD (MACD Histogram) αντιπροσωπεύει τη διαφορά μεταξύ της γρήγορης γραμμής και της αργής σήματος. Παρουσιάζεται σε στυλ ιστογράμματος σε γραφήματα, είναι μια οπτική αναπαράσταση του κενού μεταξύ των δύο γραμμών. Όταν το ιστόγραμμα διασχίζει πάνω ή κάτω από τη μηδενική γραμμή, θεωρείται σημαντικό σήμα, υποδεικνύοντας μετατοπίσεις από πτωτική σε ανοδική τάση ή αντίστροφα. Οι έμποροι συχνά χρησιμοποιούν ανάλυση ιστογράμματος σε συνδυασμό με τα επίπεδα του βασικού MACD και των γραμμών σήματος για πιο ολοκληρωμένες αξιολογήσεις τάσεων [22]. Αντιπροσωπεύεται μαθηματικά ως:

$$MACD\ Histogram = MACD\ Line - \ Signal\ Line$$

2.2.7 Δείκτης Σχετικής Ισχύος (RSI)

Ο δείκτης σχετικής ισχύος (RSI), που αναπτύχθηκε από τον Welles Wilder, είναι ένας ευρέως χρησιμοποιούμενος δείκτης για τον προσδιορισμό των συνθηκών υπεραγοράς και υπερπώλησης. Το RSI κλιμακώνει τις τιμές μεταξύ 0 και 100, παρέχοντας σταθερότητα σε σύγκριση με τους δείκτες ορμής. Υπολογίζεται ως

$$RSI = 100 - \frac{100}{1 + RS}$$

όπου

$$RS = \frac{AU}{AD}$$

με την AU να αντιπροσωπεύει το σύνολο των ανοδικών αλλαγών της τιμής και την AD να αντιπροσωπεύει το σύνολο των καθοδικών αλλαγών τιμών σε μια καθορισμένη περίοδο, συνήθως 14 ημερών.

Ο Wilder προτείνει τη χρήση 14 ημερών, που αντιπροσωπεύουν το μισό ενός φυσικού κύκλου, ως περίοδο υπολογισμού. Τα σημαντικά επίπεδα κατωφλίου ορίζονται στα 30 και 70, υποδεικνύοντας πιθανά σημεία αντιστροφής. Τιμές RSI κάτω από 30 υποδηλώνουν επικείμενη ανάκαμψη, ενώ τιμές πάνω από 70 υποδηλώνουν μια εκκρεμή ύφεση.

Η ερμηνεία του RSI περιλαμβάνει τη χάραξη γραμμών τάσης και την αναγνώριση μοτίβων γραφημάτων. Ο Wilder πρότεινε τη χρήση του RSI για τον εντοπισμό σχηματισμών πάνω και κάτω. Τα σπασίματα κάτω από το κάτω μέρος της αντίδρασης

μεταξύ των κορυφών που πέφτουν θεωρούνται σήματα πώλησης. Επιπλέον, η ταλάντευση ή η απόκλιση αστοχίας μπορεί να υποδηλώνει μια ανεπιτυχή δοκιμή πρόσφατων υψηλών ή χαμηλών τιμών RSI.

Το RSI είναι ένα άμεσο μέτρο της απόκλισης της αγοράς από μια τάση, με τιμές πάνω από 70 να σηματοδοτούν συνθήκες υπεραγοράς και τιμές κάτω από 30 να σηματοδοτούν συνθήκες υπερπώλησης. Παρέχει πιο ομοιόμορφη κατανομή σε διαφορετικές αγορές και χρονικές περιόδους. Ενώ το RSI είναι αποτελεσματικό, ορισμένοι έμποροι μπορεί να προτιμούν στοχαστικούς δείκτες για ακόμη καλύτερη απόδοση [21] [22].

2.2.8 Δείκτης Vortex

Ο δείκτης Vortex είναι ένα εργαλείο τεχνικής ανάλυσης εμπνευσμένο από τον δείκτη κατευθυντικής κίνησης (Dmi) του J. Welles Wilder, σχεδιασμένο να προσδιορίζει την κατεύθυνση των τάσεων και τις σημαντικές κινήσεις τιμών στην αγορά. Η ιδέα έχει τις ρίζες της στον ορισμό του Wilder για την κατευθυντική κίνηση, δίνοντας έμφαση στη σχέση μεταξύ των ράβδων τιμών για τη διάκριση των τάσεων της αγοράς. Η θετική κατευθυντική κίνηση, που αντιπροσωπεύει το τμήμα μιας ράβδου τιμής πάνω από το προηγούμενο υψηλό, και η αρνητική κατευθυντική κίνηση, το τμήμα κάτω από το προηγούμενο χαμηλό, είναι κρίσιμα στοιχεία για την λειτουργία του δείκτη. Ο Δείκτης Vortex εισάγει μια καινοτόμο προσέγγιση, αντλώντας έμπνευση από τις μελέτες του Viktor Schauberger για τις ρευστές δίνες (fluidic vortexes) στη φύση. Συνδέοντας διαδοχικά χαμηλά με υψηλά και αντίστροφα, εντοπίζεται ένα μοτίβο δίνης στην αγορά, που αποτελεί τη βάση για τον υπολογισμό του δείκτη. Ο δείκτης που προκύπτει αποτελείται από δύο γραμμές, +VI και -VI, που δηλώνουν θετικές και αρνητικές κινήσεις δίνης. Τα σημεία διέλευσης μεταξύ αυτών των γραμμών είναι ζωτικής σημασίας για την αναγνώριση πιθανών αλλαγών τάσης. Οι έμποροι μπορούν να χρησιμοποιήσουν τη ρύθμιση συναλλαγών του Wilder, χρησιμοποιώντας ακραία υψηλά ή χαμηλά σημεία την ημέρα της διέλευσης ως σημεία εισόδου για θέσεις στην αγορά, υποστηριζόμενη από μια στρατηγική τελικής στάσης. Ο δείκτης Vortex εξυπηρετεί τόσο τους βραχυπρόθεσμους εμπόρους όσο και τους διαχειριστές μακροπρόθεσμων κεφαλαίων, προσδιορίζοντας αποτελεσματικά την έναρξη, τη συνέχιση ή τον τερματισμό των τάσεων της αγοράς [23].

Ο υπολογισμός του δείκτη Vortex περιλαμβάνει διάφορα βήματα:

1. Υπολογισμός πραγματικού εύρους (TR)

$$TR = \max(|current\ high - current\ low|, |\ current\ high - previous\ close|, |\text{current low} - \text{previous close}|)$$

2. Υπολογισμός θετικής και αρνητικής κίνησης (VM+ και VM-)

$$VM^{+} = current\ high - previous\ low$$
$$VM^{-} = previous\ high - current\ low$$

3. Υπολογισμός του πραγματικού εύρους (TR) για μια καθορισμένη περίοδο (συνήθως 14 ημερών) και άθροισμα των θετικών και αρνητικών κινήσεων της ιδίας περιόδου

$$\sum_{i=1}^{14} TR_i$$
$$\sum_{i=1}^{14} VM_i^{+}$$
$$\sum_{i=1}^{14} VM_i^{-}$$

4. Υπολογισμός του δείκτη θετικής δίνης VI^{+} και του δείκτη αρνητικής δίνης VI^{-}

$$VI^{+} = \frac{\sum_{i=1}^{14} VM_i^{+}}{\sum_{i=1}^{14} TR_i}$$

$$VI^{-} = \frac{\sum_{i=1}^{14} VM_i^{-}}{\sum_{i=1}^{14} TR_i}$$

Αυτοί οι υπολογισμοί περιλαμβάνουν τον προσδιορισμό του πραγματικού εύρους, των θετικών και αρνητικών κινήσεων και στη συνέχεια τη συγκέντρωση αυτών των τιμών για μια καθορισμένη περίοδο. Οι θετικοί και αρνητικοί δείκτες δίνης που προκύπτουν παρέχουν πληροφορίες για τις τάσεις της αγοράς και τις πιθανές αλλαγές τάσεων με βάση τη σχέση μεταξύ θετικών και αρνητικών κινήσεων σε σχέση με το πραγματικό εύρος [23].

2.2.9 Average True Range (ATR)

Το Average True Range (ATR) είναι ένας δείκτης μεταβλητότητας που μετρά το μέσο πραγματικό εύρος τιμών για μια συγκεκριμένη περίοδο. Το αληθινό εύρος είναι το μέγιστο από τις ακόλουθες τρεις τιμές: η απόσταση μεταξύ του σημερινού υψηλού και του χαμηλού, η απόλυτη τιμή της διαφοράς μεταξύ του χθεσινού κλεισίματος και του σημερινού υψηλού και η απόλυτη τιμή της διαφοράς μεταξύ του χθεσινού κλεισίματος και του σημερινού χαμηλού [19].

Το ATR υπολογίζεται ως ένας κινητός μέσος όρος των πραγματικών περιοχών και παρέχει πολύτιμες πληροφορίες σχετικά με το επίπεδο μεταβλητότητας στην αγορά. Ένα υψηλότερο ATR υποδηλώνει μεγαλύτερη μεταβλητότητα, ενώ ένα χαμηλότερο ATR υποδηλώνει χαμηλότερη μεταβλητότητα [25]. Ο υπολογισμός πραγματικού εύρους γίνεται με τον εξής μαθηματικό τύπο:

$$ATR = \frac{1}{n}\sum_{i=1}^{n} TR_i$$

Για να ληφθούν υπόψη πιθανά κενά στα δεδομένα τιμών, ο υπολογισμός ATR προσαρμόζεται ξεκινώντας από το κλείσιμο της προηγούμενης ημέρας και τελειώνοντας στο σημερινό υψηλό (ή χαμηλό σε περίπτωση καθοδικής διαφοράς). Αυτό διασφαλίζει ότι το κενό ενσωματώνεται στη μέτρηση [19].

Το ATR μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της κατάλληλης τοποθέτησης εντολών stop-loss. Κατανοώντας τη μέση ημερήσια κίνηση της τιμής, οι έμποροι μπορούν να ορίσουν επίπεδα stop-loss που αντιστοιχούν στις κανονικές διακυμάνσεις της αγοράς. Για παράδειγμα, εάν το ATR είναι 1,75 $ για μια μετοχή, υποδηλώνει ότι, κατά μέσο όρο, η μετοχή κινείται 1,75 $ την ημέρα. Οι έμποροι μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για να ορίσουν εντολές stop-loss σε απόσταση που να ικανοποιεί αυτήν την αστάθεια [25].

2.2.10 Awesome Oscillator (AO)

Ο Awesome Oscillator (AO) είναι ένας δείκτης ορμής που χρησιμοποιείται στις συναλλαγές τόσο σε αγορές μετοχών όσο και σε αγορές εμπορευμάτων. Μετρά την άμεση ορμή των τελευταίων 5 ράβδων και τη συγκρίνει με την ορμή των τελευταίων

34 ράβδων. Αυτός ο ταλαντωτής θεωρείται ένας από τους καλύτερους και πιο ακριβείς δείκτες όταν κατανοηθεί και χρησιμοποιηθεί σωστά. Υπολογισμός του Awesome Oscillator (AO) γίνεται με έναν απλό κινητό μέσο όρο 34 ράβδων που αφαιρείται από έναν απλό κινητό μέσο όρο 5 ράβδων των μεσαίων σημείων και εμφανίζεται σε μορφή ιστογράμματος.

$$AO = \sum_{i=1}^{5} median_i - \sum_{i=1}^{34} median_i$$

Οπού,

$$median_i = \frac{High_i + Low_i}{2}$$

Τα μεσαία σημεία αντιπροσωπεύουν τον μέσο όρο των υψηλών (H) και των χαμηλών (L) τιμών κάθε ράβδου. Το ιστόγραμμα απεικονίζει τη διαφορά μεταξύ του κινητού μέσου όρου των 5 ράβδων και των 34 ράβδων.

Το AO παρέχει πληροφορίες για την άμεση δυναμική στην αγορά. Καταγράφει τη μεταβαλλόμενη ταχύτητα της τρέχουσας ορμής, η οποία συχνά προηγείται των αλλαγών τιμών. Η τιμή είναι το τελευταίο πράγμα που αλλάζει στις αγορές και το AO βοηθά στον εντοπισμό των αλλαγών στην ορμή πριν συμβούν κινήσεις τιμών. Ο AO αντικατοπτρίζει τις δραστηριότητες των εμπόρων και των επενδυτών που λαμβάνουν αποφάσεις στην αγορά. Ο AO εμφανίζεται συνήθως στο κάτω μέρος του γραφήματος τιμών. Όταν ο ταλαντωτής σβήσει, μπορεί να σηματοδοτήσει μια ευκαιρία πώλησης και όταν εμφανιστεί, θα μπορούσε να υποδεικνύει μια ευκαιρία αγοράς.

Όταν ο AO φθίνει, οι έμποροι μπορεί να εξετάσουν το ενδεχόμενο να πουλήσουν στην αγορά και περιμένουν μέχρι την επόμενη αύξηση του. Αντίθετα, όταν αύξηθει το AO, οι έμποροι μπορεί να ξεκινήσουν να επενδύουν στην αγορά. Ο AO επίσης, μπορεί να χρησιμοποιηθεί για την επιβεβαίωση αλλαγών τάσης. Για παράδειγμα, μια μετατόπιση από θετικές σε αρνητικές τιμές μπορεί να σηματοδοτήσει μια πιθανή αντιστροφή σε μια ανοδική τάση.

Συμπερασματικά, ο Awesome Oscillator [24], είναι ένας ισχυρός δείκτης ορμής που, όταν χρησιμοποιείται αποτελεσματικά, μπορεί να ενισχύσει τις αποφάσεις συναλλαγών και δυνητικά να συμβάλει στην κερδοφορία.

2.2.11 Δείκτης Καναλιών Εμπορευμάτων (CCI)

Ο Δείκτης Καναλιών Εμπορευμάτων (CCI) είναι ένα ζωτικό εργαλείο για τους ενδοημερήσιους εμπόρους, που προέρχεται ως ταλαντωτής ορμής για τα εμπορεύματα αλλά αποδεικνύεται προσαρμόσιμος σε διάφορες αγορές [26].

- *Συνθήκες Υπεραγοράς/Υπερπώλησης:* Η μέτρηση της απόκλισης του CCI από τη στατιστική μέση τιμή ενός περιουσιακού στοιχείου βοηθά τους επενδυτές εντός της ημέρας να εντοπίσουν συνθήκες υπεραγοράς ή υπερπώλησης, παρέχοντας ευκαιρίες με βάση ακραία σενάρια αγοράς.

- *Ανίχνευση τάσεων:* Το CCI σταθερά πάνω από +100 ή κάτω από -100 υποδεικνύει τάσεις, καθοδηγώντας τους ενδοημερήσιους συναλλασσόμενους σε στρατηγικές αποφάσεις εισόδου και εξόδου.

- *Αποκλίσεις:* Το CCI εντοπίζει αποκλίσεις μεταξύ τιμής και ορμής, προσφέροντας πληροφορίες για πιθανές αλλαγές τάσεων με βάση την αναγνώριση προτύπων.

- *Επιβεβαίωση τάσης:* Οι σταθερές τιμές CCI πάνω από 100 ή κάτω από 100 επιβεβαιώνουν τις υπάρχουσες τάσεις, βοηθώντας τους ενδοημερήσιους εμπόρους να ευθυγραμμίσουν τις στρατηγικές τους με την κατεύθυνση της αγοράς.

Αναγνωρίζοντας τη φύση της υστέρησης της CCI, παρέχει ένα στιγμιότυπο της τρέχουσας δυναμικής της αγοράς, που απαιτεί συμπληρωματική ανάλυση για μια ολιστική άποψη [26].

Στο "Trade the Patterns", ο K. Wood εισάγει μια ψυχολογική προοπτική για τα πρότυπα CCI, θεωρώντας κάθε πρότυπο μια μοναδική αφήγηση της συμπεριφοράς της αγοράς [27]. Η έμφαση που δίνει ο Wood στην εκμάθηση από τα πρότυπα CCI εμπλουτίζει την ικανότητα των εμπόρων να αποκρυπτογραφούν αποτελεσματικά τις κινήσεις της αγοράς.

Η πλατφόρμα του συστήματος συναλλαγών του Wood, που διαθέτει γραμμές αξίας, μηδενική γραμμή και διάφορα μοτίβα, βελτιώνει την εφαρμογή της CCI στις ενδοημερήσιες συναλλαγές. Η μηδενική γραμμή χρησιμεύει ως αναφορά για υποστήριξη και αντίσταση και οι διακριτές γραμμές τιμών (100 και 200 γραμμές) προσφέρουν ένα ολοκληρωμένο πλαίσιο για την αναγνώριση προτύπων και τη λήψη αποφάσεων [27]. Ο ορισμός του CCI γίνεται τον μαθηματικό τύπο:

$$CCI = \frac{TypicalPrice - \frac{1}{n}\sum_{i=1}^{n} TypicalPrice_i}{0.015 \times MeanDeviation}$$

Όπου,

$$TypicalPrice_i = \frac{High_i + Low_i + Close_i}{3}$$

$$MeanDeviation = \sum_{k=1}^{n} | TypicalPrice_k - \frac{1}{n}\sum_{i=1}^{n} TypicalPrice_i |$$

2.2.12 Μέσος Δείκτης Κατεύθυνσης (ADX)

Αναπτύχθηκε από τον J. Welles Wilder, το ADX είναι ένα βασικό εργαλείο για τη μέτρηση της έντασης των τάσεων. Προερχόμενο από τον δείκτη θετικής κατεύθυνσης κίνησης (+DMI) και τον δείκτη αρνητικής κατεύθυνσης κίνησης (-DMI), το ADX είναι καθοριστικό για την επιβεβαίωση των σημάτων εισόδου και εξόδου. Ο συνδυασμός ADX, +DMI και -DMI σχηματίζει έναν δείκτη τριών γραμμών, που συνήθως αναφέρεται ως ADX/DMI. Η κύρια λειτουργία του ADX είναι να ποσοτικοποιεί την ένταση της τάσης, ενώ το DMI καθορίζει την κατεύθυνση της τάσης και επικυρώνει τα σήματα εισόδου και εξόδου. Όταν το +DMI είναι πάνω από -DMI, σημαίνει μια ανοδική τάση και όταν το -DMI ξεπερνά το +DMI, υποδεικνύεται μια πτωτική τάση. Η κλίση και η τιμή της γραμμής ADX μεταδίδουν τη δύναμη της τάσης [28].

Ο μέσος δείκτης κατεύθυνσης (ADX) συνδέεται περίπλοκα με τη σχέση μεταξύ των δύο γραμμών Δείκτης Κίνησης Κατεύθυνσης (DMI), που χρησιμεύει ως κρίσιμο

στοιχείο στην ακριβή ανάλυση τάσεων. Το ADX υπολογίζεται λαμβάνοντας τη διαφορά μεταξύ των δύο τιμών DMI, διαιρώντας την με το άθροισμά τους και στη συνέχεια πολλαπλασιάζοντας το αποτέλεσμα με το 100. Αυτό έχει ως αποτέλεσμα μια ταλαντούμενη τιμή μεταξύ μηδέν και 100, που μοιάζει με ταλαντωτή όπως το DMI. Η γραμμή ADX είναι επίσης ένας κινητός μέσος όρος, παρέχοντας μια ομαλοποιημένη αναπαράσταση της ισχύος των τάσεων [28]. Όπου μαθηματικά αποτυπώνεται ως:

$$ADX = \frac{(|DI^{+} - DI^{-}|)}{(|DI^{+} + DI^{-}|)} \times 100$$

2.2.13 Williams %R

Το %R του Larry Williams, ή το ποσοστό R, είναι ένας ισχυρός τεχνικός δείκτης που μετράει τη θέση μιας τιμής κλεισίματος εντός ενός καθορισμένου εύρους τιμών σε μια δεδομένη χρονική περίοδο, συνήθως αρκετές ημέρες. Ο δείκτης αντικατοπτρίζει το ποσοστό ανατροπής της τιμής κλεισίματος από την υψηλότερη τιμή εντός της παρατηρούμενης περιόδου. Η τιμή %R, που κυμαίνεται από 0 έως 100 τοις εκατό, παίζει καθοριστικό ρόλο στη δημιουργία σημάτων για πιθανές ευκαιρίες αγοράς ή πώλησης [29].

Το %R υπολογίζεται αφαιρώντας τη σημερινή τιμή κλεισίματος από την υψηλότερη τιμή εντός της καθορισμένης χρονικής περιόδου, διαιρώντας αυτή τη διαφορά με το συνολικό εύρος (υψηλότερο - χαμηλότερο) για την ίδια περίοδο και στη συνέχεια πολλαπλασιάζοντας με το 100. Μια τιμή %R μηδέν υποδηλώνει ότι Το σημερινό κλείσιμο είναι το ίδιο με το υψηλό της περιόδου, υποδηλώνοντας πιθανές συνθήκες υπεραγοράς. Αντίθετα, μια τιμή %R κοντά στο 100 υποδηλώνει ότι η τιμή κλεισίματος βρίσκεται στο χαμηλό άκρο του παρατηρούμενου εύρους τιμών, σηματοδοτώντας πιθανές συνθήκες υπερπώλησης [29].

$$\%R = \frac{Highest\ High - Close}{Highest\ High - Low} \times -100$$

Ο Larry Williams %R έχει ομοιότητες με τους στοχαστικούς ταλαντωτές στη μέτρηση του πιο πρόσφατου πλησιέστερου σε σχέση με το εύρος τιμών του για μια συγκεκριμένη περίοδο. Η ερμηνεία περιλαμβάνει τον εντοπισμό αποκλίσεων σε περιοχές με υπεραγορές ή υπερπωλήσεις και, όπως τα στοχαστικά, το %R μπορεί να

εμφανιστεί ανάποδα. Τα πακέτα γραφημάτων παρουσιάζουν συχνά μια ανεστραμμένη έκδοση του %R για ευθυγράμμιση με τις συμβατικές αναπαραστάσεις [5].

Το %R του Larry Williams, με επίκεντρο τον εντοπισμό πιθανών αντιστροφών και την ισχύ των τάσεων, αποτελεί πολύτιμο εργαλείο για τους εμπόρους που επιδιώκουν να λάβουν τεκμηριωμένες αποφάσεις με βάση τη σχέση μεταξύ των τιμών κλεισίματος και του εύρους τιμών εντός συγκεκριμένων χρονικών πλαισίων [29].

2.2.14 Στοχαστικός Ταλαντωτής (%K και %D)

Η Στοχαστική, ένα βασικό εργαλείο στην τεχνική ανάλυση, χρησιμοποιεί δύο βασικές γραμμές: %K και %D. Αυτός ο ταλαντωτής μετράει τη θέση κλεισίματος της τιμής ενός τίτλου μέσα σε ένα προκαθορισμένο εύρος για μια καθορισμένη περίοδο, προσφέροντας πληροφορίες για πιθανές ανατροπές της αγοράς και αποκλίσεις από τις τάσεις. Οι γραμμές %K και %D διαδραματίζουν κρίσιμους ρόλους στην αξιολόγηση των συνθηκών υπεραγοράς ή υπερπώλησης, καθώς και στην ένδειξη της σχετικής ισχύος των τάσεων [22].

- *Ερμηνεία γραμμής %K:* Η γραμμή %K, που υπολογίζεται ως ποσοστό της τρέχουσας τιμής κλεισίματος σε σχέση με το εύρος μεταξύ του υψηλότερου υψηλού και του χαμηλότερου χαμηλού των τελευταίων εννέα παρατηρήσεων, κυμαίνεται από 0 έως 100. Αυτή η μέτρηση παρέχει πολύτιμα σήματα, με τιμές πάνω από 80 που υποδηλώνουν συνθήκες υπεραγοράς και δυνητικές μειώσεις, ενώ τιμές κάτω από 20 υποδηλώνουν συνθήκες υπερπώλησης και πιθανές ανακάμψεις. Η γραμμή %K προσφέρει μια αντανάκλαση σε πραγματικό χρόνο της ορμής μιας ασφάλειας [22].

$$\%K = \left(\frac{Current\ Close - Lowest\ Low}{Highest\ High - Lowest\ Low}\right) \times 100$$

- *Ερμηνεία γραμμής %D:* Η γραμμή %D, που εισήχθη για να μετριάσει τον αντίκτυπο των μεμονωμένων παρατηρήσεων στο %K, βασίζεται σε μέσους όρους τριών ημερών. Παρέχει μια ομαλή ερμηνεία της δυναμικής της αγοράς και είναι ιδιαίτερα χρήσιμο για την παροχή σταθερότητας. Παρόμοια με το %K, οι διασταυρώσεις %D και οι σχετικές θέσεις παρέχουν ενδείξεις πιθανών αλλαγών στην αγορά, συμβάλλοντας σε μια πιο λεπτή κατανόηση της ισχύος των τάσεων [22].

$$\%D = \frac{1}{3}\sum_{i=1}^{3} \%K_i$$

Ο συνδυασμός γραμμών %K και %D ενισχύει την αποτελεσματικότητα της στοχαστικής στην τεχνική ανάλυση. Οι διασταυρώσεις μεταξύ αυτών των γραμμών σηματοδοτούν πιθανά σημεία καμπής στην αγορά. Η υστέρηση του %D σε σύγκριση με το %K προσφέρει μια πολυεπίπεδη ανάλυση, με διασταυρώσεις και σχετικές θέσεις που παρέχουν στους εμπόρους πολύτιμες πληροφορίες σχετικά με τη δυναμική των τάσεων της αγοράς και τις πιθανές ανατροπές . Παρά την αποτελεσματικότητα των στοχαστικών, προκλήσεις προκύπτουν κατά τη διάρκεια παρατεταμένων τάσεων, όπου οι ταλαντωτές μπορούν να παραμείνουν σε ζώνες υπεραγοράς ή υπερπώλησης. Συνιστάται προσοχή να μην χρησιμοποιείτε ταλαντωτές έναντι ισχυρών τάσεων. Η στρατηγική του συνδυασμού ταλαντωτών με δείκτες τάσης αντιμετωπίζει δυσκολίες στην ακριβή πρόβλεψη των τάσεων και στον καθορισμό της έναρξης και της ολοκλήρωσης των τάσεων της αγοράς. Η αντιμετώπιση αυτών των προκλήσεων απαιτεί μια ολοκληρωμένη προσέγγιση για τη βελτίωση των μεθοδολογιών τεχνικής ανάλυσης [22].

2.2.15 On Balance Volume (OBV)

Το On Balance Volume (OBV) είναι ένας θεμελιώδης δείκτης όγκου που αναπτύχθηκε από τον Joseph Granville, προσφέροντας μια οπτική αναπαράσταση της πίεσης αγοράς και πώλησης σε ένα διάγραμμα τιμών. Αντί να βασίζεται σε παραδοσιακές ράβδους όγκου, η OBV παράγει μια αθροιστική γραμμή στο γράφημα τιμών, καθιστώντας ευκολότερο να διακρίνουμε τις αλλαγές στη ροή όγκου. Η κατεύθυνση της τάσης της γραμμής OBV είναι ζωτικής σημασίας, ακολουθώντας την αρχή ότι πρέπει να ευθυγραμμίζεται με την τάση των τιμών. Η απόκλιση μεταξύ της γραμμής OBV και της τιμής σηματοδοτεί πιθανές ανατροπές στην αγορά [5].

Η κατασκευή της γραμμής OBV είναι απλή, με τον συνολικό όγκο κάθε ημέρας να αποδίδεται μια τιμή συν ή πλην βάσει του αν οι τιμές κλείνουν υψηλότερα ή χαμηλότερα. Η τιμή της γραμμής OBV δεν είναι τόσο κρίσιμη όσο η κατεύθυνση της τάσης της. Εάν η τάση της τιμής παρουσιάζει υψηλότερες κορυφές και κατώτατες τιμές, η γραμμή OBV θα πρέπει να αντικατοπτρίζει αυτήν τη συμπεριφορά. Μια απόκλιση μεταξύ της γραμμής OBV και των τιμών υποδηλώνει μια πιθανή αντιστροφή

της τάσης, παρέχοντας στους εμπόρους πολύτιμες γνώσεις σχετικά με την υποκείμενη δύναμη ή αδυναμία στην αγορά [5] [6].

$$OBV = \begin{cases} OBV_{prev} + Volume & ,if\ Close > Close_{prev} \\ OBV_{prev} - Volume & ,if\ Close < Close_{prev} \\ OBV_{prev} & ,if\ Close = Close_{prev} \end{cases}$$

Ενώ το OBV είναι ένας δημοφιλής και ευρέως χρησιμοποιούμενος δείκτης, υπάρχουν προκλήσεις, ιδιαίτερα στην ικανότητά του να παρέχει ακριβή σήματα με συνέπεια. Οι θετικές και αρνητικές αποκλίσεις μπορεί να μην ευθυγραμμίζονται πάντα με τις πραγματικές κινήσεις της αγοράς, κάτι που απαιτεί προσοχή [6]. Συνιστάται να επιβεβαιώνονται τα σήματα OBV με άλλες προσεγγίσεις τεχνικής ανάλυσης για επιβεβαίωση. Η κοινή τεχνική γραμμής τάσης σημειώνεται για την αξιοπιστία της στην ερμηνεία του δείκτη OBV, προσφέροντας στους εμπόρους μια πιο λεπτή κατανόηση των πιθανών αλλαγών στην αγορά [6].

ΚΕΦΑΛΑΙΟ 3

ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Όταν μιλάμε για δεδομένα εννοούμε την κάθε πληροφορία που αποθηκεύεται σε ψηφιακή μορφή και έχει ως μονάδα μέτρησης το bit (0 ή 1). Κάθε ομάδα από 8 bits αποτελεί το byte, το οποίο αντιστοιχεί στην κωδικοποίηση ενός χαρακτήρα. Τα παραδοσιακά σύνολα δεδομένων φτάνουν σε μεγέθη μερικών gigabyte ή ακόμα και terabyte, όγκος δηλαδή που μπορεί να αποθηκευτεί κεντρικά σε κάποιον απομακρυσμένο διακομιστή.

Πλέον, στην εποχή της 4^{η} βιομηχανικής επανάστασης, του Industry 4.0, με την τεχνολογία να εξελίσσεται συνεχώς (υπολογιστές, εφαρμογές, λογισμικό, Διαδίκτυο των πραγμάτων) και την ανάπτυξη του διαδικτύου να έχει ξεπεράσει προ πολλού τα bytes, μιλάμε για μεγάλα δεδομένα (Big Data). Σύμφωνα με τον Eric Schmid της Google το 2003 η ανθρωπότητα παρήγαγε περίπου 5 Exabytes πληροφορίας ανά 2 ημέρες, δηλαδή 5Χ10^{18} bytes. Ο ίδιος όγκος πληροφορίας είχε παραχθεί συνολικά από την αρχή του πολιτισμού μέχρι τότε. Σύμφωνα με τις εκτιμήσεις της IBM, η εκθετική αύξηση παραγωγής δεδομένων ξεπερνούσε τα 2,6 Exabytes ημερησίως το 2016 και σήμερα φτάνει την τάξη του Yottabyte (10^{24} bytes), αλλά και την τάξη του Brondobyte (10^{27} bytes) και Geopbyte (10^{30}bytes).

Το 2023 η εταιρεία Domo δημοσίευσε για ακόμη μία χρονιά την έκδοση του Data Never Sleeps, από την οποία μπορούμε να αντιληφθούμε τον τεράστιο όγκο δεδομένων που παράγονται στο διαδίκτυο ανά δευτερόλεπτο σε πολύ δημοφιλείς εφαρμογές [31]. Από μια γρήγορη αναζήτηση στο Διαδίκτυο ή την αποστολή email, μέχρι τον έλεγχο των τελευταίων πρωτοσέλιδων στο δρόμο προς τη δουλειά, οι άνθρωποι συνεχώς αλληλεπιδρούν με ψηφιακές πλατφόρμες και υπηρεσίες, από το Instagram, Tik Tok έως το Amazon, και πολλές άλλες. Τα δεδομένα αυξάνονται και εξελίσσονται αλλάζοντας τον τρόπο που τα χρησιμοποιούμε και τον αντίκτυπο που έχουν στην καθημερινή μας ζωή. Για παράδειγμα, το 2020 με την εμφάνιση της πανδημίας του COVID-19 οι ψηφιακές δραστηριότητες και η συλλογή δεδομένων

αυξήθηκαν κατά 65% σε σχέση με τις προηγούμενες χρονιές. Σύμφωνα με τον ιδρυτή και διευθύνων σύμβουλο της DOMO αξιοσημείωτη αύξηση δεδομένων υπήρξε και το 2023, η οποία οφείλονταν στην ταχεία άνοδο της δημοτικότητας των μοντέλων τεχνητής νοημοσύνης, όπως το ChatGPT, που θα δούμε να μεταβάλλει με γρήγορους ρυθμούς το ψηφιακό τοπίο τις επόμενες χρονιές [31].

Σύμφωνα με τα παραπάνω, ο επιστημονικός κλάδος της Πληροφορικής έχει πλέον να διαχειριστεί καθημερινά και να προστατεύσει τεράστιους όγκους δεδομένων σε όλα τα στάδια: αποθήκευση, επεξεργασία, ανάλυση, αναζήτηση, οπτικοποίηση, μεταφορά, κοινή χρήση. Πρόκειται για δεδομένα μεγάλου όγκου, τα οποία είναι συνεχώς αυξανόμενα, βρίσκονται πολλές φορές σε αδόμητη μορφή, δεδομένα που αποθηκεύονται σε βάσεις δεδομένων και αποτελούν τα ιστορικά δεδομένα, αλλά και δεδομένα που παράγονται σε πραγματικό χρόνο, όπως αυτά των κρυπτονομισμάτων. Αυτός ο τεράστιος και συνεχώς αυξανόμενος όγκος δεδομένων δεν είναι δυνατόν να αποθηκευτεί και να διαχειριστεί από τις παραδοσιακές σχεσιακές βάσεις δεδομένων, όπως SQL Server και Oracle. Για τη διαχείριση τόσο μεγάλου όγκου πληροφορίας με ταχύτητα και ασφάλεια, χρησιμοποιούνται εργαλεία και εφαρμογές, όπως είναι το Apache Hadoop και το Apache Spark που υποστηρίζουν τον κατανεμημένο τρόπο αποθήκευσης και την παράλληλη επεξεργασία των μεγάλων δεδομένων σε συστοιχίες υπολογιστών προσφέροντας επεκτασιμότητα, αξιοπιστία και μεγάλη ανθεκτικότητα σε σφάλματα.

3.1 Τα 3Vs των Μεγάλων δεδομένων

Ο όρος Μεγάλα Δεδομένα (Big Data) ορίστηκε για πρώτη φορά το 2001 από τον όμιλο Gartner και συγκεκριμένα από τον Doug Laney, ο οποίος πρότεινε ότι η έρευνα των Μεγάλων δεδομένων πρέπει να εστιάζει στο μοντέλο 3Vs. Τα δεδομένα μεγάλου όγκου είναι άπειρα, ασύλληπτα σε μέγεθος (volume), τα οποία παράγονται και μεταβάλλονται με μεγάλη ταχύτητα στη μονάδα του χρόνου (velocity) και προέρχονται από πολλές και διαφορετικές πηγές (variety). Για την πλήρη διαχείριση των μεγάλων δεδομένων απαιτείται να δημιουργηθούν καινοτόμες και προηγμένες τεχνικές και τεχνολογίες, οι οποίες να είναι οικονομικά αποδοτικές για να επιτευχθεί η σύλληψη, αποθήκευση, διανομή, διαχείριση και ανάλυση των πληροφοριών με σκοπό να παραχθεί η νέα γνώση, να μπορούν να ληφθούν αποφάσεις και να αυτοματοποιηθούν οι διαδικασίες [32]. Η παραπάνω θεωρία του ομίλου Gartner συναντάται στη

βιβλιογραφία ως η θεωρία των 3Vs και ο όγκος, η ταχύτητα και η ποικιλία αποτελούν τους τρεις βασικούς άξονες χαρακτηριστικών της θεωρίας αυτής [33]:

- *Όγκος (Volume):* Ο όγκος των δεδομένων είναι η ασύλληπτη ποσότητα δεδομένων που παράγονται ανά δευτερόλεπτο από την ταυτόχρονη χρήση της τεχνολογίας και του διαδικτύου. Η παραγόμενη γνώση που μπορεί να εξαχθεί από την σωστή επεξεργασία τους έχει μεγάλη αξία για εταιρείες και οργανισμούς, επηρεάζοντας τις αποφάσεις τους και βελτιώνοντας τις αποδόσεις τους [33].

- *Ταχύτητα (Velocity):* Η ταχύτητα καταγράφει τον ρυθμό παραγωγής των δεδομένων μέσω συστημάτων ή αισθητήρων που χρησιμοποιούν οι εφαρμογές, την ταχύτητα επεξεργασία τους και την ανάλυσή τους ακόμη και σε πραγματικό χρόνο ή σε σχεδόν πραγματικό χρόνο. Επιπλέον, η ταχύτητα καταγράφει και τον ρυθμό αλληλεπίδρασης μεταξύ των δεδομένων που είναι αποθηκευμένα σε διαφορετικές βάσεις δεδομένων με τα δεδομένα που παράγονται σε πραγματικό χρόνο [33].

- *Ποικιλία (Variety):* Στις μέρες μας οι εφαρμογές μπορούν να επεξεργαστούν δεδομένα διαφορετικού τύπου και διαφορετικών μορφών, έναντι των δεδομένων παρόμοιου τύπου που συγκεντρώνονταν άλλοτε στις σχεσιακές βάσεις δεδομένων. Υπάρχουν δεδομένα, τα οποία έχουν δομημένη μορφή και μπορούν να αναγνωστούν εύκολα από τα υπολογιστικά συστήματα, αλλά και δεδομένα που έχουν ημιδομημένη μορφή ή αδόμητη μορφή όπου η πληροφορία δεν έχει καμία σχέση με τα προκαθορισμένα μοντέλα δεδομένων που αποθηκεύονταν στις σχεσιακές βάσεις δεδομένων. Τα δεδομένα αυτά είναι αρχεία διάφορων μορφών, όπως: κείμενο, αρχεία ήχου ή βίντεο τα οποία χρειάζονται ειδικό χειρισμό κατά την ανάλυσή τους [33].

3.2 Η επέκταση του μοντέλου 3Vs σε 5Vs

Με την αυξανόμενη ανάπτυξη της τεχνολογίας η θεωρία του ομίλου Gardner έπρεπε να αναπροσαρμοστεί και να προστεθούν δύο ακόμα χαρακτηριστικά (Vs) που θα απέδιδαν καλύτερα τον όρο των Μεγάλων δεδομένων. Η αξία (value) και η εγκυρότητα (veracity) [34].

- *Αξία (Value):* Τα δεδομένα μεγάλου όγκου για να είναι χρήσιμα σε μία εταιρεία ή οργανισμό πρέπει να επεξεργαστούν κατάλληλα για να δημιουργούν αξία, με σκοπό την διατήρησή τους στην αγορά και την αύξηση της ανταγωνιστικότητάς τους. Με τον όρο «δημιουργία αξίας» για μια εταιρεία, εννοείται η κατάλληλη επεξεργασία και ανάλυση των δεδομένων προκειμένου μέσα από την καταναλωτική συνήθεια που έχουν και τις αναζητήσεις που εκτελούν οι πελάτες, η εταιρεία να κατανοήσει τις συμπεριφορές και τις επιθυμίες τους. Με αυτό τον τρόπο μπορεί να επικεντρωθεί στοχευμένα σε επιχειρηματικές διαδικασίες και λειτουργίες με στόχο την παροχή προϊόντων και υπηρεσιών που ανταποκρίνονται καλύτερα στις τρέχουσες και μελλοντικές ανάγκες της αγοράς αποφέροντας έτσι μεγαλύτερο κέρδος. Ωστόσο, οι διαδικασίες της αποθήκευσης, συντήρησης και επεξεργασίας των δεδομένων είναι ιδιαίτερα κοστοβόρες [35].

- *Εγκυρότητα (Veracity):* Η εγκυρότητα των δεδομένων, είναι ένα χαρακτηριστικό μεγάλης σημασίας για την αξία του παραγόμενου αποτελέσματος. Πολλά από τα δεδομένα που συλλέγονται μπορεί να διαφέρουν ποιοτικά, το οποίο πρέπει να ληφθεί υπόψη κατά την επεξεργασία των δεδομένων, ώστε τα αποτελέσματα να είναι έγκυρα [35].

Υπάρχουν βέβαια και άλλα χαρακτηριστικά (Vs) που χαρακτηρίζουν τα Μεγάλα δεδομένα, όπως η μεταβλητότητα (Variability), αφού οι πηγές των δεδομένων είναι πλέον δυναμικές, εξελισσόμενες και μεταβάλλονται ανάλογα με το χρόνο και τον τόπο. Η αξιοπιστία (Validity) είναι ένα χαρακτηριστικό που σχετίζεται με τις παρεμβολές, το θόρυβο και τις διάφορες δυσκολίες που προκύπτουν κατά τη συλλογή των δεδομένων και τη μεγάλη ποικιλία των μορφών των μεγάλων δεδομένων που συλλέγονται πια.

Στις μέρες μας, τα χαρακτηριστικά των Vs των Μεγάλων δεδομένων έχουν ξεπεράσει τα 42Vs και σίγουρα θα προστεθούν και άλλα, λόγω των τεχνολογιών που θα προκύψουν για τη σωστή επεξεργασία τους, για να είναι αξιοποιήσιμα στους τομείς των οικονομικών, της διοίκησης των επιχειρήσεων καθώς και για τη λήψη αποφάσεων.

3.3 Η Δομή των Μεγάλων Δεδομένων

Τα Δεδομένα που συλλέγονται και αποθηκεύονται στις βάσεις δεδομένων, όπως έχει ήδη αναφερθεί, είναι πολύ χρήσιμα για την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων σε πολλούς τομείς της κοινωνίας. Όμως για να γίνει αυτό χρειάζεται να εξασφαλιστεί η αξιοπιστία των δεδομένων. Πολλές φορές όταν συλλέγονται τα δεδομένα μπορεί να υπάρχουν ανακολουθίες, ελλείψεις ή πλεονάζουσες εγγραφές, γι΄ αυτό τον λόγο στο στάδιο της ανάλυσης, τα δεδομένα φιλτράρονται και με διάφορους μηχανισμούς ελέγχονται έτσι ώστε να προκύψουν χρήσιμα δεδομένα για περαιτέρω ανάλυση και προβλέψεις [36]. Είναι πολύ χρήσιμο να γνωρίζουμε σε ποιες κατηγορίες διακρίνονται τα δεδομένα, τις πηγές από τις οποίες συλλέγονται και τους τύπους των δεδομένων που υπάρχουν, ώστε να εφαρμόζονται οι σωστοί μηχανισμοί ανάλυσης.

- *Κατηγορίες Δεδομένων :*

    - Τα δομημένα δεδομένα (structured data) προέρχονται από τις συναλλαγές που πραγματοποιούν οι χρήστες. Σε αυτά τα δεδομένα καταγράφονται με ακρίβεια όλα τα στοιχεία σε όλα τα πεδία της συναλλαγής προκειμένου αυτή να ολοκληρωθεί και έχουν συγκεκριμένη μορφή και δομή / σχήμα. Αυτά τα δεδομένα αποθηκεύονται σε σχεσιακές βάσεις δεδομένων (RDBMS) και σε αυτά μπορούν να γίνουν διάφοροι μαθηματικοί υπολογισμοί και άμεσοι συσχετισμοί και διασταυρώσεις [36].
    - Τα ημιδομημένα δεδομένα (semistructured data) είναι τα δεδομένα που έχουν μια χαλαρή δομή, η οποία δεν είναι συνήθως προκαθορισμένη αλλά επιτρέπει την εύκολη επεξεργασία τους και πολλές φορές την εξαγωγή δομημένης πληροφορίας. Ενδεικτικά παραδείγματα είναι πληροφορίες που έχουν αποτυπωθεί σε μορφή XML ή JSON κλπ.
    - Τα μη δομημένα δεδομένα (unstructured data) είναι τα δεδομένα που έχουν τη μορφή κειμένου ή πολυμεσικού περιεχομένου (π.χ. τα ηλεκτρονικά μηνύματα, βίντεο κ.α.) και δεν ακολουθούν κανένα σχήμα ούτε διέπονται από κάποιους κανόνες. Αυτή η κατηγορία δεδομένων απαιτεί περισσότερη επεξεργασία προτού χρησιμοποιηθούν για οποιαδήποτε ανάλυση [36].

- *Τύποι Δεδομένων:* Τα δεδομένα που συλλέγονται διακρίνονται σε δύο τύπους: τα συνεχή και τα κατηγορικά.

    - Τα συνεχή δεδομένα είναι τα δεδομένα που παίρνουν τιμές από ένα περιορισμένο ή απεριόριστο συνεχές σύνολο, για παράδειγμα τα ποσά των τραπεζικών συναλλαγών [36].

    - Τα κατηγορικά δεδομένα διακρίνονται σε υποκατηγορίες:

        - *Τα ονομαστικά δεδομένα (nominal data):* Τα δεδομένα αυτά μπορούν να λάβουν τιμές από ένα περιορισμένο σύνολο τιμών πχ η επαγγελματική ιδιότητα, η διεύθυνση μόνιμης κατοικίας κ.α. Οι τιμές του συνόλου δεν έχουν κάποια φυσική σειρά ούτε έχει έννοια η ταξινόμησή τους.

        - *Τα τακτικά δεδομένα (ordinal data):* Τα δεδομένα αυτά μπορούν να λάβουν τιμές από ένα περιορισμένο σύνολο τιμών με τη διαφορά ότι οι τιμές του συνόλου έχουν φυσική σειρά. Για παράδειγμα η επαγγελματική ιδιότητα σε κωδικοποιημένη μορφή ως ελεύθερος επαγγελματίας, δημόσιος υπάλληλος κ.α.

        - *Τα δυαδικά δεδομένα (binary data):* Τα δεδομένα αυτά μπορούν να λάβουν μόνο μία από τις δύο διαθέσιμες τιμές, όπως για παράδειγμα, το φύλο ενός ατόμου κ.α. [36].

Κατά την συλλογή δεδομένων είναι σημαντικό αυτά να κατηγοριοποιούνται σωστά προκειμένου στη συνέχεια οι αλγόριθμοι επεξεργασίας να κάνουν σωστά την ανάλυσή τους. Για παράδειγμα, αν τα δεδομένα που συλλέγονται για το πεδίο επαγγελματική ιδιότητα χαρακτηριστούν ως συνεχούς τύπου δεδομένα αντί για κατηγορικά, τότε το λογισμικό θα προχωρήσει να κάνει διάφορους μαθηματικούς υπολογισμούς (πχ υπολογισμό μέσης τυπικής απόκλισης) που δεν συνάδουν με το συγκεκριμένο τύπο δεδομένων.

- *Πηγές Δεδομένων:* Τα δεδομένα μπορούν να αντληθούν από διάφορες πηγές και έχει σημασία να τις γνωρίζουμε. Έτσι, ενδεικτικά, μπορούμε να διακρίνουμε τα δεδομένα ανάλογα με την πηγή τους σε:

    - Δεδομένα που αντλούνται από το διαδίκτυο. Τα δεδομένα αυτά είναι πολύ χρήσιμα γιατί μπορούν να δώσουν πολλές πληροφορίες στις εταιρείες σχετικά με τις καταναλωτικές συνήθειες των χρηστών (είτε με τις ηλεκτρονικές αγορές που πραγματοποιούν, είτε με τις αναζήτησεις σε διάφορες ιστοσελίδες). Η Yahoo Finance και η Bloomberg είναι εφαρμογές που αποθηκεύουν ιστορικά και δεδομένα πραγματικού χρόνου, κυρίως οικονομικού τύπου, και είναι διαθέσιμα σε όποιον ενδιαφέρεται. Όσα περισσότερα δεδομένα έχουν στη διάθεσή τους οι εταιρείες τα αξιοποιούν κατάλληλα για να επενδύσουν σε πιο στοχευμένες διαφημίσεις, βελτιώνοντας τις παροχές και τις υπηρεσίες τους με σκοπό να αυξήσουν τις πωλήσεις τους [37].

    - Δεδομένα που αντλούνται από “έξυπνα δίκτυα” και αισθητήρες (IoT). Τα δεδομένα αυτά συλλέγονται από αγωγούς πετρελαίων, ανεμογεννήτριες, περιβαλλοντικούς σταθμούς κ.α. και δίνουν πληροφορίες για την απόδοση αυτών των μηχανημάτων, τυχόν προβλήματα που μπορεί να προκύψουν σε αυτά και πώς μπορούν να επιλυθούν.

    - Δεδομένα που αντλούνται από GPS και κινητά τηλέφωνα, τα οποία οποιαδήποτε στιγμή δείχνουν την τοποθεσία των χρηστών και σε ποια χρονική στιγμή.

    - Δεδομένα που προέρχονται από τα social media. Τα social media αποτελούν μία πηγή που δίνει πολύ χρήσιμες πληροφορίες στις εταιρείες αναφορικά με τις καταναλωτικές επιθυμίες των χρηστών. Τα δεδομένα αυτά αντλούνται όχι μόνο από τις επισκέψεις των χρηστών σε κάποιες ιστοσελίδες, αλλά προκύπτουν και από τον κύκλο των “φίλων” των χρηστών στα δίκτυα αυτά και από τα διάφορα πολυμέσα που αναρτούν (εικόνες, βίντεο κ.α) [38].

- Τέλος υπάρχουν και τα ανοιχτά δεδομένα, τα οποία συλλέγονται από διάφορους δημόσιους οργανισμούς, πανεπιστήμια, στατιστικές εταιρείες και πολλές άλλες πηγές.

3.4 Τεχνικές και εργαλεία ανάλυσης των Μεγάλων Δεδομένων

Ο μεγάλος όγκος δεδομένων απαιτεί να βρεθούν τρόποι και τεχνικές διαχείρισης, όχι μόνο για την αποθήκευση, αλλά και για την επεξεργασία τους σε περιορισμένους χρόνους εκτέλεσης, προκειμένου να μετατραπούν σε χρήσιμες πληροφορίες για τις εταιρείες και τους οργανισμούς. Υπάρχουν πλέον πολλές τεχνικές και εργαλεία ανάλυσης δεδομένων, που πολλές φορές επικαλύπτονται μεταξύ τους, τα οποία βασίζονται σε τρεις επιστημονικούς τομείς: τα μαθηματικά, τη στατιστική και την πληροφορική.

Ενδεικτικά, μερικές τεχνικές ανάλυσης των μεγάλων δεδομένων είναι:

- Η Εξόρυξη Δεδομένων (Data Mining) αποτελεί μία τεχνική για την εξαγωγή πολύτιμων πληροφοριών από δεδομένα που πολλές φορές είναι φαινομενικά ασυσχέτιστα. Μέσα από αυτή την τεχνική επιτυγχάνεται ο εντοπισμός μοτίβων που οδηγεί στην ομαδοποίηση και ταξινόμηση των δεδομένων ώστε να έχουν νόημα και να μπορούν να εξαχθούν λογικά συμπεράσματα.

- Η Μηχανική Μάθηση, η οποία χρησιμοποιείται για το σχεδιασμό αλγορίθμων που επιτρέπουν στα συστήματα να παίρνουν έξυπνες αποφάσεις, οι οποίες βασίζονται σε εμπειρία που έχουν αποκτήσει μετά από εκπαίδευση και όχι βάση προγραμματισμένης λογικής. Ορισμένα μοντέλα μηχανικής μάθησης επιτρέπουν τη διαρκή αυτόματη εκπαίδευσή τους κατά την παραγωγική λειτουργία τους ώστε να μπορούν να εξελίσσουν διαφορετικές συμπεριφορές ανάλογα με τις περιστάσεις και οι επιχειρήσεις να μπορούν να λαμβάνουν έξυπνες αποφάσεις.

- Η Οπτικοποίηση, η οποία χρησιμοποιείται για τη δημιουργία πινάκων, διαγραμμάτων και άλλων αναπαραστάσεων για την κατανόηση των δεδομένων.

- Τα Τεχνητά Νευρωνικά Δίκτυά (ANN) είναι μια προηγμένη τεχνική, βασισμένη στον τρόπο που λειτουργεί ο ανθρώπινος εγκέφαλος, που βρίσκει εφαρμογή στην αναγνώριση προτύπων, τον προσαρμοστικό έλεγχο, την ανάλυση εικόνας και πολλά άλλα. Αποτελεί μια ειδική περίπτωση μηχανικής μάθησης που ξεχωρίζει λόγω του τρόπου υλοποίησής της. Χαρακτηρίζεται από αυξημένες απαιτήσεις σε υπολογιστική δύναμη αλλά έχει μεγάλα πλεονεκτήματα λόγω της αυξημένης απόδοσής της ακόμα και σε πολύπλοκα προβλήματα όπου οι υπόλοιπες τεχνικές μηχανικής μάθησης αποτυγχάνουν.

- Η Ανάλυση Κοινωνικών Δικτύων (SNA) είναι μία σημαντική τεχνική που χρησιμοποιείται στη σύγχρονη κοινωνιολογία, παρατηρεί τις κοινωνικές σχέσεις και χρησιμοποιεί κόμβους, δεσμούς. κ.α. για τη δημιουργία και ανάλυση διαγραμμάτων αλληλεπίδρασης και την εξαγωγή συμπερασμάτων [39][40].

Για τη υλοποίηση των παραπάνω τεχνικών χρησιμοποιούνται κυρίως εργαλεία όπως στοιχειώδη μαθηματικά, στατιστική και διάφορες μέθοδοι βελτιστοποίησης:

- Τα στοιχειώδη μαθηματικά αποτελούν βασικό εργαλείο στην προετοιμασία των δεδομένων πριν την επεξεργασία και ανάλυσή τους, αλλά και τη βάση για όλα σχεδόν τα εργαλεία και τις μεθόδους ανάλυσης.

- Η Βελτιστοποίηση (optimization), χρησιμοποιείται για την επίλυση ποσοτικών προβλημάτων σε τομείς όπως της βιολογίας, της οικονομίας και της μηχανικής. Στην ανάλυση μεγάλων δεδομένων χρησιμοποιείται κυρίως για εκπαίδευση των ANN καθώς και για την ανάπτυξη των μοντέλων μηχανικής μάθησης.

- Η Στατιστική, περιλαμβάνει τη συλλογή, την οργάνωση και την ερμηνεία των δεδομένων και συνήθως χρησιμοποιείται για να περιγραφεί η συσχέτιση μεταξύ διαφορετικών στόχων αλλά και για την υλοποίηση ορισμένων μοντέλων μηχανικής μάθησης.

Οι παραπάνω τεχνικές των μεγάλων δεδομένων περιλαμβάνουν κατανεμημένα υπολογιστικά συστήματα, συστήματα αρχείων, συλλογή δεδομένων και αποθήκευση που βασίζεται στο cloud και σε τοπικούς διακομιστές.

Επιπρόσθετα, υπάρχει πληθώρα εργαλείων ανάλυσης των μεγάλων δεδομένων που μπορούν να βοηθήσουν την κάθε εταιρεία να εξάγει χρήσιμες πληροφορίες και να εξοικονομήσει έτσι χρόνο και χρήματα. Μερικά από τα εργαλεία αυτά είναι:

- Το Apache Hadoop είναι μια βιβλιοθήκη-πλαίσιο κατανεμημένης επεξεργασίας και ανάλυσης μεγάλων δεδομένων ανοιχτού κώδικα βασισμένο σε Java, που χρησιμοποιείται από πολλές μεγάλες εταιρείες. Είναι γνωστό γιατί χρησιμοποιεί απλά μοντέλα προγραμματισμού για την ανάπτυξη των αλγορίθμων επεξεργασίας και για τις εξαιρετικές του ικανότητες κλιμάκωσης από ένα υπολογιστή μέχρι συστοιχίες χιλιάδων υπολογιστών, εκμεταλλευόμενο τοπικά την υπολογιστική ισχύ και τον αποθηκευτικό χώρο ανάλογα με τις ανάγκες. Επιπλέον επιτρέπει τη δημιουργία ενός αξιόπιστου συστήματος αντιμετωπίζοντας τις αστοχίες του υλικού σε επίπεδο εφαρμογής χωρίς τη χρήση εξειδικευμένου και κατά συνέπεια ακριβού εξοπλισμού.

- Το Cloudera είναι ένα εμπορικό σήμα για το Hadoop με μερικές επιπλέον υπηρεσίες. Το Cloudera είναι μια επιχειρηματική λύση που βοηθά τις εταιρείες να διαχειρίζονται καλύτερα το οικοσύστημα του Hadoop. Οι επιχειρήσεις χρησιμοποιούν το Cloudera για να δημιουργήσουν ένα αποθετήριο δεδομένων, στο οποίο μπορούν να έχουν πρόσβαση οι εταιρικοί χρήστες για διάφορους σκοπούς.

- Το MongoDB είναι μια βάση δεδομένων η οποία είναι προσαρμοσμένη στη διαχείριση μεγάλου όγκου κατανεμημένων δεδομένων που είναι αδόμητα ή ημι-δομημένα. Έχει επίσης εξελιχθεί ώστε να μπορεί να διαχειρίζεται δεδομένα που αλλάζουν συχνά, όπως χρονοσειρές, ροές δεδομένων πραγματικού χρόνου και δεδομένα από συσκευές και αισθητήρες. Υποστηρίζει τη γρήγορη αναζήτηση των δεδομένων σε εγγυημένο χρονικό πλαίσιο και τη γεωγραφική κατανομή τους ώστε να είναι άμεσα διαθέσιμα εκεί που χρειάζονται. Χρησιμοποιείται συχνά στην αποθήκευση δεδομένων από εφαρμογές για κινητά, καταλόγους προϊόντων, συστήματα διαχείρισης περιεχομένου και πολλά άλλα.

- Το Apache Cassandra επιτρέπει στους χρήστες να επεξεργάζονται δομημένα σύνολα δεδομένων που διανέμονται σε έναν τεράστιο αριθμό κόμβων παγκοσμίως. Η Cassandra είναι μια δημοφιλής βάση δεδομένων που προσφέρει υψηλή διαθεσιμότητα και επεκτασιμότητα και βελτιώνει την απόδοση του υλικού και της υποδομής cloud. Μερικά από τα κύρια πλεονεκτήματα της Cassandra περιλαμβάνουν υψηλή απόδοση, ανοχή σε σφάλματα, αποκέντρωση, ανθεκτικότητα και εξαιρετική υποστήριξη.

- Το Hive είναι γνωστό για τη διαχείριση κατανεμημένων δεδομένων για το Hadoop. Υποστηρίζει ερωτήματα τύπου SQL για πρόσβαση σε μεγάλα δεδομένα και χρησιμοποιείται για σκοπούς εξόρυξης δεδομένων.

- Το Apache Spark είναι μια εναλλακτική, ελαφρώς διαφορετική, του Hadoop. Επιτρέπει να εκτελούνται εργασίες MapReduce με μεγαλύτερη ταχύτητα, αφού τα δεδομένα διατηρούνται διαρκώς στη μνήμη σε αντίθεση με το Hadoop όπου τα δεδομένα αποθηκεύονται σε μέσο αποθήκευσης. Το Apache Spark ανοίγει νέες ευκαιρίες για διαδικασίες που επεξεργάζονται ροές δεδομένων. Ο εντοπισμός απάτης, η επεξεργασία αρχείων καταγραφής και δεδομένων συναλλαγών γίνονται ευκολότερα και ταχύτερα με το Apache Spark.

- Το Apache Storm είναι ιδανικό για δεδομένα που συλλέγονται σε πραγματικό χρόνο. Μπορεί να ενσωματωθεί στο Hadoop ή να χρησιμοποιηθεί μόνο του για να βελτιστοποιήσει τις διαδικασίες.

- Η Talend είναι μια εξαιρετική εταιρεία ανοιχτού κώδικα που είναι γνωστή για την παροχή διαφόρων προϊόντων δεδομένων ώστε να μπορεί κάθε εταιρεία να έχει το δικό της σύστημα διαχείρισης δεδομένων [40].

3.5 Η Γλώσσα Python στην ανάλυση Μεγάλων Δεδομένων

Η Python είναι μια διερμηνευόμενη (interpreted), αντικειμενοστραφής γλώσσα προγραμματισμού υψηλού επιπέδου με δυναμική σημασιολογία. Ο δημιουργός της είναι ο Guido van Rossum και για πρώτη φορά κυκλοφόρησε το 1991. Οι ενσωματωμένες δομές δεδομένων υψηλού επιπέδου, σε συνδυασμό με το δυναμικό

σύστημα τύπων (dynamic typing) και τη δυναμική δέσμευση (dynamic binding), την καθιστούν πολύ ελκυστική για ταχεία ανάπτυξη εφαρμογών, καθώς και για χρήση ως γλώσσα σεναρίου (scripting) ή για τη σύνδεση υπαρχόντων στοιχείων μεταξύ τους.

Στην παρούσα διπλωματική χρησιμοποιήθηκε η γλώσσα Python για την καλύτερη ανάλυση των μεγάλων δεδομένων που προκύπτουν από την αγορά των κρυπτονομισμάτων και για την εφαρμογή του αλγορίθμου της Μηχανικής Μάθησης. Διαθέτει εξαιρετικές βιβλιοθήκες για την επεξεργασία δεδομένων και είναι σχετικά εύκολη στην εκμάθησή της, δίνει έμφαση στην αναγνωσιμότητα και ως εκ τούτου μειώνει το κόστος συντήρησης του προγράμματος [41].

Οι πιο αξιοσημείωτες βιβλιοθήκες της για ανάλυση δεδομένων είναι οι :

- NumPy, για επιστημονικούς υπολογισμούς και κυρίως για τις πράξεις μεταξύ πινάκων
- Pandas, για ανάλυση και επεξεργασία δεδομένων με χρήση πλαισίων δεδομένων (Dataframes)
- Matplotlib, για δημιουργία γραφημάτων
- Scikit-learn, για μηχανική μάθηση.

Τέλος, όσον αφορά την ανάλυση δεδομένων προτιμάται η χρήση ως περιβάλλον ανάπτυξης (Integrated Development Environment) του IPython Notebook (γνωστό και ως Jupyter) [42].

3.6 Η Μηχανική Μάθηση στα Μεγάλα Δεδομένα

Η Μηχανική Μάθηση αποτελεί μία από τις τεχνικές ανάλυσης των μεγάλων δεδομένων. Σύμφωνα με τον Άρθουρ Σάμουελ, που επινόησε τον όρο το 1959, η μηχανική μάθηση είναι ένα πεδίο μελέτης, στον τομέα της τεχνητής νοημοσύνης, που μέσω στατιστικών τεχνικών δίνει τη δυνατότητα σε έναν υπολογιστή να μαθαίνει από τα δεδομένα, δίχως να έχει ρητά προγραμματιστεί. Η μηχανική μάθηση είναι κομμάτι τριών επιστημών: της επιστήμης των υπολογιστών, του προγραμματισμού και της στατιστικής, οι οποίες συνδυάζονται και μελετούν την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα, να τα αναλύουν και να προχωρούν σε προβλέψεις [43].

3.7 Τεχνικές Μηχανικής Μάθησης

Στη μηχανική μάθηση χρησιμοποιούνται τρεις τεχνικές εκπαίδευσης: α) επιτηρούμενη μάθηση (Supervised learning), β) μη επιτηρούμενη μάθηση (Unsupervised learning) και γ) ενισχυτική μάθηση (Reinforcement learning). Και οι τρεις αν συνδυαστούν μπορούν να επιτύχουν θεαματικά αποτελέσματα [44] [45]. Αναλυτικά:

- *Supervised learning (επιτηρούμενη μάθηση):* Είναι η τεχνική, όπου ένα υπολογιστικό σύστημα εκπαιδεύεται με τη χρήση ενός συνόλου δεδομένων εισόδου για τα οποία του δίνονται και οι αντίστοιχες επιθυμητές τιμές εξόδου. Ο αλγόριθμος χρησιμοποιεί τα ζεύγη γνωστών εισόδων-εξόδων για να προσαρμόσει τις εσωτερικές παραμέτρους του ώστε να προκύψει ένας γενικός κανόνας, ο οποίος αντιστοιχεί στις εισόδους και στα αποτελέσματα που εξάγονται. Η τεχνική αυτή είναι κατάλληλη για προβλήματα regression (παλινδρόμησης) και classification (ταξινόμησης) [44].

- *Unsupervised learning (μη επιτηρούμενη μάθηση):* Είναι η τεχνική, όπου ένα υπολογιστικό σύστημα εκπαιδεύεται μόνο του. Χρησιμοποιεί αποκλειστικά ένα σύνολο δεδομένων εισόδου χωρίς να είναι γνωστά ή να του δίνονται οι αντίστοιχες επιθυμητές τιμές εξόδου. Επομένως, ο αλγόριθμος προσπαθεί μόνος του να δημιουργήσει τον γενικό κανόνα εντοπίζοντας κρυφές συσχετίσεις στα δεδομένα εισόδου. Χρησιμοποιείται κυρίως για προβλήματα dimensionality reduction (μείωσης διαστάσεων) και clustering (ομαδοποίησης) [44][45].

- *Reinforcement learning (ενισχυτική μάθηση):* Είναι η τεχνική, όπου επιτρέπει σε ένα υπολογιστικό σύστημα να εκπαιδευτεί και να μάθει τη συμπεριφορά του μέσω της ανατροφοδότησης που λαμβάνει από το περιβάλλον. Ο αλγόριθμος προσπαθεί μόνος του να δημιουργήσει τον γενικό κανόνα από τις μεταβλητές που χρησιμοποιεί ως εισόδους. Μετά από μία σειρά αποφάσεων που παίρνει χωρίς επίβλεψη, του δίνεται μία ανταμοιβή +1 ή -1. Ανάλογα με την ανάδραση που λαμβάνει ο αλγόριθμος συνεχίζει ή επαναπρογραμματίζει τις διαδρομές που θα ακολουθήσει για να επιτύχει το τελικό αποτέλεσμα [44][45].

3.8 Βήματα που ακολουθούμε στη Μηχανική Μάθηση

Για να αναπτύξουμε και να εφαρμόσουμε στα δεδομένα μας διάφορα μοντέλα μηχανικής μάθησης πρέπει να ακολουθήσουμε κάποια βασικά στάδια, τα οποία είναι τα εξής:

Στο **πρώτο στάδιο** της μηχανικής μάθησης γίνεται η συλλογή των δεδομένων, τα οποία αντλούνται απευθείας από πηγές δομημένων δεδομένων, από δεδομένα internet (web scraping), από API, κ.λπ., καθώς η μηχανική μάθηση μπορεί να λειτουργήσει τόσο με δομημένα όσο και με μη δομημένα δεδομένα (φωνή, εικόνα και κείμενο).

Στο **δεύτερο στάδιο** γίνεται προετοιμασία των δεδομένων που έχουν συλλεχθεί και διαχείριση των ακραίων τιμών (outliers) και των ελλιπών δεδομένων (Missing values), ώστε να συμφωνούν με τις απαιτήσεις του αλγορίθμου που έχει επιλεγεί για τη συγκεκριμένη διαδικασία. Τα δεδομένα μορφοποιούνται κατάλληλα, ενώ οι προβληματικές τιμές συνήθως αντικαθίστανται με τον μέσο όρο ή την διάμεσο ή κ.α.

Στο **τρίτο στάδιο** αναλύονται τα δεδομένα προκειμένου να εντοπιστούν τυχόν κρυμμένα μοτίβα και σχέσεις μεταξύ των μεταβλητών. Αυτό γίνεται ακολουθώντας τη μηχανική χαρακτηριστικών δεδομένων (feature engineering) σε συνδυασμό με πληροφορία που προκύπτει από τη γνώση του υπό μελέτη αντικειμένου (domain knowledge) και συνήθως λύνει το 70% των προβλημάτων.

Στο **τέταρτο στάδιο** τα δεδομένα που θα χρησιμοποιηθούν χωρίζονται σε τρία υποσύνολα (εκπαίδευσης, επαλήθευσης και δοκιμών) και εκπαιδεύεται ο αλγόριθμος με το σύνολο των δεδομένων εκπαίδευσης.

Στο **πέμπτο** στάδιο, χρησιμοποιούνται τα δεδομένα επαλήθευσης για να πραγματοποιηθούν μικρορυθμίσεις στις παραμέτρους του μοντέλου και να αποφευχθεί η υπερεφαρμογή.

Στο **έκτο** στάδιο γίνεται πλέον αξιολόγηση του μοντέλου χρησιμοποιώντας το σύνολο δεδομένων ελέγχου για να μετρηθεί η ακρίβεια των προβλέψεων.

Τέλος, εφόσον η ακρίβεια των προβλέψεων είναι ικανοποιητική για το σκοπό που προορίζονται, το μοντέλο μπορεί να χρησιμοποιηθεί σε πραγματικά δεδομένα [44][45].

3.9 Αλγόριθμοι Μηχανικής Μάθησης

Η επιλογή του κατάλληλου αλγόριθμου μηχανικής μάθησης εξαρτάται από πολλούς παράγοντες, όπως το μέγεθος των δεδομένων, την ποιότητα και ποικιλομορφία τους

καθώς και το είδος των αποτελεσμάτων που θέλουμε να αποκομίζουμε από την επεξεργασία τους. Άλλοι παράγοντες που επηρεάζουν την επιλογή είναι η ταχύτητα εκπαίδευσης και η ακρίβεια των αποτελεσμάτων. Σε κάθε περίπτωση η επιλογή του αλγορίθμου θα πρέπει να γίνει με βάση τις εκάστοτε ανάγκες.

Στη συνέχεια αναφέρονται μερικοί από τους πιο διαδεδομένους αλγόριθμους μηχανικής μάθησης [45]:

Supervised learning

- Classification problems (ταξινόμηση)
    - Logistic regression (αλγοριθμική παλινδρόμηση)
- Lasso and ridge regression
- Decision trees (δέντρα αποφάσεων)
- Bagging
    - Random forest (τυχαία δέντρα αποφάσεων)
- Boosting (adaboost, gradient boost, and xgboost)
- SVM (Support Vector Machines)
- Recommendation engine (μηχανή προτάσεων)
- Linear regression (γραμμική παλινδρόμηση)

Unsupervised learning

- Principal component analysis (PCA)
- K-means clustering (συσταδοποίηση)

Reinforcement learning

- Markov decision process (αλυσίδα Μαρκόφ)
- Monte Carlo methods (Μέθοδοι Μόντε Κάρλο)
- Temporal difference learning

ΚΕΦΑΛΑΙΟ 4

ΚΑΤΑΝΕΜΗΜΕΝΑ ΣΥΣΤΗΜΑΤΑ

Με τον όρο «κατανεμημένο σύστημα» εννοούμε οποιοδήποτε υπολογιστικό σύστημα αποτελείται από περισσότερο από έναν υπολογιστές που είναι φυσικά διακριτοί, αλλά είναι διασυνδεδεμένοι και επικοινωνούν μεταξύ τους ώστε να συντονίζουν τις ενέργειές τους για την επίτευξη ενός κοινού στόχου. Υπάρχουν πολλοί διαφορετικοί λόγοι για τους οποίους μπορεί να απαιτείται η ανάπτυξη και χρήση ενός κατανεμημένου συστήματος. Στην περίπτωση όμως των μεγάλων δεδομένων, ο πρωταρχικός λόγος είναι η αύξηση της υπολογιστικής ισχύος για την αποθήκευση και τη διαχείρισή τους.

4.1 Κατανεμημένα Συστήματα για τη διαχείριση των Μεγάλων Δεδομένων

Κατά την προσπάθεια διαχείρισης και ανάλυσης των τεράστιων συνόλων δεδομένων που παράγονται ημερησίως προκύπτουν ορισμένα σημαντικά ζητήματα:

- ***Θέματα αποθήκευσης και μεταφοράς των μεγάλων δεδομένων***
  Ο μεγάλος όγκος των δεδομένων που παράγονται ημερησίως προκαλεί σημαντικές δυσκολίες όσο αφορά την αποθήκευση και τη μεταφορά τους. Οι παραδοσιακοί τρόποι αποθήκευσης δεν δύναται να χρησιμοποιηθούν, επειδή δεν έχουν επιτευχθεί ακόμα ανάλογες, του τεράστιου όγκου των δεδομένων, ταχύτητες στη διακίνηση και ακόμα περισσότερο στην καταγραφή τους στα αποθηκευτικά μέσα, γεγονός που αυξάνει δραματικά τον απαιτούμενο χρόνο. Μείωση του χρόνου αποθήκευσης και διακίνησης μπορεί να επιτευχθεί με τα κατανεμημένα υπολογιστικά συστήματα, όπως το κατανεμημένο σύστημα αρχείων της Google, τα οποία χρησιμοποιούν πολλές χιλιάδες υπολογιστές λειτουργώντας σαν ένας, διαθέτοντας πληθώρα λογισμικών και υπολογιστικών

πόρων. Για να επιτευχθεί αυτό οι υπολογιστές είναι διασυνδεδεμένοι μεταξύ τους αποτελώντας κόμβους ενός δικτύου υψηλής ταχύτητα. Ωστόσο ακόμα και η μεταφορά τόσο μεγάλου όγκου δεδομένων μεταξύ των κόμβων, απαιτεί συχνά πολλές ώρες οδηγώντας σε υπερφόρτωση των δίκτυων επικοινωνίας. Για παράδειγμα, η μεταφορά ενός Exabyte δεδομένων με ταχύτητα δικτύου 1 Gb ανά δευτερόλεπτο απαιτεί 2800 ώρες. Μια λύση σε αυτό το πρόβλημα είναι η επεξεργασία των δεδομένων στη «πηγή» ώστε να μεταδίδονται μόνο τα αποτελέσματα της έρευνας.

- **Θέματα διαχείρισης των μεγάλων δεδομένων**
  Οι υπάρχουσες τεχνολογικές λύσεις για τη διαχείριση δεδομένων δεν μπορούν να ανταποκριθούν στον όγκο αλλά και στη φύση των μεγάλων δεδομένων. Το τεράστιο μέγεθός τους σε συνδυασμό με την πολυπλοκότητα της ανάλυσής τους και την επιτακτική ανάγκη να δημιουργηθεί αξία από αυτά, έχει οδηγήσει σε μία νέα κατηγορία από τεχνολογίες και εργαλεία για τη διαχείρισή τους, όπως το MapReduce, τα RDDs κλπ., ειδικά σχεδιασμένα ώστε να λειτουργούν πάνω σε κατανεμημένα υπολογιστικά συστήματα.

Εκτός από τα παραπάνω προβλήματα που λύνουν τα κατανεμημένα συστήματα προσφέρουν πολλά ακόμη πλεονεκτήματα, έναντι των κεντρικών συστημάτων:

- ***Επεκτασιμότητα (Scalability):*** Στο ήδη κατανεμημένο σύστημα μπορεί εύκολα να αυξηθεί η υπολογιστική ισχύς του και η χωρητικότητά του, απλά με την προσθήκη επιπλέον υπολογιστών.

- ***Πλεονασμός (Redundancy):*** Κάθε υπολογιστής του κατανεμημένου συστήματος προσφέρει τις ίδιες λειτουργίες. Αυτό σημαίνει ότι στην περίπτωση που διαπιστωθεί βλάβη σε έναν υπολογιστή του δικτύου, η εργασία συνεχίζει να εκτελείται συνεργατικά από κάποιον άλλον υπολογιστή, χωρίς κάποια εξωτερική παρέμβαση.

Τα κατανεμημένα συστήματα μπορούν να λειτουργούν συνεργατικά, ανεξάρτητα από το υλικό, το λογισμικό αλλά και το λειτουργικό σύστημα που διαθέτει (Linux, Unix, Windows) ο κάθε υπολογιστής του συστήματος. Επιπρόσθετα μπορούν να

χρησιμοποιούν διάφορα πρωτόκολλα επικοινωνίας (SNA, TCP/IP σε Ethernet ή Token Ring). Δύο από τα πιο διαδεδομένα εργαλεία που είναι σχεδιασμένα να λειτουργούν σε τέτοια κατανεμημένα συστήματα είναι το Hadoop και το Spark. Πρόκειται για ισχυρές εφαρμογές κατανεμημένης ροής δεδομένων, οι οποίες είναι ικανές να διαχειριστούν μεγάλους όγκους δεδομένων τόσο κατά την αποθήκευση όσο και κατά την επεξεργασία τους. [46].

4.2 Apache Hadoop

Τα διαφορετικού τύπου και τεράστιου όγκου δεδομένα, που πλέον παράγονται ανά δευτερόλεπτο, δεν γίνεται να αξιοποιηθούν κατάλληλα και αποδοτικά με τα παραδοσιακά συστήματα διαχείρισης. Για τον λόγο αυτό, ήδη από το 2005 αναπτύχθηκε το Hadoop. Το Hadoop είναι ένα επεκτάσιμο, ανοικτού κώδικα (open source framework) πλαίσιο και ταυτόχρονα αρχιτεκτονική ανάπτυξης λογισμικού με μεγάλη ανεκτικότητα σε αστοχίες χάρη στο σύστημα αρχείων HDFS του Apache Software Foundation. Το HDFS έχει σχεδιαστεί για την αξιόπιστη αποθήκευση πολύ μεγάλων αρχείων, σε πολλαπλά συστήματα ενός πολύ μεγάλου cluster υπολογιστών. Ένα ακόμη πολύ χρήσιμο εργαλείο του Hadoop είναι η αρχιτεκτονική MapReduce για την κατανεμημένη επεξεργασία των δεδομένων η οποία λειτουργεί ορθά είτε με δομημένα είτε με αδόμητα δεδομένα.

Το Hadoop παρέχει ένα αξιόπιστο κατανεμημένο υπολογιστικό σύστημα (distributed computing), ώστε τα μεγάλα δεδομένα να μπορούν να αποθηκευτούν και να επεξεργαστούν εκτελώντας διεργασίες παράλληλα, εξασφαλίζοντας οικονομία και απόδοση. Τα δεδομένα δεν είναι απαραίτητο να μετακινούνται σε έναν κεντρικό κόμβο, μέσω του δικτύου, για την επεξεργασία τους. Αρκεί τα σύνολα μεγάλων δεδομένων να διασπαστούν σε μικρότερα σύνολα, που ονομάζονται διεργασίες, οι οποίες αποθηκεύονται και επεξεργάζονται παράλληλα και αυτόνομα σε κάθε υπολογιστικό κόμβο ξεχωριστά και εν συνεχεία τα αποτελέσματα που προκύπτουν να συνδυαστούν ώστε να δώσουν το τελικό αποτέλεσμα. Σε αυτή τη διαδικασία μπορεί να προκύψουν αστοχίες υλικού τόσο των αποθηκευτικών μέσων όσο και ολόκληρων κόμβων. Το Hadoop αντιμετωπίζει και τις δύο περιπτώσεις αποτελεσματικά σε επίπεδο λογισμικού. Τα δεδομένα αποθηκεύονται σε πολλαπλά αντίγραφα ώστε σε περίπτωση αστοχίας να χρησιμοποιηθεί κάποιο από τα αντίγραφα. Έτσι μπορούν να χρησιμοποιηθούν αποθηκευτικά μέσα χαμηλότερης αξιοπιστίας, μειώνοντας ταυτόχρονα το κόστος. Για να διαχειριστεί την απώλεια ολόκληρου κόμβου, το Hadoop

χωρίζει την συνολική διεργασία σε πολλές μικρότερες, ανεξάρτητες μεταξύ τους, που έχουν τυποποιημένη δομή και αυτές τις αναθέτει στους κόμβους του συστήματος. Όταν κάποιος κόμβος σταματήσει να ανταποκρίνεται η διεργασία που είχε αναλάβει ανατίθεται σε κάποιον άλλο και έτσι το σύστημα συνεχίζει αδιάληπτα την λειτουργία του [46].

4.2.1 Τα εργαλεία του Hadoop

Όπως προαναφέρθηκε, το Apache Hadoop δημιουργήθηκε λόγω των αυξημένων απαιτήσεων για την ορθή λειτουργία των μηχανών αναζήτησης και της σωστής διαχείρισης του μεγάλου όγκου των παραγόμενων δεδομένων. Η συνεχής συντήρηση και ενημέρωση του ευρετηρίου (indexing), αλλά και ο τρόπος διαχείρισης των κόμβων αποθήκευσης απαιτούσαν πολύ χρόνο. Ταυτόχρονα η διαδικασία ανίχνευσης του ιστού παρήγαγε τεράστιους όγκους δεδομένων που δεν μπορούσαν να αποθηκευτούν με τα παραδοσιακά μέσα. Έτσι δημιουργήθηκε η ανάγκη αυτοματοποίησης της διαδικασίας αυτής, με χρήση πολλαπλών συστημάτων, τα οποία θα διασφάλιζαν τη διαχείριση του τεράστιου όγκου πληροφοριών μέσω του διαμοιρασμού των δεδομένων και την εκμετάλλευση της συνδυαστικής επεξεργαστικής τους ισχύος [46]. Για τον λόγο αυτό η Google προχώρησε στη δημιουργία ενός νέου μοντέλου κατανεμημένης επεξεργασίας, το MapReduce. Το MapReduce αποτελείται από δύο λειτουργίες, την map και την reduce. Η λειτουργία map αναλαμβάνει να επεξεργαστεί ένα σύνολο δεδομένων μετατρέποντάς το σε ένα σύνολο από ζεύγη κλειδιών/τιμών και η λειτουργία reduce συνδυάζει τα αποτελέσματα εξόδου που προέκυψαν από την map σε ένα μόνο αποτέλεσμα. Η παραπάνω μέθοδος θεωρήθηκε ιδανική από τους προγραμματιστές του “Nutch”, το πρόγραμμα ανίχνευσης ιστού (webcrawler) του Apache, για την επίλυση των προβλημάτων που αντιμετώπιζαν και έτσι ξεκίνησε η ανάπτυξη του Hadoop [47]

Το Hadoop απαρτίζεται από τέσσερα εργαλεία: Το Hadoop distributed file system (HDFS), το Map Reduce, το Yet Another Resource Negotiator (YARN) και το Hadoop Common [46].

4.2.1.1 Hadoop distributed file system (HDFS)

Το HDFS είναι ένα σύστημα αποθήκευσης το οποίο χρησιμοποιείται από το Apache Hadoop για την κατανεμημένη αποθήκευση μεγάλων αρχείων δεδομένων. Το HDFS

σχεδιάστηκε για να μπορεί να αποθηκεύσει μεγάλου όγκου δεδομένα, όπως ένα Mainframe, μειώνοντας το κόστος. Το HDFS είναι υπεύθυνο για να χωρίσει τα δεδομένα και να τα διανείμει στους κόμβους του cluster των υπολογιστών. Το HDFS είναι σχεδιασμένο για σειριακή ανάγνωση των δεδομένων (συνήθως μεγάλο μέρος του συνόλου ή και ολόκληρο το σύνολο) και γι' αυτό δίνεται μεγαλύτερη έμφαση στην ταχύτητα ανάγνωσης των δεδομένων εις βάρος του χρόνου προσπέλασής τους. Επιπλέον επιτρέπει ταυτόχρονη ανάγνωση από πολλές διεργασίες αλλά εγγραφή μόνο από μία. Ο τρόπος λειτουργίας του στηρίζεται σε δύο βασικά δομικά στοιχεία: το NameNode και τα DataNodes. Το NameNode είναι ένα σύστημα, που συντηρεί τα metadata του filesystem, δηλαδή ποια αρχεία υπάρχουν, πως λέγονται και που βρίσκονται τα δεδομένα τους. Στα DataNodes αποθηκεύονται τα δεδομένα των αρχείων ως μία ακολουθία μπλοκ δεδομένων και είναι διαθέσιμα για ανάγνωση, εγγραφή και επεξεργασία. Όλα τα μπλοκ δεδομένων, πλην του τελευταίου, έχουν το ίδιο μέγεθος. Ο αριθμός των DataNodes δεν είναι συγκεκριμένος και εξαρτάται από τον αριθμό των συστημάτων που απαρτίζουν το cluster των υπολογιστών. Για λόγους αξιοπιστίας και αποφυγής σφαλμάτων, δημιουργούνται πολλαπλά αντίγραφα των μπλοκ που απαρτίζουν τα αρχεία σε διαφορετικά DataNodes. Επειδή το NameNode έχει κεντρικό ρόλο στο HDFS και αν χαθεί χάνονται και όλα τα αρχεία, αποθηκεύει τα βασικά μεταδεδομένα των αρχείων (όνομά, μέγεθος, θέση στην ιεραρχία του filesystem κλπ.) στον τοπικό δίσκο του και σε ένα δικτυακό αποθηκευτικό χώρο (NFS). Σε περίπτωση αστοχίας του NameNode, μπορεί να ανασυσταθεί ένα νέο από τα αποθηκευμένα μεταδεδομένα. Αντίθετα η θέση των δεδομένων των αρχείων, η οποία μεταβάλλεται συνεχώς, συντηρείται σε μία δομή στη μνήμη που δημιουργείται από τις αναφορές που στέλνει κάθε DataNode. Οι αναφορές περιλαμβάνουν κατάλογο όλων των μπλοκ που είναι αποθηκευμένα στο DataNode και ταυτόχρονα επιτρέπουν στο NameNode να επιβεβαιώσει ότι το αντίστοιχο DataNode εξακολουθεί να λειτουργεί κανονικά. Εάν ένας κόμβος αποτύχει, το HDFS θα μπορεί να ανακτήσει τα δεδομένα από τα διαθέσιμα αντίγραφα που υπάρχουν σε κάποιον άλλον κόμβο.

Για παράδειγμα, ένα αρχείο 50GB μπορεί να διασπαστεί σε μία ακολουθία από μπλοκ μικρότερης χωρητικότητας πχ των 128MB, τα οποία θα μοιραστούν σε όλα τα DataNodes. Επίσης, τα μπλοκ αυτά μπορούν να διαμοιραστούν και σε επιπλέον τυχαία DataNodes της συστάδας ώστε να υπάρχουν αντίγραφα, και σε περίπτωση που ένας κόμβος χαθεί να μπορέσει να ανακτηθεί το αρχείο.

Συνοψίζοντας, το HDFS είναι σύστημα αποθήκευσης με σκοπό να διαχειρίζεται τον μεγάλο όγκο δεδομένων που παράγεται, με τα ακόλουθα χαρακτηριστικά:

- Χρησιμοποιεί κατανεμημένη αποθήκευση δεδομένων
- Χρησιμοποιεί συνηθισμένους υπολογιστές και όχι ακριβά συστήματα υψηλής διαθεσιμότητας
- Έχει ανθεκτικότητα σε αστοχίες του υλικού
- Παρέχει εύκολη επεκτασιμότητα
- Είναι ευέλικτο
- Έχει υψηλή απόδοση
- Εξασφαλίζει ελάχιστη επιβάρυνση του δικτύου

Εναλλακτικά του HDFS, το Hadoop μπορεί να χρησιμοποιήσει και άλλα συστήματα αποθήκευσης όπως το S3 της Amazon Cloud [46].

4.2.1.2 Map Reduce

Το MapReduce είναι ένα προγραμματιστικό μοντέλο με σκοπό την επεξεργασία συνόλων μεγάλων δεδομένων παράλληλα πάνω σε μία συστάδα υπολογιστών. Το μοντέλο προτάθηκε αρχικά το 2004 από τη Google, για να καλύψει την ανάγκη επεξεργασίας μεγάλου όγκου ακατέργαστων δεδομένων όπως έγγραφα, αρχεία καταγραφής κλπ. Συνήθως οι υπολογισμοί που πρέπει να γίνουν είναι απλοί ωστόσο λόγω του μεγάλου όγκου δεδομένων, η επεξεργασία τους πρέπει να κατανεμηθεί σε χιλιάδες υπολογιστές για να υπολογίζονται τα αποτελέσματα σε εύλογο χρόνο. Για να αντιμετωπισθεί η πολυπλοκότητα του προβλήματος (που περιλαμβάνει το διαχωρισμό σε παράλληλες διεργασίες, την κατανομή τους και τη διαχείριση των αποτυχιών) επιλέχθηκε η χρήση των διαδικασιών map και reduce και της συναρτησιακής γλώσσας LISP. Πιο συγκεκριμένα, το MapReduce αποτελείται από δύο φάσεις, όπου οι χρήστες χρησιμοποιούν τις μεθόδους map και reduce αντίστοιχα. Η μέθοδος map στην πρώτη φάση επεξεργάζεται ζευγάρια κλειδί/τιμή, παράγοντας για καθένα από αυτά μία λίστα από ενδιάμεσα ζευγάρια κλειδί/τιμή. Το σύστημα ομαδοποιεί όλες τις ενδιάμεσες τιμές που σχετίζονται με το ίδιο ενδιάμεσο κλειδί και τις τροφοδοτεί στη δεύτερη φάση. Στην δεύτερη φάση, με την μέθοδο reduce, συγχωνεύονται όλες οι ενδιάμεσες τιμές που σχετίζονται με το ίδιο ενδιάμεσο κλειδί και παράγεται το συνολικό αποτέλεσμα. Το

εργαλείο MapReduce φροντίζει την συνολική επικοινωνία των συνδεδεμένων υπολογιστικών συστημάτων. Χωρίζει τη συνολική εργασία σε επιμέρους διεργασίες map ή reduce που λειτουργούν πάνω σε ένα υποσύνολο των δεδομένων και τις κατανέμει στους διαθέσιμους κόμβους. Τροφοδοτεί τα δεδομένα στις διεργασίες σταδιακά σαν μία ροή δεδομένων (stream), ώστε να μην χρειάζεται να φορτωθεί ολόκληρο το σύνολο στη μνήμη επιτρέποντας έτσι το χειρισμό μεγάλου όγκου δεδομένων. Συλλέγει τα ενδιάμεσα ζευγάρια κλειδί/τιμή, τα αναδιατάσσει ομαδοποιώντας τα ανά ενδιάμεσο κλειδί και ταξινομώντας τα (shuffle & sort) πριν τα τροφοδοτήσει στη φάση reduce. Τέλος, παρακολουθεί την πρόοδο των επιμέρους διεργασιών και επιλύει τυχόν σφάλματα που μπορεί να προκύψουν [46].

4.2.1.3 Yet Another Resource Negotiator (YARN)

Το YARN είναι ένα κατανεμημένο σύστημα διαχείρισης πόρων που αναπτύχθηκε για να αντιμετωπίσει τα προβλήματα που παρουσίαζε το κλασσικό μοντέλο MapReduce σε συστήματα με πολύ μεγάλο αριθμό κόμβων. Είναι σχεδιασμένο για να διαχειρίζεται τους κόμβους μιας συστάδας υπολογιστών, τους διαθέσιμους πόρους και τις απαραίτητες εργασίες που πρέπει να εκτελεστούν. Επιπλέον, το YARN διαχωρίζει τις λειτουργίες διαχείρισης πόρων από τις λειτουργίες επιτήρησης των εφαρμογών που εκτελούνται στη συστοιχία σε ξεχωριστά υποσυστήματα: τον Resource Manager και τον Application Master αντίστοιχα. Ο resource manager διαχειρίζεται και μοιράζει τους πόρους του συστήματος στο σύνολό του. Ο application master δεσμεύει μέσω του resource manager τους απαραίτητους πόρους για μία εφαρμογή και φροντίζει για την εκτέλεσή των διεργασιών και την επιτήρησή της. Η νέα αρχιτεκτονική είναι πιο γενική και επεκτάσιμη. Το MapReduce αποτελεί πλέον μία μόνο από τις εφαρμογές που μπορούν να εκτελεστούν πάνω στο YARN. Άλλες εναλλακτικές εφαρμογές που είναι διαθέσιμες στο YARN είναι οι εξής: Dryad, Giraph, Hoya, REEF, Spark, Storm [46].

4.2.1.4 Hadoop Common

Είναι η βιβλιοθήκη υποδομής που διαθέτει το Hadoop για να υποστηρίξει τα παραπάνω εργαλεία [46].

4.2.2 Οφέλη και περιορισμοί του Hadoop

Η χρήση του Hadoop στις εφαρμογές επεξεργασίας μεγάλων δεδομένων προσφέρει πολλά οφέλη [47]:

- Συμφέρει οικονομικά γιατί χρησιμοποιεί κοινούς υπολογιστές (commodity hardware), δηλαδή το υλικό που χρησιμοποιεί είναι σχετικά φθηνό, ευρέως διαθέσιμο και εναλλάξιμο με άλλο υλικό ίδιου τύπου.

- Είναι πολύ αποδοτικό στην επίλυση των προβλημάτων που αντιμετωπίζουν οι εφαρμογές που χειρίζονται μεγάλο όγκο δεδομένων. Αυτό το καταφέρνει χρησιμοποιώντας μεγάλο αριθμό κόμβων προκειμένου να κατανέμεται η επεξεργασία και να εκτελείται παράλληλα, αναλύοντας το πρόβλημα σε μικρότερα τμήματα, ενώ εξασφαλίζει η επεξεργασία των δεδομένων να γίνεται στους κόμβους όπου είναι αποθηκευμένα. Με αυτό τον τρόπο αποφεύγονται οι καθυστερήσεις για τη μεταφορά των δεδομένων από τον κόμβο αποθήκευσης στον κόμβο υπολογισμού καθώς και η υπερφόρτωση του δικτύου.

- Είναι εύκολα και σχεδόν απεριόριστα επεκτάσιμο. Μπορούν να προστεθούν κόμβοι οποιαδήποτε στιγμή και με αυτό τον τρόπο αυξάνεται η υπολογιστική και αποθηκευτική ικανότητα.

- Είναι ευέλικτο: Εκτός από το συνηθισμένο εργαλείο MapReduce, μπορούν να χρησιμοποιηθούν και άλλα μοντέλα προγραμματισμού ή/και εφαρμογές, όπως το Spark.

- Τέλος, έχει την ικανότητα να διαχειρίζεται οποιοδήποτε τύπο δεδομένων, δομημένο ή αδόμητο.

Παρά τα θετικά σημεία και την ευελιξία του, το Hadoop δε είναι κατάλληλο για κάθε πρόβλημα. Υπάρχουν προβλήματα με μικρότερα σύνολα δεδομένων ή/και διαφορετικές απαιτήσεις, όπου αυτός ο κατανεμημένος τρόπος επεξεργασίας δεν είναι αποδοτικός. Σε αυτές τις περιπτώσεις οι παραδοσιακές μέθοδοι είναι προτιμότερες.

Μερικοί περιορισμοί που συναντούμε στο Hadoop είναι:

- Το εργαλείο HDFS είναι σχεδιασμένο έτσι ώστε τα δεδομένα να γράφονται μόνο μία φορά και να διαβάζονται πολλαπλές (write-once read-many). Συνεπώς δεν μπορεί να χρησιμοποιηθεί για εφαρμογές που χρειάζονται συνεχή ενημέρωση των δεδομένων [47].

- Το HDFS είναι προσανατολισμένο σε διεργασίες που επεξεργάζονται σειριακά το σύνολό ή το μεγαλύτερο μέρος των δεδομένων. Δεν είναι κατάλληλο για διεργασίες που χρειάζονται τυχαία προσπέλαση σε μικρό μέρος των δεδομένων [47].

- Επίσης, το Hadoop δεν είναι κατάλληλο για επεξεργασία ευαίσθητων δεδομένων. Παρόλο που διαθέτει σύστημα ασφαλείας και ελέγχου πρόσβασης, οι ρυθμίσεις του από προεπιλογή το απενεργοποιούν. Οι διαχειριστές πρέπει να είναι πολύ προσεκτικοί και να διασφαλίζουν ότι τα δεδομένα κρυπτογραφούνται και προστατεύονται επιλέγοντας κάθε φορά τις κατάλληλες ρυθμίσεις [47].

4.2.3 Χρήση Apache Hadoop

Το Hadoop είναι μία πλατφόρμα που χρησιμοποιείται από πολλούς οργανισμούς και εταιρείες. Σύμφωνα με την Cloudera, το Hadoop χρησιμοποιείται ενδεικτικά στα παρακάτω προβλήματα μεγάλων δεδομένων, που αφορούν τον τομέα του λιανικού εμπορίου, τις τράπεζες, την υγειονομική περίθαλψη και πολλούς άλλους [47]:

- Risk modeling (μοντελοποίηση κινδύνου)
- Customer churn analysis (ανάλυση φερεγγυότητας πελάτη)
- Recommendation engine (μηχανή συστάσεων)
- Ad targeting (στόχευση διαφημίσεων)
- Transaction analysis (ανάλυση συναλλαγών)
- Analyzing network data to predict failure (ανάλυση δεδομένων δικτύου για την πρόβλεψη αστοχίας)
- Threat analysis (ανάλυση απειλών)

Στην ιστοσελίδα του Hadoop φαίνεται ότι πολλές γνωστές εταιρείες κολοσσοί χρησιμοποιούν το συγκεκριμένο κατανεμημένο σύστημα με clusters που περιέχουν έως και 4500 κόμβους. Μεταξύ αυτών των εταιρειών είναι η Amazon, το EBay, το Facebook, το LinkedIn, το Twitter και το Yahoo [47].

4.3 Apache Spark

Το Apache Spark είναι μια υπολογιστική πλατφόρμα που λειτουργεί σε συστοιχίες υπολογιστών, παρόμοια με το Hadoop. Είναι σχεδιασμένη να είναι γρήγορη και γενική. Έτσι σε αντίθεση με το Hadoop χρησιμοποιεί κατά κύριο λόγο την μνήμη RAM των κόμβων αντί για τους σκληρούς δίσκους κατά την επεξεργασία των δεδομένων, επιτυγχάνοντας πολύ μεγαλύτερες ταχύτητες, ιδιαίτερα όταν τα ενδιάμεσα δεδομένα μπορούν να χωρέσουν στη μνήμη. Επιπλέον, αντί για το μοντέλο MapReduce χρησιμοποιεί ένα κατευθυνόμενο ακυκλικό γράφημα (DAG) για να αναπαραστήσει τις διεργασίες που πρέπει να εκτελεστούν με βάση τις ροές των δεδομένων ανάμεσά τους. Σε κάθε μετάβαση από ένα κόμβο του γραφήματος σε έναν άλλο εκτελείται είτε ένας μετασχηματισμός (transformation) είτε μία ενέργεια (action). Οι μετασχηματισμοί και οι ενέργειες είναι τα βασικά δομικά στοιχεία που συνθέτουν τον υπολογισμό. Με αυτό τον τρόπο μπορεί να εκτελέσει μεγαλύτερη ποικιλία υπολογισμών που δεν θα ήταν εφικτοί, ή αποδοτικοί με το MapReduce. Παράλληλα, το μοντέλο αυτό επιτρέπει στο Spark να χωρίζει αυτόματα και να κατανέμει τα δεδομένα και τις διεργασίες στους κόμβους για εκτέλεση, να παρακολουθεί την πρόοδο των διεργασιών και να ανταποκρίνεται σε αστοχίες αναδρομολογώντας τις σχετικές διεργασίες σε υγιείς κόμβους.

Το Spark διατηρεί τα πλεονεκτήματα που προσφέρει το Apache Hadoop και ταυτόχρονα, ελαχιστοποιεί πολλούς από τους περιορισμούς και τα μειονεκτήματά του. Η σημαντική αύξηση της ταχύτητας των υπολογισμών και η υποστήριξη ενός πιο γενικού μοντέλου προγραμματισμού, επιτρέπουν στο Spark να χρησιμοποιείται σε προβλήματα που το Hadoop είναι ακατάλληλο, όπως επαναληπτικούς αλγόριθμους, διαδραστικά ερωτήματα, επεξεργασία ροών και δεδομένων πραγματικού χρόνου και άλλα.

Το Spark άρχισε να αναπτύσσεται το 2009 από τον Matei Zaharia στο AMPLab του UC Berkeley και ο κώδικάς του είναι ανοικτός από τις αρχές του 2010. Στο Apache

Software Foundation εισήχθη το 2013 και από τότε αναπτύσσεται συνεχώς από πολλούς οργανισμούς και προγραμματιστές. Στην ανάπτυξη του συμμετέχουν εταιρίες όπως η Databricks, η Yahoo και η Intel.

Το Spark σχεδιάστηκε εξ αρχής ώστε να μπορεί να αντιμετωπίσει υπολογιστικές εργασίες που περιλαμβάνουν επαναλαμβανόμενους αλγορίθμους ή απαιτούν διαδραστικότητα, στις οποίες το MapReduce του Hadoop είναι αναποτελεσματικό. Αυτό επιτυχγάνεται με τη χρήση της έννοιας του ανθεκτικού κατανεμημένου συνόλου δεδομένων, το RDD. Το RDD είναι το βασικό στοιχείο του Spark, που χρησιμοποιείται για την ανταλλαγή δεδομένων μεταξύ των διεργασιών, και επιτρέπει στο σύστημα να διαχειρίζεται τη ροή των δεδομένων, να προγραμματίζει και να κατανέμει τις διεργασίες στους κόμβους και να αντιμετωπίζει αστοχίες. Παράλληλα επιτρέπει στους χρήστες τη βελτιστοποίηση της απόδοσης των υπολογιστικών εργασιών δίνοντας τη δυνατότητα αποθήκευσης των ενδιάμεσων αποτελεσμάτων στη μνήμη ή και το δίσκο των κόμβων της συστοιχίας για επαναχρησιμοποίηση σε μετέπειτα μετασχηματισμούς ή ενέργειες.

Το Spark είναι πιο γρήγορο και ευκολότερο στη χρήση, προσφέροντας API σε πολλές γλώσσες (Scala, Java, Python, SQL, και R) και πολλές ενσωματωμένες βιβλιοθήκες. Όπως αναφέρθηκε μπορεί να τρέξει σε μία συστάδα υπολογιστών είτε ανεξάρτητα είτε πάνω στο Hadoop καθώς και να έχει πρόσβαση σε πηγές δεδομένων της Hadoop και της NoSQL βάσης Cassandra. Γενικότερα, το Spark δίνει τη δυνατότητα στον χρήστη να ορίσει RDDs, συναρτήσεις, μεταβλητές και κλάσεις και να τις χρησιμοποιεί σε παράλληλες λειτουργίες μίας συστάδας [48].

4.3.1 Τα δομικά στοιχεία του Spark

Το Apache Spark έχει σχεδιαστεί για να μπορεί να επεξεργαστεί τον μεγάλο όγκο δεδομένων στη μνήμη και να εκτελεί ερωτήματα παρέχοντας γρήγορες λύσεις στα δεδομένα μεγάλου όγκου που παράγονται σε πραγματικό χρόνο. Η ταχύτητα της επεξεργασίας και η γενικότητα του προγραμματιστικού μοντέλου επιτρέπουν την αποδοτική υλοποίηση και συνδυασμό ενός μεγάλου αριθμού υπολογιστικών αλγορίθμων, που περιλαμβάνουν μεταξύ άλλων και εφαρμογές μηχανικής μάθησης σε μεγάλη κλίμακα. Το Apache Spark αποτελείται από πέντε βασικά στοιχεία (το Spark Core, το Spark SQL, το Spark Streaming, το MLib και το GraphX) που είναι στενά συνδεδεμένα και καλύπτουν ένα ευρύ φάσμα εφαρμογών. Η στενή διασύνδεση των

δομικών στοιχείων υψηλότερου επιπέδου με τον πυρήνα του Spark, εξασφαλίζει ότι τυχόν βελτιώσεις στον πυρήνα αντικατοπτρίζονται άμεσα στα επιμέρους στοιχεία βελτιώνοντας την ταχύτητά τους [48].

4.3.1.1 Spark Core

Το Spark Core αποτελεί την βασική υποδομή του Apache Spark για την επεξεργασία κατανεμημένων δεδομένων. Μέσω αυτού του στοιχείου επιτρέπεται η υλοποίηση της παράλληλης και κατανεμημένης επεξεργασίας μεγάλου όγκου δεδομένων και περιλαμβάνει βιβλιοθήκες για τη δημιουργία και διαχείριση των RDD. Το Spark Core είναι υπεύθυνο για τη διαχείριση της μνήμης, τον εντοπισμό και τη διόρθωση τυχών σφαλμάτων, το χρονοπρογραμματισμό των διεργασιών, τη διασύνδεση με συστήματα αποθήκευσης δεδομένων, τον κατανεμημένο διαμοιρασμό και τη διαχείριση των διαδικασιών στο cluster των υπολογιστικών μηχανημάτων [48].

4.3.1.2 Spark SQL

Το Spark SQL είναι χρήσιμο εργαλείο για διαχείριση δομημένων δεδομένων με χρήση της γλώσσας SQL ή της παραλλαγής της HQL, που χρησιμοποιείται από το Apache Hive. Υποστηρίζει αναζητήσεις σε δεδομένα που είναι αποθηκευμένα RDDs του Spark καθώς και σε άλλες εξωτερικές πηγές δεδομένων, όπως σχεσιακούς πίνακες πχ πίνακες Hive, αρχεία Parquet και JSON ενοποιώντας τον τρόπο πρόσβασης στα δεδομένα και επιτρέποντας πολύπλοκες αναλύσεις σε μία εφαρμογή.

Πιο συγκεκριμένα με το εργαλείο Spark SQL οι προγραμματιστές καταφέρνουν να φέρουν κοντά τα RDDs και τους σχεσιακούς πίνακες για:

- Την εισαγωγή δεδομένων από αρχεία Parquet και πίνακες Hive
- Τη εκτέλεση ερωτημάτων SQL σε δεδομένα ήδη υπαρχόντων RDDSs
- Την εγγραφή RDDs σε πίνακες Hive ή αρχεία Parquet

Το Spark SQL, επιτυγχάνει μεγάλη ταχύτητα στην εκτέλεση των ερωτημάτων χάρη στο βελτιστοποιητή, την αποθήκευση κατά στήλες, και το υποσύστημα δημιουργίας κώδικα που διαθέτει. Επίσης, πετυχαίνει την κλιμάκωση σε χιλιάδες κόμβους και την εκτέλεση πολύωρων ερωτημάτων, παρουσιάζοντας καθολική ανοχή σε σφάλματα,

χωρίς να απαιτείται η χρήση διαφορετικής μηχανής για τη διαχείριση των ιστορικών δεδομένων [48]

4.3.1.3 Spark Streaming

Το Spark Streaming είναι ένα εργαλείο που μπορεί να επεξεργαστεί δεδομένα πραγματικού χρόνου με ασφάλεια και ταχύτητα. Επιτρέπει τη χρήση αποθηκευμένων δεδομένων και δεδομένων πραγματικού χρόνου με διαφανή τρόπο με τη χρήση RDDs. Επιτρέπει έτσι την εύκολη διασύνδεση και συνδυασμό του με τα άλλα δομικά στοιχεία του Spark, όπως το Spark SQL και το MLlib, για την υλοποίηση εξελιγμένων τεχνικών, όπως για παράδειγμα τη χρήση Μηχανικής Μάθησης για την εξαγωγή συμπερασμάτων και προβλέψεων ή προτάσεων σε πραγματικό χρόνο [48].

4.3.1.4 Spark's MLlib

Ένα πολύ χρήσιμο εργαλείο του Spark για την ανάπτυξη εφαρμογών αποτελεί η κατανεμημένη βιβλιοθήκη μηχανικής μάθησης, Spark MLlib. Πριν την εμφάνιση της MLib, το Spark δε διέθετε κάποια σουίτα ισχυρών και κλιμακούμενων αλγορίθμων μηχανικής μάθησης. Η βιβλιοθήκη MLlib είναι μία υψηλού επιπέδου βιβλιοθήκη μηχανικής μάθησης, αναπτύχθηκε το 2012 ως μέρος του project MLbase και ο κώδικας της είναι δημόσια διαθέσιμος από τον Σεπτέμβριο του 2013. Η αρχική έκδοση της MLib αναπτύχθηκε στο UC Berkley από 11 συνεργάτες ενώ παρείχε ένα περιορισμένο σύνολο τυποποιημένων μεθόδων μηχανικής μάθησης για κοινές λειτουργίες μάθησης. Πλέον, παρέχει πολλούς τύπους αλγορίθμων μηχανικής μάθησης, όπως: γραμμικά μοντέλα Naïve Bayes, σύνολα δέντρων απόφασης για ταξινόμηση ή παλινδρόμηση, K-mean Clustering ομαδοποίηση, PCA για clustering και μείωση διαστάσεων. Στο MLlib έχουν υλοποιηθεί εκείνοι μόνο οι αλγόριθμοι που μπορούν να επιταχυνθούν με χρήση κλιμάκωσης στο cluster και η βιβλιοθήκη αναλαμβάνει την πολυπλοκότητα της κλιμάκωσης. Με τον τρόπο αυτό παρέχεται στους χρήστες μία ευρεία γκάμα εργαλείων που απλοποιούν την ανάπτυξη μοντέλων μηχανικής μάθησης σε κατανεμημένα συστήματα για την επεξεργασία μεγάλου όγκου δεδομένων.

Ο κώδικάς της είναι γραμμένος σε Scala, χρησιμοποιώντας εγγενείς βιβλιοθήκες γραμμικής άλγεβρας σε κάθε κόμβο, βασισμένες σε C++ και περιλαμβάνει APIs σε Java, Scala και Python [49].

4.3.1.5 GraphX

Το GraphX, το νεότερο στοιχείο του Spark, είναι η βιβλιοθήκη επεξεργασίας γραφημάτων (π.χ. γράφημα φίλων σε ένα κοινωνικό δίκτυο) του Spark και λόγω της ευελιξίας της μπορεί να εργαστεί τόσο με γραφήματα όσο και με συλλογές δεδομένων. Το GraphX παρέχει ένα API για τη δημιουργία και επεξεργασία γραφημάτων. Έχει τη δυνατότητα να εκτελεί παράλληλους υπολογισμούς πάνω σε γραφήματα χρησιμοποιώντας ένα ευρύ φάσμα αναπαραστάσεων και δομών γραφημάτων. Η βιβλιοθήκη αυτή μέσα από διάφορους υπολογισμούς και λεπτομερή ανάλυση του δικτύου μέσα στο οποίο αποθηκεύονται τα δεδομένα, μπορεί να πραγματοποιήσει ομαδοποίηση δεδομένων, ανάλυση των δεδομένων και εύρεση της βέλτιστης διαδρομής. Επιπλέον, το GraphX μέσω της διαδικασίας ELT παρέχει τη δυνατότητα να ανακτηθούν και να αξιοποιηθούν ασυνεπή δεδομένα, μειώνοντας έτσι το χρόνο και το κόστος ανάλυσης αυτών των δεδομένων. Αποθηκεύει τις πληροφορίες στη μνήμη, εκτελεί συνεχόμενα ερωτήματα, εντοπίζει εύκολα τους αλγόριθμους μηχανικής μάθησης, που είναι απαραίτητοι να χρησιμοποιηθούν στους υπολογισμούς των δεδομένων και αναπαριστά τα δεδομένα που παράγονται σε πραγματικό χρόνο με τη βοήθεια του Spark Streaming.

4.3.2 Το εργαλείο Resilient Distributed Dataset (RDDs) του Spark

Τα αρχικά των χαρακτηριστικών των RDDs συνθέτουν και το όνομά τους. Τα RDDs είναι:

1. **Resilient**: Τα RDD χαρακτηρίζονται για την ανοχή τους σε σφάλματα. Επειδή τα RDDs υπολογίζονται όταν χρειάζονται, σε περίπτωση που ένας κόμβος χαθεί, τα δεδομένα μπορούν να ξαναυπολογιστούν σε κάποιον άλλο κόμβο. Επιπλέον αν τα δεδομένα είχαν προηγουμένως αποθηκευτεί σε δίσκο, μπορούν να ξαναδιαβαστούν σε κάποιον άλλο κόμβο (18).
2. **Distributed**: Τα δεδομένα βρίσκονται διαμοιρασμένα σε πολλούς κόμβους ενός cluster.
3. **Dataset**: Το RDD είναι μια συλλογή δεδομένων που είναι χωρισμένη σε κομμάτια (partitioned collection). Υπάρχει μόνο για όσο χρόνο απαιτείται για την συλλογή, ανάλυση και επεξεργασία των δεδομένων του dataset που περιέχει. Η έννοια του RDD εμφανίστηκε για πρώτη φορά στη δημοσίευση

"Resilient Distributed Datasets: A Tolerant Fault Abstraction for Computing Cluster In Memory" [51].

Κάθε εφαρμογή του Spark έχει έναν οδηγό προγράμματος (driver) που είναι υπεύθυνος να μοιράζει και να δρομολογεί την εκτέλεση των εργασιών (tasks) παράλληλα στο cluster. Πρωταρχική οντότητα αυτού του οδηγού αποτελεί η δομή δεδομένων RDD, που αντιπροσωπεύει τα δεδομένα που είναι κατανεμημένα σε όλα τα DataNodes του cluster. Ο οδηγός προγράμματος χρησιμοποιεί τα RDDs για να συγχρονίσει και να προγραμματίσει την εκτέλεση των εργασιών, όπως:

- Τη δημιουργία των αρχικών RDD από εξωτερικά δεδομένα,
- Την μετατροπή (transformation) των ήδη υπαρχόντων RDD σε νέα, εφαρμόζοντας συναρτήσεις στα δεδομένα του dataset ενός RDD. Το νέο σύνολο δεδομένων RDD που δημιουργείται μπορεί να περιλαμβάνει ένα υποσύνολο του αρχικού (filtering) ή ένα μετασχηματισμό του (mapping).
- Την ενέργεια (action) που πραγματοποιεί υπολογισμούς στο σύνολο δεδομένων ενός RDD και επιστρέφει τα επιθυμητά αποτελέσματα στον οδηγό, πχ η συνάρτηση first(), η οποία επιστρέφει το πρώτο στοιχείο ενός RDD ή η συνάρτηση reduce(), η οποία συνοψίζει όλα τα στοιχεία ενός RDD [52].

Σημαντικό στοιχείο της όλης διαδικασίας είναι ότι οι μετασχηματισμοί που γίνονται σε ένα RDD δεν υπολογίζουν τα αποτελέσματά τους μέχρι μία ενέργεια να τα χρειαστεί.

Επίσης σε οποιοδήποτε στάδιο των υπολογισμών ο χρήστης μπορεί να ζητήσει την αποθήκευση ενός RDD στην μνήμη, στο δίσκο ή και στα δύο. Με αυτό τον τρόπο μπορούν να υλοποιηθούν αποδοτικά επαναληπτικοί αλγόριθμοι που επαναχρησιμοποιούν σύνολα δεδομένων που έχουν υπολογιστεί σε προηγούμενα στάδια [52].

4.4 Apache Kafka

Το Apache Kafka, που αρχικά σχεδιάστηκε από το LinkedIn, έχει εξελιχθεί σε μια ισχυρή πλατφόρμα ροής κατανεμημένων δεδομένων, την οποία έχουμε επιλέξει ως κρίσιμο στοιχείο του συστήματός μας. Το Kafka υπερέχει στη συλλογή και τη μεταφορά δεδομένων με κατανεμημένο τρόπο, προσφέροντας μοντέλα παρόμοια με συστήματα ανταλλαγής μηνυμάτων. Αξιοποιεί το μοντέλο ουράς για την επεξεργασία δεδομένων

σε μια ροή μηνυμάτων και το μοντέλο δημοσίευσης/εγγραφής για την αποτελεσματική μετάδοση μηνυμάτων στους τελικούς χρήστες και τους καταναλωτές.

Το 2022, το Apache Kafka ενσωματώθηκε στο Apache Software Foundation, αποτελώντας αναπόσπαστο μέρος αυτού του διάσημου οργανισμού ανοιχτού κώδικα. Αυτή η αλλαγή ενίσχυσε περαιτέρω τη θέση του ως βασική τεχνολογία στον τομέα της επεξεργασίας και ροής δεδομένων [53].

Το Apache Kafka ως ακρογωνιαίος λίθος του συστήματος επεξεργασίας δεδομένων που αναπτύξαμε επιλέχθηκε για τους ακόλουθους λόγους:

- Σωληνώσεις δεδομένων (data pipelines): Το Kafka έχει σχεδιαστεί για τη δημιουργία σωληνώσεων δεδομένων, καθιστώντας το ιδανικό για το σχεδιασμό εφαρμογών που μπορούν να λαμβάνουν και να αναλύουν μεγάλους όγκους δεδομένων σε πραγματικό χρόνο. Η κατανεμημένη αρχιτεκτονική του είναι κατάλληλη για τον αποτελεσματικό χειρισμό των ροών δεδομένων.

- Ανοιχτός κώδικας: Ως κατανεμημένο σύστημα ανοιχτού κώδικα γραμμένο σε Scala και Java, το Kafka προσφέρει διαφάνεια και ευελιξία. Μπορεί να λειτουργήσει σε ένα τοπικό μηχάνημα ή σε ένα κατανεμημένο σύμπλεγμα.

- Υψηλή απόδοση: Το Kafka παρέχει υψηλούς ρυθμούς αποστολής και λήψης, διασφαλίζοντας την έγκαιρη μετάδοση του τεράστιου όγκου των δεδομένων. Με καθυστέρηση μετάδοσης (latency) τόσο χαμηλή όσο 5ms, υπερέχει στην επεξεργασία δεδομένων σε πραγματικό χρόνο.

- Συνεχής ροή δεδομένων: Ένα από τα βασικά πλεονεκτήματα του Kafka είναι η ικανότητά του να διατηρεί συνεχή ροή δεδομένων προς τους καταναλωτές των μηνυμάτων χωρίς μη αυτόματες διακοπές. Σε περίπτωση που κάποιος καταναλωτής δεν ανταποκριθεί σε καθορισμένο χρονικό διάστημα γίνεται αυτόματα ανακατανομή του φορτίου στους υπόλοιπους καταναλωτές. Αυτό είναι πολύτιμο για την ανάλυση και επεξεργασία δεδομένων σε πραγματικό χρόνο.

- Προσωρινή αποθήκευση μηνυμάτων: Ο Kafka παρέχει προσωρινή αποθήκευση μηνυμάτων, επιτρέποντας την εκ νέου κατανάλωση των μηνυμάτων. Υποστηρίζει επίσης ομαδοποίηση και συμπίεση μηνυμάτων, μειώνοντας την επιβάρυνση του δικτύου.

- Διατήρηση και Επεξεργασία Δεδομένων: Η πλατφόρμα είναι εξοπλισμένη με λειτουργίες όπως διατήρηση δεδομένων, επεξεργασία και οπτικοποίηση. Μπορεί να χρησιμοποιηθεί για τη δημοσίευση πληροφοριών, καθιστώντας το ένα ευέλικτο εργαλείο για διάφορες ανάγκες επεξεργασίας δεδομένων.

- Ανοχή σε σφάλματα και επεκτασιμότητα: Η αρχιτεκτονική του Kafka είναι εγγενώς ανεκτική σε σφάλματα και επεκτάσιμη. Μπορεί να διατηρήσει τη σειρά προτεραιότητας ακόμα και όταν ασχολείται με μεγάλους όγκους δεδομένων, διασφαλίζοντας την αξιοπιστία και την ακεραιότητα του συστήματος.

Στην εργασία μας, ο Apache Kafka διαδραματίζει κεντρικό ρόλο στην πρόσληψη, επεξεργασία και διάδοση δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Αποτελεί μια ζωτικής σημασίας γέφυρα που συνδέει τις πηγές δεδομένων με τα αναλυτικά και προγνωστικά μας στοιχεία. Καθώς εμβαθύνουμε στην υλοποίηση του συστήματός μας, θα δείξουμε πώς ο Kafka ενσωματώνεται απρόσκοπτα στη γραμμή επεξεργασίας δεδομένων μας, συμβάλλοντας στην έγκαιρη και αποτελεσματική ανάλυση των δεδομένων κρυπτονομισμάτων.

4.4.1 Τα κύρια στοιχεία της αρχιτεκτονικής του Apache Kafka

Στο Apache Kafka, η αρχιτεκτονική ορίζεται από ένα σύνολο βασικών στοιχείων που λειτουργούν αρμονικά για να επιτρέπουν την αποτελεσματική ροή και επεξεργασία δεδομένων. Αυτά τα στοιχεία παρέχουν τη βάση για το σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Ας εμβαθύνουμε σε καθένα από αυτά τα στοιχεία [53]:

- Topic (Θέμα): Όλα τα εισερχόμενα μηνύματα και τα αρχεία είναι οργανωμένα σε κατηγορίες γνωστές ως θέματα. Οι χρήστες μπορούν να δημοσιεύουν ή να εγγραφούν σε συγκεκριμένα θέματα. Αυτά τα θέματα, με τη σειρά τους,

υποδιαιρούνται σε διαμερίσματα (partitions) για να εξασφαλιστεί ο παραλληλισμός. Αυτή η κατάτμηση επιτρέπει την αποτελεσματική διανομή και επεξεργασία δεδομένων σε πολλούς μεσίτες (brokers) [53].

- Broker: Το Kafka λειτουργεί μέσα σε ένα σύμπλεγμα υπολογιστικών συστημάτων, που περιλαμβάνει έναν ή περισσότερους διακομιστές, καθένας από τους οποίους αναφέρεται ως μεσίτης Kafka. Αυτοί οι μεσίτες παίζουν κεντρικό ρόλο στη κατάτμηση και τη διανομή δεδομένων σε όλο το σύμπλεγμα, επιτρέποντας την παράλληλη επεξεργασία. Μέσω αυτής της διανομής επιτυγχάνονται υψηλή απόδοση και γρήγορη επεξεργασία δεδομένων. Κάθε διαμέρισμα σε ένα θέμα έχει έναν καθορισμένο αρχηγό και ένα ή περισσότερα αντίγραφα που ονομάζονται ακόλουθοι. Ο αρχηγός είναι υπεύθυνος για την επίβλεψη της ανάγνωσης και της εγγραφής και σε περίπτωση αποτυχίας του αρχηγού, ένας από τους ακόλουθους αναλαμβάνει αυτόματα, διασφαλίζοντας την αδιάλειπτη ροή δεδομένων [53].

- Παραγωγός: Ο παραγωγός είναι μια υπηρεσία που είναι υπεύθυνη για τη δημοσίευση μηνυμάτων σε ένα ή περισσότερα θέματα. Οι παραγωγοί διαδραματίζουν κρίσιμο ρόλο στην έναρξη της ροής δεδομένων εντός του συστήματος Kafka [53].

- Καταναλωτής: Οι καταναλωτές είναι υπηρεσίες που διαβάζουν μηνύματα από μεσίτες για θέματα στα οποία έχουν εγγραφεί. Είναι καθοριστικής σημασίας για την επεξεργασία και την ανάλυση των δεδομένων που προσλαμβάνει το Kafka [53].

- Δομημένη ροή: Η δομημένη ροή είναι ένα θεμελιώδες χαρακτηριστικό στην αρχιτεκτονική του Apache Kafka. Διευκολύνει την επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο διατηρώντας παράλληλα τη δομή τους. Αυτό είναι ιδιαίτερα σημαντικό για την επεξεργασία δεδομένων κρυπτονομισμάτων, όπου η ακεραιότητα των δεδομένων και η διατήρηση της δομής είναι ζωτικής σημασίας [53].

- Zookeeper: Το Zookeeper είναι μια κατανεμημένη υπηρεσία ανοιχτού κώδικα που λειτουργεί ως συντονιστής στο οικοσύστημα Kafka. Ενημερώνει τους παραγωγούς και τους καταναλωτές για τους διαθέσιμους μεσίτες, παρέχει κρίσιμες πληροφορίες και διασφαλίζει την ονομασία και το συγχρονισμό σε κατανεμημένα συστήματα. Το Zookeeper είναι ζωτικής σημασίας για τη διατήρηση της σταθερότητας και του συντονισμού του συμπλέγματος Kafka [53].

- Αρχείο καταγραφής (Log): Το Kafka περιλαμβάνει ένα σύστημα αρχείων για να αποθηκεύει τα μηνύματα που λαμβάνει. Αυτό το αρχείο καταγραφής είναι η ραχοκοκαλιά του συστήματος, διασφαλίζοντας την ανθεκτικότητα και την αξιοπιστία των δεδομένων [53].

- Ομάδες Καταναλωτών: Οι καταναλωτές είναι οργανωμένοι σε ομάδες καταναλωτών, οι οποίες διευκολύνουν την παράλληλη επεξεργασία και την κατανομή φορτίου. Αυτός ο μηχανισμός ομαδοποίησης εξασφαλίζει αποτελεσματική και κλιμακούμενη κατανάλωση δεδομένων [53].

- Αντιγραφή: Η αναπαραγωγή δεδομένων είναι μια κρίσιμη πτυχή της αρχιτεκτονικής του Kafka. Εξασφαλίζει τη διαθεσιμότητα δεδομένων και την ανοχή σε σφάλματα αντιγράφοντας δεδομένα σε ένα ή περισσότερα διαμερίσματα ακολούθων. Αυτός ο μηχανισμός αναπαραγωγής είναι απαραίτητος για την ακεραιότητα των δεδομένων και την υψηλή διαθεσιμότητα [53].

Στην αρχιτεκτονική του Apache Kafka, τα δεδομένα ρέουν μέσω του συστήματος με βάση το μοντέλο ροής δομημένο από τον παραγωγό. Οι παραγωγοί στέλνουν μηνύματα σε συγκεκριμένα θέματα και οι καταναλωτές λαμβάνουν αυτά τα εισερχόμενα μηνύματα. Αυτός ο μηχανισμός παρέχει τη βάση για επεξεργασία, ανάλυση και διανομή δεδομένων σε πραγματικό χρόνο [53].

4.5 Η Βάση Δεδομένων MongoDB

Η βάση δεδομένων MongoDB είναι μία δωρεάν και ανοιχτού κώδικα εγγραφο-κεντρική (document-databased) βάση δεδομένων που είναι διαθέσιμη για διάφορα λειτουργικά συστήματα [54]. Αν θελήσουμε να την κατηγοριοποιήσουμε ανήκει στις Μη-σχεσιακές βάσεις δεδομένων (No-SQL). Τα δεδομένα αποθηκεύονται σαν έγγραφα (documents) και παρόμοια έγγραφα ομαδοποιούνται σε συλλογές (collections). Για την αποθήκευσή τους τα έγραφα κωδικοποιούνται στην μορφή BJSON (Binary JavaScript Obect Notation) που αποτελεί μια binary εκδοχή της πολύ διαδεδομένης μορφής JSON. Παρόλο που η κάθε συλλογή της MongoDB είναι σαν ένας πίνακας (table) των RDBMS και το κάθε έγγραφο που περιέχεται σε αυτή τη συλλογή μπορούμε να το αντιστοιχίσουμε ως μία γραμμή του πίνακα, ο τρόπος που λειτουργούν είναι εντελώς διαφορετικός. Οι βασικές διαφορές είναι οι εξής:

- Κάθε έγγραφο μίας συλλογής μπορεί να διαφέρει ως προς τη μορφή με τα υπόλοιπα έγγραφα που ανήκουν στην ίδια συλλογή, σε αντίθεση με τις γραμμές του πίνακα που πρέπει να είναι της ίδιας μορφής.

- Κάθε έγγραφο είναι αυτόνομο και περιέχει ολόκληρη την πληροφορία χωρίς να ακολουθούνται οι κανόνες κανονικοποίησης δεδομένων. Αντίθετα στις RDBMS οι γραμμές ενός πίνακα μπορεί να σχετίζονται με γραμμές κάποιου άλλου πίνακα, φροντίζοντας πάντα για την κανονικοποίηση των δεδομένων και την αποφυγή του πλεονασμού.

- Αντίθετα με τις RDBMS που χρησιμοποιούν SQL για την υποβολή πολύπλοκων ερωτημάτων στη βάση που μπορεί να συνδυάζουν πολλαπλούς πίνακες και να εφαρμόζουν σύνθετες κριτήρια, στη MongoDB τα ερωτήματα είναι απλά και αφορούν πάντα μία μόνο συλλογή με τα κριτήρια να δίνονται με την μορφή ενός αντικειμένου JSON.

Τα παραπάνω χαρακτηριστικά και η πιο απλή δομή των δεδομένων, δίνουν τη δυνατότητα στη MongoDB να μπορεί εύκολα να λειτουργεί κατανεμημένα και να εξυπηρετεί τα ερωτήματα πολύ ταχύτερα από κάποια RDBMS. Δεδομένου ότι η πολυπλοκότητα των ερωτημάτων είναι περιορισμένη, σε αντίθεση με το SQL που μπορεί να εκφράσει εξαιρετικά πολύπλοκα ερωτήματα, η MongoDB μπορεί να παρέχει εγγυημένη ταχύτητα στην εκτέλεσή τους.

Η MongoDB, όπως γίνεται αντιληπτό, είναι σχεδιασμένη να χειρίζεται έγγραφα που δεν ακολουθούν ένα συγκεκριμένο σχήμα. Πιο συγκεκριμένα, δομεί τα δεδομένα της σε ζεύγη με τη μορφή “πεδίο:τιμή” (field: value) και τα ομαδοποιεί σε ένα BJSON μορφής έγγραφο (document). Οι τιμές που μπορούν να πάρουν τα πεδία δεν είναι συγκεκριμένες, αλλά χαρακτηρίζονται από πολυμορφισμό, από ευελιξία, δηλαδή κάθε έγγραφο σε μία συλλογή μπορεί να έχει διαφορετικό σύνολο πεδίων και οι τύποι δεδομένων να διαφέρουν από έγγραφο σε έγγραφο και να πάρουν ως τιμές άλλα έγγραφα ή πίνακες ή πίνακες από έγγραφα με αποτέλεσμα σε μία συλλογή να υπάρχουν διαφορετικές δομές και να προκύπτουν πολύ παραπάνω ιδιότητες. Στη συνέχεια τα παρόμοια έγγραφα αποθηκεύονται σε συλλογές (collections) και είναι έτοιμα προς εκμετάλλευση από τον τελικό χρήστη μέσω API, τα πρωτόκολλα επικοινωνίας μεταξύ του server και των clients που περιέχουν όλες τις απαραίτητες πληροφορίες [54].

4.5.1 Χαρακτηριστικά της βάσης δεδομένων MongoDB

Η MongoDB, όπως προαναφέραμε, είναι μία No-SQL βάση δεδομένων στην οποία αποθηκεύονται δεδομένα σε μία μορφή BJSON με σκοπό να διαχειρίζεται μεγάλο όγκο δεδομένων. Συνεπώς διαθέτει όλα τα πλεονεκτήματα των No-SQL βάσεων και υποστηρίζει τις εξής ιδιότητες: την οριζόντια κλιμάκωση (horizontal scaling), την υψηλή διαθεσιμότητα (high availability) και την υψηλή διεκπεραιωτική ικανότητα (throughput), οι οποίες προκύπτουν αντίστοιχα από τον θρυμματισμό (sharding), τα αντίγραφα που δημιουργεί (replica sets), αλλά και την κατανομή του φόρτου εργασίας σε όλο το σύστημα των clusters (cluster load balancing) [54].

4.5.2 Η MongoDB και οι διάφορες χρήσεις της

Η χρησιμότητα της Βάσης Δεδομένων MongoDB συναντάται στις παρακάτω περιπτώσεις:

- Για την αποθήκευση μεγάλου όγκου δεδομένων που αφορούν πχ δεδομένα συμπεριφοράς χρηστών ή διάφορες αναρτήσεις στα social media και τα οποία δεν μπορούν να αποθηκευτούν σε μορφή πίνακα.

- Για να πραγματοποιηθούν αναλύσεις και αναφορές σε πραγματικό χρόνο, λόγω της ικανότητας της MongoDB να διαβάζει και να εγγράφει με τάχιστους ρυθμούς.

- Χρησιμεύει σε συστήματα διαχείρισης περιεχομένου π.χ. σε blogs, λόγω της ευελιξίας της MongoDB να αποθηκεύει και να οργανώνει τα δεδομένα πολυμορφικά.

- Τέλος, μπορεί να χρησιμοποιηθεί μεταξύ των διαφόρων εφαρμογών για κινητές συσκευές με το διαδίκτυο που απαιτούν συγχρονισμό δεδομένων σε πραγματικό χρόνο [54].

ΚΕΦΑΛΑΙΟ 5

ΣΥΛΛΟΓΗ ΚΑΙ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Στον τομέα της χρηματοοικονομικής ανάλυσης, τα βασικά στάδια της συλλογής και της προεπεξεργασίας δεδομένων χρησιμεύουν ως το θεμέλιο για την απόκτηση ισχυρών γνώσεων και τη διευκόλυνση της καλά ενημερωμένης λήψης αποφάσεων. Η επινόηση μιας μεθοδικής προσέγγισης για την προμήθεια, τη συσσώρευση και τη βελτίωση των οικονομικών δεδομένων είναι απαραίτητη για τη δημιουργία ακριβών μοντέλων και την εξαγωγή διορατικών ερμηνειών. Αυτές οι κρίσιμες φάσεις, οπλίζουν το κατανεμημένο σύστημα επεξεργασίας δεδομένων με τα απαραίτητα εργαλεία και μεθοδολογίες για την κατάλληλη πλοήγηση στις περιπλοκές που είναι εγγενείς στην ανάλυση οικονομικών δεδομένων.

5.1 Πηγές δεδομένων

Στο τοπίο της χρηματοοικονομικής ανάλυσης, η προμήθεια αξιόπιστων και περιεκτικών δεδομένων είναι πρωταρχικής σημασίας για την ακρίβεια και την αποτελεσματικότητα κάθε αναλυτικής προσπάθειας. Η διαθεσιμότητα διαφορετικών πηγών δεδομένων δίνει τη δυνατότητα στους αναλυτές να έχουν πρόσβαση σε ένα ευρύ φάσμα χρηματοοικονομικών μέσων και δεικτών, εμπλουτίζοντας έτσι τις αναλυτικές τους ικανότητες.

Τα οικονομικά δεδομένα μπορούν να συλλεχθούν από πολλές πηγές, καθεμία από τις οποίες προσφέρει μοναδικές γνώσεις για διάφορες πτυχές της αγοράς. Αυτές οι πηγές περιλαμβάνουν παραδοσιακές χρηματοοικονομικές βάσεις δεδομένων, εξειδικευμένα API, πλατφόρμες ροής σε πραγματικό χρόνο και επιμελημένα σύνολα δεδομένων που παρέχονται από χρηματοπιστωτικά ιδρύματα και ρυθμιστικούς φορείς.

5.1.1 Βιβλιοθήκη «yfinance»

Η βιβλιοθήκη «yfinance» ξεχωρίζει ως εξέχουσα επιλογή για την πηγή δεδομένων μας, προσφέροντας πληθώρα ιστορικών και οικονομικών δεδομένων σε πραγματικό χρόνο για μια μεγάλη γκάμα περιουσιακών στοιχείων, συμπεριλαμβανομένων των κρυπτονομισμάτων. Αρκετοί επιτακτικοί λόγοι στηρίζουν την απόφασή μας να επιλέξουμε τη βιβλιοθήκη «yfinance»:

- Ολοκληρωμένη κάλυψη δεδομένων: Η βιβλιοθήκη «yfinance» ικανοποιεί εντυπωσιακά την ανάγκη μας για ολοκληρωμένα δεδομένα κρυπτονομισμάτων. Παρέχει έναν εκτενή κατάλογο κρυπτονομισμάτων, ζευγών συναλλαγών και ιστορικών δεδομένων. Αυτό το εύρος κάλυψης μας δίνει τη δυνατότητα να αποκτήσουμε πρόσβαση σε ένα ευρύ φάσμα περιουσιακών στοιχείων, ζωτικής σημασίας για την ανάλυσή μας.

- Αξιοπιστία: Στον χρηματοπιστωτικό κλάδο, η αξιοπιστία είναι αδιαπραγμάτευτη. Είναι γνωστό για την ακρίβεια και τη συνέπειά του στην παροχή δεδομένων, ένα απαραίτητο χαρακτηριστικό για την έρευνα και την ανάλυσή μας.

- Προσβασιμότητα: Η ευκολία πρόσβασης στα δεδομένα είναι βασικός παράγοντας για οποιαδήποτε πηγή δεδομένων. Συγκεκριμένα, αξιοποιούμε τη βιβλιοθήκη «yfinance», η οποία εξορθολογίζει την ανάκτηση δεδομένων και την απρόσκοπτη ενσωμάτωση στο σύστημα επεξεργασίας δεδομένων μας.

5.1.2 Χρήση της Βιβλιοθήκης «yfinance»

Στο πλαίσιο του σχεδιασμού και εφαρμογής του κατανεμημένου συστήματος επεξεργασίας κρυπτονομισμάτων σε πραγματικό χρόνο", χρησιμοποιήθηκε το η Βιβλιοθήκη «yfinance» ως βασική πηγή δεδομένων για την απόκτηση περιεκτικών ιστορικών και δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Αξιοποιώντας τις ισχυρές διεπαφές προγραμματισμού εφαρμογών (API) της Βιβλιοθήκης «yfinance», σχεδιάστηκε και εφαρμόστηκε σχολαστικά ένας μηχανισμός ανάκτησης δεδομένων προσαρμοσμένο στις συγκεκριμένες απαιτήσεις του έργου.

```
import yfinance as yf
import datetime
```

```
import pandas as pd

# Define parameters for data retrieval
symbols = ['BTC']  # List of cryptocurrency symbols
target_currency = 'USD'
start_date = '2014-09-17'
end_date = datetime.datetime.now()

# Function to retrieve historical cryptocurrency data from Yahoo
Finance
def get_yf_data(symbols, start_date, end_date):
    yf_historical_data = pd.DataFrame()
    for symbol in symbols:
        try:
            symbol_with_currency = f'{symbol}-USD'
            ticker = yf.Ticker(symbol_with_currency)
            data = ticker.history(start=start_date, end=end_date)
            data['Symbol'] = symbol_with_currency
            data['Source'] = "Yahoo Finance"
            yf_historical_data = pd.concat([yf_historical_data,
data])
        except Exception as e:
            print(f'Error getting historical data for {symbol}:
{e}')
    yf_historical_data = yf_historical_data.reset_index()
    return yf_historical_data

# Retrieve historical cryptocurrency data from Yahoo Finance
historical_data = get_yf_data(symbols, start_date, end_date)
```

Μέσω του «yfinance» API, το σύστημα έχει πρόσβαση σε ένα εκτεταμένο χώρο αποθήκευσης δεδομένων αγοράς κρυπτονομισμάτων που εκτείνεται σε διάφορα χρονικά πλαίσια και ζεύγη νομισμάτων. Παράμετροι όπως σύμβολα κρυπτονομισμάτων, νομίσματα-στόχοι και εύρη ημερομηνιών καθορίζονται μέσω προγραμματισμού ώστε να ευθυγραμμίζονται με τους ερευνητικούς στόχους και τα αναλυτικά πλαίσια του συστήματος.

Με μια δομημένη και αποτελεσματική διαδικασία ανάκτησης δεδομένων, το σύστημα απορροφά απρόσκοπτα μεγάλους όγκους δεδομένων κρυπτονομισμάτων στην κατανεμημένη αρχιτεκτονική του. Οι τυποποιημένες μορφές δεδομένων που παρέχονται από το API του «yfinance» ενισχύουν τη διαλειτουργικότητα και τη συμβατότητα με τα αναλυτικά εργαλεία και τους αγωγούς επεξεργασίας του

συστήματος, διευκολύνοντας τη συνεκτική ενοποίηση των ροών χρηματοοικονομικών δεδομένων.

Αυτή η στρατηγική χρήση του «yfinance» υπογραμμίζει τον κεντρικό ρόλο του στην προώθηση των αναλυτικών δυνατοτήτων του συστήματος και στην ενίσχυση της ικανότητάς του να περιηγείται στο δυναμικό τοπίο των αγορών κρυπτονομισμάτων.

5.2 Διαδικασία συλλογής δεδομένων

Στην προσπάθειά μας να διασφαλίσουμε την απορρόφηση και την επεξεργασία δεδομένων σε πραγματικό χρόνο, έχουμε υιοθετήσει το Apache Kafka ως τον βασικό άξονα του συστήματος συλλογής δεδομένων μας. Το Kafka, γνωστό για την στιβαρότητα και την επεκτασιμότητα του, χρησιμεύει ως μια ισχυρή πλατφόρμα για το χειρισμό ροών δεδομένων υψηλής απόδοσης. Η διαδικασία συλλογής δεδομένων καθοδηγείται από βασικά στοιχεία που ευθυγραμμίζονται στενά με τα αποσπάσματα κώδικα που περιγράφονται παρακάτω.

5.2.1 Χρήση του Kafka Producer και Kafka Structured Streaming

Για να διασφαλιστεί η απορρόφηση και η επεξεργασία δεδομένων σε πραγματικό χρόνο, χρησιμοποιούμε το Apache Kafka ως σύστημα ανταλλαγής μηνυμάτων. Το Kafka χρησιμεύει ως μια ισχυρή και επεκτάσιμη πλατφόρμα για το χειρισμό ροών δεδομένων υψηλής απόδοσης. Η διαδικασία συλλογής δεδομένων περιστρέφεται γύρω από τη χρήση των στοιχείων Kafka Producer και Kafka Structured Streaming. Αυτές οι τεχνολογίες επιτρέπουν στο σύστημα να απορροφά ροές δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο, να τις επεξεργάζεται με κατανεμημένο τρόπο και να τις ενσωματώνει απρόσκοπτα στη ροή αναλυτικών εργασιών.

- Kafka Producer: Η διαδικασία συλλογής δεδομένων ενισχύεται από την ενοποίηση του Kafka Producer, το οποίο χρησιμεύει ως μεσάζοντας μεταξύ «yfinance» και Kafka Topics. Το Kafka Producer είναι υπεύθυνο για τη συνεχή ανάκτηση δεδομένων κρυπτονομισμάτων από το **Yfinance** και τη δημοσίευσή τους σε καθορισμένα θέματα Kafka. Διαμορφωμένα για να λειτουργούν σε τακτά χρονικά διαστήματα, τα παράγωγα διασφαλίζουν ότι το σύστημα παραμένει ενήμερο για τις πιο πρόσφατες πληροφορίες της αγοράς.

```
from kafka import KafkaProducer

# Create a Kafka producer
producer = KafkaProducer(bootstrap_servers=['localhost:9092'])

# Iterate over each row of the DataFrame and send data to Kafka
topic
for index, row in streamline_data.iterrows():
    # Convert the row to a dictionary
    row_dict = row.to_dict()
    # Convert the dictionary to a JSON string
    json_string = json.dumps(row_dict)
    # Convert the JSON string to a bytes object
    message = bytes(json_string, 'utf-8')
    # Send the message to the Kafka topic
    producer.send('cryptoway', message)
```

- Structured Streaming with Kafka: Το Kafka Structured Streaming διαδραματίζει κεντρικό ρόλο στην απορρόφηση και επεξεργασία ροών δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο μέσα στο κατανεμημένο σύστημα. Με την απρόσκοπτη ενσωμάτωση με τα θέματα του Kafka, η δομημένη ροή επιτρέπει στο σύστημα να καταναλώνει, να επεξεργάζεται και να αναλύει εισερχόμενες ροές δεδομένων με τρόπο κατανεμημένο και ανεκτικό σε σφάλματα, χρησιμοποιώντας τις δομημένες δυνατότητες ροής του Apache Spark για την ενσωμάτωση του Kafka. Αυτή η προσέγγιση απλοποιεί την ανάπτυξη αγωγών δεδομένων σε πραγματικό χρόνο και επιτρέπει την απρόσκοπτη ενοποίηση με το ευρύτερο οικοσύστημα εργαλείων και βιβλιοθηκών επεξεργασίας δεδομένων του Spark, επιτρέποντας σχεδόν στιγμιαίες πληροφορίες και ενέργειες που βασίζονται στις αλλαγές της αγοράς.

```
# Read streaming data from Kafka topic
streaming_data = spark \
    .readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "localhost:9092") \
    .option("subscribe", "cryptoway") \
    .load()

# Convert Kafka message value to JSON and select required fields
streaming_data =
streaming_data.select(from_json(col("value").cast("string"),
```

```
schema).alias("data")) \
    .select("data.*")
```

5.3 Βήματα Προεπεξεργασίας Δεδομένων

Η προεπεξεργασία δεδομένων είναι μια κρίσιμη φάση στην ανάλυση των οικονομικών δεδομένων, που περιλαμβάνει μια σειρά βημάτων που στοχεύουν στον καθαρισμό, τον μετασχηματισμό και τη βελτίωση των ακατέργαστων δεδομένων για την προετοιμασία τους για επακόλουθη ανάλυση και μοντελοποίηση. Στο πλαίσιο του σχεδιασμού και εφαρμογής του κατανεμημένου συστήματος επεξεργασίας κρυπτονομισμάτων σε πραγματικό χρόνο", υλοποιούνται σχολαστικά βήματα προεπεξεργασίας για τη διασφάλιση της ακεραιότητας και της ποιότητας των δεδομένων που συλλέγονται.

5.3.1 Καθαρισμός δεδομένων

Ο καθαρισμός των δεδομένων από το **Yfinance** είναι ένα θεμελιώδες βήμα στη γραμμή προεπεξεργασίας του σχεδιασμού και εφαρμογής του κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Αυτή η διαδικασία περιλαμβάνει τη διόρθωση ασυνεπειών, την αντιμετώπιση τιμών που λείπουν και την τυποποίηση των μορφών δεδομένων για να διασφαλιστεί η ακεραιότητα και η χρηστικότητα του ληφθέντος συνόλου δεδομένων. Παρακάτω, περιγράφουμε τα βασικά βήματα που σχετίζονται με τον καθαρισμό των δεδομένων **Yfinance**, διευκρινίζοντας κάθε βήμα με βάση το παρεχόμενο παραδείγμα κώδικα Python:

```
def clean_yf_data(symbols, start_date, end_date):
    historical_data = get_yf_data(symbols, start_date, end_date)

    # Selecting relevant columns
    historical_data = historical_data[['Date', 'Open', 'High',
'Low', 'Close', 'Volume', 'Symbol', 'Source']]

    # Modifying data formats
    historical_data['Symbol'] =
historical_data['Symbol'].str.replace('-USD', '')
    historical_data['Date'] =
pd.to_datetime(historical_data['Date']).dt.strftime('%d-%m-%Y')

    return historical_data
```

```
processed_data = clean_yf_data(symbols, start_date, end_date)
```

Η διαδικασία καθαρισμού ενσωματώνεται σε μια αποκλειστική λειτουργία προσαρμοσμένη για την αντιμετώπιση ασυνεπειών δεδομένων και τη βελτίωση της αναγνωσιμότητας των δεδομένων. Βασικά βήματα καθαρισμού:

- Επιλογή στήλης: Οι άσχετες στήλες απορρίπτονται για να επικεντρωθούν σε σχετικά χαρακτηριστικά δεδομένων που είναι κρίσιμα για την ανάλυση, όπως ημερομηνία, άνοιγμα, υψηλό, χαμηλό, κλείσιμο, τόμος, σύμβολο και πηγή.

- Τυποποίηση μορφής: Το σύμβολο του κρυπτονομίσματος τυποποιείται αφαιρώντας το επίθημα «-USD», ενισχύοντας τη συνοχή και διευκολύνοντας τη συγκέντρωση και ανάλυση δεδομένων. Επιπλέον, η μορφή ημερομηνίας μετατρέπεται σε πιο ευανάγνωστη και ομοιόμορφη μορφή ('%d-%m-%Y').

- Υλοποίηση και Εκροή: Τα καθαρισμένα δεδομένα **Yfinance**, δομημένα σε ένα τυποποιημένο DataFrame, χρησιμεύουν ως βασικό σύνολο δεδομένων για μετέπειτα ανάλυση και μοντελοποίηση εντός του κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο.

5.3.2 Εφαρμογή Τεχνικών δεικτών

Η εφαρμογή τεχνικών δεικτών είναι ένα κομβικό βήμα στη φάση προεπεξεργασίας του συστήματος. Αυτό το βήμα περιλαμβάνει τον υπολογισμό και την ενσωμάτωση διαφόρων τεχνικών δεικτών στο σύνολο δεδομένων, ενισχύοντάς το με πολύτιμα χαρακτηριστικά για μετέπειτα ανάλυση και μοντελοποίηση. Παρακάτω, εμβαθύνουμε στις λεπτομέρειες της εφαρμογής τεχνικών δεικτών με βάση τον κώδικα Python:

```
def apply_technical_indicators(df):
    df['MA'] = moving_average(df, window_size=10)
    df['RSI'] = RSI(df, window_size=14)
    df['MACD'], df['MACD_signal'], df['MACD_hist'] =
MACD(df['Close'])
    df['vortex'] = vortex(df, n=14)
    df['AO'] = AO(df)
    df['ATR'] = ATR(df, n=14)
    df['CCI'] = CCI(df, n=14)
    df['ADX'] = ADX(df, n=14)
```

```
    df['Williams_%R'] = williams(df, n=14)
    df['%K'], df['%D'] = stoch(df, n=14)
    df['OBV'] = OBV(df)
    df['EMA'] = EMA(df, window_size=10)

    df = df.dropna()  # Dropping rows containing NaNs

    return df

total_data = apply_technical_indicators(processed_data)
```

- Λειτουργίες τεχνικού δείκτη: Ένα σύνολο συναρτήσεων ορίζεται για τον υπολογισμό διαφορετικών τεχνικών δεικτών. Αυτές οι λειτουργίες εφαρμόζονται στο DataFrame που περιέχει τα καθαρισμένα δεδομένα κρυπτονομισμάτων **Yfinance**.

- Βασικά βήματα για την εφαρμογή τεχνικών δεικτών:

    - Κινητός μέσος όρος (MA): Ο κινητός μέσος όρος υπολογίζεται χρησιμοποιώντας τη συνάρτηση move_average και εκχωρείται σε μια νέα στήλη 'MA' στο DataFrame.

    - Δείκτης σχετικής ισχύος (RSI): Ο RSI υπολογίζεται χρησιμοποιώντας τη συνάρτηση RSI, προσθέτοντας μια νέα στήλη «RSI» στο DataFrame.

    - Κινητός μέσος όρος απόκλισης σύγκλισης (MACD): Το MACD και τα στοιχεία του (γραμμή σήματος και ιστόγραμμα) υπολογίζονται χρησιμοποιώντας τη συνάρτηση MACD.

    - Δείκτης στροβιλισμού: Ο δείκτης στροβιλισμού υπολογίζεται χρησιμοποιώντας τη συνάρτηση στροβιλισμού.

    - Awesome Oscillator (AO): Το AO υπολογίζεται χρησιμοποιώντας τη συνάρτηση AO.

    - Average True Range (ATR): Το ATR υπολογίζεται χρησιμοποιώντας τη συνάρτηση ATR.

- Δείκτης καναλιού εμπορευμάτων (CCI): Ο CCI υπολογίζεται χρησιμοποιώντας τη συνάρτηση CCI.

- Μέσος δείκτης κατεύθυνσης (ADX): Το ADX υπολογίζεται χρησιμοποιώντας τη συνάρτηση ADX.

- Williams %R: Ο Williams %R υπολογίζεται χρησιμοποιώντας τη συνάρτηση Williams.

- Στοχαστικός Ταλαντωτής (%K και %D): Οι τιμές %K και %D υπολογίζονται χρησιμοποιώντας τη συνάρτηση stoch.

- On Balance Volume (OBV): Το OBV υπολογίζεται χρησιμοποιώντας τη συνάρτηση OBV.

- Εκθετικός κινητός μέσος όρος (EMA): Το EMA υπολογίζεται χρησιμοποιώντας τη συνάρτηση EMA.

- Αφαίρεση NaN: Οι σειρές που περιέχουν τιμές NaN που προκύπτουν από τους υπολογισμούς του δείκτη αφαιρούνται.

- Ενοποίηση και Έξοδος: Το DataFrame total_data εμπλουτίζεται με αυτούς τους τεχνικούς δείκτες, δημιουργώντας ένα σύνολο δεδομένων πλούσιο σε χαρακτηριστικά, έτοιμο για περαιτέρω ανάλυση στο κατανεμημένο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο.

ΚΕΦΑΛΑΙΟ 6

ΑΠΟΘΗΚΕΥΣΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ

Η αποθήκευση και η διαχείριση δεδομένων είναι βασικό στοιχείο του σχεδιασμού και εφαρμογής κατανεμημένου συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο. Το σύστημα υιοθετεί μια ολοκληρωμένη προσέγγιση για την αποθήκευση δεδομένων, που περιλαμβάνει την αποθήκευση επεξεργασμένων ιστορικών συνόλων δεδομένων κρυπτονομισμάτων. Η απρόσκοπτη ενοποίηση των μηχανισμών αποθήκευσης επιτρέπει αποτελεσματικές διαδικασίες ανάκτησης δεδομένων, ανάλυσης και λήψης αποφάσεων εντός του κατανεμημένου συστήματος.

6.1 Μηχανισμοί Αποθήκευσης Δεδομένων

Το σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων μας αξιοποιεί έναν προσεκτικά επιλεγμένο συνδυασμό μηχανισμών αποθήκευσης για να διασφαλίσει ισχυρή διαχείριση και ανάκτηση δεδομένων. Το Hadoop Distributed File System (HDFS) αποτελεί τον πυρήνα της υποδομής αποθήκευσης των επεξεργασμένων ιστορικών δεδομένων. Με την κατανεμημένη και ανεκτική σε σφάλματα αρχιτεκτονική του, το HDFS παρέχει μια επεκτάσιμη λύση για την αποθήκευση τεράστιου όγκου ιστορικών δεδομένων κρυπτονομισμάτων. Η ικανότητά του να αναπαράγει δεδομένα σε πολλούς κόμβους εξασφαλίζει υψηλή διαθεσιμότητα και αξιοπιστία, ζωτικής σημασίας για την αδιάλειπτη πρόσβαση και επεξεργασία δεδομένων.

Συμπληρώνοντας το HDFS, MongoDB χρησιμεύει ως η βάση για τις τελικές προβλέψεις στο σύστημα επεξεργασίας δεδομένων. Το ευέλικτο προσανατολισμένο στα έγγραφα μοντέλο της MongoDB μας επιτρέπει να αποθηκεύουμε και να αναζητούμε πολύπλοκα, ημιδομημένα δεδομένα με ευκολία. Με την υιοθέτηση του MongoDB, ενισχύουμε το σύστημά μας να χειρίζεται δυναμικά και εξελισσόμενα σχήματα δεδομένων που είναι εγγενή στις τάσεις της αγοράς κρυπτονομισμάτων. Οι

δυνατότητές του οριζόντιας κλιμάκωσης επιτρέπουν την απρόσκοπτη επέκταση καθώς αυξάνονται οι όγκοι δεδομένων μας, διασφαλίζοντας βέλτιστη απόδοση και ανταπόκριση σε ερωτήματα και ενημερώσεις σε πραγματικό χρόνο.

Μαζί, το HDFS και το MongoDB σχηματίζουν ένα ισχυρό και ευέλικτο οικοσύστημα αποθήκευσης, προσαρμοσμένο στις μοναδικές απαιτήσεις της επεξεργασίας δεδομένων κρυπτονομισμάτων. Συνδυάζοντας την επεκτασιμότητα του HDFS με την ευελιξία του MongoDB, δημιουργούμε μια ολοκληρωμένη λύση ικανή να καλύψει τις εξελισσόμενες ανάγκες των εφαρμογών που βασίζονται σε δεδομένα, διασφαλίζοντας παράλληλα αξιοπιστία, απόδοση και επεκτασιμότητα σε κάθε στάδιο του κύκλου ζωής των δεδομένων.

6.1.1 Hadoop Distributed File System (HDFS)

Το σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων μας βασίζεται στο σύστημα κατανεμημένων αρχείων Hadoop (HDFS) ως βασική υποδομή για την αποθήκευση επεξεργασμένων ιστορικών δεδομένων. Το HDFS προσφέρει μια κατανεμημένη και ανεκτική σε σφάλματα αρχιτεκτονική, καθιστώντας το μια επεκτάσιμη λύση για το χειρισμό τεράστιου όγκου ιστορικών δεδομένων κρυπτονομισμάτων.

Με το HDFS, η αναπαραγωγή δεδομένων σε πολλούς κόμβους εξασφαλίζει υψηλή αξιοπιστία και διαθεσιμότητα, ζωτικής σημασίας για την αδιάλειπτη πρόσβαση και επεξεργασία δεδομένων. Η κατανεμημένη φύση του HDFS επιτρέπει την παράλληλη επεξεργασία δεδομένων, επιτρέποντας την αποτελεσματική ανάκτηση και ανάλυση μεγάλων συνόλων δεδομένων.

Αξιοποιώντας το HDFS, το σύστημά μας μπορεί να ικανοποιήσει τις αυξανόμενες απαιτήσεις επεξεργασίας δεδομένων κρυπτονομισμάτων, διατηρώντας παράλληλα τη βέλτιστη απόδοση και αξιοπιστία.

6.1.2 MongoDB

Στο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων μας, το MongoDB χρησιμεύει ως βάση για την αποθήκευση τελικών προβλέψεων και επεξεργασμένων δεδομένων. Το ευέλικτο, προσανατολισμένο στα έγγραφα μοντέλο του MongoDB προσφέρει τη δυνατότητα αποθήκευσης και αναζήτησης πολύπλοκων, ημιδομημένων δεδομένων με ευκολία.

Η φύση του MongoDB που βασίζεται σε έγγραφα μας επιτρέπει να προσαρμοστούμε σε δυναμικά και εξελισσόμενα σχήματα δεδομένων που είναι εγγενή στις αγορές κρυπτονομισμάτων. Καθώς το τοπίο των κρυπτονομισμάτων εξελίσσεται, το MongoDB μας δίνει τη δυνατότητα να προσαρμόζουμε απρόσκοπτα τις αλλαγές στη δομή και τις απαιτήσεις δεδομένων.

Η αρχιτεκτονική κλιμάκωσης του MongoDB εξασφαλίζει απρόσκοπτη επέκταση καθώς αυξάνονται οι όγκοι δεδομένων μας. Αυτή η δυνατότητα επεκτασιμότητας επιτρέπει στο σύστημά μας να χειρίζεται αυξανόμενα φορτία δεδομένων, διατηρώντας παράλληλα τη βέλτιστη απόδοση και ανταπόκριση σε ερωτήματα και ενημερώσεις σε πραγματικό χρόνο.

Αξιοποιώντας το MongoDB, ενισχύουμε την ευελιξία και την ευελιξία του συστήματός μας, επιτρέποντάς του να εξελίσσεται παράλληλα με τη δυναμική αγορά κρυπτονομισμάτων. Ο συνδυασμός του MongoDB με το HDFS σχηματίζει ένα ισχυρό και ευέλικτο οικοσύστημα αποθήκευσης, ικανό να καλύψει τις εξελισσόμενες ανάγκες των εφαρμογών που βασίζονται σε δεδομένα...

6.2 Επεξήγηση της Διαδικασίας Αποθήκευσης και Διαχείρισης Δεδομένων

Η διαδικασία αποθήκευσης και διαχείρισης δεδομένων στο κατανεμημένο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων σε πραγματικό χρόνο περιλαμβάνει μια σειρά βημάτων που στοχεύουν στη διασφάλιση της αποτελεσματικότητας, της αξιοπιστίας και της απρόσκοπτης πρόσβασης τόσο σε ιστορικά όσο και σε επεξεργασμένα δεδομένα. Παρακάτω, παρατίθεται μια περιεκτική εξήγηση των βασικών πτυχών αυτής της διαδικασίας:

- Διαχωρισμός και αποθήκευση δεδομένων: Τα ιστορικά και τα επεξεργασμένα δεδομένα κρυπτονομισμάτων διαχωρίζονται σαφώς για να βελτιστοποιηθεί η ανάκτηση και η ανάλυση. Τα ιστορικά δεδομένα, αποθηκευμένα στο HDFS, παρέχουν ένα αξιόπιστο αρχείο για backtesting και αναφορά. Τα επεξεργασμένα δεδομένα, που στεγάζονται στο MongoDB, αποτελούν τη βάση για προβολές και αναλύσεις σε πραγματικό χρόνο.

- Αποθήκευση ιστορικών δεδομένων σε HDFS: Το Hadoop Distributed File System (HDFS) λειτουργεί ως αποθήκη για ιστορικά δεδομένα κρυπτονομισμάτων. Η κατανεμημένη αρχιτεκτονική του επιτρέπει στο σύστημα να χειρίζεται τεράστια σύνολα δεδομένων με ανοχή σφαλμάτων. Η διαδικασία περιλαμβάνει τη μεταφόρτωση ιστορικών δεδομένων στο HDFS, διασφαλίζοντας τον πλεονασμό μέσω της αναπαραγωγής σε πολλούς κόμβους.

- Αποθήκευση επεξεργασμένων δεδομένων στο MongoDB: Το MongoDB χρησιμοποιείται για την αποθήκευση τελικών προβολών και επεξεργασμένων δεδομένων. Το ευέλικτο, προσανατολισμένο στα έγγραφα μοντέλο του ενσωματώνει τη δυναμική φύση των δεδομένων κρυπτονομισμάτων. Το σύστημα εγγράφει επεξεργασμένα δεδομένα στο MongoDB, αξιοποιώντας την επεκτασιμότητα του για τη διαχείριση των εξελισσόμενων δομών και απαιτήσεων δεδομένων.

Αυτή η ολιστική προσέγγιση για την αποθήκευση και τη διαχείριση δεδομένων δίνει τη δυνατότητα στο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων μας να χειρίζεται ανεμπόδιστα την πολυπλοκότητα τόσο των ιστορικών όσο και των δεδομένων σε πραγματικό χρόνο. Εξασφαλίζει αξιοπιστία, απόδοση και προσαρμοστικότητα, συμβάλλοντας στην ικανότητα του συστήματος να λαμβάνει τεκμηριωμένες αποφάσεις με βάση τις πιο πρόσφατες και σχετικές γνώσεις.

6.2.1 Διαχωρισμός και αποθήκευση δεδομένων

Το ταξίδι διαχείρισης δεδομένων μας ξεκινά με το κομβικό βήμα του διαχωρισμού και της αποθήκευσης δεδομένων κρυπτονομισμάτων. Ο πρωταρχικός στόχος είναι να διασφαλιστεί ο αποτελεσματικός χειρισμός των δεδομένων διαχωρίζοντάς τα προσεκτικά σε δύο βασικά στοιχεία: ιστορικά δεδομένα και δεδομένα σε πραγματικό χρόνο. Αυτός ο διαχωρισμός δίνει τη δυνατότητα στο σύστημά μας να διαχειρίζεται ιστορικά δεδομένα με εξειδικευμένη προσοχή, ενώ εγγυάται πρόσβαση σε πραγματικό χρόνο στις πιο πρόσφατες πληροφορίες.

```
        # Splitting total_data into two datasets
        historical_data = total_data.iloc[:-1]  # Historical Data
```

```
        streamline_data = total_data.iloc[[-1]]  # Streaming Data

        historical_data.to_csv("historical_data.csv")
```

6.2.2 Αποθήκευση ιστορικών δεδομένων στο HDFS

Στο κατανεμημένο σύστημα, η αποθήκευση ιστορικών δεδομένων στο κατανεμημένο σύστημα αρχείων Hadoop (HDFS) είναι ένα κρίσιμο στοιχείο. Αυτή η διαδικασία περιλαμβάνει πολλά βήματα για τη διασφάλιση ενοποίησης και της αξιόπιστης αποθήκευσης εκτεταμένων ιστορικών συνόλων δεδομένων κρυπτονομισμάτων.

```
# Create an HDFS client
        hdfs_client = InsecureClient('http://localhost:9870',
user='yannis')

        # Define the local file path and HDFS destination path
        local_file_path = 'historical_data.csv'
        hdfs_destination_path =
'/user/yannis/historical_data/historical_data.csv'

        # Upload the local file to HDFS
        with open(local_file_path, 'rb') as local_file:
            hdfs_client.write(hdfs_destination_path, local_file,
overwrite=True)

        print(f"File uploaded to HDFS at: {hdfs_destination_path}")
```

- Δημιουργία προγράμματος-πελάτη HDFS: Αρχικά, δημιουργείται ένα πρόγραμμα-πελάτης HDFS, που δημιουργεί μια σύνδεση με το σύστημα HDFS. Αυτός ο πελάτης διευκολύνει την επικοινωνία και την αλληλεπίδραση με το HDFS cluster.

- Καθορισμός Διαδρομών Αρχείων: Οι τοπικές διαδρομές και οι διαδρομές αρχείων HDFS ορίζονται. Η μεταβλητή local_file_path δείχνει το τοπικό αρχείο CSV που περιέχει ιστορικά δεδομένα κρυπτονομισμάτων, ενώ το hdfs_destination_path υποδεικνύει τη θέση-στόχο στο HDFS όπου θα αποθηκευτούν τα δεδομένα.

- Μεταφόρτωση δεδομένων στο HDFS: Το τοπικό αρχείο CSV ανοίγει και διαβάζεται σε δυαδική λειτουργία. Χρησιμοποιώντας τον υπολογιστή-πελάτη HDFS, το αρχείο εγγράφεται στη συνέχεια στην καθορισμένη διαδρομή προορισμού στο HDFS. Η παράμετρος overwrite=True διασφαλίζει ότι οποιοδήποτε υπάρχον αρχείο στη διαδρομή προορισμού αντικαθίσταται.

- Μήνυμα επιβεβαίωσης: Μετά την επιτυχή μεταφόρτωση, εκτυπώνεται ένα μήνυμα επιβεβαίωσης, το οποίο υποδεικνύει την ολοκλήρωση της διαδικασίας και τη θέση του μεταφορτωμένου αρχείου στο HDFS.

Αυτή η διαδικασία διασφαλίζει ότι τα ιστορικά δεδομένα κρυπτονομισμάτων αποθηκεύονται με ασφάλεια στο HDFS, έτοιμα για πρόσβαση και επεξεργασία από το κατανεμημένο σύστημα για ανάλυση και πληροφορίες.

6.2.3 Αποθήκευση επεξεργασμένων δεδομένων στο MongoDB

Στο σύστημα επεξεργασίας δεδομένων κρυπτονομισμάτων, το MongoDB διαδραματίζει κεντρικό ρόλο στη διατήρηση των εκτιμήσεων των τιμών κρυπτονομισμάτων. Αποτελεί αναπόσπαστο μέρος του αγωγού δεδομένων σε πραγματικό χρόνο, που λειτουργεί ως κεντρικός χώρος αποθήκευσης των αποτελεσμάτων

```
# Establish a connection to MongoDB
client = pymongo.MongoClient("mongodb://localhost:27017/")
db = client["db"]
collection = db["cryptocollection"]
```

- Δημιουργία σύνδεσης με το MongoDB: Αρχικά, δημιουργείται μια σύνδεση με το MongoDB χρησιμοποιώντας τη βιβλιοθήκη pymongo. Η κλάση MongoClient διευκολύνει τη σύνδεση και το URI "mongodb://localhost:27017/" δείχνει την τοπική παρουσία MongoDB που εκτελείται στην προεπιλεγμένη θύρα.

```
# Write to MongoDB using the modified DataFrame
streaming_predictions.writeStream \
    .foreachBatch(write_to_mongodb) \
    .start() \
    .awaitTermination()
```

- Μέθοδο WriteStream: Στη συνέχεια, τα επεξεργασμένα δεδομένα από τις προβλέψεις ροής εγγράφονται στο MongoDB χρησιμοποιώντας τη μέθοδο writeStream. Η συνάρτηση foreachBatch καθορίζει μια επιστροφή κλήσης (write_to_mongodb) που θα εκτελεστεί για κάθε παρτίδα δεδομένων.

```
def write_to_mongodb(df, epoch_id):
    static_data = df.collect()  # Collect streaming data into a
static list

    # Apply your processing logic to the collected static data
    processed_data = []
    for row in static_data:
        # Convert DenseVector to a serializable format
        processed_row = {key: convert_dense_vector(value) for key,
value in row.asDict().items()}

        # Remove the DenseVector column from the dictionary
        processed_row.pop('features', None)

        processed_data.append(processed_row)

    # Insert the processed data into the MongoDB collection
    if processed_data:  # Check if the list is not empty
        collection.insert_many(processed_data)

    # Convert the processed data back to a DataFrame if necessary
    processed_df = spark.createDataFrame(processed_data,
schema=df.schema)

    # Show or perform any actions with the processed DataFrame
    processed_df.show()
```

- Εγγραφή στο MongoDB σε παρτίδες: Η συνάρτηση write_to_mongodb έχει σχεδιαστεί για να χειρίζεται την εγγραφή δεδομένων στο MongoDB σε παρτίδες. Συλλέγει τα δεδομένα ροής σε μια στατική λίστα, επεξεργάζεται κάθε σειρά, μετατρέπει στήλες DenseVector και εγγράφει τα επεξεργασμένα δεδομένα στη συλλογή MongoDB. Η συνάρτηση περιλαμβάνει επίσης λογική για το χειρισμό πιθανών ζητημάτων μετατροπής και διασφαλίζει ότι τα επεξεργασμένα δεδομένα εισάγονται στη συλλογή MongoDB. Τα επεξεργασμένα δεδομένα μπορούν στη συνέχεια να μετατραπούν ξανά σε DataFrame εάν χρειάζεται και μπορούν να εκτελεστούν πρόσθετες ενέργειες.

Αυτή η διαδικασία διασφαλίζει ότι τα επεξεργασμένα δεδομένα, εμπλουτισμένα με πολύτιμες πληροφορίες, αποθηκεύονται αποτελεσματικά στο MongoDB, διαμορφώνοντας μια ισχυρή βάση για περαιτέρω ανάλυση και λήψη αποφάσεων εντός του συστήματος επεξεργασίας δεδομένων κρυπτονομισμάτων.

ΚΕΦΑΛΑΙΟ 7

ΣΧΕΔΙΑΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Στον τομέα της λήψης αποφάσεων βάσει δεδομένων, ο σχεδιασμός και η εφαρμογή ενός ισχυρού μοντέλου μηχανικής μάθησης αποτελούν πυλώνες καινοτομίας και διορατικότητας. Εμβαθύνοντας στην περίπλοκη διαδικασία της αρχιτεκτονικής και της ανάπτυξης μιας λύσης μηχανικής μάθησης, προσαρμοσμένης για να ξετυλίξει τις πολυπλοκότητες και τις αποχρώσεις που ενσωματώνονται σε τεράστια σύνολα δεδομένων. Από την αξιοποίηση τεχνολογιών αιχμής όπως το Apache Spark μέχρι τη σχολαστική δημιουργία εποπτευόμενων αλγορίθμων εκμάθησης και τη διεξαγωγή αυστηρών αξιολογήσεων μοντέλων, το ταξίδι συνδυάζει τη σύντηξη της θεωρητικής γνώσης με την πρακτική εφαρμογή. Καθώς η εξερεύνηση ξεδιπλώνεται, αποκαλύπτονται μεθοδολογίες, προβληματισμοί και βέλτιστες πρακτικές που στηρίζουν το σχεδιασμό και την εφαρμογή ενός μοντέλου μηχανικής μάθησης που είναι έτοιμο να φωτίσει τα μοτίβα, τις τάσεις πρόβλεψης και να οδηγήσει σε τεκμηριωμένες αποφάσεις στο συνεχώς εξελισσόμενο τοπίο της ανάλυσης δεδομένων με την βοήθεια μηχανικής μάθησης.

7.1 Επισκόπηση της ανάπτυξης μοντέλου με το Apache Spark

Στον τομέα της ανάπτυξης μοντέλων μηχανικής μάθησης σε κλίμακα, το Apache Spark αναδεικνύεται ως βασικό πλαίσιο, προσφέροντας κατανεμημένες υπολογιστικές δυνατότητες και ένα ολοκληρωμένο σύνολο εργαλείων για το χειρισμό εργασιών επεξεργασίας δεδομένων μεγάλης κλίμακας. Το Apache Spark διευκολύνει την ανάπτυξη μοντέλων μηχανικής μάθησης, εστιάζοντας ιδιαίτερα σε σενάρια πρόβλεψης σε πραγματικό χρόνο.

Η διαδικασία ανάπτυξης μοντέλων μηχανικής μάθησης σε κλίμακα παρουσιάζει μια σειρά από πολύπλευρες προκλήσεις, που κυμαίνονται από επεκτασιμότητα και βελτιστοποίηση απόδοσης έως απρόσκοπτη ενσωμάτωση με πηγές δεδομένων ροής,

όλα υπογραμμισμένα από την επιτακτική ανάγκη για ισχυρούς μηχανισμούς παρακολούθησης και διαχείρισης. Το Apache Spark, μέσω του κατανεμημένου υπολογιστικού του περιβάλλοντος, πλοηγείται επιδέξια σε αυτές τις προκλήσεις, προσφέροντας ανοχή σφαλμάτων και επεκτασιμότητα, ενώ χειρίζεται με χάρη μεγάλους όγκους δεδομένων.

Το Apache Spark χρησιμεύει ως ένας τρομερός σύμμαχος σε εργασίες πρόβλεψης σε πραγματικό χρόνο. Αξιοποιώντας τις εγγενείς δυνατότητες ροής του Spark, οι οργανισμοί μπορούν να απορροφούν απρόσκοπτα δεδομένα ροής από πηγές όπως ο Kafka και να εκτελούν προβλέψεις σε πραγματικό χρόνο χρησιμοποιώντας σχολαστικά εκπαιδευμένα μοντέλα μηχανικής μάθησης. Αυτή η ενοποίηση δίνει τη δυνατότητα στις επιχειρήσεις να λαμβάνουν καλά ενημερωμένες και έγκαιρες αποφάσεις, που βασίζονται σε γνώσεις που προέρχονται από τη δυναμική εισροή ροών δεδομένων.

Επιπλέον, το πλούσιο οικοσύστημα του Apache Spark εκτείνεται πέρα από την ανάπτυξη μοντέλων, προσφέροντας μια πληθώρα βιβλιοθηκών και εργαλείων για προεπεξεργασία δεδομένων, μηχανική χαρακτηριστικών, εκπαίδευση μοντέλων και αξιολόγηση. Η ενοποιημένη μηχανή ανάλυσης του απλοποιεί τον κύκλο ζωής ανάπτυξης των μοντέλων μηχανικής μάθησης, ενισχύοντας τη συνεργασία και την καινοτομία μεταξύ επιστημόνων δεδομένων, μηχανικών και ειδικών στον τομέα.

Ουσιαστικά, το Apache Spark ενσωματώνει τη μετατόπιση παραδείγματος στις μεθοδολογίες ανάπτυξης μοντέλων, δίνοντας τη δυνατότητα στους οργανισμούς να ξεκλειδώσουν το πλήρες δυναμικό των πόρων δεδομένων τους και να οδηγήσουν σε δραστικές ιδέες σε κλίμακα. Αγκαλιάζοντας το Apache Spark, οι επιχειρήσεις ξεκινούν ένα μετασχηματιστικό ταξίδι προς τη λήψη αποφάσεων με γνώμονα τα δεδομένα, την καινοτομία και το ανταγωνιστικό πλεονέκτημα στο σημερινό δυναμικό και συνεχώς εξελισσόμενο επιχειρηματικό τοπίο.

7.2 Ρύθμιση του περιβάλλοντος ανάπτυξης

Στην ανάπτυξη μοντέλων με το Apache Spark, η διαμόρφωση του περιβάλλοντος ανάπτυξης παίζει καθοριστικό ρόλο στη διασφάλιση της απρόσκοπτης εκτέλεσης των αγωγών μηχανικής μάθησης. Η περίπλοκη διαδικασία ρύθμισης του περιβάλλοντος ανάπτυξης, διευκρινίζει τα βασικά στοιχεία που είναι απαραίτητα για την ορθή λειτουργία της εφαρμογής.

- Διαμόρφωση του περιβάλλοντος Spark: Ο ακρογωνιαίος λίθος της ανάπτυξης μοντέλων μηχανικής εκμάθησης με το Apache Spark έγκειται στη διαμόρφωση του περιβάλλοντος Spark ώστε να ανταποκρίνεται στις συγκεκριμένες απαιτήσεις του σεναρίου ανάπτυξης. Το παρεχόμενο απόσπασμα κώδικα αποτελεί παράδειγμα της διαδικασίας διαμόρφωσης, ενσωματώνοντας βασικές παραμέτρους και εξαρτήσεις που είναι απαραίτητες για την ενορχήστρωση αγωγών ανάπτυξης μοντέλων.

```
spark = SparkSession.builder \
    .appName('CryptoPrediction') \
    .config("spark.jars.packages",
            "org.apache.spark:spark-sql-kafka-0-
10_2.12:3.2.0,org.mongodb.spark:mongo-spark-connector_2.12:10.2.0")
\
    .config("spark.mongodb.output.uri",
"mongodb://127.0.0.1/cryp.new_collection") \
    .config("spark.mongodb.output.bson.typeDetection", "true") \
    .config("spark.sql.streaming.checkpointLocation", "no-
checkpoint") \
    .getOrCreate()
```

Οι παράμετροι διαμόρφωσης περιλαμβάνουν διάφορες πτυχές, συμπεριλαμβανομένου του ονόματος της εφαρμογής, των εξωτερικών εξαρτήσεων, όπως οι υποδοχές Kafka και της MongoDB, και της θέσης σημείου ελέγχου για τη διασφάλιση της ανοχής σφαλμάτων στο Spark Structured Streaming.

- Εγκατάσταση απαραίτητων εξαρτήσεων: Εκτός από τη διαμόρφωση του Spark, η εγκατάσταση και η διαμόρφωση των απαραίτητων εξαρτήσεων αποτελεί μια κρίσιμη πτυχή της ρύθμισης του περιβάλλοντος ανάπτυξης. Η συμπερίληψη εξαρτήσεων όπως οι υποδοχές Kafka και MongoDB, καθοριστικής σημασίας για την απορρόφηση δεδομένων ροής και την αποθήκευση εξόδων πρόβλεψης, αντίστοιχα.

```
.config("spark.jars.packages",
        "org.apache.spark:spark-sql-kafka-0-
10_2.12:3.2.0,org.mongodb.spark:mongo-spark-connector_2.12:10.2.0")
```

Αυτή η οδηγία διαμόρφωσης διασφαλίζει ότι το Spark μπορεί να ενσωματωθεί απρόσκοπτα με το Kafka και το MongoDB, διευκολύνοντας την απορρόφηση και αποθήκευση των ροών δεδομένων και των αποτελεσμάτων πρόβλεψης, αντίστοιχα.

Συνοπτικά, η ρύθμιση του περιβάλλοντος ανάπτυξης θέτει τα θεμέλια για την εκτέλεση αγωγών μηχανικής εκμάθησης με το Apache Spark. Διαμορφώνοντας σχολαστικά το Spark και εγκαθιστώντας τις απαραίτητες εξαρτήσεις,

7.3 Ενσωμάτωση του Spark με πηγές δεδομένων ροής

Η ενσωμάτωση του Apache Spark με πηγές δεδομένων ροής αποτελεί ένα κρίσιμο στοιχείο της υποδομής ανάπτυξης του μοντέλου. Η διαδικασία ενσωμάτωσης του Spark με πηγές δεδομένων ροής, επιτελεί βασικό στοιχείο ζωτικής σημασίας για την αδιάκοπη απορρόφηση δεδομένων και την επεξεργασία σε πραγματικό χρόνο.

- Ενσωμάτωση με τον Kafka για την απορρόφηση δεδομένων ροής: Το Apache Kafka αποτελεί μια πανταχού παρούσα επιλογή για την απορρόφηση δεδομένων ροής, προσφέροντας υψηλή απόδοση, ανοχή σφαλμάτων και επεκτασιμότητα. Το παρεχόμενο απόσπασμα κώδικα δείχνει την απρόσκοπτη ενσωμάτωση του Apache Spark με τον Kafka, επιτρέποντας την απορρόφηση ροών δεδομένων ροής για προγνωστικά αναλυτικά στοιχεία σε πραγματικό χρόνο.

```
# Streaming data from Kafka
streaming_data = spark \
    .readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "localhost:9092") \
    .option("subscribe", "cryptoway") \
    .load()
```

Αυτό το απόσπασμα κώδικα διαμορφώνει το Spark ώστε να καταναλώνει δεδομένα ροής από το θέμα "cryptoway" Kafka, αξιοποιώντας τις δυνατότητες ενσωμάτωσης Kafka που παρέχονται από το Spark Structured Streaming.

- Χειρισμός δεδομένων ροής με Spark Structured Streaming: Το Spark Structured Streaming προσφέρει ένα API υψηλού επιπέδου για τη δημιουργία

επεκτάσιμων και ανεκτικών σε σφάλματα εφαρμογών ροής. Το απόσπασμα κώδικα δείχνει τη χρήση του Spark Structured Streaming για τη διαχείριση δεδομένων ροής, επιτρέποντας την επεξεργασία και ανάλυση των ροών δεδομένων σε πραγματικό χρόνο.

```
# Convert Kafka message value to JSON and select required fields
streaming_data =
streaming_data.select(from_json(col("value").cast("string"),
schema).alias("data")) \
    .select("data.*")
```

Τα δεδομένα ροής από το Kafka μετατρέπονται σε μορφή JSON και επιλέγονται σχετικά πεδία για περαιτέρω επεξεργασία. Αυτός ο μετασχηματισμός προετοιμάζει τα δεδομένα ροής για μετέπειτα ανάλυση και συμπέρασμα μοντέλου.

7.4 Στρατηγικές ανάπτυξης μοντέλου

Η κατανόηση και η εφαρμογή αποτελεσματικών στρατηγικών ανάπτυξης είναι ζωτικής σημασίας για τη διασφάλιση της επεκτασιμότητας, της αξιοπιστίας και της απόδοσης στην ανάπτυξη μοντέλων μηχανικής εκμάθησης με το Apache Spark. Κάθε στοιχείο δικαιολογεί μια λεπτομερή εξέταση, συμπληρωμένη με αποσπάσματα κώδικα, όπου ισχύει, για την παροχή ολοκληρωμένης κατανόησης των στρατηγικών ανάπτυξης μοντέλων.

7.4.1 Κατανόηση διαφορετικών στρατηγικών ανάπτυξης

Η ανάπτυξη μοντέλων μηχανικής εκμάθησης με το Apache Spark περιλαμβάνει την αξιολόγηση δύο βασικών στρατηγικών ανάπτυξης:

- Επεξεργασία παρτίδας: Η μαζική επεξεργασία περιλαμβάνει την επεξεργασία δεδομένων σε διακριτές παρτίδες, που συνήθως συλλέγονται σε μια περίοδο. Αυτή η στρατηγική είναι κατάλληλη για σενάρια όπου δεν απαιτείται επεξεργασία σε πραγματικό χρόνο, επιτρέποντας ολοκληρωμένη ανάλυση και εκπαίδευση μοντέλων σε μεγάλα σύνολα δεδομένων.

- Σωληνώσεις πρόβλεψης σε πραγματικό χρόνο: Οι αγωγοί πρόβλεψης σε πραγματικό χρόνο επιτρέπουν την επεξεργασία δεδομένων ροής σχεδόν σε

πραγματικό χρόνο. Αυτή η στρατηγική είναι ζωτικής σημασίας για σενάρια όπου οι άμεσες απαντήσεις στις εισερχόμενες ροές δεδομένων είναι κρίσιμες, όπως ο εντοπισμός απάτης ή η δυναμική τιμολόγηση.

7.4.2 Αντισταθμίσεις μεταξύ αγωγών επεξεργασίας παρτίδας και αγωγών πρόβλεψης σε πραγματικό χρόνο

Η επιλογή μεταξύ επεξεργασίας παρτίδας και αγωγών πρόβλεψης σε πραγματικό χρόνο εξαρτάται από διάφορους παράγοντες:

- Απαιτήσεις καθυστέρησης: Οι αγωγοί πρόβλεψης σε πραγματικό χρόνο προσφέρουν χαμηλή καθυστέρηση, επιτρέποντας άμεσες αποκρίσεις στις εισερχόμενες ροές δεδομένων. Αντίθετα, η επεξεργασία κατά παρτίδες μπορεί να εισάγει υψηλότερο λανθάνοντα χρόνο λόγω των δεδομένων επεξεργασίας σε διακριτές παρτίδες.

- Όγκος δεδομένων: Η μαζική επεξεργασία είναι κατάλληλη για μεγάλους όγκους δεδομένων που μπορούν να υποβληθούν σε επεξεργασία εκτός σύνδεσης. Οι αγωγοί πρόβλεψης σε πραγματικό χρόνο, από την άλλη πλευρά, είναι βελτιστοποιημένοι για την επεξεργασία συνεχών ροών δεδομένων σε πραγματικό χρόνο.

- Υπολογιστικοί πόροι: Οι αγωγοί πρόβλεψης σε πραγματικό χρόνο απαιτούν αποκλειστικούς υπολογιστικούς πόρους για την επεξεργασία ροών δεδομένων σε πραγματικό χρόνο. Η μαζική επεξεργασία μπορεί να αξιοποιήσει τους αδρανείς υπολογιστικούς πόρους για τη μαζική επεξεργασία δεδομένων.

7.4.3 Μόχλευση του Spark MLlib για Ανάπτυξη Μοντέλου

Το MLlib του Apache Spark παρέχει ένα πλούσιο σύνολο αλγορίθμων μηχανικής μάθησης και βοηθητικών προγραμμάτων για τη δημιουργία και την ανάπτυξη μοντέλων μηχανικής μάθησης σε κλίμακα. Τα βασικά συστατικά περιλαμβάνουν:

- Επιλογή αλγορίθμου: Επιλογή του κατάλληλου αλγόριθμου μηχανικής εκμάθησης με βάση τα χαρακτηριστικά των δεδομένων και την εργασία πρόβλεψης.

- Μηχανική Χαρακτηριστικών: Προεπεξεργασία και μετατροπή ακατέργαστων δεδομένων σε διανύσματα χαρακτηριστικών κατάλληλα για εκπαίδευση μοντέλων και εξαγωγή συμπερασμάτων.

- Κατασκευή σωλήνων: Κατασκευή ισχυρών αγωγών μηχανικής εκμάθησης που ενσωματώνουν την προεπεξεργασία δεδομένων, τη μηχανική χαρακτηριστικών, την εκπαίδευση μοντέλων και την αξιολόγηση.

```
# Linear regression model
lr = LinearRegression(featuresCol="features", labelCol="Close",
predictionCol="prediction")

# Create a pipeline
pipeline = Pipeline(stages=[feature_assembler, lr])

# Train the model on historical data
trained_model = pipeline.fit(historical_data)
```

Αυτό το απόσπασμα κώδικα αποτελεί παράδειγμα της χρήσης του Spark MLlib για την εκπαίδευση ενός μοντέλου γραμμικής παλινδρόμησης σε ιστορικά δεδομένα και την εφαρμογή του σε ροές δεδομένων ροής για προβλέψεις σε πραγματικό χρόνο.

7.5 Αγωγός Πρόβλεψης σε Πραγματικό Χρόνο και Ενοποίηση MongoDB

Αναδιπλώνοντας την λεπτομερή περιγραφή της διαδικασίας από άκρο σε άκρο σχεδιασμού και υλοποίησης ενός αγωγού πρόβλεψης σε πραγματικό χρόνο χρησιμοποιώντας το Apache Spark MLlib. Γίνεται αντιληπτή η μηχανική χαρακτηριστικών, στην εκπαίδευση μοντέλων και στην εφαρμογή μοντέλων μηχανικής εκμάθησης στη ροή δεδομένων. Επιπλέον, εξετάζοντας το κρίσιμο βήμα της μετατροπής του DenseVectors σε σειριοποιήσιμη μορφή και την ενσωμάτωση του MongoDB για αποτελεσματική αποθήκευση και ανάλυση προβλέψεων σε πραγματικό χρόνο. Αυτή η ολοκληρωμένη προσέγγιση διασφαλίζει την επεκτασιμότητα, την

αξιοπιστία και την αποτελεσματικότητα του συστήματος πρόβλεψης σε πραγματικό χρόνο σε δυναμικά περιβάλλοντα δεδομένων.

7.5.1 Μηχανική Χαρακτηριστικών - Συναρμολόγηση Χαρακτηριστικών

Η μηχανική χαρακτηριστικών είναι μια κρίσιμη πτυχή της ανάπτυξης μοντέλων μηχανικής μάθησης, ιδιαίτερα σε αγωγούς πρόβλεψης σε πραγματικό χρόνο. Αναλυτικότερα, διερευνάτε πώς συναρμολογούνται οι λειτουργίες χρησιμοποιώντας το VectorAssembler στο Apache Spark MLlib, επιτρέποντας την αποτελεσματική προετοιμασία δεδομένων για εκπαίδευση μοντέλων και εξαγωγή συμπερασμάτων.

Η μηχανική χαρακτηριστικών περιλαμβάνει την επιλογή, τον μετασχηματισμό και τον συνδυασμό χαρακτηριστικών ακατέργαστων δεδομένων για τη δημιουργία ουσιαστικών εισροών για μοντέλα μηχανικής εκμάθησης. Τα σωστά σχεδιασμένα χαρακτηριστικά μπορούν να επηρεάσουν σημαντικά την απόδοση του μοντέλου και την ακρίβεια πρόβλεψης.

```
# Feature engineering - assemble features
feature_assembler = VectorAssembler(
    inputCols=["High", "Low", "Volume", "OBV"],
    outputCol="features"
)
```

Στο παρεχόμενο απόσπασμα κώδικα, το VectorAssembler δημιουργείται για να συγκεντρώνει χαρακτηριστικά από μεμονωμένες στήλες στα δεδομένα ροής. Με τις παραμέτρους να ορίζονται ως εξής:

- inputCols: Καθορίζει τη λίστα των ονομάτων στηλών από την οποία συναρμολογούνται τα χαρακτηριστικά. Σε αυτήν την περίπτωση, τα χαρακτηριστικά συγκεντρώνονται από στήλες με τα ονόματα "High", "Low", "Volume" και "OBV".

- outputCol: Καθορίζει το όνομα της στήλης εξόδου όπου θα αποθηκευτεί το συναρμολογημένο διάνυσμα χαρακτηριστικών. Εδώ, η στήλη εξόδου ονομάζεται "χαρακτηριστικά".

Το VectorAssembler ενοποιεί πολλαπλές στήλες εισόδου σε μια ενιαία διανυσματική στήλη χαρακτηριστικών, η οποία είναι απαίτηση για πολλούς αλγόριθμους μηχανικής εκμάθησης στο Apache Spark. Συνδυάζοντας τα σχετικά χαρακτηριστικά σε ένα ενιαίο διάνυσμα, απλοποιεί τον χειρισμό δεδομένων και ενισχύει την ερμηνευσιμότητα του μοντέλου.

7.5.2 Μοντέλο Γραμμικής Παλινδρόμησης

Στον τομέα των αγωγών πρόβλεψης σε πραγματικό χρόνο, η επιλογή ενός κατάλληλου μοντέλου μηχανικής εκμάθησης είναι ζωτικής σημασίας για ακριβείς και αποτελεσματικές προβλέψεις.

Η γραμμική παλινδρόμηση είναι μια θεμελιώδης στατιστική μέθοδος που χρησιμοποιείται για τη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Στο πλαίσιο της πρόβλεψης σε πραγματικό χρόνο, η γραμμική παλινδρόμηση παρέχει μια απλή και ερμηνεύσιμη προσέγγιση για την πρόβλεψη αριθμητικών αποτελεσμάτων με βάση τα χαρακτηριστικά εισόδου.

```
# Linear regression model
lr = LinearRegression(featuresCol="features", labelCol="Close",
predictionCol="prediction")
```

Το απόσπασμα κώδικα δημιουργεί ένα μοντέλο γραμμικής παλινδρόμησης χρησιμοποιώντας το Apache Spark MLlib. Με τις παραμέτρους να ορίζονται ως εξής:

- featuresCol: Καθορίζει το όνομα της στήλης χαρακτηριστικών εισόδου που περιέχει τα συναρμολογημένα διανύσματα χαρακτηριστικών. Σε αυτήν την περίπτωση, το όνομα της στήλης είναι "features", το οποίο είναι η έξοδος του VectorAssembler.

- labelCol: Καθορίζει το όνομα της στήλης που περιέχει τη μεταβλητή στόχο ή τις ετικέτες που χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Εδώ, η μεταβλητή στόχος είναι "Κλείσιμο", που αντιπροσωπεύει την τιμή κλεισίματος ενός χρηματοοικονομικού μέσου.

- predictionCol: Καθορίζει το όνομα της στήλης όπου θα αποθηκευτούν οι προβλέψεις του μοντέλου. Οι προβλεπόμενες τιμές θα αποθηκευτούν στη στήλη με το όνομα " prediction".

Το μοντέλο γραμμικής παλινδρόμησης μαθαίνει τη σχέση μεταξύ των χαρακτηριστικών εισόδου και της μεταβλητής στόχου υπολογίζοντας τους συντελεστές μιας γραμμικής εξίσωσης. Παρέχει πληροφορίες για τον αντίκτυπο κάθε χαρακτηριστικού στο προβλεπόμενο αποτέλεσμα και επιτρέπει την απλή ερμηνεία της συμπεριφοράς του μοντέλου.

7.5.3 Δημιουργία αγωγού

Στον τομέα των αγωγών μηχανικής μάθησης, η οργάνωση και η σύνδεση διαφόρων σταδίων επεξεργασίας δεδομένων και εκπαίδευσης μοντέλων είναι απαραίτητη για την εξορθολογισμένη και αποτελεσματική ανάπτυξη μοντέλων.

Οι αγωγοί παρέχουν μια συστηματική προσέγγιση στην ανάπτυξη μοντέλων μηχανικής μάθησης ενσωματώνοντας την προεπεξεργασία δεδομένων, τη μηχανική χαρακτηριστικών, την εκπαίδευση μοντέλων και την αξιολόγηση σε μια ενιαία, συνεκτική ροή εργασίας. Με την οργάνωση αυτών των σταδίων σε έναν αγωγό, γίνεται ευκολότερο η διαχείριση, η αναπαραγωγή και η επανάληψη της διαδικασίας ανάπτυξης του μοντέλου.

```
# Create a pipeline
pipeline = Pipeline(stages=[feature_assembler, lr])
```

Το παρεχόμενο απόσπασμα κώδικα δημιουργεί μια διοχέτευση στο Apache Spark MLlib. Με τις παραμέτρους να ορίζονται ως εξής:

- Στάδια: Καθορίζει μια λίστα σταδίων που περιλαμβάνουν τον αγωγό. Σε αυτήν την περίπτωση, ο αγωγός περιλαμβάνει δύο στάδια: το feature_assembler, υπεύθυνο για τη συναρμολόγηση χαρακτηριστικών εισόδου, και το lr (μοντέλο γραμμικής παλινδρόμησης), υπεύθυνο για την πρόβλεψη της μεταβλητής στόχου.

Ο αγωγός ενορχηστρώνει τη ροή δεδομένων και λειτουργιών από το ένα στάδιο στο άλλο, διασφαλίζοντας συνέπεια και επαναληψιμότητα στην ανάπτυξη μοντέλων. Κάθε στάδιο στη διοχέτευση επεξεργάζεται τα δεδομένα διαδοχικά, με την έξοδο ενός σταδίου να χρησιμεύει ως είσοδος στο επόμενο στάδιο.

7.5.4 Εφαρμογή του εκπαιδευμένου μοντέλου στη ροή δεδομένων

Οι αγωγοί πρόβλεψης σε πραγματικό χρόνο απαιτούν την εφαρμογή εκπαιδευμένων μοντέλων μηχανικής εκμάθησης στα εισερχόμενα δεδομένα ροής για έγκαιρες και ακριβείς προβλέψεις.

Η εφαρμογή ενός εκπαιδευμένου μοντέλου μηχανικής μάθησης στη ροή δεδομένων δίνει τη δυνατότητα στους οργανισμούς να αντλούν χρήσιμες πληροφορίες και να λαμβάνουν τεκμηριωμένες αποφάσεις σε πραγματικό χρόνο. Με τη συνεχή επεξεργασία των εισερχόμενων ροών δεδομένων, το μοντέλο μπορεί να δημιουργήσει προβλέψεις άμεσα, διευκολύνοντας την ταχεία απόκριση σε μεταβαλλόμενες συνθήκες και γεγονότα.

```
# Apply the trained model to streaming data
streaming_predictions = trained_model.transform(streaming_data)
```

Το παρεχόμενο απόσπασμα κώδικα εφαρμόζει το εκπαιδευμένο μοντέλο μηχανικής εκμάθησης (trained_model) σε ροή δεδομένων (streaming_data) χρησιμοποιώντας το Apache Spark. Με την λειτουργικότητα να ορίζετε ως εξής:

- Transform: Η μέθοδος μετασχηματισμού εφαρμόζει το εκπαιδευμένο μοντέλο στα δεδομένα ροής, δημιουργώντας προβλέψεις για κάθε εισερχόμενο σημείο δεδομένων.

- Το DataFrame που προκύπτει (streaming_predictions) περιέχει τα αρχικά δεδομένα μαζί με τις προβλεπόμενες τιμές.

Η εφαρμογή του εκπαιδευμένου μοντέλου στη ροή δεδομένων επιτρέπει τη δημιουργία προβλέψεων σε πραγματικό χρόνο, οι οποίες μπορούν να χρησιμοποιηθούν για διάφορους σκοπούς, όπως ανίχνευση ανωμαλιών, ανάλυση τάσεων και λήψη αποφάσεων. Με τη συνεχή επεξεργασία δεδομένων ροής, το

μοντέλο προσαρμόζεται στις μεταβαλλόμενες συνθήκες και παρέχει ενημερωμένες πληροφορίες.

7.5.5 Μετατροπή πυκνών διανυσμάτων και εγγραφή προβλέψεων σε MongoDB

Η σειριοποίηση δεδομένων και η συμβατότητα με συστήματα αποθήκευσης είναι πρωταρχικής σημασίας. Σε αυτήν την ενότητα, εμβαθύνουμε στη διαδικασία μετατροπής των DenseVectors σε σειριοποιήσιμη μορφή και στη συνέχεια εγγραφής προβλέψεων στο MongoDB. Αυτό το βήμα είναι ζωτικής σημασίας για τη διασφάλιση της συμβατότητας και της αποτελεσματικότητας στις ροές εργασίας επεξεργασίας δεδομένων στο πλαίσιο των αγωγών πρόβλεψης σε πραγματικό χρόνο.

Τα πυκνά διανύσματα χρησιμοποιούνται συνήθως για την αναπαράσταση διανυσμάτων χαρακτηριστικών σε μοντέλα μηχανικής μάθησης. Ωστόσο, είναι απαραίτητο τα DenseVectors να μετατρέπονται σε μια μορφή που μπορεί εύκολα να σειριοποιηθεί και να αποθηκευτεί, διασφαλίζοντας συμβατότητα και αποτελεσματικότητα στις ροές εργασίας επεξεργασίας δεδομένων.

```
def convert_dense_vector(vector):
    if isinstance(vector, DenseVector):
        return vector.toArray().tolist()
    return vector
```

Το παρεχόμενο απόσπασμα κώδικα ορίζει μια συνάρτηση βοηθητικού προγράμματος, convert_dense_vector, υπεύθυνη για τη μετατροπή των DenseVectors σε σειριοποιήσιμη μορφή. Με την λειτουργικότητα να ορίζετε ως εξής:

- Vector.toArray().tolist(): Μετατρέπει το DenseVector σε λίστα Python, καθιστώντας το σειριοποιήσιμο και συμβατό με διάφορα συστήματα αποθήκευσης και μορφές σειριοποίησης.

Μετά τη μετατροπή του DenseVectors, οι προβλέψεις που δημιουργούνται από το μοντέλο μπορούν να γραφτούν στο MongoDB για αποθήκευση και περαιτέρω ανάλυση. Το ακόλουθο απόσπασμα κώδικα δείχνει αυτήν τη διαδικασία:

```
# Write to MongoDB using the modified DataFrame
streaming_predictions.writeStream \
    .foreachBatch(write_to_mongodb) \
    .start() \
    .awaitTermination()
```

Η μετατροπή των DenseVectors διασφαλίζει ότι τα διανύσματα χαρακτηριστικών που δημιουργούνται από μοντέλα μηχανικής εκμάθησης μπορούν να σειριοποιηθούν και να αποθηκευτούν αποτελεσματικά. Με την ενσωμάτωση του MongoDB στον αγωγό, οι οργανισμοί μπορούν να αποθηκεύουν και να αναλύουν απρόσκοπτα προβλέψεις σε πραγματικό χρόνο. Αυτή η διπλή διαδικασία ενισχύει τη συνολική αποτελεσματικότητα και τη συμβατότητα του συστήματος πρόβλεψης σε πραγματικό χρόνο.

ΚΕΦΑΛΑΙΟ 8

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΟΠΤΙΚΕΣ

8.1 Ανακεφαλαίωση βασικών ευρημάτων

Το αποκορύφωμα του ταξιδιού της διατριβής ξετυλίγεται μέσα από μια σχολαστική σύνθεση των βασικών ευρημάτων που προέρχονται από το σχεδιασμό και την εφαρμογή του κατανεμημένου συστήματος επεξεργασίας δεδομένων για κρυπτονομίσματα σε πραγματικό χρόνο. Η διατριβή ξεκίνησε ένα διερευνητικό ταξίδι που εμβαθύνει στην ανάλυση κρυπτονομισμάτων, των τεχνικών δεικτών, της επεξεργασίας μεγάλων δεδομένων, της μηχανικής μάθησης και της αρχιτεκτονικής κατανεμημένων συστημάτων.

Μέσα από εξέταση, κατέστη προφανές ότι η συγχώνευση τεχνολογιών αιχμής όπως το Apache Spark και το Apache Kafka έχει φέρει επανάσταση στο τοπίο της απορρόφησης, επεξεργασίας και ανάλυσης δεδομένων σε πραγματικό χρόνο. Η αξιοποίηση των δυνατοτήτων αυτών των κατανεμημένων συστημάτων όχι μόνο διευκόλυνε την απρόσκοπτη ενσωμάτωση με πηγές δεδομένων ροής, αλλά έχει επίσης προικίσει το σύστημα με απαράμιλλη επεκτασιμότητα και ανοχή σφαλμάτων.

Επιπλέον, η εφαρμογή μοντέλων μηχανικής εκμάθησης για την πρόβλεψη τιμών κρυπτονομισμάτων και την ανάλυση τάσεων έχει δώσει βαθιές γνώσεις σχετικά με τη δυναμική των αγορών κρυπτονομισμάτων. Αξιοποιώντας τη δύναμη των προηγμένων αλγορίθμων, το σύστημα έχει επιδείξει αξιοσημείωτη ικανότητα στην αντίληψη των μοτίβων, στον εντοπισμό τάσεων και στην πραγματοποίηση τεκμηριωμένων προβλέψεων σε σενάρια σε πραγματικό χρόνο.

Ενώ το σύστημα παρουσιάζει ελπιδοφόρες δυνατότητες στον εντοπισμό των τάσεων και στην πραγματοποίηση τεκμηριωμένων προβλέψεων σε σενάρια πραγματικού χρόνου, είναι επιτακτική ανάγκη να σημειωθεί ότι απαιτείται περαιτέρω παρακολούθηση του συστήματος για οποιοδήποτε τελικό συμπέρασμα. Η συνεχής

αξιολόγηση και τελειοποίηση είναι απαραίτητες για τη διασφάλιση της ευρωστίας και της αποτελεσματικότητας του συστήματος σε δυναμικές αγορές κρυπτονομισμάτων.

Η σύνθεση που παρουσιάζεται σε αυτήν την ενότητα χρησιμεύει ως απόδειξη της αποτελεσματικότητας και της εφευρετικότητας του κατανεμημένου συστήματος επεξεργασίας δεδομένων, διευκρινίζοντας τον κεντρικό ρόλο του στην ενδυνάμωση των ενδιαφερομένων με αξιόπιστες ιδέες και ενημερωμένες ικανότητες λήψης αποφάσεων. Μέσα από μια ολοκληρωμένη επισκόπηση του σχεδιασμού, της υλοποίησης και της απόδοσης του συστήματος, αυτή η ενότητα περικλείει την ουσία του ταξιδιού της διατριβής, ανοίγοντας το δρόμο για μετασχηματιστικές εξελίξεις στην ανάλυση κρυπτονομισμάτων και τα παραδείγματα επεξεργασίας δεδομένων σε πραγματικό χρόνο.

8.2 Αναγνώριση περιορισμών

Καθώς περιηγούμαστε στα επιτεύγματα του έργου μας, είναι επιτακτική ανάγκη να αναγνωρίσουμε της προκλήσεις που χαρακτηρίζουν και τους περιορισμούς που διαμόρφωσαν το έργο μας:

- Προκλήσεις ποιότητας δεδομένων: Ανάμεσα στις περιπλοκές της επεξεργασίας δεδομένων, αντιμετωπίσαμε μια επίμονη πρόκληση για τη διασφάλιση της άψογης ποιότητας των δεδομένων εισόδου μας. Η αξιοπιστία των προγνωστικών μας μοντέλων εξαρτάται σε μεγάλο βαθμό από την ακρίβεια και την πληρότητα των συνόλων δεδομένων που χρησιμοποιούμε. Ανεπάρκειες, όπως τιμές που λείπουν, ακραίες τιμές και αποκλίσεις δεδομένων δημιουργούν εμπόδια, απαιτώντας σχολαστικές στρατηγικές προεπεξεργασίας δεδομένων για την ενίσχυση της πιστότητας των αναλύσεών μας.

- Αστάθεια της αγοράς: Η σφαίρα των κρυπτονομισμάτων είναι συνώνυμη με την αστάθεια, ένα χαρακτηριστικό που υπογραμμίζει τόσο τη γοητεία του όσο και τις προκλήσεις του. Η ιδιότροπη φύση των αγορών κρυπτονομισμάτων, που χαρακτηρίζεται από ξαφνικές και δραματικές διακυμάνσεις των τιμών, αποτελεί ένα τρομερό εμπόδιο στις προγνωστικές προσπάθειές μας. Τα μοντέλα μας πρέπει να παλέψουν με την εγγενή αβεβαιότητα και αστάθεια, προσπαθώντας να περιηγηθούν στη ταραχώδη δυναμική της αγοράς για να παρέχουν ουσιαστικές πληροφορίες και προβλέψεις.

Μπροστά σε αυτούς τους περιορισμούς, το έργο μας παραμένει σταθερό στη δέσμευσή του για καινοτομία και ανθεκτικότητα. Κάθε πρόκληση που αντιμετωπίζουμε χρησιμεύει ως καταλύτης για ανάπτυξη και προσαρμογή, ωθώντας μας προς νέες λύσεις και εκλεπτυσμένες μεθοδολογίες. Αγκαλιάζοντας την πολυπλοκότητα που είναι εγγενής στην ανάλυση κρυπτονομισμάτων και την πρόβλεψη σε πραγματικό χρόνο, ανοίγουμε το δρόμο για μετασχηματιστικές εξελίξεις και διαρκή αντίκτυπο στον τομέα των οικονομικών αναλυτικών στοιχείων.

8.3 Πιθανοί τομείς για μελλοντική έρευνα

Καθώς ολοκληρώνουμε την εξερεύνηση μας, στρέφουμε το βλέμμα μας προς τον ορίζοντα των δυνατοτήτων, οραματιζόμαστε δρόμους για μελλοντική έρευνα που υπόσχονται να εμπλουτίσουν και να επεκταθούν στις βάσεις που θέτει η διατριβή μας:

- Προηγμένες πηγές δεδομένων: Εξερεύνηση και ενσωμάτωση ποικίλων πηγών δεδομένων, συμπεριλαμβανομένης της ανάλυσης συναισθημάτων από τα μέσα κοινωνικής δικτύωσης και τις ειδήσεις, για την ενίσχυση των προγνωστικών δυνατοτήτων του μοντέλου μας.

- Προηγμένες τεχνικές μηχανικής μάθησης: Διερεύνηση της εφαρμογής προηγμένων τεχνικών μηχανικής μάθησης, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και η βαθιά μάθηση, για τη βελτίωση της ακρίβειας πρόβλεψης και της προσαρμοστικότητας στη δυναμική της αγοράς.

- Επεξηγηματικά μοντέλα: Η ανάπτυξη μοντέλων που όχι μόνο προβλέπουν τις τιμές, αλλά παρέχουν επίσης διαφανή λογική πίσω από τις προβλέψεις, ενισχύοντας την κατανόηση και την εμπιστοσύνη των χρηστών.

- Εκτίμηση και διαχείριση κινδύνου: Διεύρυνση του πεδίου εφαρμογής του έργου ώστε να συμπεριλάβει την αξιολόγηση και τη διαχείριση κινδύνου στις συναλλαγές κρυπτονομισμάτων, επιτρέποντας στους χρήστες να λαμβάνουν πιο ενημερωμένες και υπολογισμένες αποφάσεις.

- Ενοποίηση αγοράς και υποστήριξη συναλλαγών: Ενσωμάτωση με ανταλλακτήρια κρυπτονομισμάτων για την παροχή υποστήριξης και εκτέλεσης συναλλαγών σε πραγματικό χρόνο, δημιουργώντας ένα ολοκληρωμένο οικοσύστημα συναλλαγών κρυπτονομισμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] "BANKING AND FINANCE - Cryptocurrencies," European Commission, Apr. 27, 2018. https://ec.europa.eu/newsroom/fisma/items/624021 (accessed Jan. 16, 2024).

[2] "Research Guides: Fintech: Financial Technology Research Guide: Cryptocurrency & Blockchain Technology," Library of Congress. https://guides.loc.gov/fintech/21st-century/cryptocurrency-blockchain (accessed Jan. 16, 2024).

[3] "Ενημέρωση για τη χρήση εικονικών νομισμάτων," Τράπεζα της Ελλάδος, 2014. https://www.bankofgreece.gr/enimerosi/grafeio-typoy/anazhthsh-enhmerwsewn/enhmerwseis?announcement=daf61fe9-53d2-49f0-883d-b175607e404c (accessed Jan. 16, 2024).

[4] "European countries join Blockchain Partnership | Shaping Europe's digital future," European Commission, Apr. 10, 2018. https://digital-strategy.ec.europa.eu/en/news/european-countries-join-blockchain-partnership (accessed Jan. 16, 2024).

[5] J. J. Murphy, Technical analysis of the financial markets : a comprehensive guide to trading methods and applications. New York: New York Institute Of Finance, 1999.

[6] M. J. Pring, Technical analysis explained : the successful investor's guide to spotting investment trends and turning points. New York: Mcgraw-Hill Education, 2014.

[7] M. C. Thomsett, A Technical Approach To Trend Analysis. FT Press, 2015.

[8] W. J. O'Neil, How to Make Money in Stocks: A Winning System in Good Times and Bad, Fourth Edition. McGraw Hill Professional, 2009.

[9] "Elon Musk, Twitter and an epic case of buyer's remorse," The Economist, May 19, 2022. https://www.economist.com/business/2022/05/19/elon-musk-twitter-and-an-epic-case-of-buyers-remorse (accessed Jan. 17, 2024).

[10] "Elon Musk is buying Twitter. Really. Probably," The Economist, Oct. 05, 2022. https://www.economist.com/business/2022/10/05/elon-musk-is-buying-twitter-really-probably (accessed Jan. 17, 2024).

[11] B. Mitchell, "Council Post: What Business Leaders Can Learn From Elon Musk's Marketing Strategies," Forbes, Oct. 25, 2021. https://www.forbes.com/sites/forbescommunicationscouncil/2021/10/25/what-business-leaders-can-learn-from-elon-musks-marketing-strategies/ (accessed Jan. 17, 2024).

[12] B. Constanty, “Council Post: The Elon Musk Effect: The Timeless Power Of Disruption And Brand Authority,” Forbes, Feb. 25, 2021. https://www.forbes.com/sites/forbescommunicationscouncil/2021/02/25/the-elon-musk-effect-the-timeless-power-of-disruption-and-brand-authority/ (accessed Jan. 17, 2024).

[13] M. Bobrowsky, “Elon Musk Says Overpaying for Twitter in Short Term,” The Wall Street Journal, Oct. 19, 2022. https://www.wsj.com/livecoverage/stock-market-news-today-2022-10-19/card/elon-musk-says-overpaying-for-twitter-in-short-term-xzuG92m6dHUd3vHRwe87 (accessed Jan. 17, 2024).

[14] M. Bobrowsky, “Dogecoin Gets Lift From Elon Musk’s Twitter Action,” The Wall Street Journal, Apr. 26, 2022. https://www.wsj.com/livecoverage/twitter-elon-musk-latest-news/card/dogecoin-gets-lift-from-elon-musk-s-twitter-action-2yckr4xl09gBMP5ERefU (accessed Jan. 17, 2024).

[15] J. Stepek, “So Entertaining It Hurts: Twitter Sale May Be Tech Turning Point,” Bloomberg, Oct. 28, 2022. https://www.bloomberg.com/news/newsletters/2022-10-28/elon-musk-s-twitter-twtr-sale-may-be-tech-stock-turning-point (accessed Jan. 17, 2024).

[16] T. L. O’Brien, “Introducing Crash Course: Elon Musk vs. the Twitterverse,” Bloomberg, Jan. 10, 2023. https://www.bloomberg.com/opinion/articles/2023-01-10/crash-course-elon-musk-vs-the-twitterverse (accessed Jan. 17, 2024).

[17] S.-G. Mankarious, M. Chacón, C. Duffy, and C. Thorbecke, “Here’s what Elon Musk has tweeted over the years ... about Twitter,” CNN, Apr. 29, 2022. https://edition.cnn.com/interactive/2022/04/business/elon-musk-tweets-twitter/index.html (accessed Jan. 17, 2024).

[18] H. Murphy and T. Bradshaw, “Twitter job cuts begin as Musk warns of ‘massive’ revenue drop,” Financial Times, Nov. 05, 2022. Accessed: Jan. 17, 2024. [Online]. Available: https://www.ft.com/content/b9a2a0ec-d3fe-422d-bc62-fa9ddfd3f06c

[19] B. Rockefeller, Technical analysis for dummies. Hoboken, N.J: Wiley, 2011.

[20] The Financial Edits, 100+ Technical Indicators for Intraday Trading. By Mocktime Publication, 2023.

[21] P. J. Kaufman, Trading Systems and Methods. Hoboken, N.J.: Wiley, 2013.

[22] X. Xie, Full View Integrated Technical Analysis : a Systematic Approach to Active Stock Market Investing. Hoboken, N.J., Chichester: Wiley ; John Wiley [distributor], 2011.

[23] E. Botes and D. Siepman, “The Vortex Indicator,” Technical Analysis of Stocks & Commodities, vol. 28, no. 1, pp. 20–30, Jan. 2010.

[24] J. Gregory-Williams and B. Williams, Trading Chaos : Maximize Profits with Proven Technical Techniques. Hoboken, N.J: J. Wiley, 2004.

[25] M. Larson, 12 Simple Technical Indicators That Really Work : [Course Book]. Columbia, Md.: Marketplace Books, 2007.

[26] The Financial Edits, 70+ Technical Indicators - Mastering Intraday Trading. By Mocktime Publication, 2023.

[27] K. Wood, Trade the Patterns. Traders Press, 2009.

[28] C. B. Schaap, ADXcellence. Gibbs Smith, 2006.

[29] G. Yioryalis, Technical Analysis for Beginners. PageFree Publishing, Inc., 2005.

[30] B. Dormeier, Investing with Volume Analysis. FT Press, 2011.

[31] Data Never Sleeps, https://www.domo.com/learn/ infographic/data-never-sleeps-11, (accessed Jan. 24, 2024)

[32] Data Never Sleeps, https://www.domo.com/data-never-sleeps, (accessed Dec. 10, 2023)

[33] Ishwarappa, Anuradha, J. (2015) A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology , International Conference on Intelligent Computing,Communication & Convegence (ICCC-2015)

[34] Apache Software Foundation, The Apache™ Hadoop® project, [online]: https://hadoop.apache.org/ (accessed Mar. 21, 2023)

[35] Sarkar, A., Ghosh, A., Dr. Nath A. (2015) MapReduce: A Comprehensive Study on Applications, Scope and Challenges, International Journal of Advance Research in Computer Science and Management Studies, July 2015, 3(7)

[36] Baesens B., (2014), “Analytics in a Big Data World - The Essential Guide to Data Science and Its Applications”, Wiley

[37] Xiaomeng S., (2013), “Introduction to Big Data”, learning material for course IINI3012 Big Data – Norwegian University of Science and Technology

[38] Joshi N., (2017), “Top 5 sources of big data”, https://www.allerin.com/blog/top-5-sources-of-big-data

[39] Πολυμένη Ι., (2017), “Τα δεδομένα μεγάλης κλίμακας: Τεχνικές και εργαλεία ανάλυσης τους, και η προσφορά τους ως υπηρεσία υπολογιστικού νέφους”, Μεταπτυχιακή Διπλωματική εργασία – πανεπιστημίου Μακεδονίας

[40] New Generation Big Data Tools and Techniques for business, https://i-verve.com/blog/big- data-tools-and-techniques-for-business (accessed Dec. 10, 2023)

[41] Hooja S., (2017), “Why Python is the Right Programming Language for Data Science”/,https//datafloq.com/read/why-python-programming-language-data-science/2426 (accessed Dec. 10, 2023)

[42] Theuwissen M., (2015), “R vs Python for Data Science”, https://www.kdnuggets.com/2015/05/r-vs-python-data-science.html (accessed Nov. 29, 2023)

[43] Simon P., (2013), “Too Big to Ignore: The Business Case for Big Data”, Wiley

[44] Dangeti P., (2017), “Statistics for Machine Learning: Build supervised, unsupervised, and reinforcement learning models using both Python and R”, Packt

[45] FPGA-Acceleration of Machine Learning in Cloud Computing, a case study using Logistic Regression. Ηλίας Ν. Κορομηλάς. χ.χ.

[46] Guide, Hadoop: The Definitive. White, Tom. χ.χ.

[47] Gilbert E., (2016), “Hadoop Overview: A Big Data Toolkit”, [online]: http://www.dataversity.net/comparative-roundup-artificial-intelligence-vs-machine-learning-vs-deep-learning-2 (accessed Mar. 21, 2023)

[48] Karau H., Konwinski A., Wendell P., Zaharia M., (2015), “Learning Spark: Lighting-fast Data Analysis”, O’reilly

[49] Algorithms, Introduction to. Thomash, Charlese, Ronaldl, Clifford stein.

[50] IBM, (n.d.), “What is distributed computing”, https://www.ibm.com/support/knowledgecenter/en/SSAL2T_8.2.0/com.ibm.cics.tx.doc/concepts/c_wht_is_distd_comptg.htm (accessed Mar. 21, 2023)

[51] MapReduce, A Survey of Large-Scale Analytical Query Processing in Christos Doulkeridis Kjetil Norvag.

[52] Environment, A Review: Resource Allocation Problem in Cloud. Prof. Rupali M.Pandharpatte.

[53] A developer's guide to using Kafka with Java, Part 1, https://developers.redhat.com/articles/2022/04/05/developers-guide-using-kafka-java-part-1#kafka_architecture (accessed Dec. 10, 2023)

[54] MongoDB, https://www.mongodb.com/features, (accessed Dec. 10, 2023)