Πρόγραμμα Μεταπτυχιακών Σπουδών

στη

Φορολογική Λογιστική και Χρηματοοικονομική

Διοίκηση Στρατηγικών Αποφάσεων

<u>Διπλωματική Εργασία</u>

Big Data Analysis και Επιχειρηματικότητα

του

Κωνσταντίνου Καζακλή - Ταϊλίδη
του Κωνσταντίνου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του Μεταπτυχιακού Διπλώματος στη Φορολογική Λογιστική και Χρηματοοικονομική Διοίκηση Στρατηγικών Αποφάσεων

Φεβρουάριος 2022

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία εκπονήθηκε για την ολοκλήρωση της φοίτησης στο τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων της σχολής Επιστημών Διοίκησης Επιχειρήσεων του Πανεπιστημίου Μακεδονίας. Η έναρξη της εργασίας έγινε χρονικά τον Ιούνιο του 2022

Το αντικείμενο της πτυχιακής αυτής εργασίας είναι η μελέτη των Μεγάλων Δεδομένων (Big Data). Ο αναγνώστης ύστερα από την κατανόηση τους σε εισαγωγικό επίπεδο, είναι σε θέση να μπορεί να κατανοήσει στο ακέραιο την χρησιμότητα τους και την εφαρμογή τους. Κρίθηκε σκόπιμο να κάνουμε μια επιμέρους ανάλυση στις τεχνολογίες ανάλυσης τους καθώς και στην εξόρυξη των δεδομένων. Στο τέλος της πτυχιακής αναφέρονται κάποιες εταιρείες-οργανισμοί που έχουν εφαρμόσει τα μεγάλα δεδομένα.

Abstract

The present thesis was prepared for the completion of the course of study at the department of Business Administration of school Business Administration Sciences of the University of Macedonia. The start of the work was timed in June 2022
The subject of this thesis is the study of Big Data. After understanding them at an introductory level, the reader is able to fully understand their usefulness and application. It was deemed appropriate to make a partial analysis on their analysis technologies as well as data mining. At the end of the thesis some of the companies-organizations that have implemented big data are mentioned.

Πίνακας περιεχομένων

ΚΕΦΑΛΑΙΟ 1 : ΕΙΣΑΓΩΓΗ ΣΤΑ BIG DATA

Σε αυτό το κεφάλαιο παρουσιάζονται τα μεγάλα δεδομένα (καθώς και με τον ορισμό τους για το τι ακριβώς αφορά). Γίνεται αναφορά στους περιορισμούς που υπάρχουν στις παραδοσιακές βάσεις δεδομένων, που εξαιτίας αυτών αναπτύχθηκαν τα μεγάλα δεδομένα (big data). Οι κύριες διαφορές τους εντοπίζονται στη ταχύτητα, στον όγκο και στην ποικιλία. Με την εξέλιξη των μεγάλων δεδομένων, δεν περιοριζόμαστε πλέον στα δομημένα δεδομένα. Εξηγούνται οι διαφορετικοί τύποι δεδομένων που παράγονται από τον άνθρωπο και τη μηχανή -δηλαδή δομημένα, ημιδομημένα και αδόμητα- που μπορούν να διαχειριστούν από τα μεγάλα δεδομένα. Δίνεται μια σαφής εικόνα των διαφόρων πηγών που συμβάλλουν σε αυτόν τον τεράστιο όγκο δεδομένων. Δείχνουμε τα διάφορα στάδια του κύκλου ζωής των μεγάλων δεδομένων, ξεκινώντας από την παραγωγή δεδομένων, την απόκτηση, την προεπεξεργασία, την ενσωμάτωση, τον καθαρισμό, τον μετασχηματισμό, την ανάλυση και την οπτικοποίηση για τη λήψη επιχειρηματικών αποφάσεων.

1.1 Κατανόηση των μεγάλων δεδομένων

Με τη ραγδαία αύξηση των χρηστών του Διαδικτύου, ο όγκος των παραγόμενων δεδομένων έχει αυξηθεί εκθετικά. Τα δεδομένα παράγονται από εκατομμύρια μηνύματα που αποστέλλονται και επικοινωνούνται μέσω του WhatsApp, του Facebook ή του Twitter, τρισεκατομμύρια φωτογραφίες που λαμβάνονται κάθε λεπτό και ώρες βίντεο που ανεβαίνουν στο YouTube. Σύμφωνα με μια πρόσφατη μελέτη 2,5 x 10^18 bytes δεδομένων παράγονται κάθε μέρα (Financesonline.com, 2022). Αυτός ο τεράστιος όγκος δεδομένων που παράγεται αναφέρεται ως "μεγάλα δεδομένα". Τα μεγάλα δεδομένα δεν σημαίνουν μόνο ότι τα σύνολα δεδομένων είναι πολύ μεγάλα, είναι ένας γενικός όρος για τα δεδομένα που είναι πολύ μεγάλα σε μέγεθος, πολύπλοκα στη φύση, τα οποία μπορεί να είναι δομημένα ή μη δομημένα και φθάνουν επίσης με υψηλή ταχύτητα. Από τα δεδομένα που είναι διαθέσιμα σήμερα, το 80% έχει παραχθεί τα τελευταία χρόνια. Η ανάπτυξη των μεγάλων δεδομένων τροφοδοτείται από το γεγονός ότι παράγονται περισσότερα δεδομένα σε κάθε γωνιά του κόσμου που πρέπει να καταγραφούν. Η καταγραφή αυτών των μαζικών δεδομένων δίνει μόνο πενιχρή αξία, εκτός εάν αυτή η αξία της πληροφορικής μετατραπεί σε επιχειρηματική αξία. Η διαχείριση των δεδομένων και η ανάλυσή τους ήταν πάντα επωφελής για τους οργανισμούς, ωστόσο η μετατροπή αυτών των δεδομένων σε πολύτιμες επιχειρηματικές γνώσεις αποτελούσε πάντα τη μεγαλύτερη πρόκληση. Οι επιστήμονες δεδομένων αγωνίζονταν να βρουν ρεαλιστικές τεχνικές για την ανάλυση των συλλεχθέντων δεδομένων. Τα δεδομένα πρέπει να διαχειρίζονται με την κατάλληλη ταχύτητα και χρόνο για να αντλήσουν πολύτιμες πληροφορίες από αυτά. Αυτά τα δεδομένα είναι τόσο πολύπλοκα που κατέστη δύσκολο να τα επεξεργαστούν με τη χρήση παραδοσιακών συστημάτων διαχείρισης βάσεων δεδομένων, γεγονός που προκάλεσε την εξέλιξη της εποχής των μεγάλων δεδομένων. Επιπλέον, υπήρχαν περιορισμοί στον όγκο των δεδομένων που

μπορούσαν να διαχειριστούν οι παραδοσιακές βάσεις δεδομένων. Με την αύξηση του μεγέθους των δεδομένων είτε υπήρχε μείωση της απόδοσης και αύξηση της καθυστέρησης είτε ήταν δαπανηρή η προσθήκη πρόσθετων μονάδων μνήμης. Όλοι αυτοί οι περιορισμοί έχουν ξεπεραστεί με την εξέλιξη των τεχνολογιών μεγάλων δεδομένων που μας επιτρέπουν να συλλαμβάνουμε, να αποθηκεύουμε, να επεξεργαζόμαστε και να αναλύουμε τα δεδομένα σε ένα κατανεμημένο περιβάλλον. Παραδείγματα τεχνολογιών μεγάλων δεδομένων είναι το Hadoop, ένα πλαίσιο για όλες τις διεργασίες μεγάλων δεδομένων, το κατανεμημένο σύστημα αρχείων Hadoop (HDFS) για την κατανεμημένη αποθήκευση σε συστάδες και το MapReduce για την επεξεργασία (SAS, n.d.).

1.2 Εξέλιξη των μεγάλων δεδομένων (Big Data)

Τα μεγάλα δεδομένα εμφανίστηκαν για πρώτη φορά σε ένα ντοκιμαντέρ σε ένα άρθρο ενός επιστήμονα της NASA το 1997, το οποίο περιέγραφε προβλήματα στην οπτικοποίηση μεγάλων συνόλων δεδομένων και παρουσίαζε ενδιαφέρουσες προκλήσεις για τους επιστήμονες δεδομένων. Τα σύνολα δεδομένων ήταν αρκετά μεγάλα, επιβαρύνοντας περισσότερο τους πόρους μνήμης. Το πρόβλημα αυτό ονομάστηκε big data (μεγάλα δεδομένα). Τα μεγάλα δεδομένα, η ευρύτερη έννοια, προτάθηκε για πρώτη φορά από μια γνωστή συμβουλευτική εταιρεία: McKinsey. Οι τρεις διαστάσεις των μεγάλων δεδομένων, δηλαδή ο όγκος, η ταχύτητα και η ποικιλία, ορίστηκαν από τον αναλυτή Doug Laney. Ο κύκλος ζωής της επεξεργασίας των μεγάλων δεδομένων μπορεί να κατηγοριοποιηθεί σε απόκτηση, προεπεξεργασία, αποθήκευση και διαχείριση, προστασία της ιδιωτικής ζωής και ασφάλεια, ανάλυση και οπτικοποίηση (Rob Kitchin, 2016).
Ο ευρύτερος όρος big data περιλαμβάνει οτιδήποτε περιλαμβάνει δεδομένα του διαδικτύου, όπως δεδομένα ροής κλικ, δεδομένα υγείας ασθενών, γονιδιωματικά δεδομένα από βιολογικές έρευνες κ.ο.κ.

1.2 Χειρισμός των μεγάλων δεδομένων από παραδοσιακές βάσης δεδομένων

Μέχρι πρόσφατα, τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS), ήταν το πιο ευρέως χρησιμοποιούμενο μέσο αποθήκευσης δεδομένων για την αποθήκευση δεδομένων που παράγονται από οργανισμούς. Αυτά τα RDBMS έχουν σχεδιαστεί για να αποθηκεύουν δεδομένα που υπερβαίνουν την αποθηκευτική ικανότητα ενός μεμονωμένου υπολογιστή. Η εισαγωγή νέων τεχνολογιών οφείλεται πάντα στους περιορισμούς των παλαιών τεχνολογιών και στην ανάγκη να ξεπεραστούν. Παρακάτω παρουσιάζονται οι περιορισμοί των παραδοσιακών βάσεων δεδομένων κατά την επεξεργασία μεγάλων δεδομένων.

- Η εκθετική αύξηση του όγκου των δεδομένων, που κλιμακώνεται σε terabytes και petabytes, αποδεικνύεται πρόκληση για τα RDBMS να διαχειριστούν τέτοιες τεράστιες ποσότητες δεδομένων.
- Για την επίλυση αυτού του προβλήματος, τα RDBMS αύξησαν τον αριθμό των επεξεργαστών και πρόσθεσαν μονάδες μνήμης, αυξάνοντας έτσι το κόστος.
- Σχεδόν το 80% των δεδομένων που ανακτήθηκαν ήταν σε ημιδομημένες και μη δομημένες μορφές που τα RDBMS δεν μπορούν να διαχειριστούν.
- Το RDBMS δεν ήταν σε θέση να συλλάβει γρήγορα εισερχόμενα δεδομένα.

(Shukla, 2014)

Διαφορές εξόρυξης δεδομένων και μεγάλων δεδομένων

Χαρακτηριστικά	Rdbms	Μεγάλα δεδομένα
Όγκος Δεδομένων	Gigabytes-Terabytes	Petabytes to Zettabytes
Οργάνωση	Συγκεντρωτική	Κατανεμημένη
Τύπος Δεδομένων	Δομημένα	Μη Δομημένα και Ημιδομημένα
Τύπος Υλικού	Μοντέλο Υψηλών Προδιαγραφών	Βασικό Υλικό
Ενημερώσεις	Ανάγνωση/Εγγραφή πολλές φορές	Εγγραφή μια φορά/ Ανάγνωση πολλές φορές
Schema	Στατικό	Δυναμικό

Πίνακας 1 Διαφορές στα χαρακτηριστικά των big data και Rdbms

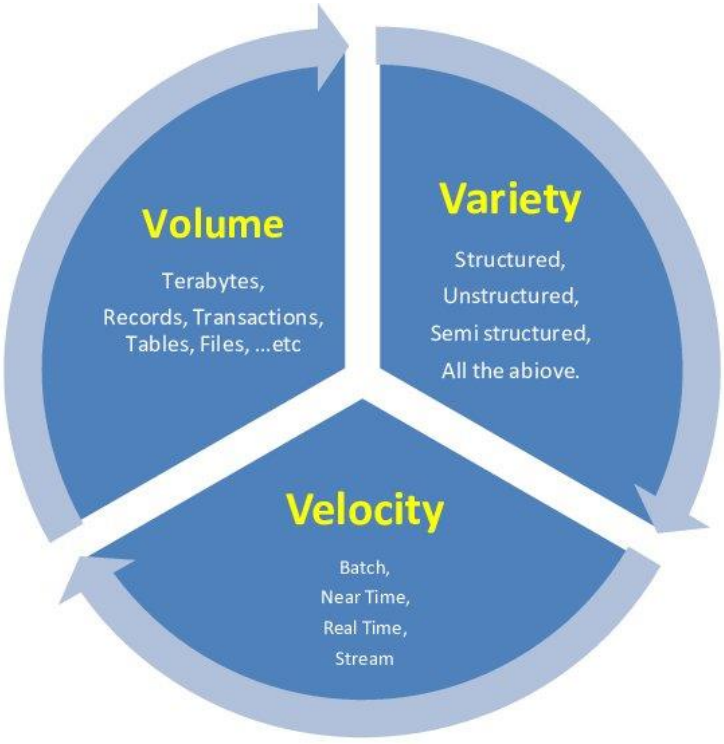
Εξόρυξη δεδομένων	Μεγάλα Δεδομένα
Η εξόρυξη δεδομένων είναι η διαδικασία ανακάλυψης της υποκείμενης γνώσης από τα σύνολα δεδομένων	Τα μεγάλα δεδομένα αναφέρονται σε τεράστιο όγκο δεδομένων που χαρακτηρίζονται από όγκο, ταχύτητα και ποικιλία
Δομημένα δεδομένα που ανακτώνται από λογιστικά φύλλα, σχεσιακές βάσεις δεδομένων κ.ά.	Δομημένα, αδόμητα ή ημιδομημένα δεδομένα που ανακτώνται από μη σχεσιακές βάσεις δεδομένων, όπως η NoSQl
Η εξόρυξη δεδομένων είναι ικανή να επεξεργάζεται μεγάλα σύνολα δεδομένων, αλλά το κόστος επεξεργασίας δεδομένων είναι υψηλό	Τα εργαλεία και οι τεχνολογίες μεγάλων δεδομένων είναι ικανά να αποθηκεύουν και να επεξεργάζονται μεγάλο όγκο δεδομένων με συγκριτικά χαμηλότερο κόστος
Η εξόρυξη δεδομένων μπορεί να επεξεργαστεί μόνο σύνολα δεδομένων που κυμαίνονται από gigabytes έως terabytes	Η τεχνολογία των μεγάλων δεδομένων είναι ικανή να αποθηκεύει και να επεξεργάζεται δεδομένα που κυμαίνονται από petabytes έως zettabytes.

Πίνακας 2 Διαφορές εξόρυξης δεδομένων και big data

(Data Mining vs Big Data , n.d.)

1.3 Τα 3Vs των Big Data

Το χαρακτηριστικό γνώρισμα των μεγάλων δεδομένων είναι τα εξαιρετικά χαρακτηριστικά τους με διάφορες διαστάσεις. Η πρώτη διάσταση είναι το μέγεθος των δεδομένων. Τα μεγέθη των δεδομένων αυξάνονται, εν μέρει επειδή η αποθήκευση σε συστάδες με χρήση υλικού βασικών προϊόντων έχει γίνει φθηνότερη. Το υλικό κοινής χρήσης είναι υλικό χαμηλών προδιαγραφών, χαμηλού κόστους, χαμηλών επιδόσεων, λειτουργικό και χωρίς ιδιαίτερα χαρακτηριστικά. Αυτό αναφέρεται με τον όρο "όγκος" στην τεχνολογία των μεγάλων δεδομένων. Η δεύτερη διάσταση είναι η ποικιλομορφία. Αυτή αντιπροσωπεύει την ετερογένεια της αποδοχής όλων των τύπων δεδομένων, είτε πρόκειται για δομημένα, είτε για μη δομημένα, είτε για ένα μείγμα και των δύο. Η τρίτη διάσταση είναι η ταχύτητα. Αναφέρεται στην ταχύτητα με την οποία παράγονται και επεξεργάζονται τα δεδομένα για την εξαγωγή των επιθυμητών τιμών από τα ακατέργαστα, μη επεξεργασμένα δεδομένα (Al-Barhamtoshy, 2014).



Εικόνα 1 Τα 3Vs των Big Data (Al-Barhamtoshy, 2014)

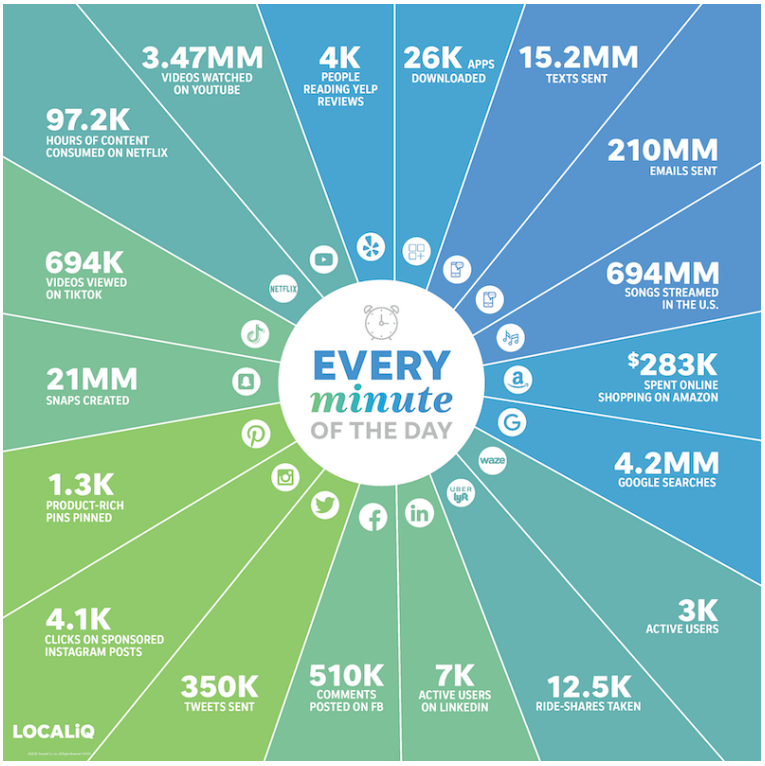
Όγκος (Volume)

Ο όγκος των δεδομένων που παράγονται και επεξεργάζονται από τα μεγάλα δεδομένα αυξάνεται συνεχώς και ταχύτερα από ποτέ. Ο όγκος αυξάνεται εκθετικά λόγω του γεγονότος ότι οι εταιρείες συλλέγουν συνεχώς δεδομένα για την ανάπτυξη καλύτερων και μεγαλύτερων επιχειρηματικών λύσεων. Τα μεγάλα δεδομένα μετρώνται από Terabyte έως Zettabyte (1024 GB = 1 Terabyte, 1024 TB = 1 Petabyte, 1024 PB = 1 Exabyte, 1024 EB = 1 Zettabyte, 1024 ZB = 1 Yottabyte). Η συλλογή αυτού του

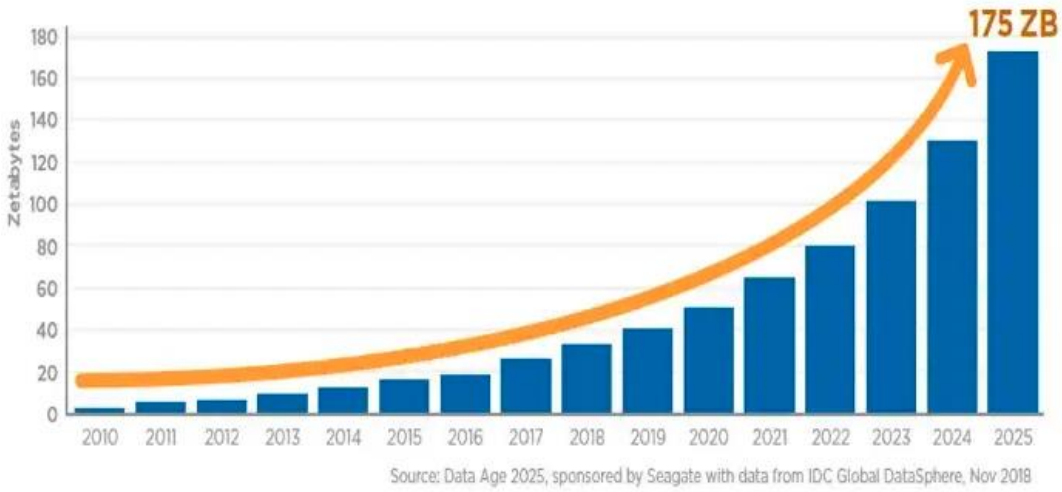
τεράστιου όγκου δεδομένων αναφέρεται ως μια μεγάλη ευκαιρία για την παροχή καλύτερης εξυπηρέτησης πελατών και επιχειρηματικού πλεονεκτήματος. Αυτός ο συνεχώς αυξανόμενος όγκος δεδομένων απαιτεί εξαιρετικά κλιμακούμενη και αξιόπιστη αποθήκευση. Τα μέσα κοινωνικής δικτύωσης, οι συναλλαγές στα σημεία πώλησης (POS), οι ηλεκτρονικές τραπεζικές συναλλαγές, οι αισθητήρες GPS και οι αισθητήρες οχημάτων είναι οι κύριες πηγές που συμβάλλουν σε αυτή την εκπληκτική αύξηση του όγκου δεδομένων. Το Facebook παράγει περίπου 500 terabytes δεδομένων κάθε μέρα. Κάθε φορά που γίνεται κλικ σε έναν σύνδεσμο ιστοτόπου, μια ηλεκτρονική αγορά, ένα βίντεο που μεταφορτώνεται στο YouTube, παράγονται δεδομένα (Longzhi Yang, 2019).

Ταχύτητα (Velocity)

Καθώς ο όγκος των δεδομένων έχει αυξηθεί δραματικά, έχει αυξηθεί και η ταχύτητα με την οποία παράγονται. Ο όρος "ταχύτητα" δεν αναφέρεται μόνο στην ταχύτητα με την οποία παράγονται τα δεδομένα, αλλά και στην ταχύτητα με την οποία επεξεργάζονται και αναλύονται. Στην εποχή των μεγάλων δεδομένων, παράγονται τεράστιες ποσότητες δεδομένων με μεγάλη ταχύτητα, και τα δεδομένα μπορεί να φτάνουν γρήγορα και να είναι δύσκολο να συλλεχθούν, αλλά να εξακολουθούν να χρειάζονται ανάλυση (Khalid Adam Ismail Hammad, 2015).



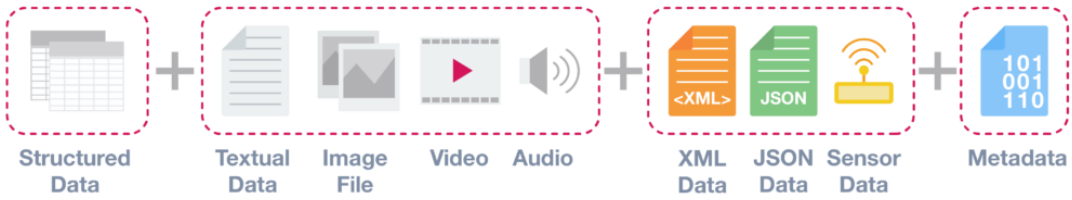
Εικόνα 2 Δεδομένα που παράγονται από γνωστές εταιρείες (έτος 2021) [πηγή : https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/ 01/12/2022]



Εικόνα 3 Πρόγνωση παραγωγής δεδομένων έως το έτος 2025 [πηγή : https://medium.com/analytics-vidhya/the-5-vs-of-big-data-2758bfcc51d 01/12/2022]

Ποικιλία (Variety)

Η ποικιλία αναφέρεται στη μορφή των δεδομένων που υποστηρίζονται από τα μεγάλα δεδομένα. Τα δεδομένα φθάνουν σε δομημένη, ημιδομημένη και αδόμητη μορφή. Τα δομημένα δεδομένα αναφέρονται στα δεδομένα που επεξεργάζονται από τα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων, όπου τα δεδομένα είναι οργανωμένα σε πίνακες, όπως τα στοιχεία των εργαζομένων και τα στοιχεία των τραπεζικών πελατών. Τα ημιδομημένα δεδομένα είναι ένας συνδυασμός δομημένων και μη δομημένων δεδομένων, όπως τα XML. Τα δεδομένα XML είναι ημιδομημένα, καθώς δεν ταιριάζουν στο επίσημο μοντέλο δεδομένων (πίνακας) που σχετίζεται με τις παραδοσιακές βάσεις δεδομένων- αντίθετα, περιέχουν ετικέτες για την οργάνωση των πεδίων εντός των δεδομένων. Τα αδόμητα δεδομένα αναφέρονται σε δεδομένα χωρίς συγκεκριμένη δομή, όπως μηνύματα ηλεκτρονικού ταχυδρομείου, φωτογραφίες και ιστοσελίδες. Τα δεδομένα που φτάνουν από το Facebook, τις ροές Twitter, τους αισθητήρες των οχημάτων και τα μαύρα κουτιά των αεροπλάνων είναι όλα μη δομημένα, τα οποία η παραδοσιακή βάση δεδομένων δεν μπορεί να επεξεργαστεί, και εδώ είναι που μπαίνουν στο προσκήνιο τα μεγάλα δεδομένα (Motashim Rasool, 2015).



Εικόνα 4 Ποικιλία δεδομένων στα big data [πηγή https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/ 02/12/2022]

1.4 Πηγές των Big Data

Μέσα μαζικής ενημέρωσης

Τα μέσα μαζικής ενημέρωσης είναι η πιο κοινή πηγή μεγάλων δεδομένων, καθώς παρέχουν πολύτιμες πληροφορίες για τις μεταβαλλόμενες προτιμήσεις και τάσεις των καταναλωτών. Επειδή αυτοεκπέμπονται και καταρρίπτουν όλα τα φυσικά και δημογραφικά εμπόδια, είναι ο ταχύτερος τρόπος για τις επιχειρήσεις να αποκτήσουν μια ολοκληρωμένη εικόνα του κοινού τους, να εξαγάγουν μοτίβα και συμπεράσματα και να βελτιώσουν τη λήψη αποφάσεων.
Περιλαμβάνονται τα μέσα κοινωνικής δικτύωσης και οι διαδραστικές πλατφόρμες, όπως το Google, το Facebook, το Twitter, το YouTube και το Instagram, καθώς και γενικά μέσα που έχουν εικόνες, βίντεο, ήχο και podcasts, παρέχοντας ποσοτική και ποιοτική εικόνα όλων των πτυχών της αλληλεπίδρασης των χρηστών.

Cloud (νέφος)

Επίσης σήμερα οι επιχειρήσεις προχωρούν μπροστά από τις παραδοσιακές πηγές δεδομένων, μεταφέροντας τα δεδομένα τους στο cloud. Η αποθήκευση αυτή απορροφά δομημένα και μη δομημένα δεδομένα και παρέχει στους οργανισμούς πληροφορίες σε πραγματικό χρόνο και πληροφορίες κατά παραγγελία. Ένα βασικό χαρακτηριστικό του υπολογιστικού νέφους είναι η ευελιξία και η επεκτασιμότητά του. Τα μεγάλα δεδομένα μπορούν να αποθηκευτούν σε δημόσια ή ιδιωτικά νέφη και να αποκτήσουν πρόσβαση μέσω δικτύων και διακομιστών, καθιστώντας το νέφος μια αποτελεσματική και οικονομική πηγή δεδομένων.

Διαδίκτυο (Web)

Ο δημόσιος ιστός αντιπροσωπεύει ευρέως κατανεμημένα και εύκολα προσβάσιμα μεγάλα δεδομένα. Τα δεδομένα στον ιστό ή στο "διαδίκτυο" είναι γενικά διαθέσιμα τόσο σε ιδιώτες όσο και σε επιχειρήσεις. Επιπλέον, υπηρεσίες ιστού όπως η Wikipedia παρέχουν δωρεάν και γρήγορες πληροφορίες για όλους. Το τεράστιο μέγεθος του ιστού εξασφαλίζει ευέλικτη χρηστικότητα και είναι ιδιαίτερα επωφελές για μια νεοσύστατη επιχείρηση ή τη μικρομεσαία επιχείρησή του, καθώς δεν χρειάζεται να περιμένουν να αναπτύξουν τη δική τους υποδομή και αποθετήριο μεγάλων δεδομένων πριν χρησιμοποιήσουν τα μεγάλα δεδομένα.

Internet of Things

Το περιεχόμενο που παράγεται από μηχανές και τα δεδομένα που παράγονται από το IoT αποτελούν πολύτιμες πηγές μεγάλων δεδομένων. Τα δεδομένα αυτά παράγονται συνήθως από αισθητήρες που συνδέονται με ηλεκτρονικές συσκευές. Η ικανότητα απόκτησης εξαρτάται από την ικανότητα του αισθητήρα να παρέχει ακριβείς πληροφορίες σε πραγματικό χρόνο. Το IoT κερδίζει σήμερα έδαφος και περιλαμβάνει μεγάλα δεδομένα που παράγονται από κάθε συσκευή που μπορεί να μεταδίδει δεδομένα, όχι μόνο από υπολογιστές και smartphones. Το IoT έχει καταστήσει δυνατή την απόκτηση δεδομένων από ιατρικές συσκευές, διαδικασίες οχημάτων, βιντεοπαιχνίδια, εξοπλισμό μετρήσεων, κάμερες, ηλεκτρονικά είδη ευρείας κατανάλωσης και πολλά άλλα.

Βάσεις Δεδομένων

Οι σημερινές επιχειρήσεις προτιμούν τη συγχώνευση παραδοσιακών και σύγχρονων βάσεων δεδομένων για την απόκτηση σχετικών μεγάλων δεδομένων. Αυτή η ενοποίηση ανοίγει το δρόμο για ένα υβριδικό μοντέλο δεδομένων, διατηρώντας το κόστος κεφαλαίου και υποδομής IT σε χαμηλά επίπεδα. Επιπλέον, αυτές οι βάσεις δεδομένων χρησιμοποιούνται επίσης για διάφορους σκοπούς επιχειρηματικής ευφυΐας. Αυτές οι βάσεις δεδομένων μπορούν να χειριστούν την εξαγωγή πληροφοριών που χρησιμοποιούνται για την αύξηση των επιχειρηματικών κερδών. Οι κοινές βάσεις δεδομένων περιλαμβάνουν διάφορες πηγές δεδομένων, όπως η MS Access, η DB2, η Oracle, η SQL και η Amazon Simple. Η εξαγωγή και η ανάλυση δεδομένων από μεγάλες πηγές δεδομένων μεγάλης κλίμακας είναι μια πολύπλοκη διαδικασία που μπορεί να είναι αγχωτική και χρονοβόρα. Αυτά τα πολύπλοκα προβλήματα μπορούν να επιλυθούν εάν ένας οργανισμός κάνει όλες τις απαραίτητες εκτιμήσεις για τα μεγάλα δεδομένα, εξετάζει τις σχετικές πηγές δεδομένων και τα χρησιμοποιεί με τρόπο που ευθυγραμμίζεται με τους επιχειρηματικούς στόχους.

(Naveen, 2017)

Κλάδος Υγείας

Τα μεγάλα δεδομένα στην υγειονομική περίθαλψη προέρχονται από ηλεκτρονικά αρχεία υγείας μεγάλης κλίμακας. Τα αρχεία αυτά είναι πολύ δύσκολο να διαχειριστούν με το παραδοσιακό υλικό και λογισμικό. Οι παραδοσιακές μέθοδοι και τα εργαλεία διαχείρισης δεδομένων εμποδίζουν επίσης την ορθή αξιοποίηση όλων αυτών των δεδομένων. Τα μεγάλα δεδομένα στην υγειονομική περίθαλψη είναι μια συντριπτική έννοια, όχι μόνο λόγω του όγκου των δεδομένων, αλλά και λόγω της ταχύτητας με την οποία πρέπει να γίνεται η διαχείριση των διαφόρων τύπων δεδομένων και η διαχείριση των ιατρικών δεδομένων. Το άθροισμα των δεδομένων που περιβάλλουν τους ασθενείς και την ευημερία τους αποτελεί το πρόβλημα των "μεγάλων δεδομένων" στην υγειονομική περίθαλψη. Η επιστήμη συμβάλλει επίσης στην ανάπτυξη τεχνολογιών ανάλυσης μεγάλων δεδομένων θέτοντας νέες προκλήσεις που σχετίζονται με την αναπαράσταση δεδομένων, το σχεδιασμό βάσεων δεδομένων, την αναζήτηση δεδομένων και την υποστήριξη κλινικών αποφάσεων.

Τα περισσότερα ιατρικά δεδομένα αποθηκεύονται σε έντυπη μορφή, αλλά η τρέχουσα τάση είναι η ταχεία ψηφιοποίηση αυτών των δεδομένων. Τα μεγάλα δεδομένα στην ιατρική βιομηχανία υπόσχονται να υποστηρίξουν διάφορες λειτουργίες διαχείρισης ιατρικών δεδομένων, όπως η Φροντίδα υγείας του πληθυσμού, υποστήριξη κλινικών αποφάσεων και επιτήρηση ασθενειών. Ο κλάδος της υγειονομικής περίθαλψης βρίσκεται ακόμη στα αρχικά στάδια της ενσωμάτωσης και ανάλυσης των μεγάλων δεδομένων σε μεγάλη κλίμακα.

Με το 80% των δεδομένων υγειονομικής περίθαλψης να είναι αδόμητα, η κατανόηση όλων αυτών των δεδομένων και η αποτελεσματική χρήση τους για την κλινική πρακτική, την ιατρική έρευνα και τις θεραπευτικές αγωγές αποτελεί πρόκληση για τον κλάδο της υγειονομικής περίθαλψης.

Ο όγκος των μεγάλων δεδομένων στην υγειονομική περίθαλψη αναμένεται να αυξηθεί τα επόμενα χρόνια και ο κλάδος της υγειονομικής περίθαλψης αναμένεται να αναπτυχθεί με τις αλλαγές στα μοντέλα αποζημίωσης της υγειονομικής περίθαλψης, θέτοντας σημαντικές προκλήσεις για την υγειονομική περίθαλψη. Παρόλο που το κέρδος δεν είναι το μοναδικό κίνητρο, είναι πολύ σημαντικό για τις εταιρείες

μεγάλων δεδομένων στον τομέα της υγειονομικής περίθαλψης να υιοθετήσουν τις καλύτερες τεχνολογίες και εργαλεία που μπορούν να χρησιμοποιήσουν αποτελεσματικά τα μεγάλα δεδομένα στον τομέα της υγειονομικής περίθαλψης (Project Pro, 2022).

1.5 Τύποι Δεδομένων

Δομημένα δεδομένα (structured data)

Τα δεδομένα αυτά είναι ιδιαίτερα οργανωμένα και εύκολα αποκρυπτογραφούνται από αλγορίθμους μηχανικής μάθησης. Η SQL (Structured Query Language), που αναπτύχθηκε από την IBM το 1974, είναι μια γλώσσα προγραμματισμού για τη διαχείριση δομημένων δεδομένων. Οι χρήστες μπορούν να εισάγουν, να αναζητούν και να επεξεργάζονται γρήγορα δομημένα δεδομένα χρησιμοποιώντας σχεσιακές βάσεις δεδομένων. Ημερομηνίες, ονόματα, διευθύνσεις, αριθμοί πιστωτικών καρτών αποτελούν παραδείγματα τέτοιων δεδομένων (δομημένων).

Τα πλεονεκτήματα τους είναι :

- Ευκολία χρήσης με αλγορίθμους μηχανικής μάθησης (ML[1]): Μια συγκεκριμένη και οργανωμένη αρχιτεκτονική δομημένων δεδομένων διευκολύνει τον χειρισμό και την αναζήτηση δεδομένων μηχανικής μάθησης
- Ευκολία για τους επιχειρηματικούς χρήστες: Τα δομημένα δεδομένα δεν απαιτούν βαθιά κατανόηση των διαφόρων τύπων δεδομένων και του τρόπου λειτουργίας τους. Τα δεδομένα μπορούν εύκολα να προσπελαστούν και να ερμηνευτούν από χρήστες με βασική κατανόηση του αντικειμένου με το οποίο σχετίζονται τα δεδομένα
- Προσβάσιμα από περισσότερα εργαλεία: Καθώς τα δομημένα δεδομένα προηγούνται των μη δομημένων δεδομένων, υπάρχουν περισσότερα εργαλεία για τη χρήση και την ανάλυση δομημένων δεδομένων

Τα μειονεκτήματα τους είναι :

- Περιορισμοί χρήσης : Τα δεδομένα με προκαθορισμένη δομή μπορούν να χρησιμοποιηθούν μόνο για τον προβλεπόμενο σκοπό τους, περιορίζοντας την ευελιξία και τη χρηστικότητα
- Περιορισμένες επιλογές αποθήκευσης: Τα δομημένα δεδομένα αποθηκεύονται συνήθως σε συστήματα αποθήκευσης δεδομένων με άκαμπτα σχήματα (όπως οι "αποθήκες δεδομένων[2]"). Ως εκ τούτου, η αλλαγή των απαιτήσεων

[1] Η μηχανική μάθηση (ML) είναι μια μορφή τεχνητής νοημοσύνης (AI) που επιτρέπει στις εφαρμογές λογισμικού να προβλέπουν τα αποτελέσματα με μεγαλύτερη ακρίβεια. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν ιστορικά δεδομένα ως είσοδο για να προβλέψουν νέες τιμές εξόδου.
Οι μηχανές συστάσεων είναι μια κοινή περίπτωση χρήσης της μηχανικής μάθησης. Άλλες συνήθεις χρήσεις περιλαμβάνουν την ανίχνευση απάτης, το φιλτράρισμα ανεπιθύμητης αλληλογραφίας, την ανίχνευση απειλών κακόβουλου λογισμικού, την αυτοματοποίηση επιχειρηματικών διαδικασιών (BPA) και την προγνωστική συντήρηση (Burns, n.d.).

[2] Μια αποθήκη δεδομένων είναι η συγκέντρωση δεδομένων από πολλές πηγές σε ένα ενιαίο, κεντρικό αποθετήριο που ενοποιεί τις ιδιότητες και τη μορφή των δεδομένων, καθιστώντας τα χρήσιμα για τους επιστήμονες δεδομένων που τα χρησιμοποιούν στην εξόρυξη δεδομένων, την τεχνητή νοημοσύνη (AI),

δεδομένων απαιτεί την ενημέρωση όλων των δομημένων δεδομένων, γεγονός που απαιτεί χρόνο και πόρους.

Τα εργαλεία δομημένων δεδομένων είναι :

- OLAP : Εκτελεί πολυδιάστατη ανάλυση δεδομένων υψηλής ταχύτητας από ενοποιημένες, κεντρικές αποθήκες δεδομένων
- SQLite : Υλοποιεί μια αυτοδύναμη, χωρίς διακομιστή, χωρίς ρυθμίσεις, συναλλακτική μηχανή σχεσιακής βάσης δεδομένων
- MySQL : Ενσωματώνει δεδομένα σε λογισμικό μαζικής ανάπτυξης, ιδίως σε κρίσιμα συστήματα παραγωγής με μεγάλο φορτίο
- PostgreSQL: Υποστηρίζει ερωτήματα SQL[3] και JSON[4], καθώς και γλώσσες προγραμματισμού υψηλού επιπέδου (C/C+, Java, Python κ.ά)

Περιπτώσεις χρήσης για δομημένα δεδομένα :

- Διαχείριση πελατειακών σχέσεων (CRM): Το λογισμικό CRM τρέχει δομημένα δεδομένα μέσω αναλυτικών εργαλείων για τη δημιουργία συνόλων δεδομένων που αποκαλύπτουν πρότυπα και τάσεις συμπεριφοράς πελατών
- Διαδικτυακές κρατήσεις: Τα δεδομένα κρατήσεων ξενοδοχείων και εισιτηρίων (π.χ. ημερομηνίες, τιμές, προορισμοί κ.ά.) ταιριάζουν στη μορφή "γραμμών και στηλών" που είναι ενδεικτική του προκαθορισμένου μοντέλου δεδομένων
- Λογιστική: Τα λογιστικά γραφεία ή τμήματα χρησιμοποιούν δομημένα δεδομένα για την επεξεργασία και την καταγραφή των οικονομικών συναλλαγών

(IBM Cloud Education, 2021)

Μη Δομημένα Δεδομένα

Τα μη δομημένα δεδομένα, τα οποία συνήθως ταξινομούνται ως ποιοτικά δεδομένα, δεν μπορούν να επεξεργαστούν και να αναλυθούν με τη χρήση παραδοσιακών εργαλείων και μεθόδων δεδομένων. Τα δεδομένα αυτά δεν έχουν προκαθορισμένο μοντέλο δεδομένων και είναι καλύτερο να διαχειρίζονται σε μια μη σχεσιακή βάση

τη μηχανική μάθηση και, τελικά, την επιχειρηματική ανάλυση και την επιχειρηματική ευφυΐα (Sinha, 2021).

[3] Η SQL είναι Δομημένη Γλώσσα Ερωτήσεων, η οποία είναι μια γλώσσα υπολογιστών για την αποθήκευση, τον χειρισμό και την ανάκτηση δεδομένων που είναι αποθηκευμένα σε μια σχεσιακή βάση δεδομένων. Είναι η τυποποιημένη γλώσσα για το σχεσιακό σύστημα βάσεων δεδομένων. Όλα τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (RDMS), όπως η MySQL, η MS Access, η Oracle, η Sybase, η Informix, η Postgres και ο SQL Server, χρησιμοποιούν την SQL ως τυποποιημένη γλώσσα βάσεων δεδομένων (www.tutorialspoint.com, n.d.).

[4] Η JSON (JavaScript Object Notation) είναι μια ελαφριά μορφή ανταλλαγής δεδομένων. Είναι εύκολο για τον άνθρωπο να το διαβάσει και να το γράψει και είναι εύκολο για τις μηχανές να το αναλύσουν και να το παράγουν. Βασίζεται σε ένα υποσύνολο του προτύπου γλώσσας προγραμματισμού JavaScript ECMA-262 και αποτελεί μια μορφή κειμένου που είναι εντελώς ανεξάρτητη από τη γλώσσα, αλλά χρησιμοποιεί συμβάσεις που είναι οικείες στους προγραμματιστές της οικογένειας γλωσσών C, συμπεριλαμβανομένων των C, C++, C#, Java, JavaScript, Perl, Python και πολλών άλλων. Αυτές οι ιδιότητες καθιστούν τη JSON μια ιδανική γλώσσα ανταλλαγής δεδομένων (www.json.org, n.d.)

δεδομένων (NoSQL[5]). Ένας άλλος τρόπος διαχείρισης μη δομημένων δεδομένων είναι η χρήση μιας λίμνης δεδομένων[6] για τη φύλαξη των ακατέργαστων δεδομένων. Η σημασία των μη δομημένων δεδομένων αυξάνεται ραγδαία. Σύμφωνα με πρόσφατες προβλέψεις, τα μη δομημένα δεδομένα αποτελούν πάνω από το 80% όλων των επιχειρηματικών δεδομένων και το 95% των οργανισμών θέτουν τη διαχείριση των μη δομημένων δεδομένων ως προτεραιότητα (Dialani, 2020).
Παραδείγματα μη δομημένων δεδομένων είναι τα κείμενα, η δραστηριότητα κινητών τηλεφώνων, οι αναρτήσεις στα μέσα κοινωνικής δικτύωσης, το email κ.ά. Τα πλεονεκτήματά τους αφορούν τη μορφή, την ταχύτητα και την αποθήκευση, ενώ οι υποχρεώσεις τους περιστρέφονται γύρω από την τεχνογνωσία και τους διαθέσιμους πόρους:

Τα πλεονεκτήματα τους είναι :

- Μητρική μορφή: Τα μη δομημένα δεδομένα, αποθηκευμένα στη μητρική τους μορφή, παραμένουν απροσδιόριστα μέχρι να χρειαστούν. Η προσαρμοστικότητά της αυξάνει τις μορφές αρχείων στη βάση δεδομένων, γεγονός που διευρύνει τη δεξαμενή δεδομένων και επιτρέπει στους επιστήμονες δεδομένων να προετοιμάζουν και να αναλύουν μόνο τα δεδομένα που χρειάζονται
- Γρήγοροι ρυθμοί συσσώρευσης: Δεδομένου ότι δεν υπάρχει ανάγκη προκαθορισμού των δεδομένων, τα δεδομένα μπορούν να συγκεντρωθούν γρήγορα και εύκολα.
- Αποθήκευση στη λίμνη δεδομένων: Επιτρέπει μαζική αποθήκευση και τιμολόγηση κατά χρήση, γεγονός που μειώνει το κόστος και διευκολύνει την επεκτασιμότητα

Τα μειονεκτήματα τους είναι :

- Απαιτεί τεχνογνωσία : Λόγω της απροσδιόριστης φύσης τους, η προετοιμασία και η ανάλυση μη δομημένων δεδομένων απαιτεί τεχνογνωσία στην επιστήμη των δεδομένων. Αν και αυτό είναι χρήσιμο για τους αναλυτές δεδομένων, αποξενώνει τους εξειδικευμένους επιχειρηματικούς χρήστες, οι οποίοι ενδέχεται να μην κατανοούν πλήρως τα εξειδικευμένα θέματα δεδομένων ή τον τρόπο χρήσης των δεδομένων
- Εξειδικευμένα εργαλεία : Η εργασία με αδόμητα δεδομένα απαιτεί εξειδικευμένα εργαλεία, περιορίζοντας τις επιλογές προϊόντων για τους διαχειριστές δεδομένων
- Χρονοβόρα και ακριβά : Η επεξεργασία μη δομημένων δεδομένων μπορεί να διαρκέσει πολύ. Μπορεί επίσης να είναι δαπανηρή η μετατροπή τους σε

[5] Οι βάσεις δεδομένων NoSQL (ή αλλιώς "όχι μόνο SQL") είναι μη-ταμπλετικές βάσεις δεδομένων και αποθηκεύουν δεδομένα διαφορετικά από τους σχεσιακούς πίνακες. Οι βάσεις δεδομένων NoSQL κυκλοφορούν σε διάφορους τύπους με βάση το μοντέλο δεδομένων τους. Οι κύριοι τύποι είναι έγγραφο, κλειδί-τιμή, ευρεία στήλη και γράφος. Παρέχουν ευέλικτα σχήματα και κλιμακώνονται εύκολα με μεγάλες ποσότητες δεδομένων και υψηλά φορτία χρηστών (MongoDB, n.d.)
[6] Μια λίμνη δεδομένων είναι ένα κεντρικό αποθετήριο που έχει σχεδιαστεί για την αποθήκευση, επεξεργασία και ασφάλεια μεγάλων ποσοτήτων δομημένων, ημιδομημένων και αδόμητων δεδομένων. Μπορεί να αποθηκεύει δεδομένα στη φυσική τους μορφή και να επεξεργάζεται οποιαδήποτε ποικιλία τους, αγνοώντας τα όρια μεγέθους (google, n.d.).

χρήσιμες, πρακτικές πληροφορίες, καθώς θα χρειαστεί τεχνητή νοημοσύνη και επιστήμονες δεδομένων για να τα δομήσουν

- Δύσκολη αποθήκευση : Λόγω του τεράστιου μεγέθους τους, γενικά αποθηκεύουμε τα μη δομημένα δεδομένα σε λίμνες δεδομένων. Οι λίμνες δεδομένων είναι χώροι αποθήκευσης με τεράστια αποθηκευτική ικανότητα

(Accern, 2022)

Διαφορές μεταξύ δομημένων και μη δομημένων δεδομένων

- Τα δομημένα δεδομένα είναι ένας σαφώς καθορισμένος τύπος δεδομένων εντός μιας δομής. Τα μη δομημένα δεδομένα αποθηκεύονται συνήθως στη φυσική τους μορφή, ενώ τα δομημένα δεδομένα αποτελούνται από γραμμές και στήλες που μπορούν να αντιστοιχιστούν σε προκαθορισμένα πεδία. Σε αντίθεση με τα δομημένα δεδομένα, τα οποία είναι οργανωμένα και εύκολα προσβάσιμα σε σχεσιακές βάσεις δεδομένων, τα μη δομημένα δεδομένα δεν έχουν καθορισμένο μοντέλο δεδομένων και θεωρούνται απροσδιόριστα.
- Τα δομημένα δεδομένα είναι συχνά ποσοτικά δεδομένα και αποτελούνται από αριθμούς ή πράγματα που μπορούν να μετρηθούν. Οι μέθοδοι ανάλυσης περιλαμβάνουν την παλινδρόμηση (για την πρόβλεψη των σχέσεων μεταξύ μεταβλητών), την ταξινόμηση (για την εκτίμηση πιθανοτήτων), την ομαδοποίηση των δεδομένων (με βάση διάφορα χαρακτηριστικά). Τα μη δομημένα δεδομένα συχνά ταξινομούνται ως ποιοτικά δεδομένα και δεν μπορούν να επεξεργαστούν και να αναλυθούν με τη χρήση παραδοσιακών εργαλείων και μεθόδων. Σε επιχειρηματικό πλαίσιο, τα ποιοτικά δεδομένα προέρχονται από έρευνες πελατών, συνεντεύξεις, αλληλεπιδράσεις στα μέσα κοινωνικής δικτύωσης κ.ά. Η εξαγωγή συμπερασμάτων από τα ποιοτικά δεδομένα απαιτεί προηγμένες αναλυτικές τεχνικές, όπως η εξόρυξη δεδομένων και η στοίβαξη δεδομένων
- Τα δομημένα δεδομένα αποθηκεύονται συχνά σε αποθήκες δεδομένων και τα μη δομημένα δεδομένα σε λίμνες δεδομένων. Και οι δύο τύποι έχουν τη δυνατότητα χρήσης του νέφους. Τα δομημένα δεδομένα καταλαμβάνουν λιγότερο αποθηκευτικό χώρο, ενώ τα μη δομημένα δεδομένα καταλαμβάνουν περισσότερο. Για παράδειγμα μια μικρή εικόνα καταλαμβάνει περισσότερο χώρο από το κείμενο αρκετών εκατοντάδων σελίδων
- Όπως και οι βάσεις δεδομένων, τα δομημένα δεδομένα αποθηκεύονται συνήθως σε σχεσιακές βάσεις δεδομένων (RDBMS), ενώ τα μη δομημένα δεδομένα αποθηκεύονται καλύτερα στις λεγόμενες μη σχεσιακές ή NoSQL βάσεις δεδομένων
- Μία από τις κύριες διαφορές μεταξύ δομημένων και μη δομημένων δεδομένων είναι το πόσο καλά προσφέρονται για ανάλυση. Τα δομημένα δεδομένα είναι εύκολα αναζητήσιμα τόσο για τους ανθρώπους όσο και για τους αλγορίθμους. Τα μη δομημένα δεδομένα είναι εγγενώς δύσκολο να αναζητηθούν και πρέπει να υποστούν επεξεργασία για να γίνουν κατανοητά. Είναι δύσκολο να αναλυθούν επειδή δεν έχουν καθορισμένο μοντέλο δεδομένων και δεν χωράνε σε σχεσιακές βάσεις δεδομένων
- Οι πιο συνηθισμένες μορφές δομημένων δεδομένων είναι το κείμενο και οι αριθμοί. Τα δομημένα δεδομένα ορίζονταν προηγουμένως σε μοντέλα δεδομένων. Τα μη δομημένα δεδομένα, από την άλλη πλευρά, υπάρχουν σε πολλές μορφές και μεγέθη. Μπορεί να αποτελούνται από οτιδήποτε, από ήχο,

βίντεο και εικόνες μέχρι ηλεκτρονικό ταχυδρομείο και δεδομένα αισθητήρων. Δεν υπάρχει μοντέλο δεδομένων για τα μη δομημένα δεδομένα. Αποθηκεύονται εγγενώς ή σε μια λίμνη δεδομένων που δεν απαιτεί μετασχηματισμό

(Smallcombe, 2022)

Ημιδομημένα δεδομένα

Τα ημιδομημένα δεδομένα είναι δεδομένα που δεν συμμορφώνονται με κάποιο μοντέλο δεδομένων, αλλά έχουν κάποια δομή. Δεν υπάρχουν σταθερά ή άκαμπτα πρότυπα. Πρόκειται για δεδομένα που δεν υπάρχουν σε μια λογική βάση δεδομένων, αλλά έχουν κάποιες οργανωτικές ιδιότητες που διευκολύνουν την ανάλυσή τους. Ορισμένες διαδικασίες μπορούν να αποθηκευτούν σε μια σχεσιακή βάση δεδομένων.

Χαρακτηριστικά των ημιδομημένων δεδομένων:

- Τα δεδομένα δεν αντιστοιχούν σε ένα μοντέλο δεδομένων, αλλά έχουν μια συγκεκριμένη δομή
- Τα δεδομένα δεν μπορούν να αποθηκευτούν σε γραμμές και στήλες όπως μια βάση δεδομένων
- Τα ημιδομημένα δεδομένα περιέχουν ετικέτες και στοιχεία (metadata[7]) που χρησιμοποιούνται για την ομαδοποίηση των δεδομένων και την περιγραφή του τρόπου αποθήκευσης των δεδομένων
- Παρόμοιες οντότητες ομαδοποιούνται και οργανώνονται σε ιεραρχίες
- Οι οντότητες εντός της ίδιας ομάδας μπορεί να έχουν ή να μην έχουν τα ίδια χαρακτηριστικά ή ιδιότητες
- Δύσκολο να αυτοματοποιηθούν και να διαχειριστούν τα δεδομένα επειδή δεν περιέχουν αρκετά μεταδεδομένα
- Το μέγεθος και ο τύπος των ίδιων χαρακτηριστικών σε μια ομάδα μπορεί να διαφέρουν
- Δεν έχει σαφώς καθορισμένη δομή, οπότε δεν μπορεί να χρησιμοποιηθεί εύκολα σε προγράμματα υπολογιστών

Πηγές των ημιδομημένων δεδομένων :

- Email
- XML και γλώσσες σήμανσης
- Binary αρχεία
- Πακέτα TCP/IP[8]

[7] Τα μεταδεδομένα είναι ένα σύνολο δεδομένων που παρέχουν πληροφορίες σχετικά με άλλα δεδομένα. Τα μεταδεδομένα πλαισιώνουν άλλα δεδομένα - παρέχοντας πληροφορίες όπως πότε και πώς συλλέχθηκαν - γεγονός που καθιστά τα δεδομένα ευκολότερα στην εύρεση, κατανόηση, χρήση και διαχείριση. Τα μεταδεδομένα μπορούν να πουν πότε στάλθηκε ένα μήνυμα, αλλά όχι το πραγματικό κείμενο του μηνύματος (Avast, n.d.)

[8] TCP/IP, πλήρης ονομασία Transmission Control Protocol/Internet Protocol, είναι τυποποιημένα πρωτόκολλα επικοινωνιών στο Διαδίκτυο που επιτρέπουν στους υπολογιστές να επικοινωνούν σε μεγάλες αποστάσεις. Το Διαδίκτυο είναι ένα δίκτυο μεταγωγής πακέτων, στο οποίο οι πληροφορίες αναλύονται σε μικρά πακέτα, αποστέλλονται μεμονωμένα σε πολλές διαφορετικές διαδρομές

- Συμπιεσμένα αρχεία
- Ενσωμάτωση δεδομένων από διαφορετικές πηγές
- Ιστοσελίδες

Πλεονεκτήματα των ημιδομημένων δεδομένων :

- Τα δεδομένα δεν περιορίζονται από ένα σταθερό σχήμα (schema[9])
- Ευέλικτα και εύκολα στην αλλαγή των schema
- Τα δεδομένα είναι φορητά
- Τα δομημένα δεδομένα μπορούν να θεωρηθούν ως ημιδομημένα δεδομένα
- Υποστηρίζει χρήστες που δεν μπορούν να εκφράσουν τις ανάγκες τους σε SQL
- Μπορεί εύκολα να χειριστεί την ετερογένεια των πηγών

Μειονεκτήματα των ημιδομημένων δεδομένων :

- Η έλλειψη ενός σταθερού και άκαμπτου σχήματος καθιστά δύσκολη την αποθήκευση δεδομένων
- Το σχήμα και τα δεδομένα δεν διαχωρίζονται, γεγονός που καθιστά δύσκολη την ερμηνεία των σχέσεων μεταξύ τους
- Τα ερωτήματα είναι λιγότερο αποτελεσματικά από τα δομημένα δεδομένα

Προβλήματα με την αποθήκευση ημιδομημένων δεδομένων :

- Τα δεδομένα έχουν συνήθως ακανόνιστη και μερική δομή. Ορισμένες πηγές έχουν έμμεσες δομές δεδομένων που καθιστούν δύσκολη την ερμηνεία των σχέσεων μεταξύ των δεδομένων
- Η διάκριση μεταξύ σχήματος και δεδομένων είναι εξαιρετικά αβέβαιη ή ασαφής. Αυτό καθιστά δύσκολο το σχεδιασμό δομών δεδομένων
- Υψηλότερο κόστος αποθήκευσης σε σύγκριση με τα δομημένα δεδομένα

(www.geeksforgeeks.org, 2021)

Ορισμένα παραδείγματα που γίνεται χρήση των ημιδομημένων δεδομένων είναι :

- Ταξινόμηση εικόνων και ήχων. Η βαθιά μάθηση (Deep Learning[10]) μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός συστήματος για την αναγνώριση

ταυτόχρονα και στη συνέχεια επανασυντίθενται στο τέλος του παραλήπτη. Το TCP είναι το στοιχείο που συλλέγει και επανασυνθέτει τα πακέτα δεδομένων, ενώ το IP είναι υπεύθυνο για τη διασφάλιση της αποστολής των πακέτων στον σωστό προορισμό. Το TCP/IP αναπτύχθηκε τη δεκαετία του 1970 και υιοθετήθηκε ως πρότυπο πρωτοκόλλου για το ARPANET (τον προκάτοχο του Internet) το 1983 (Britannica, 2022)

[9] Ένα σχήμα βάσης δεδομένων είναι μια δομή που αναπαριστά τη λογική αποθήκευση των δεδομένων σε μια βάση δεδομένων. Αντιπροσωπεύει την οργάνωση των δεδομένων και παρέχει πληροφορίες σχετικά με τις σχέσεις μεταξύ των πινάκων σε μια δεδομένη βάση δεδομένων (www.javatpoint.com, n.d.)

[10] Η βαθιά μάθηση είναι μια τεχνική μηχανικής μάθησης που διδάσκει στους υπολογιστές να κάνουν αυτό που είναι φυσικό για τους ανθρώπους: να μαθαίνουν από το παράδειγμα. Η βαθιά μάθηση είναι μια βασική τεχνολογία πίσω από τα αυτοκίνητα χωρίς οδηγό, που τους επιτρέπει να αναγνωρίζουν ένα σήμα στοπ ή να διακρίνουν έναν πεζό από μια κολόνα. Είναι το κλειδί για τον φωνητικό έλεγχο σε

εικόνων και ήχων. Το σύστημα μαθαίνει από τα σημειωμένα παραδείγματα για να ταξινομεί με ακρίβεια νέες εικόνες και ήχους. Για παράδειγμα, ένας υπολογιστής μπορεί να εκπαιδευτεί να αναγνωρίζει ορισμένους ήχους που υποδεικνύουν ότι ένας κινητήρας παρουσιάζει βλάβη. Αυτός ο τύπος εφαρμογής χρησιμοποιείται στην αυτοκινητοβιομηχανία και την αεροδιαστημική. Η τεχνολογία χρησιμοποιείται επίσης για την ταξινόμηση επαγγελματικών φωτογραφιών για διαδικτυακές πωλήσεις αυτοκινήτων και για την αναγνώριση άλλων προϊόντων. Για παράδειγμα, οι φωτογραφίες αντικειμένων που πωλούνται σε διαδικτυακές δημοπρασίες μπορούν να επισημανθούν αυτόματα με λεζάντες. Η αναγνώριση εικόνας χρησιμοποιείται στην ιατρική για την ταξινόμηση μαστογραφιών ως πιθανών καρκίνων και στη γονιδιωματική για την κατανόηση δεικτών ασθενειών

- Ως είσοδος για μοντέλα πρόβλεψης. Η ανάλυση κειμένου με χρήση επεξεργασίας φυσικής γλώσσας (NLP[11]) ή μηχανικής μάθησης χρησιμοποιείται για τη δόμηση αδόμητου κειμένου. Για παράδειγμα, οι επιχειρήσεις μπορούν να εξάγουν οντότητες (άτομα, μέρη ή πράγματα), θέματα ή συναισθήματα από σημειώσεις του τηλεφωνικού κέντρου. Μπορείτε να συνδυάσετε αυτές τις πληροφορίες με άλλες πληροφορίες σχετικά με τους πελάτες σας για να δημιουργήσετε μοντέλα πρόβλεψης. Για παράδειγμα, οι οντότητες, οι έννοιες και τα θέματα μπορούν να ομαδοποιηθούν χρησιμοποιώντας στατιστικές τεχνικές
- Chatbots[12] στην εμπειρία πελατών. Τα chatbots υπάρχουν εδώ και αρκετά χρόνια, αλλά τα νεότερα έχουν καλύτερη γλωσσική κατανόηση και είναι πιο διαδραστικά κάνοντας χρήση του NLP

(Halper, 2018)

καταναλωτικές συσκευές όπως τηλέφωνα, tablet, τηλεοράσεις και ηχεία hands-free. Η βαθιά εκμάθηση λαμβάνει μεγάλη προσοχή τελευταία και για καλό λόγο, καθώς επιτυγχάνει αποτελέσματα που δεν ήταν εφικτά πριν. Στη βαθιά μάθηση, ένα μοντέλο υπολογιστή μαθαίνει να εκτελεί εργασίες ταξινόμησης απευθείας από εικόνες, κείμενο ή ήχο. Τα μοντέλα βαθιάς μάθησης μπορούν να επιτύχουν κορυφαία ακρίβεια, που μερικές φορές ξεπερνά τις επιδόσεις σε ανθρώπινο επίπεδο. Τα μοντέλα εκπαιδεύονται χρησιμοποιώντας ένα μεγάλο σύνολο επισημασμένων δεδομένων και αρχιτεκτονικές νευρωνικών δικτύων που περιέχουν πολλά επίπεδα (www.mathworks.com, n.d.)

[11] Η επεξεργασία φυσικής γλώσσας (NLP) είναι μια μορφή τεχνητής νοημοσύνης που βοηθά τις μηχανές να "διαβάζουν" κείμενο προσομοιώνοντας την ανθρώπινη ικανότητα κατανόησης της γλώσσας. Οι τεχνικές NLP ενσωματώνουν μια ποικιλία μεθόδων, συμπεριλαμβανομένης της γλωσσολογίας, της σημασιολογίας, της στατιστικής και της μηχανικής μάθησης για την εξαγωγή οντοτήτων, σχέσεων και την κατανόηση του πλαισίου, γεγονός που επιτρέπει την κατανόηση του τι λέγεται ή γράφεται, με ολοκληρωμένο τρόπο. Αντί να κατανοεί μεμονωμένες λέξεις ή συνδυασμούς τους, η NLP βοηθά τους υπολογιστές να κατανοούν τις προτάσεις όπως αυτές προφέρονται ή γράφονται από έναν άνθρωπο. Χρησιμοποιεί διάφορες μεθοδολογίες για να αποκρυπτογραφήσει τις ασάφειες της γλώσσας, συμπεριλαμβανομένης της αυτόματης περίληψης, της επισήμανσης μέρους του λόγου, της αποσαφήνισης, της εξαγωγής οντοτήτων και σχέσεων, καθώς και της αποσαφήνισης και της κατανόησης και αναγνώρισης φυσικής γλώσσας (Allouche, 2014)

[12] Τα chatbots διευκολύνουν τους χρήστες να βρουν τις πληροφορίες που χρειάζονται, απαντώντας σε ερωτήσεις και αιτήματα των χρηστών μέσω εισαγωγής κειμένου, φωνής ή και των δύο χωρίς την ανάγκη ανθρώπινης παρέμβασης (IBM, n.d.)

1.6 Υποδομή Μεγάλων Δεδομένων

Τα βασικά στοιχεία των τεχνολογιών μεγάλων δεδομένων είναι τα εργαλεία και οι τεχνολογίες που παρέχουν τη δυνατότητα αποθήκευσης, επεξεργασίας και ανάλυσης των δεδομένων. Η μέθοδος αποθήκευσης των δεδομένων σε πίνακες δεν ήταν πλέον υποστηρικτική με την εξέλιξη των δεδομένων με 3 Vs, δηλαδή όγκο, ταχύτητα και ποικιλία. Το εύρωστο RBDMS δεν ήταν πλέον οικονομικά αποδοτικό. Η κλιμάκωση των RDBMS για την αποθήκευση και την επεξεργασία τεράστιου όγκου δεδομένων έγινε ακριβή. Αυτό οδήγησε στην εμφάνιση νέας τεχνολογίας, η οποία ήταν εξαιρετικά επεκτάσιμη με πολύ χαμηλό κόστος. Οι κύριες τεχνολογίες περιλαμβάνουν :

Apache Hadoop

Το Apache Hadoop είναι ένα πλαίσιο λογισμικού ανοικτού κώδικα που παρέχει εξαιρετικά αξιόπιστη κατανεμημένη επεξεργασία μεγάλων συνόλων δεδομένων με τη χρήση απλών μοντέλων προγραμματισμού. Το Hadoop, γνωστό για την επεκτασιμότητά του, είναι κατασκευασμένο σε συστάδες υπολογιστών κοινής χρήσης, παρέχοντας μια οικονομικά αποδοτική λύση για την αποθήκευση και επεξεργασία τεράστιων ποσοτήτων δομημένων, ημιδομημένων και αδόμητων δεδομένων χωρίς απαιτήσεις μορφοποίησης. Μια αρχιτεκτονική λίμνης δεδομένων που περιλαμβάνει το Hadoop μπορεί να προσφέρει μια ευέλικτη λύση διαχείρισης δεδομένων για τις πρωτοβουλίες σας για την ανάλυση μεγάλων δεδομένων. Επειδή το Hadoop είναι ένα έργο λογισμικού ανοικτού κώδικα και ακολουθεί ένα μοντέλο κατανεμημένου υπολογισμού, μπορεί να προσφέρει χαμηλότερο συνολικό κόστος ιδιοκτησίας για μια λύση λογισμικού και αποθήκευσης μεγάλων δεδομένων.Το Hadoop μπορεί επίσης να εγκατασταθεί σε διακομιστές cloud για την καλύτερη διαχείριση των πόρων υπολογισμού και αποθήκευσης που απαιτούνται για μεγάλα δεδομένα (IBM, n.d.)

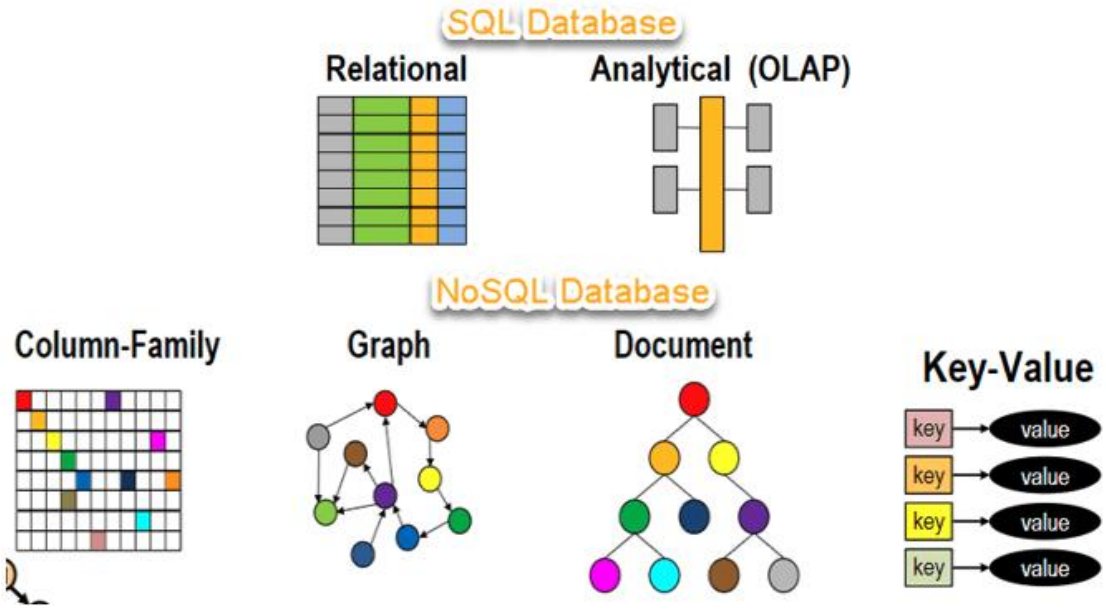
Το οικοσύστημα Apache Hadoop περιλαμβάνει :

- HDFS : το HDFS είναι ένα κατανεμημένο σύστημα αρχείων που διαχειρίζεται μεγάλα σύνολα δεδομένων και εκτελείται σε βασικό υλικό. Χρησιμοποιείται για την κλιμάκωση μιας συστάδας Apache Hadoop σε εκατοντάδες (ακόμη και χιλιάδες) κόμβους. Το HDFS είναι ένα από τα κύρια συστατικά του Apache Hadoop, τα άλλα είναι το MapReduce και το YARN. Το HDFS δεν πρέπει να συγχέεται ή να αντικαθίσταται από το Apache HBase, το οποίο είναι ένα σύστημα διαχείρισης μη σχεσιακών βάσεων δεδομένων με προσανατολισμό προς τις στήλες που βρίσκεται πάνω στο HDFS και μπορεί να υποστηρίξει καλύτερα τις ανάγκες δεδομένων σε πραγματικό χρόνο με τη μηχανή επεξεργασίας στη μνήμη (ibm, n.d.)
- Hadoop Common: Τα κοινά βοηθητικά προγράμματα και οι βιβλιοθήκες που υποστηρίζουν τις άλλες ενότητες του Hadoop. Επίσης γνωστό ως Hadoop Core
- Hadoop YARN : Ένα πλαίσιο για τη διαχείριση πόρων συστάδας και τον προγραμματισμό εργασιών. Το YARN σημαίνει Yet Another Resource Negotiator. Υποστηρίζει περισσότερους φόρτους εργασίας, όπως διαδραστική SQL, προηγμένη μοντελοποίηση και ροή σε πραγματικό χρόνο
- Hadoop Ozone: σχεδιασμένο για εφαρμογές μεγάλων δεδομένων.

- MapReduce : το MapReduce είναι ένα παράδειγμα προγραμματισμού (programming paradigm) που επιτρέπει τεράστια κλιμάκωση σε εκατοντάδες ή χιλιάδες διακομιστές σε ένα σύμπλεγμα Hadoop. Ως συστατικό επεξεργασίας, το MapReduce είναι η καρδιά του Apache Hadoop. Ο όρος "MapReduce" αναφέρεται σε δύο ξεχωριστές και διακριτές εργασίες που εκτελούν τα προγράμματα Hadoop. Η πρώτη είναι η εργασία map, η οποία λαμβάνει ένα σύνολο δεδομένων και το μετατρέπει σε ένα άλλο σύνολο δεδομένων, όπου τα επιμέρους στοιχεία αναλύονται σε πλειάδες (ζεύγη κλειδιών/τιμών). Η εργασία reduce λαμβάνει την έξοδο από έναν χάρτη ως είσοδο και συνδυάζει αυτές τις πλειάδες δεδομένων σε ένα μικρότερο σύνολο πλειάδων. Όπως υποδηλώνει η ακολουθία του ονόματος MapReduce, η εργασία μείωσης εκτελείται πάντα μετά την εργασία χαρτογράφησης (ibm, n.d.)

NoSQL

Μια βάση δεδομένων NoSQL (NoSQL, n.d.) παρέχει έναν μηχανισμό αποθήκευσης και ανάκτησης δεδομένων που μοντελοποιείται με μέσα διαφορετικά από τις σχέσεις σε μορφή πίνακα που χρησιμοποιούνται στις σχεσιακές βάσεις δεδομένων. Τέτοιες βάσεις δεδομένων υπάρχουν από τα τέλη της δεκαετίας του 1960, αλλά η ονομασία "NoSQL" επινοήθηκε μόλις στις αρχές του 21ου αιώνα. Οι βάσεις δεδομένων NoSQL χρησιμοποιούνται όλο και περισσότερο σε εφαρμογές μεγάλου όγκου δεδομένων και διαδικτυακών εφαρμογών πραγματικού χρόνου και τα συστήματα NoSQL αποκαλούνται επίσης μερικές φορές Not only SQL για να τονιστεί ότι μπορούν να υποστηρίζουν γλώσσες ερωτημάτων που μοιάζουν με SQL ή να βρίσκονται δίπλα σε βάσεις δεδομένων SQL σε πολυγλωσσικές-διάρκεις αρχιτεκτονικές (Fowler, 2012). Τα κίνητρα για αυτή την προσέγγιση περιλαμβάνουν την απλότητα του σχεδιασμού, την απλούστερη "οριζόντια" κλιμάκωση σε συστάδες μηχανών, τον λεπτότερο έλεγχο της διαθεσιμότητας και τον περιορισμό της αναντιστοιχίας αντικειμενικής-σχεσιακής εμπλοκής. Οι δομές δεδομένων που χρησιμοποιούνται από τις βάσεις δεδομένων NoSQL (π.χ. ζεύγος κλειδιού-τιμής, ευρεία στήλη, γράφος ή έγγραφο) είναι διαφορετικές από αυτές που χρησιμοποιούνται εξ ορισμού στις σχεσιακές βάσεις δεδομένων, καθιστώντας ορισμένες λειτουργίες ταχύτερες στην NoSQL. Η ιδιαίτερη καταλληλότητα μιας δεδομένης βάσης δεδομένων NoSQL εξαρτάται από το πρόβλημα που πρέπει να επιλύσει. Μερικές φορές οι δομές δεδομένων που χρησιμοποιούνται από τις βάσεις δεδομένων NoSQL θεωρούνται επίσης ως "πιο ευέλικτες" από τους πίνακες των σχεσιακών βάσεων δεδομένων (allthingsdistributed, 2012).
Τα παραδοσιακά RDBMS χρησιμοποιούν σύνταξη SQL για την αποθήκευση και ανάκτηση δεδομένων για περαιτέρω πληροφορίες. Αντίθετα, ένα σύστημα βάσεων δεδομένων NoSQL περιλαμβάνει ένα ευρύ φάσμα τεχνολογιών βάσεων δεδομένων που μπορούν να αποθηκεύουν δομημένα, ημιδομημένα, αδόμητα και πολυμορφικά δεδομένα.



Εικόνα 5 Διαφορές στην αποθήκευση μεταξύ SQL και NoSQL (Deepashree Karanjkar, 2019)

Η έννοια των βάσεων δεδομένων NoSQL έγινε δημοφιλής με εταιρείες του Διαδικτύου όπως η Google, το Facebook, η Amazon κ.λπ. που διαχειρίζονται τεράστιους όγκους δεδομένων. Ο χρόνος απόκρισης του συστήματος γίνεται αργός όταν γίνεται χρήση RDBMS για τεράστιους όγκους δεδομένων. Για να επιλυθεί αυτό το ζήτημα, θα μπορούσε να γίνει αναβάθμιση στο ήδη υπάρχων hardware, κάτι το οποίο ωστόσο είναι υπερβολικά δαπανηρό.
Ως εναλλακτική λύση για αυτό το ζήτημα είναι να γίνει κατανομή του φορτίου της βάσης δεδομένων σε πολλούς κεντρικούς υπολογιστές κάθε φορά που το φορτίο αυξάνεται, και η μέθοδος είναι γνωστή ως "κλιμάκωση".

Χαρακτηριστικά της NoSQL :

Είναι μη σχεσιακή (Non-relational)

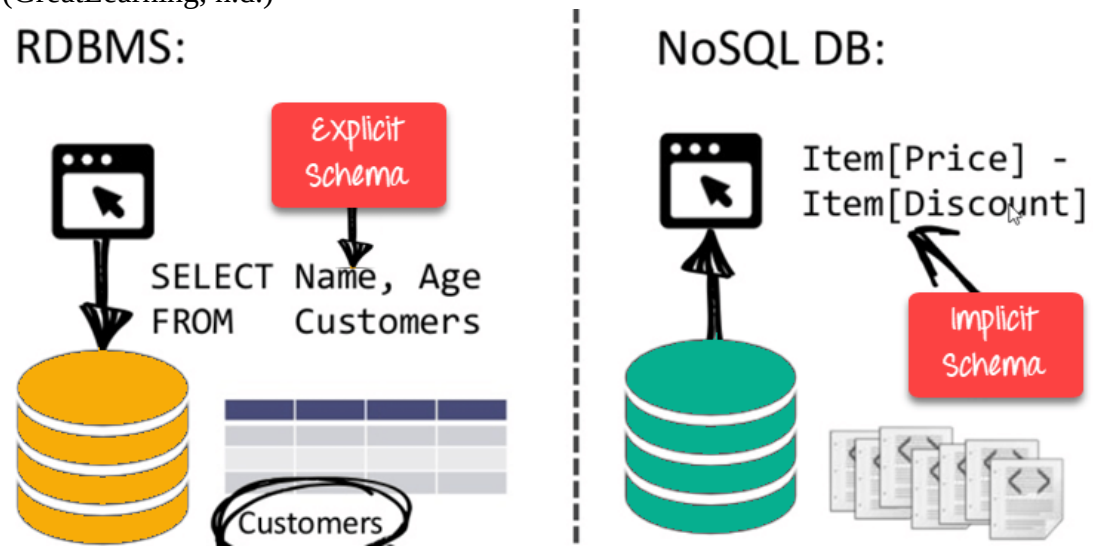
- Οι βάσεις δεδομένων NoSQL δεν ακολουθούν ποτέ το σχεσιακό μοντέλο
- Ποτέ δεν παρέχουν πίνακες με επίπεδες εγγραφές σταθερών στηλών
- Δουλεύουν με αυτοτελή σύνολα ή BLOBs[13]
- Δεν απαιτούν αντιστοίχιση αντικειμένου-σχεσιακής απεικόνισης και κανονικοποίηση δεδομένων
- Δεν διαθέτουν πολύπλοκα χαρακτηριστικά

[13] BLOB σημαίνει "Binary Large Object", ένας τύπος δεδομένων που αποθηκεύει δυαδικά δεδομένα. Τα δυαδικά μεγάλα αντικείμενα (BLOB) μπορούν να είναι σύνθετα αρχεία όπως εικόνες ή βίντεο, σε αντίθεση με άλλες συμβολοσειρές δεδομένων που αποθηκεύουν μόνο γράμματα και αριθμούς. Ένα BLOB θα περιέχει αντικείμενα πολυμέσων για να προστεθούν σε μια βάση δεδομένων, ωστόσο δεν υποστηρίζουν όλες οι βάσεις δεδομένων την αποθήκευση BLOB. Λόγω της σύνθετης φύσης τους, τα BLOBs δεν θα είναι επίσης εύκολα αναγνώσιμα από τις περισσότερες βάσεις δεδομένων. Αυτοί οι τύποι αρχείων είναι καλύτερα κατανοητοί από ανθρώπους αντί για λογισμικό. Η πολυπλοκότητα ενός BLOB του δίνει την αξία του, αλλά μπορεί επίσης να καταστήσει δύσκολη τη χρήση του (Microsoft, 2022)

Είναι χωρίς σχήμα βάσης δεδομένων (Schema-free)

- Οι βάσεις δεδομένων NoSQL είναι είτε χωρίς σχήματα είτε έχουν χαλαρά σχήματα
- Δεν απαιτούν κανενός είδους ορισμό του σχήματος των δεδομένων
- Προσφέρει ετερογενείς δομές δεδομένων στον ίδιο τομέα

(GreatLearning, n.d.)



Εικόνα 6 Η NoSQL είναι χωρίς σχήμα βάσης δεδομένων [πηγή : (GreatLearning, n.d.)]

Απλή Διεπαφή API[14]

- Προσφέρει εύχρηστες διεπαφές για την αποθήκευση και την αναζήτηση των παρεχόμενων δεδομένων
- Τα API επιτρέπουν μεθόδους χειρισμού και επιλογής δεδομένων χαμηλού επιπέδου
- Πρωτόκολλα βασισμένα σε κείμενο που χρησιμοποιούνται κυρίως με HTTP REST με JSON

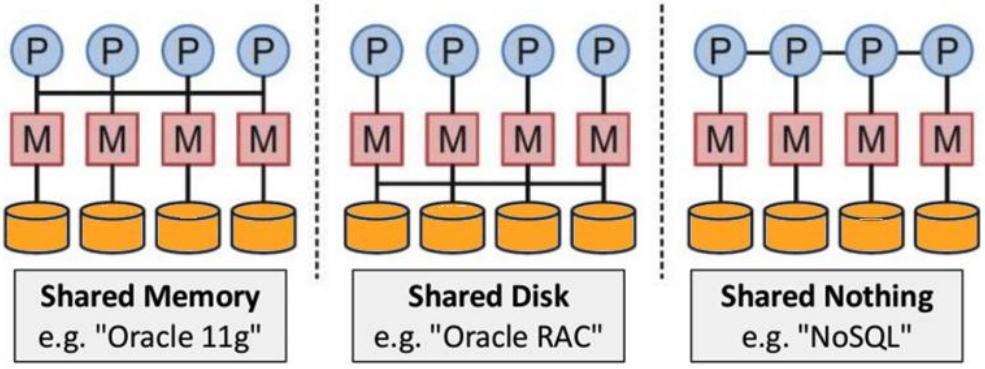
[14] Τα API είναι μηχανισμοί που επιτρέπουν σε δύο στοιχεία λογισμικού να επικοινωνούν μεταξύ τους χρησιμοποιώντας ένα σύνολο ορισμών και πρωτοκόλλων. Για παράδειγμα, το σύστημα λογισμικού της μετεωρολογικής υπηρεσίας περιέχει καθημερινά δεδομένα καιρού. Η εφαρμογή καιρού σε ένα smartphone "συνομιλεί" με αυτό το σύστημα μέσω APIs και εμφανίζει καθημερινές ενημερώσεις για τον καιρό στην οθόνη του τηλεφώνου.
Το API σημαίνει Application Programming Interface (διεπαφή προγραμματισμού εφαρμογών). Στο πλαίσιο των API, η λέξη Application (Εφαρμογή) αναφέρεται σε οποιοδήποτε λογισμικό με διακριτή λειτουργία. Η διεπαφή μπορεί να θεωρηθεί ως σύμβαση παροχής υπηρεσιών μεταξύ δύο εφαρμογών. Αυτή η σύμβαση ορίζει τον τρόπο με τον οποίο οι δύο επικοινωνούν μεταξύ τους χρησιμοποιώντας αιτήματα και απαντήσεις. Η τεκμηρίωση του API τους περιέχει πληροφορίες σχετικά με τον τρόπο με τον οποίο οι προγραμματιστές πρέπει να δομούν αυτά τα αιτήματα και τις απαντήσεις. Η αρχιτεκτονική API εξηγείται συνήθως με όρους πελάτη και διακομιστή. Η εφαρμογή που στέλνει το αίτημα ονομάζεται πελάτης και η εφαρμογή που στέλνει την απάντηση ονομάζεται διακομιστής. Έτσι, στο παράδειγμα του καιρού, η βάση δεδομένων καιρού του γραφείου είναι ο διακομιστής και η εφαρμογή για κινητά είναι ο πελάτης (Apple, n.d.)

- Διαδικτυακές βάσεις δεδομένων που εκτελούνται ως υπηρεσίες που βλέπουν στο διαδίκτυο

Κατανεμημένη

- Πολλαπλές βάσεις δεδομένων NoSQL μπορούν να εκτελεστούν με κατανεμημένο τρόπο
- Προσφέρει δυνατότητες αυτόματης κλιμάκωσης και fail-over
- Συχνά η έννοια ACID[15] μπορεί να θυσιαστεί για την επεκτασιμότητα και την απόδοση
- Ως επί το πλείστον δεν υπάρχει σύγχρονη αντιγραφή μεταξύ κατανεμημένων κόμβων Ασύγχρονη Multi-Master αντιγραφή, peer-to-peer, HDFS αντιγραφή
- Παρέχει μόνο ενδεχόμενη συνέπεια
- Αρχιτεκτονική κοινόχρηστου τίποτα. Αυτό επιτρέπει λιγότερο συντονισμό και υψηλότερη κατανομή

(bigdatapath, n.d.)



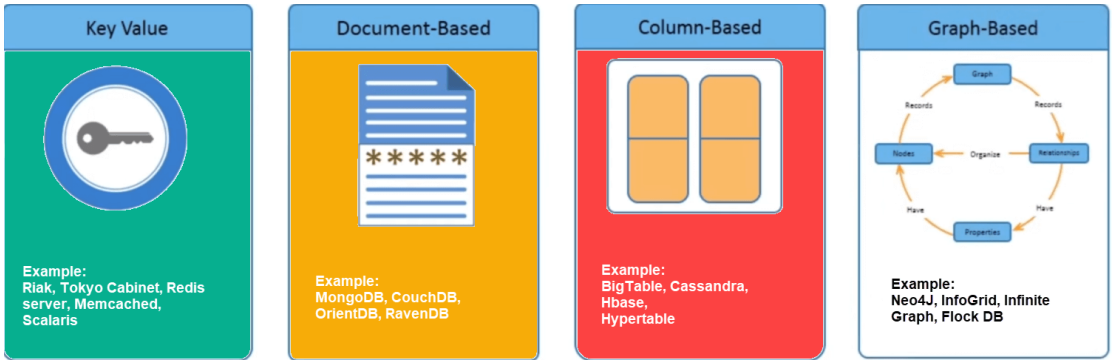
Εικόνα 7 Η NoSQL είναι αρχιτεκτονική κοινόχρηστου τίποτα [πηγή : (bigdatapath, n.d.)]

Τύποι NoSQL βάσεων δεδομένων

Οι βάσεις δεδομένων NoSQL κατηγοριοποιούνται κυρίως σε τέσσερις τύπους :

- Ζεύγος κλειδιού-τιμής
- Γράφημα προσανατολισμένο σε στήλες
- Με βάση το γράφημα
- Προσανατολισμένο σε έγγραφα

[15] Στα συστήματα βάσεων δεδομένων, ο όρος ACID (Atomicity, Consistency, Isolation, Durability) αναφέρεται σε ένα τυποποιημένο σύνολο ιδιοτήτων που εγγυώνται την αξιόπιστη επεξεργασία των συναλλαγών της βάσης δεδομένων. Το ACID ασχολείται ιδιαίτερα με τον τρόπο με τον οποίο μια βάση δεδομένων ανακάμπτει από οποιαδήποτε αποτυχία που μπορεί να συμβεί κατά την επεξεργασία μιας συναλλαγής. Ένα Σύστημα διαχείρισης βάσεων δεδομένων συμβατό με ACID εξασφαλίζει ότι τα δεδομένα στη βάση δεδομένων παραμένουν ακριβή και συνεπή παρά τις όποιες αποτυχίες (https://database.guide, 2016)



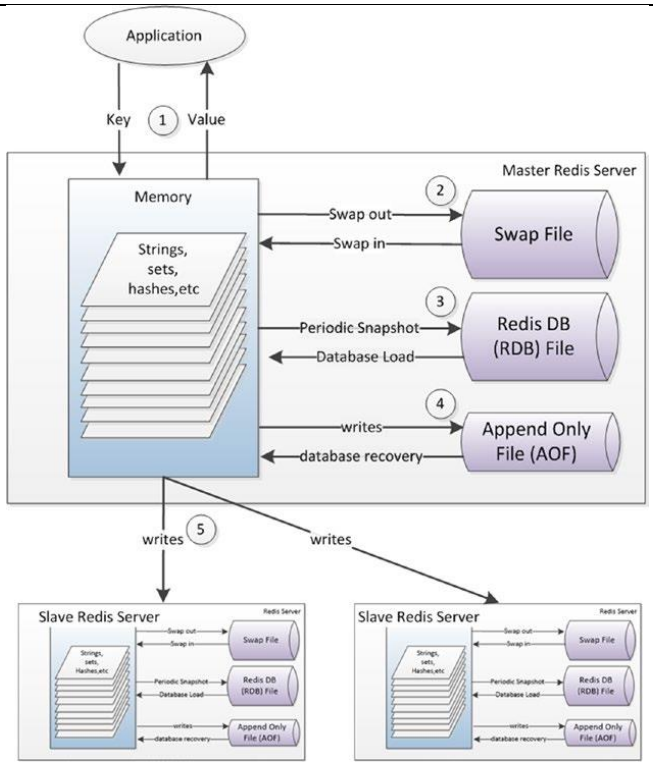
Εικόνα 8 τύποι της NoSQL [πηγή : (guru99, 2022)]

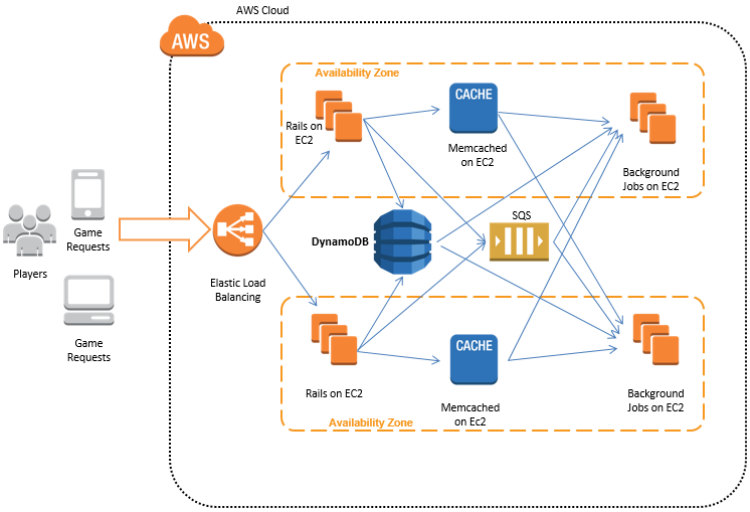
Ζεύγος τιμών κλειδιού

Τα δεδομένα αποθηκεύονται σε ζεύγη κλειδιών/τιμών. Έχει σχεδιαστεί με τέτοιο τρόπο ώστε να διαχειρίζεται πολλά δεδομένα και μεγάλο φορτίο.
Οι βάσεις δεδομένων αποθήκευσης ζεύγους κλειδιού-τιμής αποθηκεύουν δεδομένα ως πίνακα κατακερματισμού όπου κάθε κλειδί είναι μοναδικό και η τιμή μπορεί να είναι JSON, BLOB(Binary Large Objects), συμβολοσειρά κ.ά.
Οι Redis, Dynamo, Riak είναι μερικά παραδείγματα NoSQL για βάσεις δεδομένων με αποθήκευση κλειδιών-τιμών (guru99, 2022).

Redis	Το Redis ουσιαστικά είναι ένας αποθηκευτής κλειδιών-τιμών στη μνήμη. Το Redis (Remote Dictionary Server) σχεδιάστηκε αρχικά ως ένα απλό σύστημα στη μνήμη ικανό να διατηρήσει πολύ υψηλούς ρυθμούς συναλλαγών σε συστήματα με χαμηλή ισχύ, όπως εικόνες εικονικών μηχανών. Το Redis δημιουργήθηκε από τον Salvatore Sanfilippo το 2009. Το Redis ακολουθεί μια γνωστή αρχιτεκτονική αποθήκευσης κλειδιών-τιμών στην οποία τα κλειδιά δείχνουν σε αντικείμενα. Στο Redis, τα αντικείμενα αποτελούνται κυρίως από συμβολοσειρές και διάφορους τύπους συλλογών συμβολοσειρών (λίστες, ταξινομημένες λίστες, χάρτες κατακερματισμού κ.ά.). Υποστηρίζονται μόνο πρωτογενείς αναζητήσεις κλειδιών και το Redis δεν διαθέτει μηχανισμό δευτερογενούς ευρετηρίασης. Παρόλο που το Redis σχεδιάστηκε για να κρατάει όλα τα δεδομένα στη μνήμη, είναι δυνατό για το Redis να λειτουργεί σε σύνολα δεδομένων μεγαλύτερα από τη διαθέσιμη μνήμη, χρησιμοποιώντας τη λειτουργία εικονικής μνήμης. Όταν αυτό είναι ενεργοποιημένο, το Redis θα "ανταλλάσσει" παλαιότερες τιμές κλειδιών σε ένα αρχείο δίσκου. Εάν τα κλειδιά χρειαστούν, θα επανέλθουν στη μνήμη. Αυτή η επιλογή προφανώς συνεπάγεται σημαντική επιβάρυνση της απόδοσης, καθώς ορισμένες αναζητήσεις κλειδιών θα οδηγήσουν σε IO στο δίσκο. Η Redis χρησιμοποιεί αρχεία δίσκου για τη διατήρηση: • Τα αρχεία στιγμιότυπων (snapshot) αποθηκεύουν αντίγραφα ολόκληρου του συστήματος Redis σε μια

	συγκεκριμένη χρονική στιγμή. Τα στιγμιότυπα μπορούν να δημιουργηθούν κατά παραγγελία ή μπορούν να ρυθμιστούν ώστε να εμφανίζονται σε προγραμματισμένα χρονικά διαστήματα ή μετά την επίτευξη ενός ορίου εγγραφών. Ένα στιγμιότυπο εμφανίζεται επίσης όταν ο διακομιστής τερματίζεται. • Το Append Only File (AOF) διατηρεί ένα ημερολόγιο αλλαγών που μπορεί να χρησιμοποιηθεί για την "επαναφορά" της βάσης δεδομένων από ένα στιγμιότυπο σε περίπτωση αποτυχίας. Οι επιλογές διαμόρφωσης επιτρέπουν στο χρήστη να ρυθμίζει τις εγγραφές στο AOF μετά από κάθε λειτουργία, ανά ένα δευτερόλεπτο ή με βάση τα καθορισμένα από το λειτουργικό σύστημα διαστήματα. Επιπροσθέτως, η Redis υποστηρίζει ασύγχρονη αντιγραφή master/slave. Εάν η απόδοση είναι πολύ κρίσιμη και κάποια απώλεια δεδομένων είναι αποδεκτή, τότε ένα αντίγραφο μπορεί να χρησιμοποιηθεί ως εφεδρική βάση δεδομένων και ο κύριος να ρυθμιστεί με ελάχιστη επιμονή στο δίσκο. Ωστόσο, δεν υπάρχει τρόπος να περιοριστεί το μέγεθος της πιθανής απώλειας δεδομένων- κατά τη διάρκεια υψηλών φορτίων. Στην κάτωθι εικόνα έχουμε αυτά τα αρχιτεκτονικά στοιχεία. Η εφαρμογή αλληλεπιδρά με το Redis μέσω αναζητήσεων πρωτογενούς κλειδιού που επιστρέφουν "τιμές" συμβολοσειρές, σύνολα συμβολοσειρών, κατακερματισμούς συμβολοσειρών κ.ά (1). Οι τιμές των κλειδιών θα βρίσκονται σχεδόν πάντα στη μνήμη, αν και είναι δυνατόν να διαμορφωθεί το Redis με ένα σύστημα εικονικής μνήμης, οπότε οι τιμές των κλειδιών μπορεί να πρέπει να αντικατασταθούν (swap in or out) (2). Περιοδικά, το Redis μπορεί να κάνει απόρριψη ενός αντιγράφου ολόκληρου του χώρου μνήμης στο δίσκο (3). Επιπλέον, το Redis μπορεί να ρυθμιστεί ώστε να γράφει τις αλλαγές σε ένα αρχείο ημερολογίου μόνο με προσθήκη είτε σε μικρά χρονικά διαστήματα είτε μετά από κάθε λειτουργία (4). Τέλος, το Redis μπορεί να αναπαράγει την κατάσταση της κύριας βάσης δεδομένων ασύγχρονα σε δευτερεύοντες διακομιστές Redis (5) (Harrison, 2012).

	 *Εικόνα 9 Αρχιτεκτονική Redis [πηγή : (Harrison, 2012)]*
DynamoDB	Η Amazon DynamoDB είναι μια cloud-native βάση δεδομένων NoSQL κλειδιών-τιμών. • Η DynamoDB είναι cloud-native, δεδομένου ότι δεν εκτελείται στις εγκαταστάσεις ή ακόμη και σε υβριδικό νέφος- εκτελείται μόνο στις υπηρεσίες Amazon Web Services (AWS). Αυτό της επιτρέπει να κλιμακώνεται ανάλογα με τις ανάγκες, χωρίς να απαιτείται η κεφαλαιακή επένδυση του πελάτη σε υλικό. Διαθέτει επίσης χαρακτηριστικά κοινά με άλλες cloud-native εφαρμογές, όπως η ελαστική ανάπτυξη υποδομής. • Η DynamoDB είναι NoSQL, δεδομένου ότι δεν υποστηρίζει ANSI Structured Query Language (SQL). Αντ' αυτού, χρησιμοποιεί ένα ιδιόκτητο API που βασίζεται σε JavaScript Object Notation (JSON). Αυτό το API γενικά δεν καλείται απευθείας από τους προγραμματιστές-χρήστες, αλλά καλείται μέσω των AWS Software Developer Kits (SDKs) για το DynamoDB που είναι γραμμένα σε διάφορες γλώσσες προγραμματισμού. • Το DynamoDB είναι πρωτίστως ένα key-value store, υπό την έννοια ότι το μοντέλο δεδομένων του αποτελείται από ζεύγη τιμής-κλειδιού σε έναν χωρίς σχήμα, πολύ μεγάλο, μη σχεσιακό πίνακα γραμμών (εγγραφών). Δεν υποστηρίζει τις μεθόδους των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) για την ένωση πινάκων μέσω ξένων κλειδιών. Μπορεί επίσης να υποστηρίξει ένα μοντέλο δεδομένων αποθήκευσης εγγράφων με χρήση

	JavaScript Object Notation (JSON). Ο σχεδιασμός NoSQL της DynamoDB είναι προσανατολισμένος προς την απλότητα και την επεκτασιμότητα, οι οποίες απευθύνονται σε προγραμματιστές και ομάδες devops αντίστοιχα. Μπορεί να χρησιμοποιηθεί για μια ευρεία ποικιλία εφαρμογών που βασίζονται σε ημιδομημένα δεδομένα και είναι διαδεδομένες σε σύγχρονες και αναδυόμενες περιπτώσεις χρήσης πέρα από τις παραδοσιακές βάσεις δεδομένων, από το Διαδίκτυο των πραγμάτων (IoT) έως κοινωνικές εφαρμογές ή παιχνίδια μαζικών multiplayer. Με την ευρεία υποστήριξη γλωσσών προγραμματισμού, είναι εύκολο για τους προγραμματιστές να ξεκινήσουν και να δημιουργήσουν πολύ εξελιγμένες εφαρμογές με τη χρήση της DynamoDB (Scylladb, n.d.).  *Εικόνα 10 Αρχιτεκτονική DynamoDB [πηγή : (Scylladb, n.d.)]*

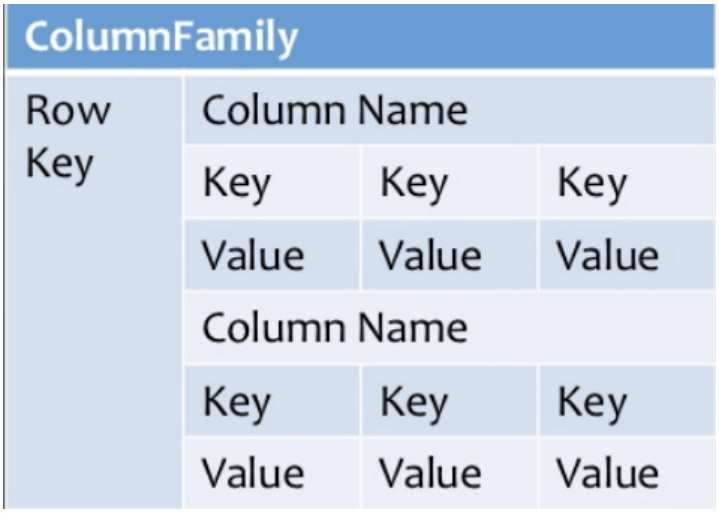
Πίνακας 3 Παραδείγματα NoSQL για βάσεις δεδομένων με αποθήκευση κλειδιού-τιμής

Key	Value
Name	Joe Bloggs
Age	42
Occupation	Stunt Double
Height	175cm
Weight	77kg

Εικόνα 11 Ζεύγος τιμών κλειδιού [πηγή : (guru99, 2022)]

Γράφημα προσανατολισμένο σε στήλες

Οι βάσεις δεδομένων προσανατολισμένες στις στήλες λειτουργούν με βάση τις στήλες και βασίζονται στο έγγραφο BigTable[16] της Google. Κάθε στήλη αντιμετωπίζεται ξεχωριστά. Οι τιμές των βάσεων δεδομένων μίας στήλης αποθηκεύονται συνεχόμενα. Παρέχουν υψηλές επιδόσεις σε ερωτήματα συνάθροισης όπως SUM, COUNT, AVGMIN κ.λπ. καθώς τα δεδομένα είναι άμεσα διαθέσιμα σε μια στήλη. Οι βάσεις δεδομένων NoSQL που βασίζονται σε στήλες χρησιμοποιούνται ευρέως για τη διαχείριση αποθηκών δεδομένων, επιχειρηματικής ευφυΐας, CRM και καταλόγους καρτών βιβλιοθηκών.
Η HBase, η Cassandra και η Hypertable είναι παραδείγματα ερωτημάτων NoSQL βάσης δεδομένων με βάση τις στήλες.

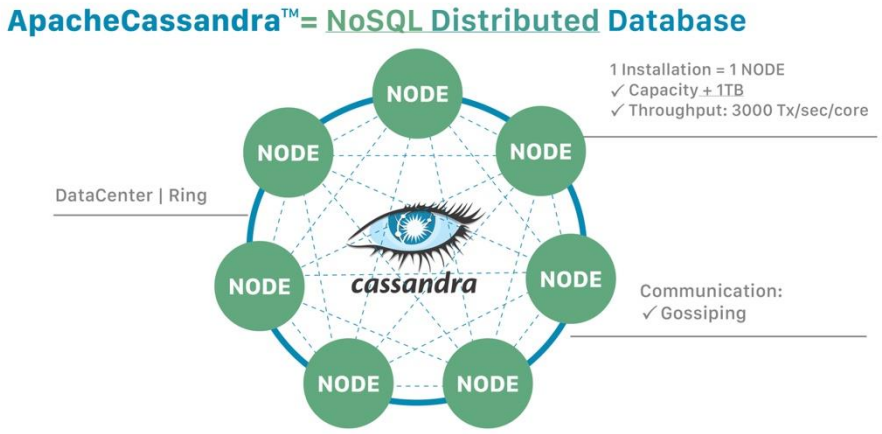


Εικόνα 12 Βάσεις Δεδομένων Προσανατολισμένες στις στήλες [πηγή : (guru99, 2022)]

Apache HBase	Η Apache HBase είναι ένας κατανεμημένος αποθηκευτικός χώρος μεγάλων δεδομένων ανοικτού κώδικα NoSQL. Επιτρέπει την τυχαία, αυστηρά συνεπή, σε πραγματικό χρόνο πρόσβαση σε petabytes δεδομένων. Η HBase είναι πολύ αποτελεσματική για το χειρισμό μεγάλων, αραιών συνόλων δεδομένων. Η HBase ενσωματώνεται απρόσκοπτα με το Apache Hadoop και το οικοσύστημα Hadoop και εκτελείται πάνω από το κατανεμημένο σύστημα αρχείων Hadoop (HDFS) ή το Amazon S3 χρησιμοποιώντας το σύστημα αρχείων Amazon Elastic MapReduce (EMR) ή EMRFS. Η HBase χρησιμεύει ως άμεση είσοδος και έξοδος στο πλαίσιο Apache MapReduce για το Hadoop και συνεργάζεται με τον Apache Phoenix για να επιτρέπει ερωτήματα τύπου SQL πάνω σε πίνακες HBase. Η HBase είναι μια μη-σχεσιακή βάση δεδομένων προσανατολισμένη σε στήλες. Αυτό σημαίνει ότι τα δεδομένα αποθηκεύονται σε μεμονωμένες στήλες και ευρετηριάζονται με ένα μοναδικό κλειδί γραμμής. Αυτή η αρχιτεκτονική επιτρέπει την ταχεία ανάκτηση

[16] Το Bigtable είναι μια πλήρως διαχειρίσιμη υπηρεσία βάσης δεδομένων NoSQL ευρείας στήλης και κλειδιών-τιμών για μεγάλους αναλυτικούς και επιχειρησιακούς φόρτους εργασίας ως μέρος του χαρτοφυλακίου Google Cloud (www.google.com, n.d.)

	μεμονωμένων γραμμών και στηλών και την αποτελεσματική σάρωση μεμονωμένων στηλών εντός ενός πίνακα. Τόσο τα δεδομένα όσο και τα αιτήματα κατανέμονται σε όλους τους διακομιστές σε μια συστάδα HBase και χρησιμοποιείται αποτελεσματικότερα για την αποθήκευση μη σχεσιακών δεδομένων, στα οποία η πρόσβαση γίνεται μέσω του API της HBase. Ο Apache Phoenix χρησιμοποιείται συνήθως ως στρώμα SQL πάνω από την HBase, επιτρέποντάς να γίνει χρήση τη γνωστής σύνταξης SQL για την εισαγωγή, τη διαγραφή και την υποβολή ερωτημάτων σε δεδομένα που είναι αποθηκευμένα στην HBase. Τα πλεονεκτήματα είναι : • Κλιμάκωση : Η HBase έχει σχεδιαστεί για να χειρίζεται την κλιμάκωση σε χιλιάδες διακομιστές και τη διαχείριση της πρόσβασης σε petabytes δεδομένων και είναι σε θέση να διαχειριστεί την online πρόσβαση σε τεράστια σύνολα δεδομένων. • Ταχύτητα : Η HBase παρέχει πρόσβαση τυχαίας ανάγνωσης και εγγραφής με χαμηλή καθυστέρηση σε petabytes δεδομένων, κατανέμοντας τα αιτήματα των εφαρμογών σε ένα σύμπλεγμα κεντρικών υπολογιστών. Κάθε κεντρικός υπολογιστής έχει πρόσβαση σε δεδομένα και εξυπηρετεί αιτήματα ανάγνωσης και εγγραφής σε χιλιοστά του δευτερολέπτου. • Ανοχή σε σφάλματα : Η HBase κατανέμει τα δεδομένα που είναι αποθηκευμένα σε πίνακες σε πολλούς κεντρικούς υπολογιστές στο σμήνος και είναι κατασκευασμένη για να αντέχει σε αποτυχίες μεμονωμένων κεντρικών υπολογιστών. Επειδή τα δεδομένα αποθηκεύονται σε HDFS, οι υγιείς κεντρικοί υπολογιστές επιλέγονται αυτόματα για να φιλοξενήσουν τα δεδομένα μόλις εξυπηρετηθούν από τον αποτυχημένο κεντρικό υπολογιστή και τα δεδομένα τίθενται αυτόματα σε λειτουργία (AMAZON, n.d.).
Cassandra	Η Cassandra είναι μια κατανεμημένη βάση δεδομένων NoSQL. Από το σχεδιασμό τους, οι βάσεις δεδομένων NoSQL είναι ελαφριές, ανοιχτού κώδικα, μη σχεσιακές και σε μεγάλο βαθμό κατανεμημένες. Στα πλεονεκτήματά τους συγκαταλέγονται η οριζόντια επεκτασιμότητα, οι κατανεμημένες αρχιτεκτονικές και η ευέλικτη προσέγγιση στον ορισμό του σχήματος. Οι βάσεις δεδομένων NoSQL επιτρέπουν την ταχεία, ad-hoc οργάνωση και ανάλυση εξαιρετικά μεγάλου όγκου, διαφορετικών τύπων δεδομένων. Αυτό έχει γίνει πιο σημαντικό τα τελευταία χρόνια, με την έλευση των Big Data και την ανάγκη ταχείας κλιμάκωσης των βάσεων δεδομένων στο cloud. Η Cassandra συγκαταλέγεται στις βάσεις δεδομένων NoSQL που έχουν αντιμετωπίσει τους περιορισμούς των προηγούμενων τεχνολογιών διαχείρισης δεδομένων, όπως οι βάσεις δεδομένων SQL. Ένα σημαντικό χαρακτηριστικό της Cassandra είναι ότι οι βάσεις δεδομένων της είναι κατανεμημένες. Αυτό αποφέρει τόσο τεχνικά όσο και επιχειρηματικά πλεονεκτήματα. Οι βάσεις δεδομένων Cassandra

	κλιμακώνονται εύκολα όταν μια εφαρμογή βρίσκεται υπό υψηλή πίεση και η κατανομή αποτρέπει επίσης την απώλεια δεδομένων από την αποτυχία υλικού οποιουδήποτε συγκεκριμένου κέντρου δεδομένων. Η κατανεμημένη αρχιτεκτονική προσφέρει επίσης τεχνική ισχύ (για παράδειγμα, ένας προγραμματιστής μπορεί να ρυθμίσει την απόδοση των ερωτημάτων ανάγνωσης ή εγγραφής σε απομόνωση). "Κατανεμημένη" σημαίνει ότι η Cassandra μπορεί να εκτελείται σε πολλαπλά μηχανήματα, ενώ εμφανίζεται στους χρήστες ως ένα ενιαίο σύνολο. Για να υπάρχει το μέγιστο όφελος από την Cassandra, θα πρέπει να την τρέχει σε πολλαπλά μηχανήματα. Δεδομένου ότι πρόκειται για μια κατανεμημένη βάση δεδομένων, η Cassandra μπορεί να έχει πολλαπλούς κόμβους. Ένας κόμβος αντιπροσωπεύει μια μεμονωμένη περίπτωση του Cassandra. Αυτοί οι κόμβοι επικοινωνούν μεταξύ τους μέσω ενός πρωτοκόλλου που ονομάζεται gossip, το οποίο είναι μια διαδικασία ομότιμης επικοινωνίας μεταξύ υπολογιστών. Η Cassandra έχει επίσης μια αρχιτεκτονική χωρίς master (οποιοσδήποτε κόμβος της βάσης δεδομένων μπορεί να παρέχει την ίδια ακριβώς λειτουργικότητα με οποιονδήποτε άλλο κόμβο), συμβάλλοντας στην ευρωστία και την ανθεκτικότητα της Cassandra. Πολλαπλοί κόμβοι μπορούν να οργανωθούν λογικά σε ένα σύμπλεγμα ή "δακτύλιο". Μπορούν επίσης να υπάρχουν πολλαπλά κέντρα δεδομένων (Apache, n.d.).  *Εικόνα 13 ApacheCassandra και Nodes [πηγή : (Apache, n.d.)]*
Hypertable	Το Hypertable είναι μια μαζικά επεκτάσιμη βάση δεδομένων υψηλής απόδοσης ανοικτού κώδικα (με πρότυπο το Bigtable). Σύγκριση με μια σχεσιακή βάση δεδομένων Το Hypertable είναι παρόμοιο με μια σχεσιακή βάση δεδομένων στο ότι αναπαριστά τα δεδομένα ως πίνακες πληροφοριών, με γραμμές και στήλες, αλλά μέχρι εκεί φτάνει η αναλογία. Ακολουθεί ένας κατάλογος με ορισμένες από τις κύριες διαφορές :

- Τα κλειδιά των γραμμών είναι συμβολοσειρές UTF-8[17]
- Καμία υποστήριξη για τύπους δεδομένων, οι τιμές αντιμετωπίζονται ως αδιαφανείς ακολουθίες byte
- Καμία υποστήριξη για ενώσεις και συναλλαγές (joins[18],transactions[19])

Οι πίνακες στο Hypertable μπορούν να θεωρηθούν ως μαζικοί πίνακες δεδομένων, ταξινομημένοι με βάση ένα μόνο πρωτεύον κλειδί, το κλειδί γραμμής.

Φυσική Διάταξη

Μια σχεσιακή βάση δεδομένων υποθέτει ότι κάθε στήλη που ορίζεται στο σχήμα του πίνακα θα έχει μια τιμή για κάθε γραμμή που υπάρχει στον πίνακα. Οι τιμές NULL αναπαρίστανται συνήθως με έναν ειδικό δείκτη (π.χ. \N). Το πρωτεύον κλειδί και το αναγνωριστικό στήλης συσχετίζονται σιωπηρά με κάθε κελί με βάση τη φυσική του θέση στη διάταξη.

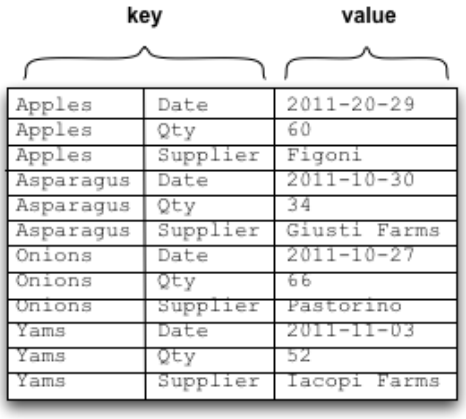
Item	Date	Qty	Supplier
Apples	2011-20-29	60	Figoni
Asparagus	2011-10-30	34	Giusti Farms
Bananas	\N	\N	\N
Cantelope	\N	\N	\N
Grapes	\N	\N	\N
Onions	2011-10-27	66	Pastorino
Oranges	\N	\N	\N
Peaches	\N	\N	\N
Pears	\N	\N	\N
Pineapples	\N	\N	\N
Plums	\N	\N	\N
Strawberries	\N	\N	\N
Yams	2011-11-03	52	Iacopi Farms

Εικόνα 14 Πίνακας σχεσιακής βάσης δεδομένων και τοποθέτηση του στον δίσκο [πηγή : (HyperTable Inc., n.d.)]

Το Hypertable βασίζεται στο σχεδιασμό του Log Structured Merge Treepdf. Επιπεδώνει τη δομή του πίνακα σε μια ταξινομημένη λίστα από ζεύγη κλειδιών/τιμών, καθένα από τα οποία αντιπροσωπεύει ένα κελί του πίνακα. Το κλειδί περιλαμβάνει το πλήρες αναγνωριστικό γραμμής και στήλης, πράγμα που σημαίνει ότι σε κάθε κελί

[17] Ένας χαρακτήρας στο UTF8 μπορεί να έχει μήκος από 1 έως 4 bytes και μπορεί να αναπαραστήσει οποιονδήποτε χαρακτήρα του προτύπου Unicode. Ο UTF-8 είναι συμβατός με τον ASCII και είναι η προτιμώμενη κωδικοποίηση για e-mail και ιστοσελίδες (w3schools, χ.χ.)
[18] Η κατασκευή της SQL που συνδυάζει δεδομένα από δύο ή περισσότερους πίνακες ονομάζεται σύνδεση (join) (oreilly., 2004)
[19] Μια συναλλαγή είναι μια ενιαία μονάδα εργασίας. Εάν μια συναλλαγή είναι επιτυχής, όλες οι τροποποιήσεις δεδομένων που πραγματοποιούνται κατά τη διάρκεια της συναλλαγής δεσμεύονται και γίνονται μόνιμο μέρος της βάσης δεδομένων. Εάν μια συναλλαγή αντιμετωπίσει σφάλματα και πρέπει να ακυρωθεί ή να ανακληθεί, τότε όλες οι τροποποιήσεις δεδομένων διαγράφονται (microsoft, 2022)

παρέχονται πλήρεις πληροφορίες διευθυνσιοδότησης. Τα κελιά που είναι NULL απλώς δεν περιλαμβάνονται στη λίστα, γεγονός που καθιστά αυτόν τον σχεδιασμό ιδιαίτερα κατάλληλο για αραιά δεδομένα.

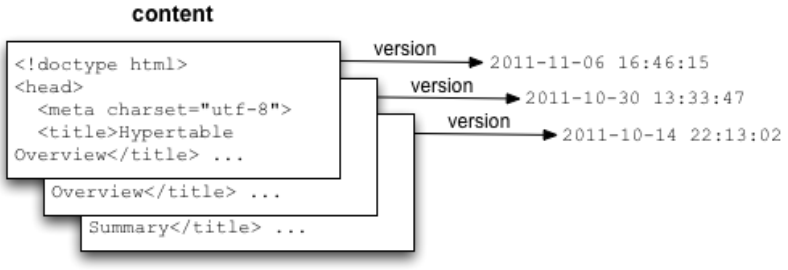


Εικόνα 15 Απεικόνιση του τρόπου που ο Hypertable αποθηκεύει δεδομένα πινάκων στον δίσκο [πηγή : (HyperTable Inc., n.d.)]

Αν και μπορεί να υπάρχει αρκετός πλεονασμός στα κλειδιά των γραμμών και τα αναγνωριστικά των στηλών, ο Hypertable χρησιμοποιεί συμπίεση δεδομένων κλειδιών-προθέματος και μπλοκ, η οποία μετριάζει σημαντικά αυτό το πρόβλημα.

Εκδόσεις κελιών

Ο Hypertable επεκτείνει το παραδοσιακό μοντέλο πίνακα δύο διαστάσεων προσθέτοντας μια τρίτη διάσταση: τη χρονοσφραγίδα. Αυτή η διάσταση timestamp μπορεί να θεωρηθεί ότι αντιπροσωπεύει διαφορετικές εκδόσεις κάθε κελιού του πίνακα, όπως απεικονίζεται στο ακόλουθο διάγραμμα.
Κατά την αναζήτηση, επιστρέφεται πρώτα η πιο πρόσφατη έκδοση κελιού. Από προεπιλογή, όλες οι εκδόσεις κελιών διατηρούνται για κάθε στήλη, αλλά ο αριθμός των εκδόσεων που διατηρούνται μπορεί να περιοριστεί καθορίζοντας την επιλογή MAX_VERSIONS στην προδιαγραφή της στήλης στην εντολή CREATE TABLE. Η χρονοσφραγίδα μπορεί να παρέχεται από την εφαρμογή κατά την εισαγωγή ή μπορεί να δημιουργείται αυτόματα (προεπιλογή).



Εικόνα 16 Hypertable και χρονοσφραγίδα [πηγή : (HyperTable Inc., n.d.)]

Προσδιορισμός Στήλης

Αυτή η λειτουργία παρέχει στους χρήστες τη δυνατότητα να εισάγουν αραιά δεδομένα στηλών που μπορούν εύκολα να επιλεγούν με τη γλώσσα ερωτημάτων υπερπίνακα (Hypertable Query Language - HQL) ή με οποιαδήποτε άλλη διεπαφή ερωτημάτων.
Μια προδιαγραφή στήλης στη δήλωση Hypertable CREATE TABLE ορίζει στην πραγματικότητα ένα σύνολο σχετικών στηλών, γνωστό ως οικογένεια στηλών. Οι χρήστες μπορούν να παρέχουν έναν προαιρετικό προσδιοριστή στήλης και να καθορίζουν την προσδιορισμένη στήλη ως family:qualifier. Το προσδιοριστικό είναι μια συμβολοσειρά με τερματισμό NUL. Για παράδειγμα, εάν σε μια δήλωση CREATE TABLE ορίζεται ένα tag οικογένειας στηλών, όπως φαίνεται παρακάτω

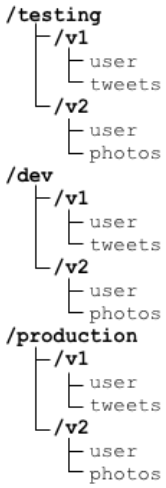
```
CREATE TABLE Info (
  tag
);
```

τότε μπορούν να δημιουργηθούν/εισάγονται στον πίνακα ειδικές στήλες όπως οι ακόλουθες

```
tag:bigtable
tag:nosql
tag:bigdata
```

Ονομαστικά Στοιχεία

Οι χώροι ονομάτων παρέχουν έναν τρόπο για τη λογική ομαδοποίηση πινάκων και είναι ανάλογοι με την ιεραρχία καταλόγων σε ένα σύγχρονο σύστημα αρχείων. Οι χώροι ονομάτων επιτρέπουν να οργανωθούν οι πίνακες σε σχετικές ομάδες, διατηρώντας τα ονόματα των πινάκων απλά, καθώς τα ονόματα των πινάκων χρειάζεται να είναι μοναδικά μόνο μέσα στο χώρο ονομάτων στον οποίο δημιουργούνται. Όλες οι περιπτώσεις Hypertable έχουν ένα ενσωματωμένο προεπιλεγμένο χώρο ονομάτων ρίζας "/". Το ακόλουθο διάγραμμα απεικονίζει ένα παράδειγμα ιεραρχίας χώρων ονομάτων.



Εικόνα 17 Ιεραρχία ονομαστικών στοιχείων Hypertable [πηγή : (HyperTable Inc., n.d.)]

Κλιμάκωση

Αυτή η ενότητα απεικονίζει τον τρόπο κλιμάκωσης του Hypertable. Το σύστημα έχει φορτωθεί με τους ακόλουθους δύο πίνακες, έναν πίνακα ID συνεδρίας και έναν πίνακα βάσης δεδομένων crawl.

Εικόνα 18 Πίνακες session και crawl [πηγή : (HyperTable Inc., n.d.)]

Με την πάροδο του χρόνου, το Hypertable θα σπάσει αυτούς τους πίνακες σε περιοχές και θα τους διανείμει σε διεργασίες που είναι γνωστές ως RangeServer. Αυτές οι διεργασίες διαχειρίζονται περιοχές δεδομένων του πίνακα και εκτελούνται σε όλες τις μηχανές του slave server στο cluster. Για παράδειγμα, υποθέτοντας ότι υπάρχουν τρεις slave servers, το ακόλουθο διάγραμμα δείχνει πώς

μπορεί να μοιάζει το σύστημα με την πάροδο του χρόνου. Όπως φαίνεται από το διάγραμμα, οι τρεις διακομιστές είναι γεμάτοι.

Εικόνα 19 Εξυπηρετητές με γεμάτη χωρητικότητα [πηγή : (HyperTable Inc., n.d.)]

Η προσθήκη μεγαλύτερης χωρητικότητας είναι ένα απλό θέμα προσθήκης νέων διακομιστών κατηγορίας commodity και εκκίνησης των διεργασιών RangeServer στα νέα μηχανήματα. Το Hypertable θα ανιχνεύσει ότι υπάρχουν διαθέσιμοι νέοι διακομιστές με άφθονη πλεονάζουσα χωρητικότητα και θα μεταφέρει αυτόματα τις σειρές από τα υπερφορτωμένα μηχανήματα στα νέα μηχανήματα.

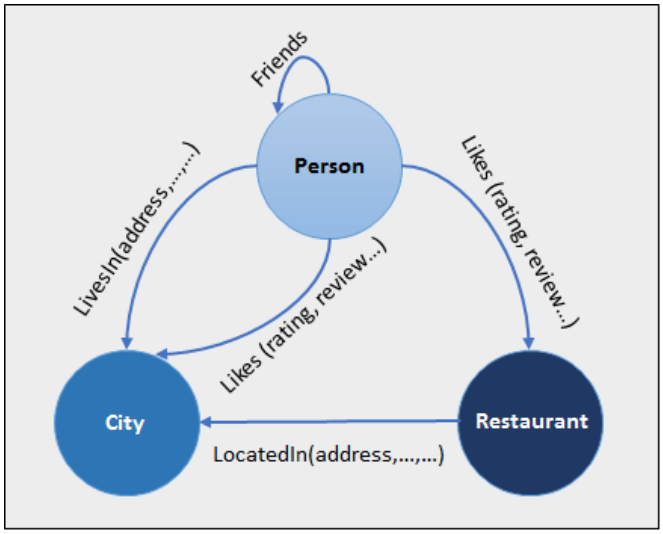
Εικόνα 20 Διαδικασία migration range (μετανάστευση εύρους) [πηγή : (HyperTable Inc., n.d.)]

Αυτή η διαδικασία μετανάστευσης εύρους (migration) έχει ως αποτέλεσμα την εξισορρόπηση του φορτίου σε ολόκληρη τη συστάδα και το άνοιγμα πρόσθετης χωρητικότητας (HyperTable Inc., n.d.).

Πίνακας 4 Παραδείγματα βάσεων δεδομένων με βάση τις στήλες

Βασισμένο σε γράφημα

Μια βάση δεδομένων τύπου γράφου αποθηκεύει οντότητες καθώς και τις σχέσεις μεταξύ αυτών των οντοτήτων. Η οντότητα αποθηκεύεται ως κόμβος με τις σχέσεις ως ακμές. Μια ακμή δίνει μια σχέση μεταξύ των κόμβων. Κάθε κόμβος και ακμή έχει ένα μοναδικό αναγνωριστικό. Σε σύγκριση με μια σχεσιακή βάση δεδομένων όπου οι πίνακες συνδέονται χαλαρά, μια βάση δεδομένων τύπου γράφου είναι πολυσχεσιακή στη φύση της. Η διέλευση των σχέσεων είναι γρήγορη, καθώς έχουν ήδη καταγραφεί στη ΒΔ και δεν χρειάζεται να υπολογιστούν. Η βάση δεδομένων γράφων χρησιμοποιείται κυρίως για κοινωνικά δίκτυα, logistics και χωρικά δεδομένα. Οι Neo4J, Infinite Graph, OrientDB, FlockDB είναι ορισμένες δημοφιλείς βάσεις δεδομένων με βάση τους γράφους.

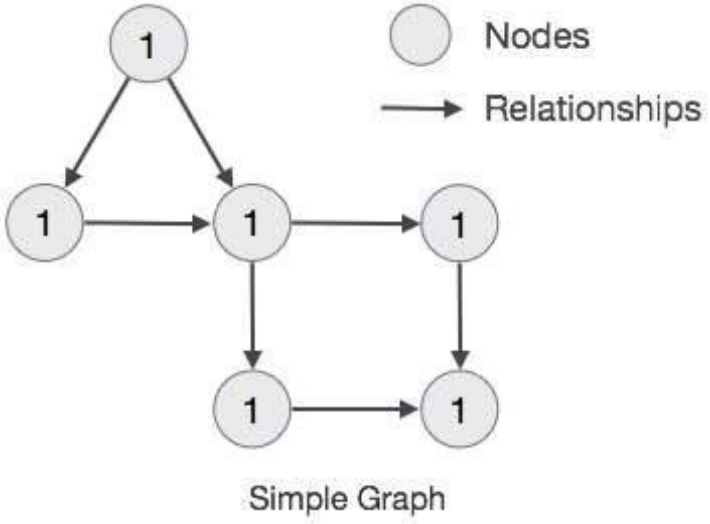


Εικόνα 21 βάση δεδομένων τύπου γράφου [πηγή : (guru99, 2022)]

Neo4j	Το Neo4j είναι ένα σύστημα διαχείρισης βάσεων δεδομένων γράφων που αναπτύχθηκε από τη Neo4j Inc. Περιγράφεται από τους προγραμματιστές του ως μια βάση δεδομένων συναλλαγών συμβατή με ACID με εγγενή αποθήκευση και επεξεργασία γράφων και είναι διαθέσιμο σε μια μη ανοιχτού κώδικα έκδοση. Το Neo4j υλοποιείται σε Java και είναι προσβάσιμο από λογισμικό γραμμένο σε άλλες γλώσσες χρησιμοποιώντας τη γλώσσα ερωτημάτων Cypher μέσω ενός συναλλακτικού τελικού σημείου HTTP ή μέσω του δυαδικού πρωτοκόλλου "Bolt". Πλεονεκτήματα του Neo4j • Ευέλικτο μοντέλο δεδομένων. Το Neo4j παρέχει ένα ευέλικτο, απλό και ταυτόχρονα ισχυρό μοντέλο δεδομένων, το οποίο μπορεί εύκολα να αλλάξει ανάλογα με τις εφαρμογές και τους κλάδους

	• Διαπιστώσεις σε πραγματικό χρόνο. Το Neo4j παρέχει αποτελέσματα με βάση δεδομένα σε πραγματικό χρόνο • Υψηλή διαθεσιμότητα. Το Neo4j είναι εξαιρετικά διαθέσιμο για μεγάλες επιχειρησιακές εφαρμογές πραγματικού χρόνου με εγγυήσεις συναλλαγών • Συνδεδεμένα και ημιδομημένα δεδομένα - Χρησιμοποιώντας το Neo4j γίνεται αναπαράσταση σε συνδεδεμένα και ημιδομημένα δεδομένα • Εύκολη ανάκτηση.Χρησιμοποιώντας τη Neo4j γίνεται εύκολα ανάκτηση των συνδεδεμένων δεδομένων ταχύτερα σε σύγκριση με άλλες βάσεις δεδομένων • Γλώσσα ερωτημάτων Cypher. Η Neo4j παρέχει μια δηλωτική γλώσσα ερωτημάτων για την οπτική αναπαράσταση του γράφου, χρησιμοποιώντας μια σύνταξη ascii-art. Οι εντολές αυτής της γλώσσας είναι σε μορφή αναγνώσιμη από τον άνθρωπο και πολύ εύκολα στην εκμάθηση • Καμία σύνδεση. Χρησιμοποιώντας το Neo4j, δεν απαιτούνται πολύπλοκες συνδέσεις για την ανάκτηση συνδεδεμένων/σχετικών δεδομένων, καθώς είναι πολύ εύκολο να γίνει ανάκτηση των λεπτομερειών των γειτονικών κόμβων ή των σχέσεων χωρίς συνδέσεις ή ευρετήρια Χαρακτηριστικά του Neo4j • Μοντέλο δεδομένων (ευέλικτο σχήμα). Το Neo4j ακολουθεί ένα μοντέλο δεδομένων που ονομάζεται native property graph model. Εδώ, ο γράφος περιέχει κόμβους (οντότητες) και αυτοί οι κόμβοι συνδέονται μεταξύ τους (απεικονίζονται με σχέσεις). Οι κόμβοι και οι σχέσεις αποθηκεύουν δεδομένα σε ζεύγη κλειδιών-τιμών γνωστά ως ιδιότητες. Στο Neo4j, δεν υπάρχει ανάγκη να ακολουθηθεί ένα σταθερό σχήμα • Ιδιότητες ACID. Το Neo4j υποστηρίζει πλήρεις κανόνες ACID (Atomicity, Consistency, Isolation, and Durability / Ατομικότητα, συνέπεια, απομόνωση και ανθεκτικότητα) • Επεκτασιμότητα και αξιοπιστία. Μπορεί να επεκταθεί η βάση δεδομένων αυξάνοντας τον αριθμό των αναγνώσεων/εγγραφών και τον όγκο χωρίς να επηρεάζεται η ταχύτητα επεξεργασίας ερωτημάτων και η ακεραιότητα των δεδομένων. Το Neo4j παρέχει επίσης υποστήριξη για αντιγραφή για την ασφάλεια και την αξιοπιστία των δεδομένων • Γλώσσα ερωτημάτων Cypher. Η Neo4j παρέχει μια ισχυρή δηλωτική γλώσσα ερωτημάτων γνωστή ως Cypher. Χρησιμοποιεί ASCII-art για την απεικόνιση γραφημάτων. Η Cypher είναι εύκολη στην εκμάθηση και μπορεί να χρησιμοποιηθεί για τη δημιουργία και ανάκτηση σχέσεων μεταξύ δεδομένων χωρίς τη χρήση πολύπλοκων ερωτημάτων όπως τα Joins • Ενσωματωμένη διαδικτυακή εφαρμογή. Η Neo4j παρέχει μια ενσωματωμένη διαδικτυακή εφαρμογή Neo4j Browser • REST API για τη συνεργασία με γλώσσες προγραμματισμού όπως

	Java, Spring, Scala κ.ά. • Indexing. Το Neo4j υποστηρίζει ευρετήρια με τη χρήση του Apache Lucence Μοντέλο δεδομένων γραφήματος ιδιοτήτων Neo4j Η Neo4j Βάση δεδομένων γραφημάτων (Graph Database) ακολουθεί το Μοντέλο γραφήματος ιδιοτήτων (Property Graph Model) για την αποθήκευση και τη διαχείριση των δεδομένων της • Το μοντέλο αναπαριστά τα δεδομένα σε κόμβους, σχέσεις και ιδιότητες • Οι ιδιότητες είναι ζεύγη κλειδιών-τιμών • Οι κόμβοι αναπαρίστανται με τη χρήση κύκλου και οι σχέσεις με τη χρήση βέλους • Οι σχέσεις έχουν κατευθύνσεις: μονοκατευθυντικές και αμφίδρομες • Κάθε σχέση περιέχει "Κόμβο έναρξης" ή "Από κόμβο" και "Προς κόμβο" ή "Κόμβο τέλους" • Τόσο οι κόμβοι όσο και οι σχέσεις περιέχουν ιδιότητες • Οι σχέσεις συνδέουν κόμβους Στο μοντέλο δεδομένων Property Graph, οι σχέσεις πρέπει να είναι κατευθυνόμενες. Αν δημιουργηθούν σχέσεις χωρίς κατεύθυνση, τότε θα εμφανιστεί μήνυμα σφάλματος. Η Neo4j Graph Database αποθηκεύει όλα τα δεδομένα της σε κόμβους και σχέσεις. Δεν χρειαζόμαστε ούτε κάποια πρόσθετη βάση δεδομένων RRBMS ούτε κάποια βάση δεδομένων SQL για να αποθηκευτούν τα δεδομένα της βάσης δεδομένων Neo4j. Αποθηκεύει τα δεδομένα της σε όρους Γραφημάτων στη μητρική της μορφή. Η Neo4j χρησιμοποιεί την εγγενή GPE (Graph Processing Engine) για να εργαστεί με την εγγενή μορφή αποθήκευσης γράφων της. Τα κύρια δομικά στοιχεία του μοντέλου δεδομένων Graph DB είναι : • Κόμβοι • Σχέσεις • Ιδιότητες

	 *Πίνακας 5 Δομικά στοιχεία μοντέλου δεδομένων Graph DB [πηγή : (tutorialspoint, n.d.)]* Στο άνωθι σχήμα έχουν αναπαρισταθεί οι κόμβους χρησιμοποιώντας κύκλους. Οι σχέσεις αναπαρίστανται με τη χρήση βελών. Οι σχέσεις είναι κατευθυνόμενες. Μπορούν να αναπαρισταθούν τα δεδομένα του Κόμβου με όρους Ιδιοτήτων (ζεύγη κλειδιών-τιμών) (tutorialspoint, χ.χ.).
InfiniteGraph	Το InfiniteGraph είναι μια κατανεμημένη βάση δεδομένων γράφων που υλοποιείται σε Java και C++ και ανήκει σε μια κατηγορία τεχνολογιών βάσεων δεδομένων NOSQL που εστιάζουν στις δομές δεδομένων γράφων. Οι προγραμματιστές χρησιμοποιούν το InfiniteGraph για να βρίσκουν χρήσιμες και συχνά κρυμμένες σχέσεις σε ιδιαίτερα συνδεδεμένα, πολύπλοκα σύνολα μεγάλων δεδομένων. Το InfiniteGraph είναι cross-platform, κλιμακούμενο, με δυνατότητα χρήσης cloud και έχει σχεδιαστεί για να διαχειρίζεται πολύ υψηλή απόδοση. Το InfiniteGraph μπορεί να εκτελέσει εύκολα και αποτελεσματικά ερωτήματα που είναι δύσκολο να εκτελεστούν, όπως η εύρεση όλων των διαδρομών ή της συντομότερης διαδρομής μεταξύ δύο στοιχείων. Το InfiniteGraph είναι κατάλληλο για εφαρμογές και υπηρεσίες που επιλύουν προβλήματα γράφων σε επιχειρησιακά περιβάλλοντα. Η γλώσσα ερωτημάτων "DO" του InfiniteGraph επιτρέπει τόσο ερωτήματα βασισμένα σε τιμές όσο και σύνθετα ερωτήματα γραφημάτων. Το InfiniteGraph υπερβαίνει τις βάσεις δεδομένων γράφων και υποστηρίζει επίσης πολύπλοκα ερωτήματα αντικειμένων. Η υιοθέτηση παρατηρείται στην ομοσπονδιακή κυβέρνηση, τις τηλεπικοινωνίες, την υγειονομική περίθαλψη, την ασφάλεια στον κυβερνοχώρο, τη μεταποίηση, τη χρηματοδότηση και τις εφαρμογές δικτύωσης (dbpedia, n.d.).
OrientDB	Το OrientDB είναι ένα NoSQL DBMS πολλαπλών μοντέλων που υποστηρίζει γραφήματα, έγγραφα, κλειδιά-τιμές και αντικειμενοστραφή αποθήκευση. Αντί να υλοποιεί απλώς ένα άλλο επίπεδο με ένα API, η OrientDB ενσωματώνει αυτά τα μοντέλα. Υποστηρίζει επίσης τόσο δισκοστραφείς όσο και αποθηκεύσεις στη μνήμη. Επιπλέον, η OrientDB υποστηρίζει σύνταξη SQL με λίγες

	διαφορές από την τυπική SQL και επεκτείνει τη σύνταξη SQL για την υποστήριξη εννοιών γράφων. Είναι επίσης ένα συμβατό με το ACID DBMS και ικανό να διαχειρίζεται συναλλακτικά φορτία εργασίας. Ιστορία Η OrientDB αναπτύχθηκε αρχικά από τον Luca Garulli το 2010. Ο Luca ξαναέγραψε το γρήγορο μόνιμο στρώμα του OrientDB ODBMS σε Java ως OrientDB. Από το 2012, το OrientDB χρηματοδοτείται από την OrientDB LTD, της οποίας ιδρυτής και διευθύνων σύμβουλος είναι ο Luca. Η OrientDB LTD είναι μια κερδοσκοπική εταιρεία, της οποίας η προηγούμενη ονομάζεται Orient Technologies LTD. Ο Andrey Lomakin ανέπτυξε εκ νέου τη μηχανή αποθήκευσης της OrientDB, που ονομάζεται plocal, από το 2012 έως το 2014. Το 2013, ο Andrey εντάχθηκε στην εταιρεία ως συνιδιοκτήτης και επικεφαλής του τμήματος E&A της OrientDB LTD. Στις 19 Σεπτεμβρίου 2017, η Callidus Software Inc. που ονομάζεται επίσης CallidusCloud εξαγόρασε την OrientDB LTD. Στις 30 Ιανουαρίου 2018, η CallidusCloud και κατά συνέπεια η OrientDB εξαγοράστηκε από τη SAP SE. Σημεία ελέγχου(checkpoint) Η OrientDB υποστηρίζει πλήρες checkpointing. Πρόκειται για ένα απλό flush cache δίσκου, το οποίο σημαίνει ότι ξεπλένει όλο το περιεχόμενο στην κρυφή μνήμη δίσκου στο δίσκο όταν καλείται το πλήρες checkpointing. Οι χρήστες μπορούν να ορίσουν προσαρμοσμένες χρονικές σφραγίδες για την εκτέλεση πλήρους checkpointing σε αυτά τα σενάρια κατά τη διάρκεια της διαμόρφωσης της μηχανής αποθήκευσης. Συμπίεση Η OrientDB υποστηρίζει συμπίεση σε επίπεδο εγγραφής. Οι εγγραφές θα αποσυμπιεστούν όταν φορτωθούν από τη μηχανή αποθήκευσης. Η συμπίεση περιλαμβάνει δύο τύπους αλγορίθμων: gzip και snappy. Η προεπιλογή είναι η μη συμπίεση. Οι χρήστες μπορούν να ορίσουν τις επιλογές συμπίεσης χρησιμοποιώντας τη σύνταξη SQL ή στη διαμόρφωση της μηχανής αποθήκευσης. Οι χρήστες μπορούν επίσης να ορίσουν προσαρμοσμένους αλγορίθμους συμπίεσης. Έλεγχος συγχρονικότητας Η OrientDB εφαρμόζει έλεγχο ταυτόχρονης χρήσης πολλαπλών εκδόσεων και ελέγχει τους περιορισμούς ακεραιότητας κατά τη δέσμευση. Είναι αισιόδοξη και η OrientDB δεν υποστηρίζει απαισιόδοξες συναλλαγές. Όταν μια συναλλαγή έχει σύγκρουση με μια άλλη, η OrientDB θα πετάξει μια εξαίρεση και η εφαρμογή μπορεί να καθορίσει αν θα τη διακόψει ή όχι. Με το Graph, η OrientDB παρέχει τρεις λειτουργίες συνέπειας. Η πρώτη λειτουργία,

	η οποία είναι η προεπιλεγμένη, διατηρεί τη συνοχή χρησιμοποιώντας συναλλαγές, ενώ οι άλλες δύο δεν χρησιμοποιούν συναλλαγές. Χρησιμοποιούν μια λειτουργία επιδιόρθωσης της βάσης δεδομένων. Η μία εκτελεί τη λειτουργία επιδιόρθωσης συγχρονισμένα με την εφαρμογή, ενώ η άλλη εκτελεί τη λειτουργία επιδιόρθωσης ασύγχρονα με την εφαρμογή. Μοντέλο δεδομένων Η OrientDB είναι ένα DBMS πολλαπλών μοντέλων. Υποστηρίζει μοντέλα γραφημάτων, εγγράφων, κλειδιών-τιμών και αντικειμενοστραφή μοντέλα. Συνδυάζει όλα τα χαρακτηριστικά των τεσσάρων μοντέλων στη μηχανή και όχι απλώς υλοποιεί ένα πρόσθετο επίπεδο API για την υποστήριξή τους. Το μοντέλο γράφου αναπαριστά δομές δικτύου που περιλαμβάνουν κορυφές που αναπαριστούν οντότητες και ακμές που δείχνουν συνδέσεις μεταξύ των κορυφών. Εκτός από τις απαραίτητες ιδιότητες για τον ορισμό των κορυφών και των ακμών, η OrientDB επιτρέπει ιδιότητες που ορίζονται από τον χρήστη τόσο για τις κορυφές όσο και για τις ακμές. Για το μοντέλο εγγράφων, η OrientDB εισάγει την έννοια "LINK" ως σχέση μεταξύ εγγράφων. Ως εκ τούτου, όταν οι χρήστες αναφέρονται σε ένα έγγραφο, όλα τα "LINK" που ορίζονται με το έγγραφο αυτό επιλύονται αυτόματα από την OrientDB αντί να επιλύονται από τους προγραμματιστές στα περισσότερα συστήματα βάσεων δεδομένων εγγράφων. Για το μοντέλο κλειδιού/τιμής, η OrientDB οργανώνει τα ζεύγη κλειδιών-τιμών παρόμοια με τα κοινά μοντέλα κλειδιών-τιμών. Η διαφορά είναι ότι η OrientDB υποστηρίζει πλουσιότερους τύπους τιμών: επιτρέπει στοιχεία γραφημάτων και έγγραφα ως τιμές. Το αντικειμενοστραφές μοντέλο προέρχεται από την έννοια του αντικειμενοστραφούς προγραμματισμού. Η OrientDB χρησιμοποιεί άμεσα την έννοια της κλάσης στον αντικειμενοστραφή προγραμματισμό για τον ορισμό εγγραφών. Υποστηρίζει την κληρονομικότητα και τον πολυμορφισμό μεταξύ των κλάσεων. Ευρετήρια Η OrientDB υποστηρίζει πέντε αλγόριθμους ευρετηρίου, οι οποίοι ανήκουν σε τρεις κατηγορίες. Επιπλέον, η OrientDB επιτρέπει στους χρήστες να ορίζουν προσαρμοσμένες μηχανές ευρετηρίου ζητώντας τους να υλοποιήσουν συγκεκριμένες κλάσεις. Ευρετήριο SBTree Το ευρετήριο SBtree είναι μια παραλλαγή του ευρετηρίου Btree με βελτιστοποιήσεις που εστιάζουν στην εισαγωγή δεδομένων και σε ερωτήματα μεγάλης εμβέλειας. Είναι ο προεπιλεγμένος τύπος ευρετηρίου της OrientDB. Ευρετήριο κατακερματισμού Η OrientDB υποστηρίζει δύο αλγορίθμους ευρετηρίου κατακερματισμού, το κανονικό ευρετήριο κατακερματισμού και το ευρετήριο αυτόματης διαχωρισμού, μια υλοποίηση κατανεμημένου πίνακα κατακερματισμού που βασίζεται στη συνάρτηση κατακερματισμού Murmur3. Και τα δύο ευρετήρια εφαρμόζουν επεκτάσιμο αλγόριθμο κατακερματισμού και δεν υποστηρίζουν ερωτήματα εύρους. Ο

	Apache Lucene Core είναι μια υλοποίηση ανεστραμμένου ευρετηρίου. Η OrientDB παρέχει ευρετήριο πλήρους κειμένου και χωρικό ευρετήριο με τη χρήση μηχανής Lucene. Η OrientDB χρησιμοποιεί SQL σύνταξη για τη διαχείριση των ευρετηρίων χρησιμοποιώντας ένα συγκεκριμένο πρόθεμα που αντιπροσωπεύει τα ευρετήρια. Η OrientDB μπορεί να ενημερώνει τα ευρετήρια αυτόματα και χειροκίνητα. Η προεπιλογή είναι χειροκίνητη. Επίπεδα απομόνωσης Η OrientDB υποστηρίζει δύο επίπεδα απομόνωσης: Read Committed και Repeatable Reads. Το προεπιλεγμένο επίπεδο απομόνωσης είναι Read Committed. Το Read Committed είναι το μόνο διαθέσιμο επίπεδο απομόνωσης όταν εκτελούνται συναλλαγές σε απομακρυσμένες βάσεις δεδομένων. Το Repeatable Reads επιτρέπεται μόνο όταν οι συναλλαγές εκτελούνται σε τοπικές βάσεις δεδομένων και καταναλώνει περισσότερη μνήμη από το Read Committed. Οι χρήστες μπορούν να αλλάξουν το επίπεδο απομόνωσης χρησιμοποιώντας το API της Java. Συνδέσεις Η OrientDB δεν υποστηρίζει τη σύνταξη join. Εισάγει την έννοια LINKS για την αναπαράσταση των σχέσεων μεταξύ οντοτήτων. Η έννοια LINKS αναφέρεται στο αναγνωριστικό της εγγραφής και ορίζεται ως δείκτης στην εγγραφή. Οι χρήστες μπορούν να διασχίσουν τα LINKS για να επιτύχουν τον ίδιο στόχο με το join. Καταγραφή Η OrientDB εφαρμόζει Write Ahead Logging (WAL). Πραγματοποιεί φυσική καταγραφή καταγράφοντας τις αλλαγές που πραγματοποιούνται στις σελίδες. Για κάθε αλλαγή σε κάθε σελίδα, η OrientDB καταγράφει το offset και το μήκος των bytes που άλλαξαν με τιμές πριν και μετά στο αρχείο καταγραφής. Σύνταξη ερωτημάτων Ο προγραμματιστής εκτέλεσης της μηχανής ερωτημάτων στην OrientDB παράγει σχέδια εκτέλεσης που αποτελούνται από στοιχεία (αντικείμενα) σε Java. Δεν μεταγλωττίζει απευθείας τα ερωτήματα σε bytecode Java. Στη συνέχεια, η OrientDB χρησιμοποιεί τη μεταγλώττιση JVM JIT. Εκτός αυτού, τα σχέδια εκτέλεσης αποθηκεύονται στην προσωρινή μνήμη για να αποφεύγεται η αναγέννηση για το ίδιο ερώτημα. Εκτέλεση ερωτήματος Η OrientDB έχει σχεδιαστεί αρχικά για να χρησιμοποιεί το μοντέλο των επαναληπτών. Ωστόσο, η OrientDB επιτρέπει σε ορισμένες

	στρατηγικές ανάκτησης να χρησιμοποιούν διανυσματικό μοντέλο. Ορισμένα στοιχεία στα σχέδια εκτέλεσης προανακτούν εγγραφές με μία μόνο κλήση και στη συνέχεια κάνουν επεξεργασία δέσμης, π.χ. αθροίσεις και ORDER BY. Αυτό το μοτίβο μπορεί να θεωρηθεί ως διανυσματικό μοντέλο. Διεπαφή ερωτημάτων Η OrientDB υποστηρίζει σύνταξη SQL με ορισμένες διαφορές από το πρότυπο SQL. Επεκτείνει επίσης την SQL ώστε να υποστηρίζει τη λειτουργικότητα γραφημάτων. Για παράδειγμα, δεν υποστηρίζει joins ή τη λέξη-κλειδί HAVING. Η OrientDB έχει επίσης τη δική της έννοια παρόμοια με τις αποθηκευμένες διαδικασίες των RDBMS. Υποστηρίζει επίσης πολλά άλλα API για την εκτέλεση ερωτημάτων για άλλα μοντέλα δεδομένων. Αρχιτεκτονική αποθήκευσης Η OrientDB υποστηρίζει βάσεις δεδομένων στη μνήμη και προσανατολισμένες στο δίσκο. Διαθέτει αντίστοιχες αφαιρέσεις για τη μνήμη και τη δισκογραφική αποθήκευση, ώστε να υποστηρίζει και τις δύο αρχιτεκτονικές αποθήκευσης. Η OrientDB υποστηρίζει επίσης βάσεις δεδομένων μεγαλύτερες από τη μνήμη. Η JVM είναι υπεύθυνη για τη διάθεση επιπλέον χώρου από την swap. Μοντέλο αποθήκευσης Η OrientDB χρησιμοποιεί τη σελίδα ως βασική μονάδα για την αποθήκευση εγγραφών. Πρόκειται για το μοντέλο αποθήκευσης N-ary. Οι εγγραφές αποθηκεύονται συνήθως σε δύο είδη σελίδων. Το πρώτο είδος σελίδων αποθηκεύει μεταδεδομένα σχετικά με τις εγγραφές, συμπεριλαμβανομένου του αναγνωριστικού εγγραφής και των δεικτών στο πραγματικό περιεχόμενο. Κάθε εγγραφή στο πρώτο είδος σελίδων έχει σταθερό μέγεθος. Το άλλο είδος σελίδων αποθηκεύει το πραγματικό περιεχόμενο των εγγραφών. Κάθε εγγραφή αποθηκεύεται ως ζεύγη κλειδιών/τιμών στο δεύτερο είδος σελίδων. Οργάνωση αποθήκευσης Οι σελίδες δεν είναι ταξινομημένες και το μέγεθος μιας σελίδας είναι 64KB. Το πραγματικό περιεχόμενο των εγγραφών αποθηκεύεται σε σελίδες. Εάν το μέγεθος μιας εγγραφής υπερβαίνει το μέγεθος μιας σελίδας, θα αποθηκευτεί σε πολλαπλές σελίδες. Αποθηκευμένες διαδικασίες Η OrientDB εισάγει την έννοια Functions παρόμοια με τη Stored Procedure. Οι χρήστες μπορούν να γράψουν συναρτήσεις σε SQL και JavaScript. Η OrientDB μπορεί να εκτελεί Functions σε SQL, Java

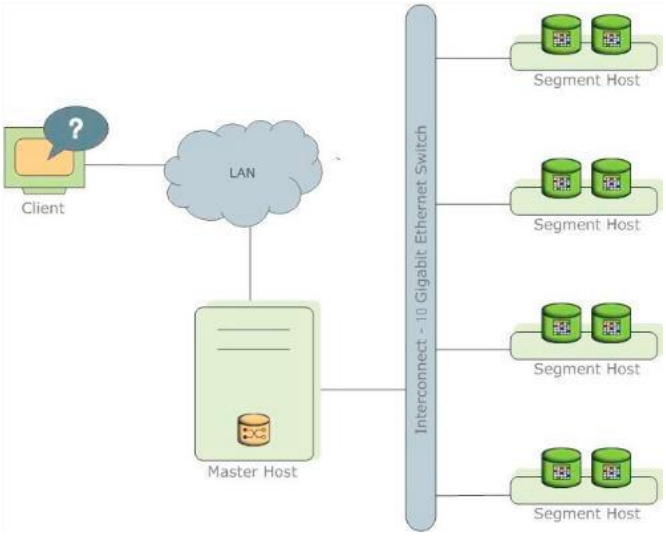
	και REST API. Αρχιτεκτονική συστήματος Η OrientDB υποστηρίζει κατανεμημένη αρχιτεκτονική multi-master και shared-nothing. Η OrientDB ενσωματώνει το έργο Hazelcast στην κατανεμημένη αρχιτεκτονική της. Χρησιμοποιεί το Hazelcast για τη διατήρηση του κύκλου ζωής κάθε κόμβου στο κατανεμημένο σύστημα. Η OrientDB χρησιμοποιεί επίσης το πρόσθετο Hazelcast για τη διαμόρφωση του κατανεμημένου συστήματος. Προβολές Η OrientDB υποστηρίζει υλοποιημένες προβολές. Χρησιμοποιεί τη σύνταξη της SQL για τη δημιουργία ή την απόρριψη προβολών. Η OrientDB υποστηρίζει materialized views μόνο για ανάγνωση και με δυνατότητα ενημέρωσης. Η προεπιλογή είναι μόνο για ανάγνωση. Για τις updatable materialized views, οι χρήστες μπορούν να ορίσουν χρονικά διαστήματα για την ενημέρωση των προβολών κάθε ορισμένο χρονικό διάστημα. Οι χρήστες μπορούν επίσης να τροποποιούν τις προβολές χειροκίνητα και η τροποποίηση θα αντικατοπτρίζεται στις αντίστοιχες εγγραφές. Οι ανανεώσιμες προβολές δεν μπορούν να δημιουργηθούν από τη συνάθροιση (OrientDB LTD, χ.χ.).

Πίνακας 6 βάσεις δεδομένων με βάση τους γράφους

(guru99, 2022)

MPP (Massively Parallel Processing/ Μαζική παράλληλη επεξεργασία)

Το MPP (Massively Parallel Processing) παρέχει μια οικονομικά αποδοτική σε ένα περιβάλλον Αποθήκης Δεδομένων που επιτρέπει στις εταιρείες να εκμεταλλευτούν την απόδοση του κόστους επισκευής και την αναλογία κάθε ενός από τους επεξεργαστές που χρησιμοποιούν το δικό τους λειτουργικό σύστημα και μνήμη, ώστε να μπορούν να εργάζονται σε διαφορετικά τμήματα του προγράμματος. Κάθε τμήμα επικοινωνεί μέσω διεπαφής ανταλλαγής μηνυμάτων. Ένα σύστημα MPP είναι επίσης γνωστό ως χαλαρά συνδεδεμένο ή κοινόχρηστο (S. Bansal, 2014). Το MPP αναφέρεται σε συστήματα με δύο ή περισσότερους επεξεργαστές που συνεργάζονται για την εκτέλεση μιας λειτουργίας, με κάθε επεξεργαστή να διαθέτει τη δική του μνήμη, το δικό του λειτουργικό σύστημα και τους δικούς του δίσκους. Το MPP αναφέρεται σε συστήματα με δύο ή περισσότερους επεξεργαστές που συνεργάζονται για την εκτέλεση μιας λειτουργίας, με κάθε επεξεργαστή να διαθέτει τη δική του μνήμη, το δικό του λειτουργικό σύστημα και τους δικούς του δίσκους. Στο κάτωθι σχήμα, το Greenplum χρησιμοποιεί αυτή την αρχιτεκτονική συστήματος υψηλής απόδοσης για να κατανέμει το φορτίο των αποθηκών δεδομένων πολλαπλών terabyte και μπορεί να χρησιμοποιεί παράλληλα όλους τους πόρους ενός συστήματος για την επεξεργασία ενός ερωτήματος.



Εικόνα 22 Αρχιτεκτονική βάσης δεδομένων υψηλού επιπέδου Greenplum [πηγή : (Suharjito, 2018)]

Η βάση δεδομένων Greenplum βασίζεται στην τεχνολογία ανοικτού κώδικα PostgreSQL. Πρόκειται ουσιαστικά για πολλές περιπτώσεις βάσεων δεδομένων PostgreSQL που δρουν μαζί ως ένα συνεκτικό σύστημα διαχείρισης βάσεων δεδομένων (DBMS). Βασίζεται στην PostgreSQL και στις περισσότερες περιπτώσεις μοιάζει πολύ με την PostgreSQL όσον αφορά την υποστήριξη SQL, τα χαρακτηριστικά, τις επιλογές διαμόρφωσης και τη λειτουργικότητα του τελικού χρήστη. Οι χρήστες της βάσης δεδομένων αλληλεπιδρούν με τη Greenplum Database όπως θα έκαναν με ένα κανονικό DBMS PostgreSQL. Τα εσωτερικά της PostgreSQL έχουν τροποποιηθεί ή συμπληρωθεί για να υποστηρίξουν την παράλληλη δομή της Greenplum Database. Για παράδειγμα, οι συνιστώσες του καταλόγου συστήματος, του βελτιστοποιητή, του εκτελεστή ερωτημάτων και του διαχειριστή συναλλαγών έχουν τροποποιηθεί και ενισχυθεί ώστε να είναι σε θέση να εκτελούν ερωτήματα ταυτόχρονα σε όλες τις παράλληλες περιπτώσεις βάσεων δεδομένων PostgreSQL. Η διασύνδεση Greenplum (το επίπεδο δικτύωσης) επιτρέπει την επικοινωνία μεταξύ των διαφορετικών περιπτώσεων PostgreSQL και επιτρέπει στο σύστημα να συμπεριφέρεται ως μία λογική βάση δεδομένων. Η βάση δεδομένων Greenplum αποθηκεύει και επεξεργάζεται μεγάλες ποσότητες δεδομένων κατανέμοντας το φόρτο εργασίας δεδομένων και επεξεργασίας σε διάφορους διακομιστές ή κεντρικούς υπολογιστές. Η Greenplum Database είναι μια συστοιχία μεμονωμένων βάσεων δεδομένων που βασίζονται στην PostgreSQL και συνεργάζονται για να παρουσιάσουν μια ενιαία εικόνα βάσης δεδομένων. Ο master είναι το σημείο εισόδου στο σύστημα Greenplum Database. Είναι η περίπτωση βάσης δεδομένων στην οποία οι πελάτες συνδέονται και υποβάλλουν δηλώσεις SQL. Ο "master" συντονίζει την εργασία του με τις άλλες περιπτώσεις βάσεων δεδομένων στο σύστημα, που ονομάζονται τμήματα, τα οποία αποθηκεύουν και επεξεργάζονται τα δεδομένα (Greenplum Database, n.d.).

Μετρητική κλιμάκωση της μαζικής παράλληλης επεξεργασίας (MPP)

Στη μέτρηση της επεκτασιμότητας της μαζικά παράλληλης επεξεργασίας (MPP) είναι η αρχιτεκτονική πολλαπλών επεξεργαστών όπου κάθε επεξεργαστής είναι μέρος ενός πλήρους συστήματος που διαθέτει μνήμη και δίσκο (Rouse, 2015). Αυτή η βάση

δεδομένων θα δημιουργήσει κατατμήσεις σε όλους τους δίσκους σε κάθε σύστημα που σχετίζεται με τη βάση δεδομένων και στη συνέχεια τα δεδομένα παρέχονται με διαφανή τρόπο για όλους τους χρήστες χρησιμοποιώντας το Greenplum έτσι ώστε να μπορεί να υπολογιστεί η επεκτασιμότητα υποστηρίζει εύκολα μεγάλους όγκους δεδομένων. Οι παράγοντες που επηρεάζουν την απόδοση της απόδοσης της βάσης δεδομένων είναι :

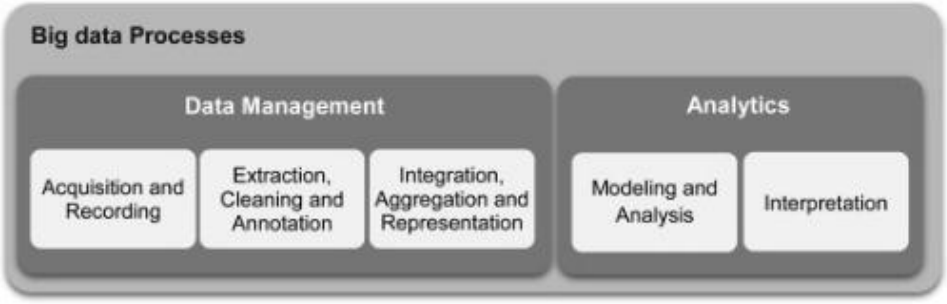
- Πόροι συστήματος. Η απόδοση της βάσης δεδομένων βασίζεται σε μεγάλο βαθμό στη χρήση δίσκου I/O και μνήμης. Για να οριστούν με ακρίβεια οι προσδοκίες απόδοσης, πρέπει να υπάρχει γνώση του υλικού στο οποίο έχει αναπτυχθεί το DBMS. Η απόδοση των στοιχείων υλικού, όπως οι CPUs, οι σκληροί δίσκοι, οι ελεγκτές δίσκων, η μνήμη RAM και οι διεπαφές δικτύου, θα επηρεάσουν σημαντικά την ταχύτητα απόδοσης της βάσης δεδομένων
- Φόρτος εργασίας. Ο φόρτος εργασίας ισούται με τη συνολική ζήτηση από το DBMS και μεταβάλλεται με την πάροδο του χρόνου. Ο συνολικός φόρτος εργασίας είναι ένας συνδυασμός ερωτημάτων χρηστών, εφαρμογών, εργασιών δέσμης, συναλλαγών και εντολών συστήματος που κατευθύνονται μέσω του DBMS ανά πάσα στιγμή. Π.χ. μπορεί να αυξάνεται όταν εκτελούνται οι αναφορές στο τέλος του μήνα ή να μειώνεται τα Σαββατοκύριακα, όταν οι περισσότεροι χρήστες λείπουν από το γραφείο. Ο φόρτος εργασίας επηρεάζει έντονα την απόδοση της βάσης δεδομένων. Η γνώση του φόρτου εργασίας σας και των περιόδων αιχμής της ζήτησης βοηθά στον προγραμματισμό για την αποδοτικότερη χρήση των πόρων του συστήματός και επιτρέπει την επεξεργασία του μεγαλύτερου δυνατού φόρτου εργασίας
- Απόδοση. Η απόδοση ενός συστήματος καθορίζει τη συνολική ικανότητά του να επεξεργάζεται δεδομένα. Η απόδοση του DBMS μετριέται σε ερωτήματα ανά δευτερόλεπτο, συναλλαγές ανά δευτερόλεπτο ή μέσους χρόνους απόκρισης. Η απόδοση DBMS συνδέεται στενά με την ικανότητα επεξεργασίας των υποκείμενων συστημάτων (είσοδος/έξοδος δίσκου, ταχύτητα CPU, εύρος ζώνης μνήμης κ.ά.), επομένως είναι σημαντικό να υπάρχει γνώση της ικανότητας απόδοσης του υλικού όταν θέτονται στόχοι απόδοσης DBMS
- Διαμάχη. Διαμάχη είναι η κατάσταση κατά την οποία δύο ή περισσότερες συνιστώσες του φόρτου εργασίας προσπαθούν να χρησιμοποιήσουν το σύστημα με αντικρουόμενο τρόπο. Π.χ. πολλαπλά ερωτήματα που προσπαθούν να ενημερώσουν το ίδιο κομμάτι δεδομένων ταυτόχρονα ή πολλαπλοί μεγάλοι φόρτοι εργασίας που ανταγωνίζονται για τους πόρους του συστήματος. Καθώς αυξάνεται ο ανταγωνισμός, μειώνεται η απόδοση
- Βελτιστοποίηση. Οι βελτιστοποιήσεις DBMS μπορούν να επηρεάσουν τη συνολική απόδοση του συστήματος. Η διατύπωση SQL, οι παράμετροι διαμόρφωσης της βάσης δεδομένων, ο σχεδιασμός των πινάκων, η κατανομή των δεδομένων κ.ά. επιτρέπουν στον βελτιστοποιητή ερωτημάτων βάσης δεδομένων να δημιουργεί τα πιο αποδοτικά σχέδια πρόσβασης

(Suharjito, 2018)

ΚΕΦΑΛΑΙΟ 2 : ΕΦΑΡΜΟΓΕΣ, ΕΡΓΑΛΕΙΑ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΤΩΝ BIG DATA

2.1 Τεχνολογίες Ανάλυσης Μεγάλων Δεδομένων

Τα μεγάλα δεδομένα είναι πολύτιμα εάν μπορούν να αναλυθούν και να χρησιμοποιηθούν για τη λήψη αποφάσεων. Για να μπορέσουν να το κάνουν αυτό αποτελεσματικά, οι οργανισμοί χρειάζονται διαδικασίες που μπορούν να μετατρέψουν πολλά γρήγορα και ποικίλα δεδομένα σε χρήσιμες πληροφορίες. Τα πέντε στάδια της ανάλυσης μεγάλων δεδομένων είναι: απόκτηση δεδομένων, αποθήκευση δεδομένων, προετοιμασία δεδομένων για ανάλυση, ανάλυση δεδομένων και απόκτηση γνώσεων. Η ανάλυση μεγάλων δεδομένων είναι μια υποδιεργασία της συνολικής διαδικασίας εξαγωγής πληροφοριών από μεγάλα δεδομένα. Υπάρχουν διάφορες διαθέσιμες τεχνικές για την ανάλυση μεγάλων δεδομένων, αλλά αυτές είναι απλώς ένα υποσύνολο των διαθέσιμων εργαλείων (Alexandros Labrinidis, 2012).



Εικόνα 23 Διαδικασίες Big Data [πηγή : (Alexandros Labrinidis, 2012)]

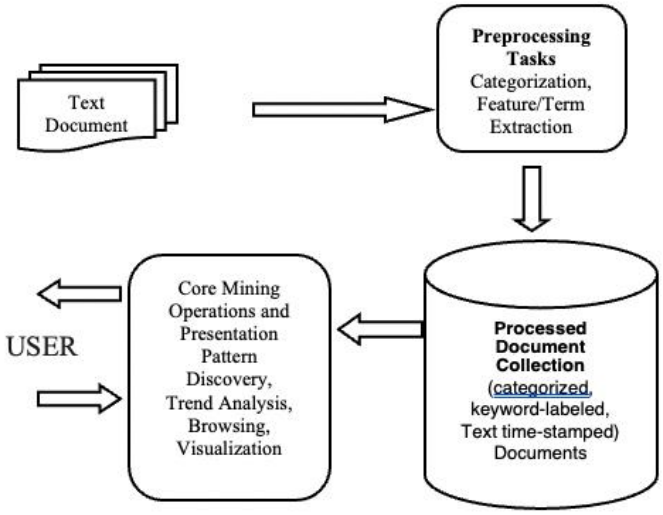
2.1.1 Ανάλυση κειμένου (Text Analytics)

Τα μεγάλα δεδομένα πρόκειται να μεταμορφώσουν τις επιχειρήσεις, αλλά η ανάλυση κειμένου θα διαδραματίσει τεράστιο ρόλο σε αυτόν τον μετασχηματισμό. Τα αδόμητα δεδομένα είναι κειμενικής φύσης, οπότε αποτελεί μεγάλο παράγοντα για την ανάλυση κειμένου στα μεγάλα δεδομένα. Η ικανότητα εξαγωγής μεταδεδομένων από μη δομημένα δεδομένα αποτελεί σημαντική ευθύνη της ανάλυσης κειμένου και παίζει τεράστιο ρόλο στο μετασχηματισμό. Η ανάλυση κειμένου είναι η διαδικασία ανάλυσης μη δομημένου κειμένου, εξαγωγής σχετικών πληροφοριών και μετασχηματισμού τους σε δομημένες πληροφορίες που μπορούν στη συνέχεια να αξιοποιηθούν με διάφορους τρόπους (Alan Nugent, 2013).

Εξόρυξη κειμένου

Η εξόρυξη είναι η ανακάλυψη μοτίβων στα δεδομένα, η εξόρυξη κειμένου είναι ένα υποπεδίο της εξόρυξης δεδομένων, που επικεντρώνεται στην ανακάλυψη γνώσης από μη δομημένα δεδομένα κειμένου. Από τα μη δομημένα δεδομένα, η διαδικασία εξαγωγής πληροφοριών και γεγονότων ονομάζεται εξόρυξη κειμένου ή ανακάλυψη γνώσης σε κείμενο (KDT). Πρόκειται για τον ανερχόμενο τομέα της ανάκτησης πληροφοριών, της στατιστικής, της μηχανικής μάθησης και της υπολογιστικής γλωσσολογίας . Οδηγεί σε εφαρμογές όπως η ανάλυση αθέτησης δανείων, η ανάλυση συναισθήματος, η εξόρυξη γνώμης, η ιατρική διάγνωση, η ηλεκτρονική ανακάλυψη κ.ά.
Η εξόρυξη κειμένου μπορεί να βοηθήσει στη λήψη δυνητικά πολύτιμων βιομηχανικών πληροφοριών από περιεχόμενο που βασίζεται σε κείμενο, όπως έγγραφα κειμένου, ηλεκτρονικό ταχυδρομείο και δημοσιεύσεις σε ροές κοινωνικών

μέσων όπως το Facebook και το Twitter. Λόγω των ασυνάρτητων δεδομένων, η εξόρυξη κειμένου με τεχνικές επεξεργασίας φυσικής γλώσσας (NLP), στατιστικής μοντελοποίησης και μηχανικής μάθησης μπορεί να είναι δύσκολη. Σε λειτουργικό επίπεδο, τα συστήματα εξόρυξης κειμένου ακολουθούν το γενικό μοντέλο που παρέχουν οι κλασικές εφαρμογές εξόρυξης δεδομένων. Οι εργασίες προεπεξεργασίας και οι βασικές λειτουργίες εξόρυξης αποτελούν τους δύο πιο σημαντικούς τομείς για κάθε σύστημα εξόρυξης κειμένου και συνήθως περιγράφουν σειριακές διαδικασίες στο πλαίσιο μιας γενικευμένης θεώρησης της αρχιτεκτονικής του συστήματος εξόρυξης κειμένου (Feldman R., 2006).



Εικόνα 24 Λειτουργική αρχιτεκτονική εξόρυξης κειμένου [πηγή : (Feldman R., 2006)]

Εξαγωγή πληροφοριών

Η εξαγωγή πληροφοριών είναι το έργο της εύρεσης δομημένων πληροφοριών από αδόμητο ή ημιδομημένο κείμενο. Αποτελεί ζωτικής σημασίας εργασία στην εξόρυξη κειμένου και έχει μελετηθεί ευρέως σε διάφορες ερευνητικές κοινότητες, όπως η επεξεργασία φυσικής γλώσσας, η ανάκτηση πληροφοριών και η εξόρυξη στον Παγκόσμιο Ιστό. Τα βασικά μέτρα για την ανάκτηση κειμένου είναι η ακρίβεια και η ανάκληση (Jiawei Han, 2011).

Προσεγγίσεις εξόρυξης κειμένου, εργαλεία και τεχνικές

Υπάρχουν πολλές προσεγγίσεις για την εξόρυξη κειμένου, οι οποίες μπορούν να ταξινομηθούν από διαφορετικές οπτικές γωνίες, με βάση τις εισροές που λαμβάνονται στο σύστημα εξόρυξης κειμένου.
Σε γενικές γραμμές, οι κυριότερες προσεγγίσεις είναι η προσέγγιση με βάση τις λέξεις-κλειδιά και η προσέγγιση εξόρυξης πληροφοριών. Οι λύσεις ανάλυσης κειμένου χρησιμοποιούν έναν συνδυασμό στατιστικής ανάλυσης και ανάλυσης περιεχομένου για την εξαγωγή πληροφοριών από μη δομημένα δεδομένα. Στατιστική ανάλυση σε κείμενο σε διάφορες διαστάσεις, όπως συχνότητα όρων, συχνότητα

εγγράφων, εγγύτητα όρων, μήκος εγγράφων. Ανάλυση περιεχομένου σε κείμενο σε διάφορα επίπεδα, όπως :

- λεξιλογική ή συντακτική επεξεργασία : αναγνώριση σημείων, κανονικοποίηση λέξεων, γλωσσικές κατασκευές, δηλαδή προτάσεις, μέρη του λόγου και παράγραφοι
- Σημασιολογική επεξεργασία : εξαγωγή νοημάτων, εξαγωγή οντοτήτων ονομάτων (κατηγοριοποίηση, σύνοψη, επέκταση ερωτημάτων και εξόρυξη κειμένου
- Πρόσθετο σημασιολογικό χαρακτηριστικό : Προσδιορισμός συναισθημάτων ή αισθημάτων (συναισθήματα, συγκινήσεις και διάθεση)
- Στόχος : Μείωση των διαστάσεων

Η υποδομή μεγάλων δεδομένων ασχολείται με το :

- Hadoop : Είναι ένα πλαίσιο προγράμματος βασισμένο σε Java που υποστηρίζει την επεξεργασία μεγάλων συνόλων δεδομένων σε ένα κατανεμημένο υπολογιστικό περιβάλλον
- Hive : Πρόκειται για μια υποδομή αποθήκευσης δεδομένων που είναι χτισμένη πάνω στο apache hadoop
- MapRaduce : Αυτό το πλαίσιο από την google και χρησιμοποιείται σε κατανεμημένο σύστημα
- Mahout : Γλώσσα μηχανής και άλλα συναφή λογισμικά όπως το Storm, το HPCC, το GridGain

Καθώς και περισσότερα διαθέσιμα εργαλεία. Διάφοροι τύποι συνόλων δεδομένων είναι διαθέσιμοι, όπως αρχεία, γραφήματα (διαδίκτυο, κοινωνικές επιστήμες και δίκτυα, μοριακή δομή), διατεταγμένα δεδομένα (ακολουθία εικόνας, ακολουθίες συναλλαγών, δεδομένα γενετικής ακολουθίας), χωρική εικόνα (χάρτης) και δεδομένα πολυμέσων (εικόνα, βίντεο). Οι βάσεις δεδομένων υπάρχουν επίσης σε ανοικτό κώδικα όπως η Cassandra από το facebook , η hbase από την apache, η mongoDB, η neo4j, η couchDB, η flockDB από το twitter και η hypertable από την Nosql κ.ο.κ. Το ανοικτό λογισμικό όπως το kttcoder, το Corrot2, το Natural language Toolkit και το GATE ενώνονται με την ανάλυση μεγάλων δεδομένων για να καλύψουν την ανάγκη των αδόμητων δεδομένων (Merlin Packiam, 2015).

Τομείς εφαρμογής στην εξόρυξη κειμένου

Οι γενικές εφαρμογές των τομέων εξόρυξης κειμένου είναι η ανάλυση σχέσεων, η ανάλυση τάσεων , οι μικτές εφαρμογές και οι επιχειρηματικές εφαρμογές των τομέων εξόρυξης κειμένου, όπως η υποστήριξη αποφάσεων στο CRM, η διαχείριση γνώσης και η εξατομίκευση στο ηλεκτρονικό εμπόριο. Υπάρχουν πολλές ευκαιρίες και προκλήσεις στην ανάλυση κειμένου μεγάλων δεδομένων, οι πιο συχνά χρησιμοποιούμενες στους ακόλουθους τομείς: Ανάλυση συναισθήματος, πρόσβαση στην αναζήτηση μη δομημένων δεδομένων, παρακολούθηση κοινωνικών μέσων ενημέρωσης, ανταγωνιστική νοημοσύνη, ηλεκτρονική ανακάλυψη, διαχείριση αρχείων, επιστημονική ανακάλυψη, ιδίως βιοεπιστήμες, εκδόσεις και μέσα μαζικής ενημέρωσης, φαρμακευτικές, ερευνητικές εταιρείες και υγειονομική περίθαλψη (Merlin Packiam, 2015).

2.1.2 Ανάλυση ήχου (Audio analytics)

Ο υπάλληλος της υπηρεσίας εξυπηρέτησης πελατών στην άλλη άκρη του τηλεφώνου μπορεί να μην αντιλαμβάνεται την αυξανόμενη οργή του καλούντος, αλλά ο υπολογιστής που καταγράφει την κλήση δεν το κάνει. Όλο και περισσότερες εταιρείες χρησιμοποιούν την επιστήμη της φωνητικής ανάλυσης για να αποκτήσουν εικόνα των αλληλεπιδράσεων με τους πελάτες, αναγνωρίζοντας ακόμη και την ανίχνευση ψεύδους. Μέχρι σχετικά πρόσφατα, ένας υπολογιστής που θα μπορούσε να κατανοεί με ακρίβεια τις προφορικές λέξεις φαινόταν σαν επιστημονική φαντασία. Σήμερα, η ανάλυση φωνής μπορεί να προχωρήσει πολύ πέρα από την κατανόηση όσων λέμε - πέρα ακόμη και από αυτά που οι άνθρωποι είναι σε θέση να ανιχνεύσουν αξιόπιστα. Αυτό που είναι εφικτό σήμερα δεν είναι μόνο η κατανόηση και η μετάφραση των προφορικών λέξεων σε κείμενο, αλλά και η ανάλυση για πράγματα όπως τα επίπεδα άγχους, τα ψέματα και πολλά άλλα.

Φωνητική αναγνώριση

Όπως το δακτυλικό αποτύπωμα, έτσι και η φωνή ενός ατόμου είναι μοναδική και δημιουργείται μόνο από τον ίδιο με βάση το σχήμα του κεφαλιού του και άλλους παράγοντες. Όταν του δίνεται ένα δείγμα ελέγχου προς ανάλυση, ένας υπολογιστής μπορεί να συγκρίνει δύο ή περισσότερες ηχογραφήσεις και να καθορίσει αν έχουν ειπωθεί από το ίδιο άτομο.
Η τεχνολογία αυτή έχει χρησιμοποιηθεί στην ασφάλεια, με κλειδαριές και συστήματα ασφαλείας που ενεργοποιούνται με τη φωνή, καθώς και στην επιβολή του νόμου και την εγκληματολογία. Ένας εκπαιδευμένος εμπειρογνώμονας ιατροδικαστής ήχου μπορεί να διαπιστώσει με μεγάλη ακρίβεια αν μια επίμαχη ηχογράφηση είναι πράγματι η φωνή ενός κατηγορούμενου ή όχι.

Customer service

Ένας σημαντικός αναπτυσσόμενος τομέας της φωνητικής ανάλυσης είναι η εξυπηρέτηση πελατών. Χρησιμοποιώντας τεχνικές μεγάλων δεδομένων σε συνδυασμό με την ανάλυση φωνής για την ανάλυση ενός τεράστιου όγκου δεδομένων κλήσεων, μια εταιρεία μπορεί να αντλήσει σημαντικές επιχειρηματικές πληροφορίες. Η φωνητική ανάλυση μπορεί να συμβάλει στη βελτίωση της απόδοσης των τηλεφωνικών κέντρων παρέχοντας πληροφορίες που μειώνουν τον χρόνο κλήσης και τις επαναλαμβανόμενες κλήσεις, παρέχουν πληροφορίες σχετικά με την ικανοποίηση των πελατών και ανταγωνιστικές πληροφορίες, μειώνουν την απομάκρυνση με την πρόβλεψη των πελατών που διατρέχουν κίνδυνο, βελτιώνουν την παρακολούθηση της ποιότητας και παρέχουν στοχευμένη καθοδήγηση σε μεμονωμένους υπαλλήλους αναλύοντας τη συγκεκριμένη απόδοσή τους.
Τα συστήματα αυτά χρησιμοποιούνται σε όλα τα σημεία επικοινωνίας με τους πελάτες. Τα βλέπουμε να χρησιμοποιούνται για την καθοδήγηση κλήσεων, ώστε να κατευθύνουν τους καλούντες στον κατάλληλο σύμβουλο μέσω της αυτόματης κατανόησης του προβλήματος (και όχι μέσω λέξεων-κλειδιών,οι οποίες οδηγούσαν στην απογοήτευση με τα παλαιότερα συστήματα). Στο άλλο άκρο της αλληλεπίδρασης με τον πελάτη, μπορούν να χρησιμοποιηθούν για τη δημιουργία αυτοματοποιημένων συστημάτων ερευνών μετά την κλήση ή μετά το συμβάν, τα οποία επιτρέπουν στον πελάτη να αφήσει ανατροφοδότηση με φυσική ομιλία, η οποία μπορεί αργότερα να αναλυθεί.

Σε συνδυασμό με τη φωνητική αναγνώριση, οι καλούντες μπορούν να αναγνωρίζονται αυτόματα από τη φωνή τους, χωρίς να χρειάζεται να εισάγουν πρόσθετες πληροφορίες αναγνώρισης. Η ανάλυση μπορεί επίσης να χρησιμοποιηθεί για την εξατομίκευση των διαδικασιών πωλήσεων, παρέχοντας πολύτιμες πληροφορίες σχετικά με τη συμπεριφορά, τις προθέσεις και τις απαιτήσεις του πελάτη.

Ανίχνευση αλήθειας

Η αναζήτηση ενός αξιόπιστου μηχανικού ανιχνευτή ψεύδους έχει ξεκινήσει από το 1900, αλλά η Nemesysco, μια εταιρεία με έδρα το Ισραήλ που ειδικεύεται σε λύσεις ανάλυσης φωνής, πιστεύει ότι μπορεί να έχει βρει μια απάντηση. Η διαδικασία τους χρησιμοποιεί την πολυεπίπεδη ανάλυση φωνής (LVA) για να εντοπίσει διαφορετικούς τύπους επιπέδων στρες, γνωστικών διαδικασιών και συναισθηματικών αντιδράσεων που, όπως λένε, αντικατοπτρίζονται στις διάφορες ιδιότητες της φωνής ενός ατόμου.
Οι περισσότεροι από τους χρήστες της εταιρείας είναι στην επιβολή του νόμου, τον στρατό, τις ασφάλειες, τις φυλακές, τα τελωνεία και τα σύνορα και την πρόληψη κλοπών, αλλά υπάρχουν και επιχειρηματικές περιπτώσεις για αυτό το είδος ανίχνευσης αλήθειας (ή ψεύδους). Η εταιρεία διαθέτει προϊόντα που μπορούν να χρησιμοποιηθούν για την επισήμανση "κακών" κλήσεων εξυπηρέτησης πελατών, όταν ο καλών αρχίζει να απογοητεύεται ή να θυμώνει, ώστε να μπορεί να επέμβει ένας επόπτης. Ένα άλλο προϊόν έχει σχεδιαστεί για την ανίχνευση ασφαλιστικών και χρηματοοικονομικών απάτης, αναλύοντας τα μοτίβα φωνής των καλούντων για ενδείξεις κινδύνου (Marr, 2016).

Επίσης τα αναλυτικά στοιχεία ήχου χρησιμοποιούνται για τη διάγνωση και τη θεραπεία ιατρικών παθήσεων, καθώς και για την ανάλυση των κραυγών ενός βρέφους για να μάθουν περισσότερα για την υγεία και τη συναισθηματική του κατάσταση (Patil, 2010).

2.1.3 Ανάλυση βίντεο (Video analytics)

Ο κύριος στόχος της ανάλυσης βίντεο είναι η αυτόματη αναγνώριση χρονικών και χωρικών γεγονότων σε βίντεο. Ένα άτομο που κινείται ύποπτα, κυκλοφοριακές πινακίδες που δεν τηρούνται, η ξαφνική εμφάνιση φλογών και καπνού, αυτά είναι μερικά μόνο παραδείγματα για το τι μπορεί να ανιχνεύσει μια λύση ανάλυσης βίντεο.

Ανάλυση βίντεο σε πραγματικό χρόνο και εξόρυξη βίντεο

Συνήθως, αυτά τα συστήματα εκτελούν παρακολούθηση σε πραγματικό χρόνο κατά την οποία ανιχνεύονται αντικείμενα, χαρακτηριστικά αντικειμένων, μοτίβα κίνησης ή συμπεριφορά που σχετίζονται με το περιβάλλον που παρακολουθείται. Ωστόσο, η ανάλυση βίντεο μπορεί επίσης να χρησιμοποιηθεί για την ανάλυση ιστορικών δεδομένων για την εξόρυξη πληροφοριών. Αυτή η εργασία εγκληματολογικής ανάλυσης μπορεί να ανιχνεύσει τάσεις και μοτίβα που απαντούν σε επιχειρηματικά ερωτήματα όπως:

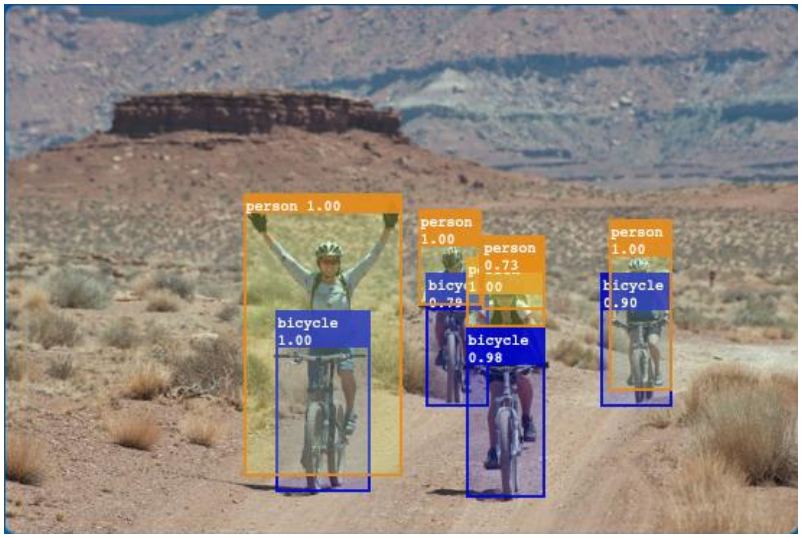
- Πότε κορυφώνεται η παρουσία των πελατών στο κατάστημά μου και ποια είναι η ηλικιακή τους κατανομή;
- Πόσες φορές παραβιάζεται ένας κόκκινος σηματοδότης και ποιες είναι οι συγκεκριμένες πινακίδες κυκλοφορίας των οχημάτων που το κάνουν;

Γνωστές εφαρμογές

Ορισμένες εφαρμογές στον τομέα της ανάλυσης βίντεο είναι ευρέως γνωστές στο ευρύ κοινό. Ένα τέτοιο παράδειγμα είναι η βιντεοεπιτήρηση, μια εργασία που υπάρχει εδώ και περίπου 50 χρόνια. Η ιδέα είναι απλή, εγκατάσταση καμερών σε στρατηγικό σημείο, ώστε οι ανθρώπινοι χειριστές να μπορούν να ελέγχουν τι συμβαίνει σε ένα δωμάτιο, μια περιοχή ή έναν δημόσιο χώρο.
Στην πράξη, ωστόσο, πρόκειται για ένα έργο που κάθε άλλο παρά απλό είναι. Ένας χειριστής είναι συνήθως υπεύθυνος για περισσότερες από μία κάμερες και, όπως έχουν δείξει αρκετές μελέτες, η αύξηση του αριθμού των καμερών που πρέπει να παρακολουθούνται επηρεάζει αρνητικά την απόδοση του χειριστή. Με άλλα λόγια, ακόμη και αν είναι διαθέσιμος μεγάλος όγκος υλικού που παράγει σήματα, δημιουργείται συμφόρηση όταν έρχεται η ώρα να επεξεργαστεί τα σήματα αυτά λόγω των ανθρώπινων περιορισμών.
Το λογισμικό ανάλυσης βίντεο μπορεί να συμβάλει σημαντικά παρέχοντας ένα μέσο για την ακριβή επεξεργασία του όγκου των πληροφοριών.

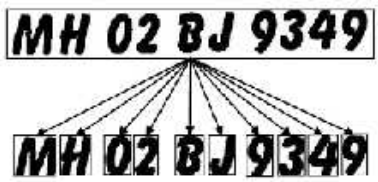
Ανάλυση βίντεο με Deep Learning

Η μηχανική μάθηση και, ειδικότερα, η θεαματική ανάπτυξη των προσεγγίσεων βαθιάς μάθησης, έχει φέρει επανάσταση στην ανάλυση βίντεο.
Η χρήση των βαθιών νευρωνικών δικτύων (DNNs) κατέστησε δυνατή την εκπαίδευση συστημάτων ανάλυσης βίντεο που μιμούνται την ανθρώπινη συμπεριφορά, με αποτέλεσμα την αλλαγή προτύπων (paradigm). Η αρχή έγινε με συστήματα που βασίζονται σε κλασικές τεχνικές όρασης υπολογιστών (π.χ. ενεργοποίηση συναγερμού αν η εικόνα της κάμερας γίνει πολύ σκοτεινή ή αλλάξει δραστικά) και μετακινήθηκε σε συστήματα ικανά να αναγνωρίζουν συγκεκριμένα αντικείμενα σε μια εικόνα και να παρακολουθούν τη διαδρομή τους.



Εικόνα 25 Ανίχνευση ποδηλάτων με την εργαλειοθήκη βαθιάς μάθησης Luminoth

Για παράδειγμα, η οπτική αναγνώριση χαρακτήρων (OCR) χρησιμοποιείται εδώ και δεκαετίες για την εξαγωγή κειμένου από εικόνες. Κατ' αρχήν, θα μπορούσε να αρκεί η εφαρμογή αλγορίθμων OCR απευθείας στην εικόνα μιας πινακίδας κυκλοφορίας για να διακρίνει τον αριθμό της. Στο προηγούμενο παράδειγμα, αυτό θα μπορούσε να λειτουργήσει εάν η κάμερα ήταν τοποθετημένη με τέτοιο τρόπο ώστε, κατά τη στιγμή της εκτέλεσης του OCR, να είμαστε σίγουροι ότι βιντεοσκοπούσαμε μια πινακίδα κυκλοφορίας.
Μια πραγματική εφαρμογή αυτού του τρόπου θα ήταν η αναγνώριση πινακίδων σε χώρους στάθμευσης, όπου η κάμερα βρίσκεται κοντά στις πύλες και θα μπορούσε να κινηματογραφήσει την πινακίδα όταν το αυτοκίνητο σταματήσει. Ωστόσο, η συνεχής εκτέλεση του OCR σε εικόνες από μια κάμερα κυκλοφορίας δεν είναι αξιόπιστη, καθώς εάν το OCR επιστρέψει ένα αποτέλεσμα είναι αβέβαιο εάν αυτό αντιστοιχεί σε πινακίδα κυκλοφορίας.
Τα μοντέλα ωστόσο που βασίζονται στη βαθιά μάθηση είναι σε θέση να προσδιορίσουν την ακριβή περιοχή μιας εικόνας στην οποία εμφανίζονται πινακίδες κυκλοφορίας. Με αυτές τις πληροφορίες, το OCR εφαρμόζεται μόνο στην ακριβή περιοχή που αφορά, οδηγώντας σε αξιόπιστα αποτελέσματα.

Αυτόματη αναγνώριση πινακίδων κυκλοφορίας με χρήση βαθιάς μάθησης
 *Εικόνα 26 Ανίχνευση αυτοκινήτων στο πλαίσιο της εικόνας [πηγή : (V. Gnanaprakash, 2021)*  *Εικόνα 27 Ανίχνευση πινακίδας κυκλοφορίας σε εικόνα αυτοκινήτου [πηγή : (V. Gnanaprakash, 2021)]*  *Εικόνα 28 Τμηματοποίηση χαρακτήρων [πηγή : (V. Gnanaprakash, 2021)]*

Πίνακας 7 Αυτόματη αναγνώριση πινακίδων κυκλοφορίας με χρήση βαθιάς μάθησης

Υγειονομική περίθαλψη	Ιστορικά, τα ιδρύματα υγειονομικής περίθαλψης έχουν επενδύσει μεγάλα χρηματικά ποσά σε λύσεις βιντεοεπιτήρησης για να διασφαλίσουν την ασφάλεια των ασθενών, του προσωπικού και των επισκεπτών τους, σε επίπεδα που συχνά διέπονται από αυστηρή νομοθεσία. Η κλοπή, η απαγωγή βρεφών και η εκτροπή φαρμάκων είναι μερικά από τα πιο συνηθισμένα προβλήματα που αντιμετωπίζονται από τα συστήματα επιτήρησης. Εκτός από τη διευκόλυνση των εργασιών επιτήρησης, η ανάλυση βίντεο μας επιτρέπει να προχωρήσουμε παραπέρα, αξιοποιώντας τα δεδομένα που συλλέγονται για την επίτευξη επιχειρηματικών στόχων. Για παράδειγμα, μια λύση ανάλυσης βίντεο θα μπορούσε να ανιχνεύσει πότε ένας ασθενής δεν έχει ελεγχθεί σύμφωνα με τις ανάγκες του και να ειδοποιήσει το προσωπικό. Η ανάλυση της κίνησης ασθενών και επισκεπτών μπορεί να είναι εξαιρετικά πολύτιμη για τον καθορισμό τρόπων μείωσης των χρόνων αναμονής, εξασφαλίζοντας παράλληλα την ελεύθερη πρόσβαση στον χώρο έκτακτης ανάγκης. Η κατ' οίκον παρακολούθηση ηλικιωμένων ενηλίκων ή ατόμων με προβλήματα υγείας είναι ένα άλλο παράδειγμα εφαρμογής που παρέχει μεγάλη αξία. Για παράδειγμα, οι πτώσεις αποτελούν σημαντική αιτία τραυματισμών και θανάτων σε ηλικιωμένους. Παρόλο που οι προσωπικές ιατρικές συσκευές μπορούν να ανιχνεύσουν τις πτώσεις, πρέπει να φοριούνται και συχνά αγνοούνται από τον καταναλωτή. Μια λύση ανάλυσης βίντεο μπορεί να επεξεργαστεί τα σήματα των οικιακών καμερών για να ανιχνεύσει σε πραγματικό χρόνο αν ένα άτομο έχει πέσει. Με την κατάλληλη ρύθμιση, ένα τέτοιο σύστημα θα μπορούσε επίσης να προσδιορίσει αν ένα άτομο πήρε ένα συγκεκριμένο φάρμακο όταν έπρεπε (briefcam, n.d.). Η ψυχική περίθαλψη είναι ένας άλλος τομέας στον οποίο η ανάλυση βίντεο μπορεί να συμβάλει σημαντικά. Μπορούν να αναπτυχθούν συστήματα που αναλύουν τις εκφράσεις του προσώπου, τη στάση του σώματος και το βλέμμα για να βοηθήσουν τους κλινικούς ιατρούς στην αξιολόγηση των ασθενών. Ένα τέτοιο σύστημα είναι σε θέση να ανιχνεύει συναισθήματα από τη γλώσσα του σώματος και τις μικροεκφράσεις, προσφέροντας στους κλινικούς γιατρούς αντικειμενικές πληροφορίες που μπορούν να επιβεβαιώσουν τις υποθέσεις τους ή να τους δώσουν νέα στοιχεία. Το Πανεπιστήμιο του Μπάφαλο ανέπτυξε μια εφαρμογή για smartphone που έχει σχεδιαστεί για να βοηθήσει στην ανίχνευση της διαταραχής του φάσματος του αυτισμού (ASD) στα παιδιά. Χρησιμοποιώντας μόνο την κάμερα του smartphone, η εφαρμογή παρακολουθεί την έκφραση του προσώπου και την προσοχή του βλέμματος ενός παιδιού που κοιτάζει εικόνες κοινωνικών σκηνών (που δείχνουν πολλά άτομα). Η εφαρμογή παρακολουθεί τις κινήσεις των ματιών και μπορεί να ανιχνεύσει με ακρίβεια τα παιδιά με ASD, καθώς οι κινήσεις των ματιών τους είναι διαφορετικές από εκείνες ενός ατόμου χωρίς αυτισμό (Nealon,

	2018).
Έξυπνες πόλεις / Μεταφορές	Η ανάλυση video έχει αποδειχθεί τεράστια βοήθεια στον τομέα των μεταφορών, βοηθώντας στην ανάπτυξη έξυπνων πόλεων. Η αύξηση της κυκλοφορίας, ιδίως στις αστικές περιοχές, μπορεί να οδηγήσει σε αύξηση των ατυχημάτων και των κυκλοφοριακών συμφορήσεων, εάν δεν ληφθούν κατάλληλα μέτρα διαχείρισης της κυκλοφορίας. Οι έξυπνες λύσεις ανάλυσης βίντεο μπορούν να διαδραματίσουν καθοριστικό ρόλο σε αυτό το σενάριο. Η ανάλυση της κυκλοφορίας μπορεί να χρησιμοποιηθεί για τη δυναμική προσαρμογή των συστημάτων ελέγχου των φωτεινών σηματοδοτών και για την παρακολούθηση των κυκλοφοριακών συμφορήσεων. Μπορεί επίσης να είναι χρήσιμη για τον εντοπισμό επικίνδυνων καταστάσεων σε πραγματικό χρόνο, όπως ένα όχημα που έχει σταματήσει σε μη εξουσιοδοτημένο χώρο στον αυτοκινητόδρομο, κάποιος που οδηγεί σε λάθος κατεύθυνση, ένα όχημα που κινείται ακανόνιστα ή οχήματα που έχουν υποστεί ατύχημα. Σε περίπτωση ατυχήματος, τα συστήματα αυτά είναι χρήσιμα για τη συλλογή αποδεικτικών στοιχείων σε περίπτωση δικαστικής διαμάχης. Η καταμέτρηση οχημάτων, ή η διαφοροποίηση μεταξύ αυτοκινήτων, φορτηγών, λεωφορείων, ταξί κ.ά., παράγει στατιστικά στοιχεία υψηλής αξίας που χρησιμοποιούνται για την απόκτηση πληροφοριών σχετικά με την κυκλοφορία. Η εγκατάσταση καμερών ταχύτητας επιτρέπει τον ακριβή έλεγχο των οδηγών μαζικά (Axians, n.d.). Η αυτόματη αναγνώριση πινακίδων κυκλοφορίας εντοπίζει τα αυτοκίνητα που διαπράττουν παράβαση ή, χάρη στην αναζήτηση σε πραγματικό χρόνο, εντοπίζει ένα όχημα που έχει κλαπεί ή έχει χρησιμοποιηθεί σε έγκλημα. Αντί να χρησιμοποιούνται αισθητήρες σε κάθε θέση στάθμευσης, ένα έξυπνο σύστημα στάθμευσης που βασίζεται στην ανάλυση βίντεο βοηθά τους οδηγούς να βρουν μια κενή θέση αναλύοντας εικόνες από κάμερες ασφαλείας. Η πόλη Ashdod στο Ισραήλ εφάρμοσε ένα έξυπνο σύστημα παρακολούθησης της κυκλοφορίας από την εταιρεία viisights για τη βελτίωση της οδικής ασφάλειας. Το σύστημα ανάλυσης βίντεο σε πραγματικό χρόνο, το οποίο αξιοποιεί το πλαίσιο ευφυούς ανάλυσης βίντεο NVIDIA Metropolis, παρακολουθεί ζωντανές ροές βίντεο από διασταυρώσεις, διαβάσεις, δρόμους και οδούς για να παρέχει εξελιγμένη συμπεριφορική κατανόηση των ενεργειών και των συμβάντων της κυκλοφορίας. Αυτό επιτρέπει στην πόλη να αντιμετωπίζει γρήγορα περιστατικά όπως ατυχήματα και κυκλοφοριακές παραβάσεις. Εκτός από την ενίσχυση της οδικής ασφάλειας, η λύση βοηθά επίσης το Ashdod να διατηρεί την κυκλοφορία σε ροή και να αποτρέπει την κυκλοφοριακή συμφόρηση (Scheyer, 2021).

2.1.4 Ανάλυση κοινωνικών δικτύων (Social Media Analytics)

Ο τύπος της ανάλυσης στα μέσα κοινωνικής δικτύωσης ποικίλλει ανάλογα με τις πηγές δεδομένων και τα μοτίβα αναζήτησης που τροφοδοτούν κάθε διαδικασία. Οι κατηγορίες που μπορούμε να χωρίσουμε τους τύπους Analytics είναι :

- Ανάλυση
- Ακρόαση
- Αναλυτικά στοιχεία διαφήμισης
- CMS Analytics
- CRM Analytics

Αρκετές εταιρείες δημιουργούν τη δική τους δομή για την ανάλυση και αυτές οι διαδικασίες περιλαμβάνουν την ενσωμάτωση δεδομένων από διάφορους τομείς μιας επιχείρησης και από συγκεκριμένα σημεία δεδομένων μέσα σε ψηφιακά περιουσιακά στοιχεία που έχουν δημιουργηθεί για συγκεκριμένους σκοπούς. Οποιοδήποτε ψηφιακό σημείο επαφής της εταιρείας μπορεί να βελτιστοποιηθεί ώστε να παρέχει πληροφορίες μέσω της ανάλυσης και να γίνει ακόμη και μέρος ενός αυτοματοποιημένου βρόχου ανατροφοδότησης που βελτιστοποιεί τις διαδικασίες και τις προσφορές με βάση τα δεδομένα που συλλέγονται. Αυτό σημαίνει ότι ορισμένα αναλυτικά στοιχεία ενσωματώνονται στο λογισμικό με τρόπο ώστε το πρόγραμμα να μαθαίνει τι είναι καλύτερο και να αρχίζει να εφαρμόζει αυτή τη μάθηση σε αυτό που κάνει.

Analytics

Όταν η πηγή των δεδομένων είναι αποκλειστικά τα κοινωνικά κανάλια που προσθέτουμε σε ένα εργαλείο ανάλυσης, τα δεδομένα αυτά εμπίπτουν στην κατηγορία των analytics. Οι πηγές δεδομένων που περιλαμβάνονται εδώ είναι το περιεχόμενο που δημοσιεύει το κανάλι, οι αλληλεπιδράσεις που σχετίζονται με το περιεχόμενο που δημοσιεύεται, ο αριθμός των ακολούθων και κάποιες πληροφορίες σχετικά με αυτούς. Ωστόσο, είναι σημαντικό να καταστεί σαφές ότι ο όρος "analytics" χρησιμοποιείται μόνο ως ονομασία αυτής της κατηγορίας ελλείψει καλύτερου όρου. Είναι επίσης ο όρος που χρησιμοποιείται στην αγορά από τα εργαλεία που προσφέρουν το συγκεκριμένο σύνολο δεδομένων. Όλοι οι τύποι analytics που εξετάζουμε αναφέρονται στους επίσημους όρους που χρησιμοποιούνται από την αγορά για την περιγραφή ή τον ορισμό τους. Ο λόγος πίσω από τους διαφορετικούς τύπους analytics που κυκλοφορούν στην αγορά σχετίζεται κυρίως με τον τρόπο που είναι δομημένα τα δίκτυα κοινωνικής δικτύωσης. Υπάρχουν διαφορετικά σημεία δεδομένων στα κοινωνικά δίκτυα που τροφοδοτούν διαφορετικά εργαλεία και πλατφόρμες στην αγορά, ή ακόμη και διαφορετικά χαρακτηριστικά ενός ίδιου εργαλείου, όπως βλέπουμε να συμβαίνει με τα υβριδικά εργαλεία. Τα κοινωνικά δίκτυα προσφέρουν πολλά διαφορετικά σημεία σύνδεσης για τα δεδομένα τους. Ορισμένα εργαλεία συνδέονται μόνο σε ένα ή λίγα από αυτά τα σημεία σύνδεσης. Έτσι, παρακολουθώντας τις ετικέτες που δίνει η αγορά σε αυτά τα εργαλεία, φιλτράρουμε επίσης τα εργαλεία ανάλογα με το είδος των δεδομένων που χρειαζόμαστε, ώστε να μπορούμε να βρούμε γρήγορα το καλύτερο εργαλείο.

Δεδομένα που μπορούμε να περιμένουμε να βρούμε σε ένα εργαλείο ανάλυσης:

- Μέγεθος κοινού και ανάπτυξη ενός καναλιού
- Όλο το περιεχόμενο που δημοσιεύεται από ένα κανάλι
- Όλες οι αλληλεπιδράσεις με το δημοσιευμένο περιεχόμενο
- Οι κορυφαίοι αλληλεπιδρώντες ακόλουθοι από το κοινό του συγκεκριμένου καναλιού
- Χρονική προβολή των μετρήσεων: ωριαία, ημερήσια, εβδομαδιαία, μηνιαία
- Συγκριτική αξιολόγηση των δεδομένων με τα κανάλια των ανταγωνιστών

Με ένα εργαλείο ανάλυσης, ένας αναλυτής μπορεί εύκολα να κατανοήσει σημαντικά σημεία, όπως το τι ενδιαφέρει τους χρήστες και τον καλύτερο τρόπο για να τους παραδώσει αυτό το περιεχόμενο. Τα επαγγελματικά εργαλεία ανάλυσης επιτρέπουν να γίνετε συγκριτική αξιολόγηση του ανταγωνισμού, χαρακτηριστικά που μας επιτρέπουν να συγκρίνουμε τις αναλύσεις μας με άλλες σελίδες και κανάλια.

Ακρόαση κοινωνικών μέσων: Ανάλυση με βάση λέξεις-κλειδιά και αναφορές

Η ακρόαση μέσω κοινωνικών μέσων έλαβε αυτό το όνομα επειδή σχετίζεται με το ότι ο αναλυτής είναι σε θέση να ακροαστεί για την άποψη του κόσμου σχετικά με το εμπορικό σήμα μέσω των καναλιών κοινωνικών μέσων. Η διαδικασία ξεκινά με μια αναζήτηση με βάση λέξεις-κλειδιά και η πηγή δεδομένων σε αυτή την περίπτωση είναι κάθε πιθανή πηγή στην οποία εντοπίζεται μια αναφορά της μάρκας (ή οποιαδήποτε λέξη-κλειδί που αναζητείται). Η ακρόαση είναι μια διαδικασία που μπορεί εύκολα να συσχετιστεί με αυτό που μπορεί να κάνει μια αναζήτηση στο Google. Είναι παρόμοια με τη Google ως προς τον τρόπο με τον οποίο εκτελεί μια αναζήτηση σε όλο το Διαδίκτυο, αλλά αντ' αυτού επικεντρώνεται στην εύρεση πληροφοριών από τα κανάλια των μέσων κοινωνικής δικτύωσης. Ορισμένα εργαλεία ακρόασης υπερβαίνουν τα μέσα κοινωνικής δικτύωσης και συλλέγουν πληροφορίες από κανάλια ειδήσεων και από μη κοινωνικά κανάλια, όπως ιστότοποι. Ο στόχος αυτών των πρόσθετων πηγών είναι να ενισχύσουν το πλαίσιο της ανάλυσης. Η διαδικασία στην ακρόαση βασίζεται στη χρήση λέξεων-κλειδιών ή εκφράσεων για την αναζήτηση. Μετά την ενεργοποίηση από τα αποτελέσματα αναζήτησης, τα περισσότερα εργαλεία ακρόασης προσθέτουν στη συνέχεια μερικές διεργασίες για να εμπλουτίσουν τα δεδομένα με περισσότερες λεπτομέρειες.
Ορισμένα από τα χαρακτηριστικά αυτών των διεργασιών σχετίζονται με :

- Δημογραφικά στοιχεία: φύλο, ηλικία, τοποθεσία
- Ενδιαφέροντα των ατόμων που αναφέρονται
- Συναίσθημα των αναφορών: θετικό, ουδέτερο ή αρνητικό
- Παράγοντες επιρροής, όπως ο αριθμός των ακολούθων των ανθρώπων που κάνουν αναφορές ή το πόσο σχετικοί είναι με το εμπορικό σήμα με βάση το τι συζητούν ή τον αριθμό των ανθρώπων που αλληλεπιδρούν με το περιεχόμενό τους

Αναλυτικά στοιχεία διαφήμισης

Τα μέσα κοινωνικής δικτύωσης είναι αυστηρά ένα κανάλι διαφήμισης για πολλές μάρκες. Πολλοί έμποροι το αντιμετωπίζουν απλά ως τέτοιο και είναι πολύ προσανατολισμένοι στις μετατροπές και στην απόδοση της επένδυσης (ROI) των εκστρατειών τους.
Η πληρωμένη προώθηση στα μέσα κοινωνικής δικτύωσης είναι χρήσιμη για όλους όσοι προσπαθούν να έχουν μεγαλύτερη προβολή από αυτήν που προσφέρει το κανάλι οργανικά. Η πληρωμένη προώθηση μπορεί να αποτελέσει μια καλή προσθήκη σε μια οργανική στρατηγική. Μπορεί να χρησιμοποιηθεί σε κρίσιμες στιγμές για να συμβάλει στην αύξηση της αναγνωρισιμότητας του καναλιού. Τα κοινωνικά δίκτυα διαθέτουν μια σειρά επιλογών για να επενδύσουν οι έμποροι και ορισμένες από αυτές τις επιλογές λειτουργούν καλύτερα από άλλες για κάθε συγκεκριμένη περίπτωση. Σε αυτό το σημείο έρχεται η διαφημιστική ανάλυση, οι πληροφορίες που παρέχει επικεντρώνονται στο να δείξουν στον έμπορο τι λειτουργεί καλύτερα και γιατί, με βάση τα αποτελέσματα από την άμεση επένδυση σε συγκεκριμένο περιεχόμενο.

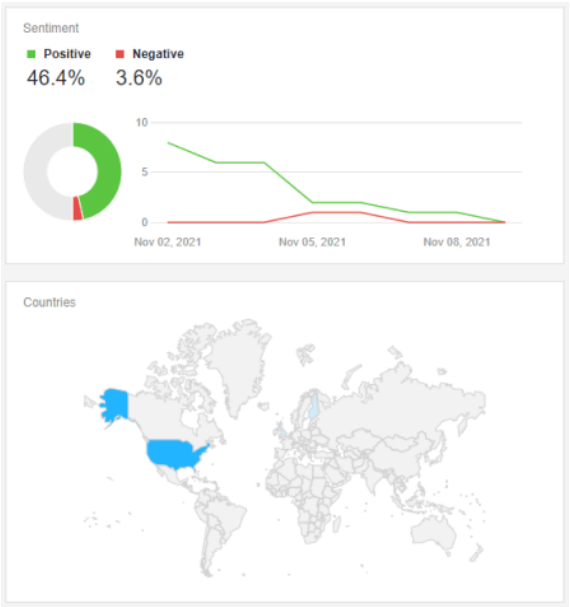
CMS Analytics

Μια άλλη μεγάλη πτυχή των μέσων κοινωνικής δικτύωσης είναι η διαχείριση του περιεχομένου σε επαγγελματικό επίπεδο. Όταν έχουμε να διατηρήσουμε μια ταυτότητα μάρκας, πολλά διαφορετικά κανάλια κοινωνικής δικτύωσης στα οποία πρέπει να δημοσιεύσουμε και μια συνεχή σημαντική ποσότητα πολύ συγκεκριμένου περιεχομένου που επιθυμούμε να δημοσιεύσουμε, χρειαζόμαστε εργαλεία που θα μας βοηθήσουν να τα καταφέρουμε.
Αυτά τα συστήματα διαχείρισης περιεχομένου (CMS) συνοδεύονται επίσης από τις δικές τους μετρήσεις.
Τα εργαλεία CMS, που αποτελούν τόσο βασικό στοιχείο για μια επαγγελματική ομάδα περιεχομένου κοινωνικών μέσων, συνήθως περιλαμβάνουν χαρακτηριστικά από διαφορετικούς τύπους εργαλείων ανάλυσης κοινωνικών μέσων. Είναι συνήθως αυτό που αποκαλούμε "υβριδικό" εργαλείο, σε αντίθεση με ένα αποκλειστικό εργαλείο που έχει μία μοναδική εστίαση όταν πρόκειται για αναλύσεις.

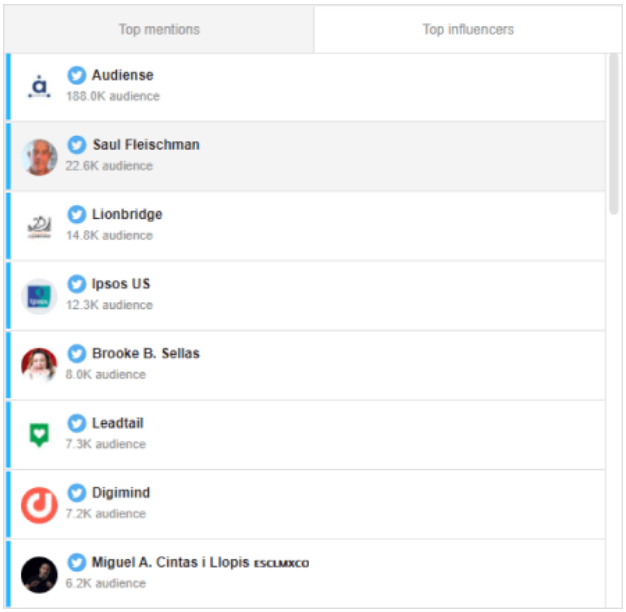
CRM Analytics

Συνήθως, όταν σκεφτόμαστε το CRM ή τη διαχείριση των πελατειακών σχέσεων, δεν σκεφτόμαστε αμέσως τα μέσα κοινωνικής δικτύωσης. Αυτό μπορεί να συμβαίνει λόγω του πόσο δύσκολο είναι να διατηρήσουμε μια μακρά συνομιλία μέσω ενός καναλιού κοινωνικών μέσων και πόσο εύκολο είναι να χάσουμε τα ίχνη των συνομιλιών που είχαμε. Συνήθως, οι συζητήσεις που σχετίζονται με την υποστήριξη πελατών και τις πωλήσεις που ξεκινούν στα μέσα κοινωνικής δικτύωσης μεταφέρονται γρήγορα σε άλλα κανάλια προκειμένου να συνεχιστούν.
Πολλά εμπορικά σήματα χρησιμοποιούν ήδη τα μέσα κοινωνικής δικτύωσης για μεγαλύτερες συνομιλίες και προχωρούν τη συζήτηση μέχρι την επιτυχή μετατροπή ή την επίλυση ενός προβλήματος πριν καλέσουν τον πελάτη να χρησιμοποιήσει ένα διαφορετικό κανάλι υποστήριξης.
Οι εταιρείες επίσης δημιουργούν bots για να αλληλεπιδρούν με τους ανθρώπους μέσω των καναλιών κοινωνικής δικτύωσης. Αυτά τα bots μπορούν να κάνουν πολλά πράγματα, από την πώληση προϊόντων μέχρι την κατεύθυνση αιτημάτων υποστήριξης πελατών και ακόμη και να ψυχαγωγούν τους ανθρώπους με παιχνίδια. Γνωστά και ως chat bots, αυτές οι εικονικές οντότητες βασίζονται σε διαφορετικά είδη τεχνολογιών

για να αλληλεπιδρούν. Κυμαίνονται από απλά και βασικά σύνολα δυνατοτήτων έως συστήματα που μαθαίνουν από τις αλληλεπιδράσεις.
Ένα πολύ συνηθισμένο κανάλι κοινωνικής δικτύωσης για τη ροή εργασιών υποστήριξης πελατών είναι το Twitter. Πολλές μάρκες διαθέτουν ειδικές ομάδες που ασχολούνται με την επικοινωνία με τους πελάτες μέσω Twitter. Το Facebook είναι επίσης ένα δίκτυο στο οποίο πολλές μάρκες προσπαθούν να αφιερώσουν μια ομάδα για να συνομιλεί με τους ανθρώπους μέσα στην ενότητα σχολίων του περιεχομένου τους ή για να απαντά σε ερωτήσεις που δημοσιεύονται στο χρονολόγιο της μάρκας ή μέσω συνομιλίας. Η πρόκληση στο Facebook είναι ότι το περιεχόμενο μπορεί να είναι μεγαλύτερο, όχι τόσο περιορισμένο όσο στο Twitter, και ο όγκος των σχολίων εντός του περιεχομένου του Facebook είναι συνήθως υπερβολικά μεγάλος για να αντιμετωπιστεί. Σε γενικές γραμμές, ο όγκος της αλληλεπίδρασης της κοινότητας είναι αυτό που περιορίζει πρωτίστως το CRM στα μέσα κοινωνικής δικτύωσης.
Το Facebook προσπαθεί να βοηθήσει σε αυτή την πτυχή της δέσμευσης, δημιουργώντας καρτέλες μηνυμάτων για συζητήσεις εντός των τμημάτων σχολίων των αναρτήσεων, έτσι ώστε οι άνθρωποι να μπορούν να συνεχίσουν τη συζήτηση χωρίς να εγκαταλείψουν το τμήμα σχολίων μιας ανάρτησης.
Ορισμένα εργαλεία CRM προσπαθούν να ενσωματώσουν τα μέσα κοινωνικής δικτύωσης με διαφορετικούς τρόπους. Τα χρησιμοποιούν για να ενισχύσουν τις πληροφορίες που έχει μια μάρκα για τους πελάτες της, για να ανοίξουν μια νέα επαφή με την ομάδα πωλήσεων, για να συνεχίσουν τη συζήτηση απευθείας από το CRM και για να διατηρήσουν τις ιστορικές πληροφορίες της σχέσης καταχωρημένες.
Είναι πιθανό οι τεχνολογίες CRM να προσθέσουν περισσότερα από τα μέσα κοινωνικής δικτύωσης καθώς εξελίσσονται. Μία από τις κύριες προκλήσεις που αντιμετωπίζουν είναι η συλλογή δεδομένων. Επειδή τα προσωπικά προφίλ υπόκεινται συνήθως σε κανονισμούς προστασίας της ιδιωτικής ζωής, μόνο λίγα κοινωνικά δίκτυα, όπως το Twitter, προσφέρουν προσωπικά δεδομένα σε δημόσιο επίπεδο (Gonçalves, 2017).



Πίνακας 8 Παρακολούθηση στατιστικών σε ζωντανό χρόνο με χρήση εργαλείου social media analytics [πηγή : (awario, n.d.)]



Πίνακας 9 Εποπτεία των ατόμων με επιρροή στα μέσα κοινωνικής δικτύωσης που έχουν αναφέρει έστω και μία επιθυμητή λέξη κλειδί [πηγή : (awario, n.d.)]

2.1.5 Προγνωστική ανάλυση (predictive analytics)

Η προγνωστική ανάλυση είναι ένας κλάδος της προηγμένης ανάλυσης που κάνει προβλέψεις για μελλοντικά αποτελέσματα χρησιμοποιώντας ιστορικά δεδομένα σε συνδυασμό με στατιστική μοντελοποίηση, τεχνικές εξόρυξης δεδομένων και μηχανική μάθηση. Οι εταιρείες χρησιμοποιούν την προγνωστική ανάλυση για να βρουν μοτίβα σε αυτά τα δεδομένα για να εντοπίσουν κινδύνους και ευκαιρίες. Η προγνωστική ανάλυση συνδέεται συχνά με τα μεγάλα δεδομένα και την επιστήμη των δεδομένων.
Σήμερα, οι εταιρείες σήμερα κατακλύζονται από δεδομένα, από αρχεία καταγραφής έως εικόνες και βίντεο, και όλα αυτά τα δεδομένα βρίσκονται σε ανομοιογενή αποθετήρια δεδομένων σε ολόκληρο τον οργανισμό. Για να αποκτήσουν πληροφορίες από αυτά τα δεδομένα, οι επιστήμονες δεδομένων χρησιμοποιούν αλγόριθμους βαθιάς μάθησης και μηχανικής μάθησης για να βρουν μοτίβα και να κάνουν προβλέψεις για μελλοντικά γεγονότα. Ορισμένες από αυτές τις στατιστικές τεχνικές περιλαμβάνουν μοντέλα λογιστικής και γραμμικής παλινδρόμησης, νευρωνικά δίκτυα και δέντρα αποφάσεων. Ορισμένες από αυτές τις τεχνικές μοντελοποίησης χρησιμοποιούν τις αρχικές προγνωστικές μαθήσεις για να κάνουν πρόσθετες προγνωστικές γνώσεις.

Τύποι προγνωστικών μοντέλων

Τα μοντέλα προβλεπτικής ανάλυσης έχουν σχεδιαστεί για να αξιολογούν ιστορικά δεδομένα, να ανακαλύπτουν μοτίβα, να παρατηρούν τάσεις και να χρησιμοποιούν αυτές τις πληροφορίες για να προβλέπουν μελλοντικές τάσεις. Τα δημοφιλή μοντέλα προβλεπτικής ανάλυσης περιλαμβάνουν την ταξινόμηση, την ομαδοποίηση και τα μοντέλα χρονοσειρών.

- Μοντέλα ταξινόμησης: Τα μοντέλα ταξινόμησης ανήκουν στον κλάδο των μοντέλων μηχανικής μάθησης με επίβλεψη. Αυτά τα μοντέλα κατηγοριοποιούν δεδομένα με βάση ιστορικά δεδομένα, περιγράφοντας σχέσεις εντός ενός

δεδομένου συνόλου δεδομένων. Για παράδειγμα, αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για την ταξινόμηση πελατών ή υποψήφιων πελατών σε ομάδες για σκοπούς τμηματοποίησης. Εναλλακτικά, μπορεί επίσης να χρησιμοποιηθεί για την απάντηση ερωτήσεων με δυαδικές εξόδους, όπως η απάντηση ναι ή όχι ή αληθές και ψευδές. Δημοφιλείς περιπτώσεις χρήσης για αυτό είναι η ανίχνευση απάτης και η αξιολόγηση πιστωτικού κινδύνου. Οι τύποι μοντέλων ταξινόμησης περιλαμβάνουν τη λογιστική παλινδρόμηση, τα δέντρα αποφάσεων, το τυχαίο δάσος, τα νευρωνικά δίκτυα και το Naïve Bayes

- Μοντέλα ομαδοποίησης: Τα μοντέλα ομαδοποίησης ανήκουν στη μάθηση χωρίς επίβλεψη. Ομαδοποιούν δεδομένα με βάση παρόμοια χαρακτηριστικά. Για παράδειγμα, ένας ιστότοπος ηλεκτρονικού εμπορίου μπορεί να χρησιμοποιήσει το μοντέλο για να διαχωρίσει τους πελάτες σε παρόμοιες ομάδες με βάση κοινά χαρακτηριστικά και να αναπτύξει στρατηγικές μάρκετινγκ για κάθε ομάδα. Οι συνήθεις αλγόριθμοι ομαδοποίησης περιλαμβάνουν την ομαδοποίηση k-means, την ομαδοποίηση με μετατόπιση μέσων όρων, τη χωρική ομαδοποίηση εφαρμογών με θόρυβο βάσει πυκνότητας (DBSCAN), την ομαδοποίηση με μεγιστοποίηση προσδοκίας (EM) με χρήση μοντέλων μίξης Gauss (GMM) και την ιεραρχική ομαδοποίηση
- Μοντέλα χρονοσειρών: Τα μοντέλα χρονοσειρών χρησιμοποιούν διάφορα δεδομένα εισόδου σε μια συγκεκριμένη χρονική συχνότητα, όπως ημερήσια, εβδομαδιαία, μηνιαία κτλ. Είναι σύνηθες να σχεδιάζεται η εξαρτημένη μεταβλητή με την πάροδο του χρόνου για την αξιολόγηση των δεδομένων ως προς την εποχικότητα, τις τάσεις και την κυκλική συμπεριφορά, γεγονός που μπορεί να υποδεικνύει την ανάγκη για συγκεκριμένους μετασχηματισμούς και τύπους μοντέλων. Τα μοντέλα αυτοπαλίνδρομης (AR), κινητού μέσου (MA), ARMA και ARIMA είναι συχνά χρησιμοποιούμενα μοντέλα χρονοσειρών. Ως παράδειγμα, ένα τηλεφωνικό κέντρο μπορεί να χρησιμοποιήσει ένα μοντέλο χρονοσειράς για να προβλέψει πόσες κλήσεις θα δέχεται ανά ώρα σε διαφορετικές ώρες της ημέρας

Περιπτώσεις χρήσης της προβλεπτικής ανάλυσης

Η προγνωστική ανάλυση μπορεί να αναπτυχθεί σε διάφορους κλάδους για διαφορετικά επιχειρηματικά προβλήματα. Παρακάτω παρατίθενται μερικές περιπτώσεις χρήσης στον κλάδο για να καταδείξουν πώς η προγνωστική ανάλυση μπορεί να ενημερώσει για τη λήψη αποφάσεων σε πραγματικές καταστάσεις.

- Τραπεζικές υπηρεσίες: Οι χρηματοπιστωτικές υπηρεσίες χρησιμοποιούν μηχανική μάθηση και ποσοτικά εργαλεία για την πρόβλεψη του πιστωτικού κινδύνου και τον εντοπισμό της απάτης. Ως παράδειγμα, η BondIT είναι μια εταιρεία που ειδικεύεται σε υπηρεσίες διαχείρισης περιουσιακών στοιχείων σταθερού εισοδήματος. Η ανάλυση πρόβλεψης τους επιτρέπει να υποστηρίζουν δυναμικές αλλαγές στην αγορά σε πραγματικό χρόνο, εκτός από τους στατικούς περιορισμούς της αγοράς. Αυτή η χρήση της τεχνολογίας της επιτρέπει τόσο την προσαρμογή των προσωπικών υπηρεσιών για τους πελάτες όσο και την ελαχιστοποίηση του κινδύνου
- Υγειονομική περίθαλψη: Η προγνωστική ανάλυση στην υγειονομική περίθαλψη χρησιμοποιείται για τον εντοπισμό και τη διαχείριση της φροντίδας χρονίως πασχόντων ασθενών, καθώς και για την παρακολούθηση συγκεκριμένων

λοιμώξεων, όπως η σήψη. Η Geisinger Health χρησιμοποίησε την προγνωστική ανάλυση για να εξορύξει αρχεία υγείας για να μάθει περισσότερα σχετικά με τον τρόπο διάγνωσης και θεραπείας της σήψης. Η Geisinger δημιούργησε ένα προγνωστικό μοντέλο με βάση τα αρχεία υγείας για περισσότερους από 10.000 ασθενείς που είχαν διαγνωστεί με σήψη στο παρελθόν. Το μοντέλο απέδωσε εντυπωσιακά αποτελέσματα, προβλέποντας σωστά ασθενείς με υψηλό ποσοστό επιβίωσης

- Ανθρώπινοι πόροι (HR): Οι ομάδες HR χρησιμοποιούν προγνωστικές αναλύσεις και μετρήσεις ερευνών εργαζομένων για την αντιστοίχιση των υποψήφιων υποψηφίων για εργασία, τη μείωση του κύκλου εργασιών των εργαζομένων και την αύξηση της δέσμευσης των εργαζομένων. Αυτός ο συνδυασμός ποσοτικών και ποιοτικών δεδομένων επιτρέπει στις επιχειρήσεις να μειώσουν το κόστος πρόσληψης και να αυξήσουν την ικανοποίηση των εργαζομένων, κάτι που είναι ιδιαίτερα χρήσιμο όταν οι αγορές εργασίας είναι ασταθείς
- Μάρκετινγκ και πωλήσεις: Ενώ οι ομάδες μάρκετινγκ και πωλήσεων είναι πολύ εξοικειωμένες με τις αναφορές επιχειρηματικής ευφυΐας για την κατανόηση των ιστορικών επιδόσεων των πωλήσεων, η προγνωστική ανάλυση επιτρέπει στις εταιρείες να είναι πιο προληπτικές στον τρόπο με τον οποίο εμπλέκονται με τους πελάτες τους σε όλο τον κύκλο ζωής του πελάτη. Για παράδειγμα, οι προβλέψεις αποχώρησης μπορούν να επιτρέψουν στις ομάδες πωλήσεων να εντοπίσουν νωρίτερα τους δυσαρεστημένους πελάτες, επιτρέποντάς τους να ξεκινήσουν συζητήσεις για την προώθηση της διατήρησης. Οι ομάδες μάρκετινγκ μπορούν να αξιοποιήσουν την ανάλυση προγνωστικών δεδομένων για στρατηγικές διασταυρούμενων πωλήσεων, και αυτό συνήθως εκδηλώνεται μέσω μιας μηχανής συστάσεων στον ιστότοπο μιας μάρκας
- Αλυσίδα εφοδιασμού: Οι επιχειρήσεις χρησιμοποιούν συνήθως την προγνωστική ανάλυση για τη διαχείριση των αποθεμάτων προϊόντων και τον καθορισμό στρατηγικών τιμολόγησης. Αυτός ο τύπος προγνωστικής ανάλυσης βοηθά τις εταιρείες να ανταποκριθούν στη ζήτηση των πελατών χωρίς να υπερφορτώνουν τις αποθήκες. Επιτρέπει επίσης στις εταιρείες να αξιολογούν το κόστος και την απόδοση των προϊόντων τους με την πάροδο του χρόνου. Εάν ένα μέρος ενός συγκεκριμένου προϊόντος γίνει πιο ακριβό στην εισαγωγή, οι εταιρείες μπορούν να προβλέψουν τον μακροπρόθεσμο αντίκτυπο στα έσοδα, εάν μετακυλήσουν ή όχι το πρόσθετο κόστος στην πελατειακή τους βάση.

Οφέλη της προβλεπτικής μοντελοποίησης

Ένας οργανισμός που γνωρίζει τι να περιμένει με βάση τα πρότυπα του παρελθόντος έχει επιχειρηματικό πλεονέκτημα στη διαχείριση των αποθεμάτων, του εργατικού δυναμικού, των εκστρατειών μάρκετινγκ και των περισσότερων άλλων πτυχών της λειτουργίας.

- Ασφάλεια: Κάθε σύγχρονος οργανισμός πρέπει να ενδιαφέρεται για τη διατήρηση της ασφάλειας των δεδομένων. Ο συνδυασμός αυτοματοποίησης και προγνωστικής ανάλυσης βελτιώνει την ασφάλεια. Συγκεκριμένα μοτίβα που σχετίζονται με ύποπτη και ασυνήθιστη συμπεριφορά των τελικών χρηστών μπορούν να ενεργοποιήσουν συγκεκριμένες διαδικασίες ασφαλείας
- Μείωση του κινδύνου: Εκτός από τη διατήρηση της ασφάλειας των δεδομένων, οι περισσότερες επιχειρήσεις εργάζονται για τη μείωση του προφίλ κινδύνου τους. Για παράδειγμα, μια εταιρεία που χορηγεί πιστώσεις μπορεί να χρησιμοποιήσει

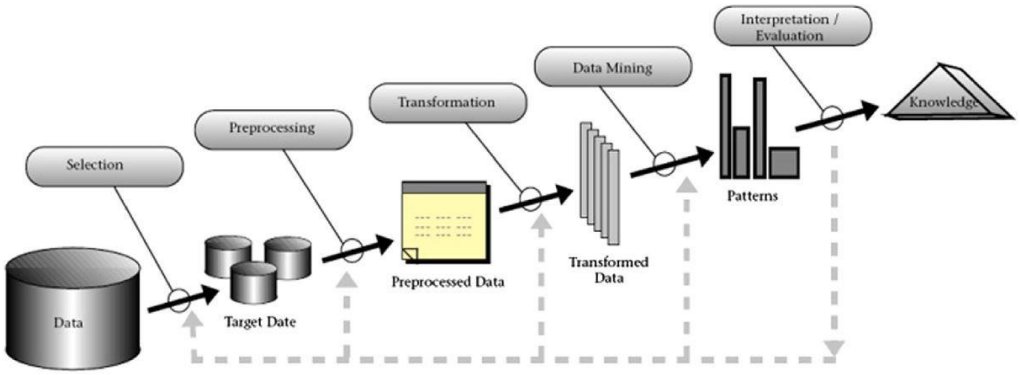
την ανάλυση δεδομένων για να κατανοήσει καλύτερα εάν ένας πελάτης ενέχει υψηλότερο από το μέσο όρο κίνδυνο αθέτησης. Άλλες εταιρείες μπορούν να χρησιμοποιούν προγνωστικά αναλυτικά στοιχεία για να κατανοήσουν καλύτερα αν η ασφαλιστική τους κάλυψη είναι επαρκής

- Λειτουργική αποτελεσματικότητα: Οι πιο αποτελεσματικές ροές εργασίας μεταφράζονται σε βελτιωμένα περιθώρια κέρδους. Για παράδειγμα, η κατανόηση του πότε ένα όχημα σε ένα στόλο που χρησιμοποιείται για διανομή θα χρειαστεί συντήρηση πριν χαλάσει στην άκρη του δρόμου σημαίνει ότι οι παραδόσεις γίνονται στην ώρα τους, χωρίς το πρόσθετο κόστος της ρυμούλκησης του οχήματος και της προσκόμισης άλλου υπαλλήλου για να ολοκληρωθεί η παράδοση
- Βελτιωμένη λήψη αποφάσεων: Η λειτουργία οποιασδήποτε επιχείρησης περιλαμβάνει τη λήψη υπολογισμένων αποφάσεων. Οποιαδήποτε επέκταση ή προσθήκη σε μια σειρά προϊόντων ή άλλη μορφή ανάπτυξης απαιτεί εξισορρόπηση του εγγενούς κινδύνου με το πιθανό αποτέλεσμα. Η ανάλυση πρόβλεψης μπορεί να παρέχει πληροφορίες για την ενημέρωση της διαδικασίας λήψης αποφάσεων και να προσφέρει ανταγωνιστικό πλεονέκτημα

(IBM, χ.χ.)

2.2 Εξόρυξη Δεδομένων (Data Mining)

Η εξόρυξη δεδομένων (Data Mining) μπορεί να οριστεί ως η διαδικασία εξερεύνησης πολύτιμων και κρυμμένων γνώσεων και νόμων από άγνωστα δεδομένα. Από επιχειρηματική άποψη, η εξόρυξη δεδομένων είναι η εξαγωγή, η επεξεργασία και η ανάλυση μεγάλου όγκου δεδομένων, που παράγει κάποιες αξίες πρόσβασης σε κρίσιμες πληροφορίες και γνώσεις που μπορούν να υποστηρίξουν τη λήψη επιχειρηματικών αποφάσεων (J. Wang, 2017). Η εξόρυξη δεδομένων είναι επίσης γνωστή ως ανακάλυψη γνώσης στα δεδομένα (KDD, Knowledge Discovery in Data). Η εξόρυξη δεδομένων μπορεί να οριστεί ως η διαδικασία ανακάλυψης μοτίβων από τεράστια δεδομένα και η πραγματοποίηση προβλέψεων για την απόκτηση νέων δεδομένων (Adeel Shiraz Hashmi, 2016).



Πίνακας 10 Διαδικασία KDD [πηγή : (Petar Ristoski, 2016)]

Βήματα εξόρυξης δεδομένων :

- Επιλογή δεδομένων: επιλογή των κατάλληλων δεδομένων και των σχετικών μεταβλητών, στις οποίες πρέπει να γίνει η ανακάλυψη

- Επεξεργασία δεδομένων: το βήμα αυτό αποσκοπεί στο να καταστήσει τα δεδομένα καθαρά, αντικαθιστώντας τις τιμές που λείπουν, αφαιρώντας το θόρυβο και τις ακραίες τιμές
- Μετασχηματισμός δεδομένων: αναγωγή και προβολή των δεδομένων προκειμένου να αποκτήσουν κατάλληλη μορφή ώστε να μπορούν να εφαρμοστούν οι αλγόριθμοι εξόρυξης δεδομένων
- Εξόρυξη δεδομένων: επιλογή κατάλληλης μεθόδου εξόρυξης δεδομένων (ταξινόμηση, ομαδοποίηση ή παλινδρόμηση), κατάλληλου αλγορίθμου για την εκτέλεση της εργασίας και εξαγωγή των προτύπων
- Αξιολόγηση και ερμηνεία: αυτό είναι το τελευταίο βήμα, τα μοτίβα εξάγονται και τώρα ο χρήστης ερμηνεύει και εξάγει τη γνώση από τα μοτίβα. Το βήμα αυτό περιλαμβάνει την οπτικοποίηση των εξαγόμενων προτύπων και μοντέλων ή την οπτικοποίηση των δεδομένων με τη χρήση των εξαγόμενων μοντέλων

(Petar Ristoski, 2016)

2.2.1 Αλγόριθμοι εξόρυξης δεδομένων

Στον σημερινό κόσμο των μεγάλων δεδομένων, μια μεγάλη βάση δεδομένων καθίσταται αναγκαία. Π.χ. μόνο το Facebook διαχειρίζεται 600 terabytes νέων δεδομένων κάθε μέρα και επίσης η πρωταρχική πρόκληση των μεγάλων δεδομένων είναι το πώς να αξιοποιηθούν. Τα μεγάλα δεδομένα είναι ποικίλα, αδόμητα και γρήγορα μεταβαλλόμενα όπως δεδομένα ήχου και βίντεο, αναρτήσεις στα μέσα κοινωνικής δικτύωσης, τρισδιάστατα δεδομένα ή γεωχωρικά δεδομένα.

Προσέγγιση με βάση τη στατιστική διαδικασία

Περιγραφικός τύπος στατιστικής ανάλυσης (Descriptive Type)
Η περιγραφική στατιστική ανάλυση βοηθά στην περιγραφή των δεδομένων. Αποκτά τη σύνοψη των δεδομένων με τρόπο ώστε να μπορούν να ερμηνευθούν σημαντικές πληροφορίες από αυτά. Χρησιμοποιώντας την περιγραφική ανάλυση μαθαίνουμε τι υπάρχει στα δεδομένα όσον αφορά την ποσοτική περιγραφή τους. Ένα απλό παράδειγμα είναι το πόσο καλά απέδωσε ο φοιτητής κατά τη διάρκεια του εξαμήνου υπολογίζοντας τον μέσο όρο. Αυτός ο μέσος όρος δεν είναι τίποτε άλλο παρά το άθροισμα της βαθμολογίας σε όλα τα μαθήματα του εξαμήνου επί τον συνολικό αριθμό των μαθημάτων. Κάθε φορά ωστόσο που προσπαθούμε να περιγράψουμε ένα μεγάλο σύνολο παρατηρήσεων με μία μόνο τιμή, διατρέχουμε τον κίνδυνο είτε να παραμορφώσουμε τα αρχικά δεδομένα είτε να χάσουμε κάποια άλλη σημαντική πληροφορία. Υπάρχουν δύο τύποι που χρησιμοποιούνται για την περιγραφή δεδομένων: • Μέτρα κεντρικής τάσης: Σε αυτό, μια ενιαία τιμή προσπαθεί να περιγράψει τα δεδομένα χρησιμοποιώντας την κεντρική τους θέση στο δεδομένο σύνολο. Ταξινομούνται επίσης ως συνοπτικό σύνολο. Για να πάρουν την κεντρική τιμή, χρησιμοποιούν τον μέσο όρο (mean), τη διάμεσο (median) ή τον τρόπο (mode) • Το μέτρο της διασποράς: Σε αυτό, τα δεδομένα συνοψίζονται περιγράφοντας πόσο καλά τα δεδομένα είναι διασκορπισμένα. Για παράδειγμα, αν η μέση βαθμολογία 100 μαθητών είναι 55, τότε θα υπάρχουν μαθητές των οποίων η

βαθμολογία θα είναι μικρότερη από 55 ή μεγαλύτερη από 55. Πράγμα που σημαίνει ότι η βαθμολογία τους θα είναι διασκορπισμένη με τρόπο ώστε ο μέσος όρος τους να είναι 55. Για να περιγράψουμε την εξάπλωση, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε από τις στατιστικές τεχνικές, δηλαδή εύρος, τεταρτημόρια, διακύμανση, τυπική απόκλιση και απόλυτη απόκλιση
Επαγωγικά στατιστικά στοιχεία (Inferial Statistics)
Η ομάδα δεδομένων που περιέχει τις πληροφορίες που μας ενδιαφέρουν είναι γνωστή ως πληθυσμός. Η επαγωγική στατιστική χρησιμοποιείται για να γίνει μια γενίκευση του πληθυσμού χρησιμοποιώντας τα δείγματα. Όπου το δείγμα προέρχεται από τον ίδιο τον πληθυσμό. Είναι απαραίτητο τα δείγματα να καταδεικνύουν σωστά τον πληθυσμό και να μην είναι μεροληπτικά. Η διαδικασία επίτευξης τέτοιων δειγμάτων ονομάζεται δειγματοληψία. Η επαγωγική στατιστική προέρχεται από το γεγονός ότι η δειγματοληψία προκαλεί φυσικά δειγματοληπτικά σφάλματα και συνεπώς δεν αναμένεται να αντιπροσωπεύει τέλεια τον πληθυσμό. Υπάρχουν δύο τύποι μεθόδων επαγωγικής στατιστικής που χρησιμοποιούνται για τη γενίκευση των δεδομένων: • Εκτίμηση των παραμέτρων • Έλεγχος στατιστικών υποθέσεων
Προδιαγραφική ανάλυση (Prescriptive Analysis)
¨Τι πρέπει να γίνει;¨. Η προδιαγραφική ανάλυση εργάζεται πάνω στα δεδομένα θέτοντας αυτή την ερώτηση. Είναι ο κοινός τομέας της επιχειρηματικής ανάλυσης για τον προσδιορισμό της καλύτερης δυνατής δράσης για μια κατάσταση. Η όλη ιδέα της είναι να παρέχει συμβουλές που αποσκοπούν στην εύρεση της βέλτιστης σύστασης για μια διαδικασία λήψης αποφάσεων. Σχετίζεται με την περιγραφική και την προγνωστική ανάλυση. Η περιγραφική ανάλυση περιγράφει τα δεδομένα, δηλαδή τι έχει συμβεί, και η προγνωστική ανάλυση προβλέπει τι μπορεί να συμβεί η προδιαγραφική ανάλυση βρίσκει την καλύτερη επιλογή μεταξύ των διαθέσιμων επιλογών. Οι τεχνικές που χρησιμοποιούνται στην προδιαγραφική ανάλυση είναι η προσομοίωση, η ανάλυση γραφημάτων, οι επιχειρηματικοί κανόνες, οι αλγόριθμοι, η επεξεργασία σύνθετων γεγονότων και η μηχανική μάθηση.
Ανάλυση πρόβλεψης
"Τι μπορεί να συμβεί;" Η ανάλυση πρόβλεψης χρησιμοποιείται για να γίνει πρόβλεψη μελλοντικών γεγονότων. Βασίζεται στα τρέχοντα και ιστορικά γεγονότα. Χρησιμοποιεί στατιστικό αλγόριθμο και τεχνικές μηχανικής μάθησης για τον προσδιορισμό της πιθανότητας μελλοντικών αποτελεσμάτων, τάσεων με βάση ιστορικά και νέα δεδομένα και συμπεριφορές. Οι επιχειρήσεις εφαρμόζουν την προγνωστική ανάλυση για να αυξήσουν το ανταγωνιστικό πλεονέκτημα και να μειώσουν τον κίνδυνο που σχετίζεται με ένα απρόβλεπτο μέλλον. Οι κύριοι χρήστες της προβλεπτικής ανάλυσης είναι το μάρκετινγκ, οι χρηματοπιστωτικές υπηρεσίες, οι πάροχοι ηλεκτρονικών υπηρεσιών και οι ασφαλιστικές εταιρείες. Οι τεχνικές που χρησιμοποιούνται στην προγνωστική ανάλυση είναι η εξόρυξη δεδομένων, η μοντελοποίηση, η A.I.
Αιτιώδης ανάλυση
"Γιατί;" Η αιτιώδης ανάλυση βοηθά στον προσδιορισμό του γιατί τα πράγματα είναι όπως είναι. Δεδομένου ότι ο σημερινός επιχειρηματικός κόσμος είναι γεμάτος από γεγονότα που μπορεί να οδηγήσουν σε αποτυχία, η Αιτιώδης Ανάλυση επιδιώκει να εντοπίσει την αιτία. Προσπαθεί να βρει τη βασική αιτία, δηλαδή τον βασικό λόγο για τον οποίο μπορεί να συμβεί κάτι. Πρόκειται για μια κοινή τεχνική

που χρησιμοποιείται στον κλάδο της πληροφορικής για τη διασφάλιση της ποιότητας του λογισμικού καθώς και σε βιομηχανίες που αντιμετωπίζουν μεγάλες καταστροφές.
Διερευνητική ανάλυση δεδομένων
Είναι μια εκθετική της επαγωγικής στατιστικής και χρησιμοποιείται κυρίως από τους επιστήμονες δεδομένων. Πρόκειται για μια αναλυτική προσέγγιση που επικεντρώνεται στον εντοπισμό μοτίβων στα δεδομένα και στην εξεύρεση των άγνωστων σχέσεων. Σκοπός της διερευνητικής ανάλυσης δεδομένων είναι να ελέγξει τα δεδομένα που λείπουν, να βρει άγνωστες σχέσεις και να ελέγξει υποθέσεις και παραδοχές. Είναι το πρώτο βήμα στην ανάλυση δεδομένων που πρέπει να εκτελείται πριν από τις άλλες επίσημες στατιστικές τεχνικές.

Πίνακας 11 Προσέγγιση με βάση τη στατιστική διαδικασία [πηγή : (Pedamkar, n.d.)

Προσέγγιση βασισμένη στη μηχανική μάθηση

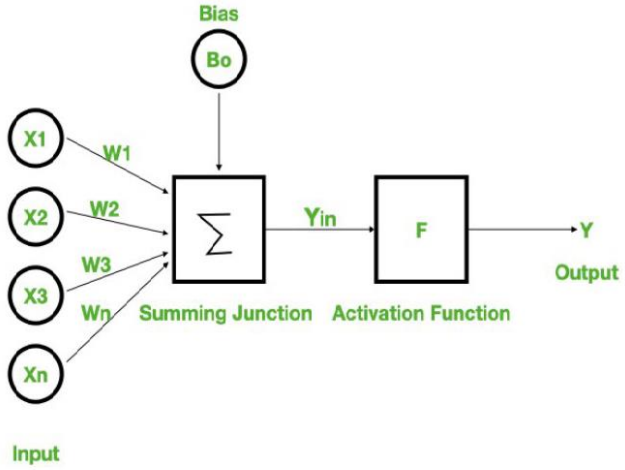
Υπάρχουν τρία βασικά σύνολα αλγορίθμων μηχανικής μάθησης: Υπό επίβλεψη και χωρίς επίβλεψη, συμπεριλαμβανομένου του συνεχώς αυξανόμενου αριθμού των υποτύπων τους, και οι αλγόριθμοι ενισχυτικής μάθησης.
Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν αλγορίθμους μάθησης με επίβλεψη, οι οποίοι υποδεικνύονται από τη χρήση επισημασμένων δεδομένων (όπως ο χρόνος και ο καιρός) που συνεπάγονται τόσο μεταβλητές εισόδου (x) όσο και μεταβλητές εξόδου (y). Ο χρήστης ως δάσκαλος γνωρίζει τη σωστή απάντηση (ή τις σωστές απαντήσεις) και επιβλέπετε τον αλγόριθμο καθώς κάνει προβλέψεις με βάση τα δεδομένα εκπαίδευσης. Εάν είναι απαραίτητο, γίνονται διορθώσεις έως ότου ο αλγόριθμος επιτύχει ένα επαρκές επίπεδο εκτέλεσης. Αν και υπάρχει μια ποικιλία αλγορίθμων μηχανικής μάθησης με επίβλεψη, οι πιο συχνά χρησιμοποιούμενοι περιλαμβάνουν:

- Γραμμική παλινδρόμηση
- Λογιστική παλινδρόμηση
- Δέντρο αποφάσεων
- Αλγόριθμος ταξινόμησης τυχαίου δάσους

Οι αλγόριθμοι μηχανικής μάθησης χωρίς επίβλεψη χρησιμοποιούνται για μη δομημένα δεδομένα για την εύρεση κοινών χαρακτηριστικών και διακριτών μοτίβων στο σύνολο δεδομένων. Επειδή αυτός ο τύπος αλγορίθμου ML δεν απαιτεί προηγούμενη εκπαίδευση ή επισημασμένα δεδομένα, είναι ελεύθερος να εξερευνήσει τη δομή των πληροφοριών (Master's in Data Science, n.d.).

Νευρωνικά Δίκτυα

Το νευρωνικό δίκτυο είναι ένα παράδειγμα επεξεργασίας πληροφοριών που είναι εμπνευσμένο από το ανθρώπινο νευρικό σύστημα. Όπως και στο ανθρώπινο νευρικό σύστημα, έχουμε βιολογικούς νευρώνες με τον ίδιο τρόπο στα νευρωνικά δίκτυα έχουμε τεχνητούς νευρώνες που είναι μια μαθηματική συνάρτηση που προέρχεται από βιολογικούς νευρώνες. Ο ανθρώπινος εγκέφαλος εκτιμάται ότι έχει περίπου 10 δισεκατομμύρια νευρώνες ο καθένας από τους οποίους συνδέεται κατά μέσο όρο με 10.000 άλλους νευρώνες. Κάθε νευρώνας λαμβάνει σήματα μέσω συνάψεων που ελέγχουν τις επιδράσεις του σήματος στο νευρώνα (V, 2021).



Εικόνα 29 Λειτουργία νευρωνικού δικτύου [πηγή : (Geeksforgeeks, 2022)]

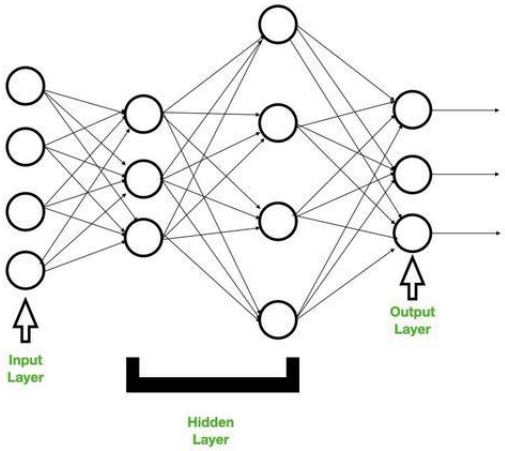
Ας υποθέσουμε ότι υπάρχουν n είσοδοι όπως X1,X2,...,Xn σε έναν νευρώνα.

- Το βάρος που συνδέει n αριθμό εισόδων σε έναν νευρώνα παριστάνεται με [W]=[W1,W2,...,Wn]
- Η λειτουργία του αθροιστικού κόμβου ενός τεχνητού νευρώνα είναι να συλλέγει τις σταθμισμένες εισόδους και να τις αθροίζει Yin=[X1*W1+X2*W2+….+Xn*Wn]
- Η έξοδος του αθροιστικού κόμβου μπορεί μερικές φορές να γίνει ίση με το μηδέν και για να αποφευχθεί μια τέτοια κατάσταση, προστίθεται σε αυτήν μια προκατάληψη σταθερής τιμής Bo Yin =[X1*W1+X2*W2+....+Xn*Wn] + Bo
- Η Yin κινείται στη συνέχεια προς τη συνάρτηση ενεργοποίησης.
- Η έξοδος Y ενός νευρώνα εξαρτάται σε μεγάλο βαθμό από τη Συνάρτηση Ενεργοποίησης (γνωστή και ως συνάρτηση μεταφοράς)

Υπάρχουν διάφοροι τύποι λειτουργίας ενεργοποίησης που χρησιμοποιούνται, όπως :

- Λειτουργία ταυτότητας
- Συνάρτηση δυαδικού βήματος με κατώφλι
- Διπολική βηματική συνάρτηση με κατώφλι
- Δυαδική σιγμοειδής συνάρτηση
- Διπολική σιγμοειδής συνάρτηση

Ενώ υπάρχουν πολλές διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων που έχουν δημιουργηθεί από ερευνητές, οι πιο επιτυχημένες εφαρμογές στα νευρωνικά δίκτυα εξόρυξης δεδομένων είναι τα πολυστρωματικά δίκτυα τροφοδότησης. Πρόκειται για δίκτυα στα οποία υπάρχει ένα στρώμα εισόδου που αποτελείται από κόμβους που απλά δέχονται τις τιμές εισόδου και διαδοχικά στρώματα κόμβων που είναι νευρώνες. Οι έξοδοι των νευρώνων σε ένα στρώμα αποτελούν εισόδους για τους νευρώνες στο επόμενο στρώμα. Το τελευταίο στρώμα ονομάζεται στρώμα εξόδου. Τα στρώματα μεταξύ των στρωμάτων εισόδου και εξόδου είναι γνωστά ως κρυφά στρώματα.



Εικόνα 30 Τεχνητός Νευρώνας [πηγή : (Geeksforgeeks, 2022)]

Υπάρχουν δύο τύποι επιβλεπόμενης μάθησης, ο ένας είναι η παλινδρόμηση και ο άλλος η ταξινόμηση. Έτσι, στο πρόβλημα τύπου παλινδρόμησης το νευρωνικό δίκτυο χρησιμοποιείται για την πρόβλεψη ενός αριθμητικού μεγέθους, υπάρχει ένας νευρώνας στο στρώμα εξόδου και η έξοδός του είναι η πρόβλεψη. Ενώ από την άλλη πλευρά στο πρόβλημα τύπου ταξινόμησης το στρώμα εξόδου έχει τόσους κόμβους όσες και οι κλάσεις και ο κόμβος του στρώματος εξόδου με τις μεγαλύτερες τιμές εξόδου δίνει την εκτίμηση του δικτύου για την κλάση για μια δεδομένη είσοδο. Στην ειδική περίπτωση των δύο κλάσεων, είναι σύνηθες να υπάρχει μόνο ένας κόμβος στο στρώμα εξόδου, ενώ η ταξινόμηση μεταξύ των δύο κλάσεων γίνεται με την εφαρμογή μιας αποκοπής στην τιμή εξόδου στον κόμβο.
Τα νευρωνικά δίκτυα βοηθούν στην εξόρυξη μεγάλου όγκου δεδομένων σε διάφορους τομείς όπως το λιανικό εμπόριο, οι τράπεζες (ανίχνευση απάτης), η βιοπληροφορική (αλληλουχία γονιδιώματος) κ.ά. Η εύρεση χρήσιμων πληροφοριών για μεγάλα δεδομένα που είναι κρυμμένα είναι πολύ δύσκολη και πολύ απαραίτητη. Η εξόρυξη δεδομένων χρησιμοποιεί νευρωνικά δίκτυα για τη συγκομιδή πληροφοριών από μεγάλα σύνολα δεδομένων από οργανισμούς αποθήκευσης δεδομένων. Τα οποία βοηθούν τον χρήστη στη λήψη αποφάσεων.
Ορισμένες από τις εφαρμογές των νευρωνικών δικτύων στην Εξόρυξη Δεδομένων είναι :

- Ανίχνευση απάτης: Στον σύγχρονο κόσμο λόγω της προόδου της τεχνολογίας, η απάτη είναι εύκολη στη διάπραξη, αλλά από την άλλη πλευρά η τεχνολογία βοηθά επίσης στην ανίχνευση της, και το νευρωνικό δίκτυο βοηθάει αρκετά
- Υγειονομική περίθαλψη: Στην υγειονομική περίθαλψη, τα νευρωνικά δίκτυα μας βοηθούν στη διάγνωση ασθενειών, καθώς γνωρίζουμε ότι υπάρχουν πολλές ασθένειες και ότι υπάρχουν μεγάλα σύνολα δεδομένων που έχουν αρχεία αυτών των ασθενειών. Με τα νευρωνικά δίκτυα και αυτά τα αρχεία, διαγνώσαμε αυτές τις ασθένειες σε πρώιμο στάδιο το συντομότερο δυνατό

Η μέθοδος των νευρωνικών δικτύων χρησιμοποιείται για ταξινόμηση, ομαδοποίηση, εξόρυξη χαρακτηριστικών, πρόβλεψη και αναγνώριση προτύπων. Το μοντέλο McCulloch-Pitts θεωρείται το πρώτο νευρωνικό δίκτυο και ο κανόνας μάθησης Hebbian είναι ένας από τους πρώτους και απλούστερους κανόνες μάθησης για το

νευρωνικό δίκτυο. Το μοντέλο νευρωνικού δικτύου μπορεί να χωριστεί σε γενικές γραμμές στους ακόλουθους τρεις τύπους:

- Νευρωνικά δίκτυα τροφοδότησης προς τα εμπρός (Feed Forward): Στο δίκτυο αυτό, εάν οι τιμές εξόδου δεν μπορούν να αναχθούν στις τιμές εισόδου και εάν για κάθε κόμβο εισόδου υπολογίζεται ένας κόμβος εξόδου, τότε υπάρχει ροή πληροφοριών προς τα εμπρός και δεν υπάρχει ανατροφοδότηση μεταξύ των επιπέδων. Η πληροφορία κινείται προς μία μόνο κατεύθυνση (προς τα εμπρός) από τους κόμβους εισόδου, μέσω των κρυφών κόμβων (εάν υπάρχουν) και προς τους κόμβους εξόδου. Ένας τέτοιος τύπος δικτύου είναι γνωστός ως δίκτυο τροφοδότησης
- Νευρωνικό δίκτυο ανατροφοδότησης: Τα σήματα μπορούν να ταξιδεύουν και προς τις δύο κατευθύνσεις σε ένα δίκτυο ανατροφοδότησης. Τα νευρωνικά δίκτυα ανατροφοδότησης είναι πολύ ισχυρά και μπορούν να γίνουν πολύ πολύπλοκα. Οι "καταστάσεις" σε ένα τέτοιο δίκτυο αλλάζουν συνεχώς μέχρι να επιτευχθεί ένα σημείο ισορροπίας (είναι δυναμικές). Παραμένουν σε ισορροπία μέχρι να αλλάξει η είσοδος και να πρέπει να βρεθεί μια νέα ισορροπία. Οι αρχιτεκτονικές νευρωνικών δικτύων ανατροφοδότησης είναι επίσης γνωστές ως διαδραστικές ή αναδρομικές. Στα δίκτυα αυτά επιτρέπονται βρόχοι ανατροφοδότησης. Χρησιμοποιούνται για τη διευθυνσιοδοτούμενη μνήμη περιεχομένου
- Νευρωνικό δίκτυο αυτοοργάνωσης: Το Νευρωνικό Δίκτυο Αυτοοργάνωσης (SONN) είναι ένας τύπος τεχνητού νευρωνικού δικτύου, αλλά εκπαιδεύεται χρησιμοποιώντας ανταγωνιστική μάθηση και όχι μάθηση διόρθωσης σφαλμάτων (π.χ. οπισθοδιάδοση με κάθοδο κλίσης) που χρησιμοποιείται από άλλα τεχνητά νευρωνικά δίκτυα. Το αυτοοργανωτικό νευρωνικό δίκτυο (SONN) είναι ένα μοντέλο μάθησης χωρίς επίβλεψη στα τεχνητά νευρωνικά δίκτυα που ονομάζονται αυτοοργανωτικοί χάρτες χαρακτηριστικών ή χάρτες Kohonen. Χρησιμοποιείται για την παραγωγή μιας χαμηλής διάστασης (συνήθως δισδιάστατης) αναπαράστασης ενός συνόλου δεδομένων υψηλότερης διάστασης, διατηρώντας παράλληλα την τοπολογική δομή των δεδομένων

(Wei-SenChen, 2009)

Αλγόριθμοι Ταξινόμησης

Οι ασφαλισμένοι πελάτες ανήκουν ήδη σε μια ορισμένη κλάση και ταξινομούνται ως "έχει λήξει" ή "δεν έχει λήξει". Η εταιρεία μπορεί να χρησιμοποιήσει τη λειτουργία εξόρυξης ταξινόμησης για να δημιουργήσει ένα προφίλ ομάδας κινδύνου με τη μορφή ενός μοντέλου εξόρυξης δεδομένων. Αυτό το προφίλ, ή μοντέλο, περιέχει τις κοινές τιμές χαρακτηριστικών των πελατών που έχουν παραγραφεί, σε σύγκριση με τους άλλους πελάτες. Η ασφαλιστική εταιρεία μπορεί στη συνέχεια να εφαρμόσει αυτό το προφίλ σε νέους πελάτες (που δεν έχουν ακόμη "ταξινομηθεί") για να διαπιστώσει αν ανήκουν στην ομάδα κινδύνου.

Η ροή εργασιών έχει ως εξής:

- Η ασφαλιστική εταιρεία χρησιμοποιεί μια εκπαιδευτική εκτέλεση ταξινόμησης για να εντοπίσει τυπικούς συνδυασμούς τιμών χαρακτηριστικών κάθε καθορισμένης κατηγορίας κινδύνου πελατών και να δημιουργήσει ένα μοντέλο
- Η ασφαλιστική εταιρεία ελέγχει την ακρίβεια αυτού του μοντέλου εφαρμόζοντας το μοντέλο σε δοκιμαστικά δεδομένα με γνωστές κατηγορίες κινδύνου πελατών

- Ο ασφαλιστής εφαρμόζει το δοκιμασμένο μοντέλο σε νέα δεδομένα και προβλέπει τους πελάτες που είναι πιθανό να αφήσουν την ασφάλισή τους να λήξει στο μέλλον

Οι τύποι ταξινόμησης είναι :

- Γενική ταξινόμηση. Η ταξινόμηση είναι η διαδικασία αυτόματης δημιουργίας ενός μοντέλου κλάσεων από ένα σύνολο εγγραφών που περιέχουν ετικέτες κλάσεων.
- Ταξινόμηση με δέντρο αποφάσεων. Υλοποίηση ταξινόμησης με δέντρα αποφάσεων. Για τον υπολογισμό ενός δέντρου αποφάσεων χρησιμοποιείται ένας αλγόριθμος ταξινόμησης δέντρων. Τα δέντρα αποφάσεων είναι εύκολο να κατανοηθούν και να τροποποιηθούν και το μοντέλο που αναπτύσσεται μπορεί να εκφραστεί ως ένα σύνολο κανόνων απόφασης. Αυτός ο αλγόριθμος κλιμακώνεται καλά, ακόμη και όταν υπάρχουν ποικίλοι αριθμοί παραδειγμάτων εκπαίδευσης και σημαντικός αριθμός χαρακτηριστικών σε μεγάλες βάσεις δεδομένων
- Ταξινόμηση Naive Bayes. Ο αλγόριθμος ταξινόμησης Naive Bayes είναι ένας πιθανοτικός ταξινομητής. Βασίζεται σε μοντέλα πιθανοτήτων που ενσωματώνουν ισχυρές υποθέσεις ανεξαρτησίας
- Ταξινόμηση με λογιστική παλινδρόμηση.
- Ο αλγόριθμος αυτός υπολογίζει δυαδικές προβλέψεις. Για παράδειγμα, αν είναι πιθανό να εμφανιστεί απάτη. Οι τιμές για αυτή την πρόβλεψη μπορεί να είναι Ναι ή Όχι

(IBM, 2021)

Αλγόριθμος ID3

Στην εκμάθηση δέντρων απόφασης, ο ID3 (Iterative Dichotomiser 3) είναι ένας αλγόριθμος που επινοήθηκε από τον Ross Quinlan και χρησιμοποιείται για τη δημιουργία ενός δέντρου απόφασης από το σύνολο δεδομένων. Ο ID3 χρησιμοποιείται συνήθως στους τομείς της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας. Η τεχνική του δέντρου απόφασης περιλαμβάνει την κατασκευή ενός δέντρου για τη μοντελοποίηση της διαδικασίας ταξινόμησης. Μόλις κατασκευαστεί ένα δέντρο, εφαρμόζεται σε κάθε πλειάδα στη βάση δεδομένων και οδηγεί σε ταξινόμηση για την εν λόγω πλειάδα. Τα ακόλουθα ζητήματα αντιμετωπίζουν οι περισσότεροι αλγόριθμοι δέντρων απόφασης:

- Επιλογή χαρακτηριστικών διαχωρισμού
- Διάταξη των χαρακτηριστικών διάσπασης
- Αριθμός διαχωρισμών
- Ισορροπία της δομής του δέντρου και του κλαδέματος
- Κριτήρια διακοπής

Ο αλγόριθμος ID3 είναι ένας αλγόριθμος ταξινόμησης που βασίζεται στην εντροπία πληροφοριών και η βασική ιδέα του είναι ότι όλα τα παραδείγματα αντιστοιχίζονται σε διαφορετικές κατηγορίες σύμφωνα με τις διαφορετικές τιμές του συνόλου χαρακτηριστικών συνθηκών. Ο αλγόριθμος επιλέγει το κέρδος πληροφορίας ως κριτήριο επιλογής του χαρακτηριστικού, και συνήθως το χαρακτηριστικό που έχει το μεγαλύτερο κέρδος πληροφορίας επιλέγεται ως το χαρακτηριστικό διαχωρισμού του

τρέχοντος κόμβου, προκειμένου η εντροπία πληροφορίας που χρειάζονται τα διαχωρισμένα υποσύνολα να είναι η μικρότερη. Σύμφωνα με τις διαφορετικές τιμές του χαρακτηριστικού, μπορούν να δημιουργηθούν κλάδοι και η παραπάνω διαδικασία καλείται αναδρομικά σε κάθε κλάδο για τη δημιουργία άλλων κόμβων και κλάδων, έως ότου όλα τα δείγματα σε έναν κλάδο ανήκουν στην ίδια κατηγορία. Για την επιλογή των χαρακτηριστικών διαχωρισμού χρησιμοποιούνται οι έννοιες της εντροπίας και του κέρδους πληροφορίας

Δεδομένων των πιθανοτήτων p1 , p2 , ..., ps , όπου , η εντροπία ορίζεται ως :

Η εντροπία βρίσκει την ποσότητα της τάξης σε μια δεδομένη κατάσταση της βάσης δεδομένων. Μια τιμή H = 0 προσδιορίζει ένα απόλυτα ταξινομημένο σύνολο. Όσο υψηλότερη είναι η εντροπία, τόσο μεγαλύτερη είναι η δυνατότητα βελτίωσης της διαδικασίας ταξινόμησης.
Το ID3 επιλέγει το χαρακτηριστικό διάσπασης με το μεγαλύτερο κέρδος σε πληροφορίες, όπου το κέρδος ορίζεται ως η διαφορά μεταξύ των πληροφοριών που απαιτούνται μετά τη διάσπαση. Αυτό υπολογίζεται με τον προσδιορισμό των διαφορών μεταξύ των εντροπιών του αρχικού συνόλου δεδομένων και του σταθμισμένου αθροίσματος των εντροπιών από κάθε ένα από τα υποδιαιρεμένα σύνολα δεδομένων. Ο τύπος που χρησιμοποιείται για το σκοπό αυτό είναι ο εξής: (Kalpesh Adhatrao, 2013).

C4.5

Ο C4.5 είναι ένας γνωστός αλγόριθμος που χρησιμοποιείται για τη δημιουργία δέντρων αποφάσεων. Είναι μια επέκταση του αλγορίθμου ID3 που χρησιμοποιείται για να ξεπεραστούν τα μειονεκτήματά του. Τα δέντρα απόφασης που παράγονται από τον αλγόριθμο C4.5 μπορούν να χρησιμοποιηθούν για ταξινόμηση και για το λόγο αυτό, ο C4.5 αναφέρεται επίσης ως στατιστικός ταξινομητής. Ο αλγόριθμος C4.5 έκανε ορισμένες αλλαγές για να βελτιώσει τον αλγόριθμο ID3. Ορισμένες από αυτές είναι οι εξής:

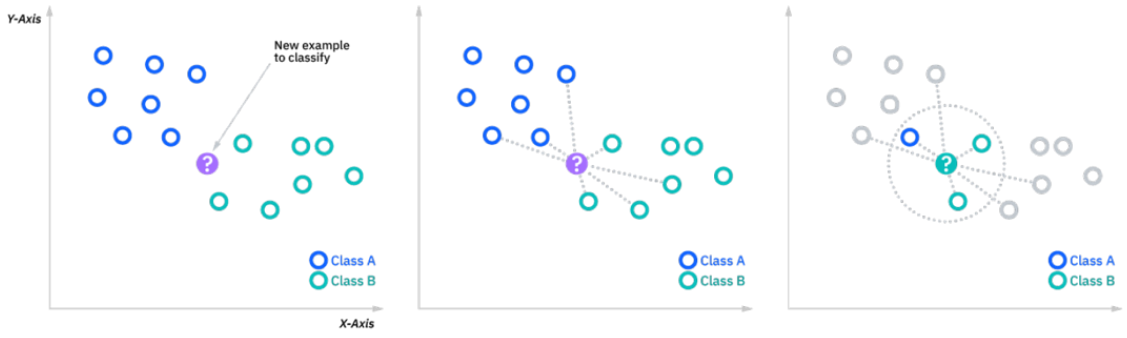
- Χειρισμός δεδομένων εκπαίδευσης με ελλείπουσες τιμές χαρακτηριστικών
- Χειρισμός χαρακτηριστικών διαφορετικού κόστους
- Κλάδεμα του δέντρου απόφασης μετά τη δημιουργία του
- Χειρισμός χαρακτηριστικών με διακριτές και συνεχείς τιμές

Σε κάθε κόμβο του δέντρου, το C4.5 επιλέγει ένα χαρακτηριστικό των δεδομένων που διαχωρίζει αποτελεσματικότερα το σύνολο δεδομένων των δειγμάτων S σε υποσύνολα που μπορούν να ανήκουν στη μία ή στην άλλη κλάση.
Έστω ότι τα δεδομένα εκπαίδευσης είναι ένα σύνολο S = s1 , s 2 ... ήδη ταξινομημένων δειγμάτων. Κάθε δείγμα είναι ένα διάνυσμα που τα αντιπροσωπεύουν χαρακτηριστικά ή γνωρίσματα του δείγματος. Τα δεδομένα εκπαίδευσης είναι ένα διάνυσμα όπου τα αντιπροσωπεύουν την κλάση στην οποία ανήκει κάθε δείγμα.
Σε κάθε κόμβο του δέντρου, το C4.5 επιλέγει ένα χαρακτηριστικό των δεδομένων που διαχωρίζει αποτελεσματικότερα το σύνολο δεδομένων των δειγμάτων S σε υποσύνολα που μπορούν να ανήκουν στη μία ή την άλλη κλάση. Είναι το

κανονικοποιημένο κέρδος πληροφορίας (διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Ο παράγοντας του χαρακτηριστικού με το μεγαλύτερο κανονικοποιημένο κέρδος πληροφορίας λαμβάνεται υπόψη για τη λήψη της απόφασης. Στη συνέχεια, ο αλγόριθμος C4.5 συνεχίζει στις μικρότερες υπολίστες που έχουν το αμέσως υψηλότερο κανονικοποιημένο κέρδος πληροφορίας (Kalpesh Adhatrao, 2013).

K-Nearest Neighbors Algorithm

Ο αλγόριθμος k-κοντινότερων γειτόνων, γνωστός και ως KNN ή k-NN, είναι ένας μη παραμετρικός ταξινομητής μάθησης με επίβλεψη, ο οποίος χρησιμοποιεί την εγγύτητα για να κάνει ταξινομήσεις ή προβλέψεις σχετικά με την ομαδοποίηση ενός μεμονωμένου σημείου δεδομένων. Αν και μπορεί να χρησιμοποιηθεί είτε για προβλήματα παλινδρόμησης είτε για προβλήματα ταξινόμησης, συνήθως χρησιμοποιείται ως αλγόριθμος ταξινόμησης, βασιζόμενος στην υπόθεση ότι παρόμοια σημεία μπορούν να βρεθούν κοντά το ένα στο άλλο.
Για προβλήματα ταξινόμησης, μια ετικέτα κλάσης αποδίδεται βάσει ψηφοφορίας πλειοψηφίας, δηλαδή χρησιμοποιείται η ετικέτα που αντιπροσωπεύεται συχνότερα γύρω από ένα δεδομένο σημείο δεδομένων. Αν και τεχνικά αυτό θεωρείται "ψηφοφορία πολλαπλότητας", ο όρος "ψήφος πλειοψηφίας" χρησιμοποιείται συχνότερα στη βιβλιογραφία. Η διάκριση μεταξύ αυτών των ορολογιών είναι ότι η "ψηφοφορία πλειοψηφίας" τεχνικά απαιτεί πλειοψηφία μεγαλύτερη του 50%, η οποία λειτουργεί κυρίως όταν υπάρχουν μόνο δύο κατηγορίες. Όταν υπάρχουν πολλές κατηγορίες, π.χ. τέσσερις κατηγορίες, δεν είναι απαραίτητο το 50% των ψήφων για να βγει ένα συμπέρασμα σχετικά με μια κατηγορία. Θα μπορούσε να αποδωθεί μια ετικέτα κατηγορίας με ψήφο μεγαλύτερη από 25%.



Εικόνα 31 Διάγραμμα KNN

Τα προβλήματα παλινδρόμησης χρησιμοποιούν μια παρόμοια έννοια με το πρόβλημα ταξινόμησης, αλλά σε αυτή την περίπτωση, ο μέσος όρος των k πλησιέστερων γειτόνων λαμβάνεται για να γίνει μια πρόβλεψη σχετικά με μια ταξινόμηση. Η κύρια

διάκριση εδώ είναι ότι η ταξινόμηση χρησιμοποιείται για διακριτές τιμές, ενώ η παλινδρόμηση χρησιμοποιείται με συνεχείς τιμές. Ωστόσο, προτού γίνει ταξινόμηση, πρέπει να οριστεί η απόσταση. Συνηθέστερα χρησιμοποιείται η ευκλείδεια απόσταση. Αξίζει επίσης να σημειωθεί ότι ο αλγόριθμος KNN ανήκει επίσης σε μια οικογένεια μοντέλων "τεμπέλικης μάθησης", που σημαίνει ότι αποθηκεύει μόνο ένα σύνολο δεδομένων εκπαίδευσης αντί να υποβάλλεται σε στάδιο εκπαίδευσης. Αυτό σημαίνει επίσης ότι όλοι οι υπολογισμοί πραγματοποιούνται όταν γίνεται μια ταξινόμηση ή πρόβλεψη. Δεδομένου ότι βασίζεται σε μεγάλο βαθμό στη μνήμη για την αποθήκευση όλων των δεδομένων εκπαίδευσης, αναφέρεται επίσης ως μέθοδος μάθησης που βασίζεται σε περιπτώσεις ή σε μνήμη.

Πλεονεκτήματα του KNN:

- Εύκολη εφαρμογή: Δεδομένης της απλότητας και της ακρίβειας του αλγορίθμου, είναι ένας από τους πρώτους ταξινομητές που θα μάθει ένας νέος επιστήμονας δεδομένων
- Προσαρμόζεται εύκολα: Καθώς προστίθενται νέα δείγματα εκπαίδευσης, ο αλγόριθμος προσαρμόζεται ώστε να λαμβάνει υπόψη του κάθε νέο δεδομένο, δεδομένου ότι όλα τα δεδομένα εκπαίδευσης αποθηκεύονται στη μνήμη
- Λίγες υπερπαραμέτρους: Ο KNN απαιτεί μόνο μια τιμή k και μια μετρική απόστασης, η οποία είναι χαμηλή σε σύγκριση με άλλους αλγορίθμους μηχανικής μάθησης

Μειονεκτήματα του KNN:

- Δεν κλιμακώνεται καλά: Δεδομένου ότι ο KNN είναι ένας τεμπέλης αλγόριθμος, καταλαμβάνει περισσότερη μνήμη και αποθήκευση δεδομένων σε σύγκριση με άλλους ταξινομητές. Αυτό μπορεί να είναι δαπανηρό τόσο από άποψη χρόνου όσο και από άποψη χρημάτων. Περισσότερη μνήμη και αποθήκευση θα αυξήσει τα επιχειρηματικά έξοδα και περισσότερα δεδομένα μπορεί να χρειαστούν περισσότερο χρόνο για τον υπολογισμό. Ενώ έχουν δημιουργηθεί διαφορετικές δομές δεδομένων, όπως η Ball-Tree, για την αντιμετώπιση της υπολογιστικής αναποτελεσματικότητας, ένας διαφορετικός ταξινομητής μπορεί να είναι ιδανικός ανάλογα με το επιχειρηματικό πρόβλημα
- Πρόβλημα της διαστατικότητας: Ο αλγόριθμος KNN τείνει να πέφτει θύμα της κατάρας της διαστατικότητας, πράγμα που σημαίνει ότι δεν αποδίδει καλά με δεδομένα εισόδου υψηλής διάστασης. Αυτό αναφέρεται μερικές φορές και ως φαινόμενο κορύφωσης όπου αφού ο αλγόριθμος επιτύχει τον βέλτιστο αριθμό χαρακτηριστικών, τα πρόσθετα χαρακτηριστικά αυξάνουν το ποσό των σφαλμάτων ταξινόμησης, ιδίως όταν το μέγεθος του δείγματος είναι μικρότερο
- Επιρρεπής σε υπερπροσαρμογή: ο KNN είναι επίσης πιο επιρρεπής στην υπερπροσαρμογή. Ενώ οι τεχνικές επιλογής χαρακτηριστικών και μείωσης της διαστατικότητας αξιοποιούνται για την αποτροπή αυτού του φαινομένου, η τιμή του k μπορεί επίσης να επηρεάσει τη συμπεριφορά του μοντέλου. Χαμηλότερες τιμές του k μπορούν να υπερπροσαρμόσουν τα δεδομένα, ενώ υψηλότερες τιμές του k τείνουν να "εξομαλύνουν" τις τιμές πρόβλεψης, δεδομένου ότι ο μέσος όρος των τιμών υπολογίζεται σε μια μεγαλύτερη περιοχή ή γειτονιά. Ωστόσο, εάν η τιμή του k είναι πολύ υψηλή, τότε μπορεί να υποπροσαρμόσει τα δεδομένα

(IBM, n.d.)

Αλγόριθμος Naive Bayes

Πρόκειται για έναν αλγόριθμο που μαθαίνει την πιθανότητα κάθε αντικειμένου, τα χαρακτηριστικά του και τις ομάδες στις οποίες ανήκουν. Είναι επίσης γνωστός ως πιθανοτικός ταξινομητής. Ο αλγόριθμος Naive Bayes ανήκει στην επιβλεπόμενη μάθηση και χρησιμοποιείται κυρίως για την επίλυση προβλημάτων ταξινόμησης. Η πιθανότητα είναι η βάση για τον αλγόριθμο Naive Bayes. Ο αλγόριθμος αυτός είναι κατασκευασμένος με βάση τα αποτελέσματα των πιθανοτήτων που μπορεί να προσφέρει για άλυτα προβλήματα με τη βοήθεια της πρόβλεψης. Η πιθανότητα βοηθά στην πρόβλεψη της εμφάνισης ενός γεγονότος από όλα τα πιθανά αποτελέσματα. Η μαθηματική εξίσωση της πιθανότητας έχει ως εξής:

Πιθανότητα ενός γεγονότος = Αριθμός ευνοϊκών γεγονότων/ Συνολικός αριθμός αποτελεσμάτων

0 < = πιθανότητα ενός γεγονότος < = 1. Το ευνοϊκό αποτέλεσμα δηλώνει το γεγονός που προκύπτει από την πιθανότητα. Η πιθανότητα είναι πάντα μεταξύ 0 και 1, όπου 0 σημαίνει ότι δεν υπάρχει πιθανότητα να συμβεί και 1 σημαίνει ότι το ποσοστό επιτυχίας του συγκεκριμένου γεγονότος είναι πιθανό.
Η θεωρία Bayes λειτουργεί με βάση την εξαγωγή μιας υπόθεσης (H) από ένα δεδομένο σύνολο αποδείξεων (E). Σχετίζεται με δύο πράγματα: την πιθανότητα της υπόθεσης πριν από τα αποδεικτικά στοιχεία P(H) και την πιθανότητα μετά τα αποδεικτικά στοιχεία P(H|E). Η θεωρία Bayes εξηγείται από την ακόλουθη εξίσωση:

P(H|E) = (P(E|H} * P(H))/P(E)

Στην παραπάνω εξίσωση :

- P(H|E) δηλώνει πώς συμβαίνει το γεγονός H όταν λαμβάνει χώρα το γεγονός E
- Το P(E|H) δηλώνει πόσο συχνά συμβαίνει το γεγονός E όταν το γεγονός H λαμβάνει χώρα πρώτο
- Η P(H) αντιπροσωπεύει την πιθανότητα να συμβεί το γεγονός X από μόνο του
- Η P(E) αντιπροσωπεύει την πιθανότητα να συμβεί από μόνο του το γεγονός Y
- Ο κανόνας Bayes είναι μια μέθοδος για τον προσδιορισμό της P(H|E) από την P(E|H)

Παρέχει έναν τρόπο υπολογισμού της πιθανότητας μιας υπόθεσης με τα παρεχόμενα αποδεικτικά στοιχεία.

Υπάρχουν τέσσερις τύποι του μοντέλου Naive Bayes, οι οποίοι εξηγούνται παρακάτω:

- Gaussian Naive Bayes. Είναι ένας απλός αλγόριθμος που χρησιμοποιείται όταν τα χαρακτηριστικά είναι συνεχή. Τα χαρακτηριστικά που υπάρχουν στα δεδομένα πρέπει να ακολουθούν τον κανόνα της κατανομής Gauss ή της κανονικής κατανομής. Επιταχύνει αξιοσημείωτα την αναζήτηση και υπό επιεικείς συνθήκες, το σφάλμα θα είναι δύο φορές μεγαλύτερο από το βέλτιστο Naive Bayes
- Βέλτιστο Naive Bayes. Το Optimal Naive Bayes επιλέγει την κλάση που έχει τη μεγαλύτερη εκ των υστέρων πιθανότητα να συμβεί. Σύμφωνα με το όνομά του, είναι βέλτιστη. Αλλά θα εξετάσει όλες τις πιθανότητες, πράγμα που είναι πολύ αργό και χρονοβόρο

- Bernoulli Naive Bayes. Ο Bernoulli Naive Bayes είναι ένας αλγόριθμος που είναι χρήσιμος για δεδομένα που έχουν δυαδικά ή boolean χαρακτηριστικά. Τα χαρακτηριστικά θα έχουν την τιμή ναι ή όχι, χρήσιμα ή όχι, χορηγούνται ή απορρίπτονται κ.ά.
- Multinominal Naive Bayes. Ο Multinominal Naive Bayes χρησιμοποιείται σε θέματα ταξινόμησης τεκμηρίωσης. Τα χαρακτηριστικά που απαιτούνται για αυτόν τον τύπο είναι η συχνότητα των λέξεων που μετατρέπονται από το έγγραφο

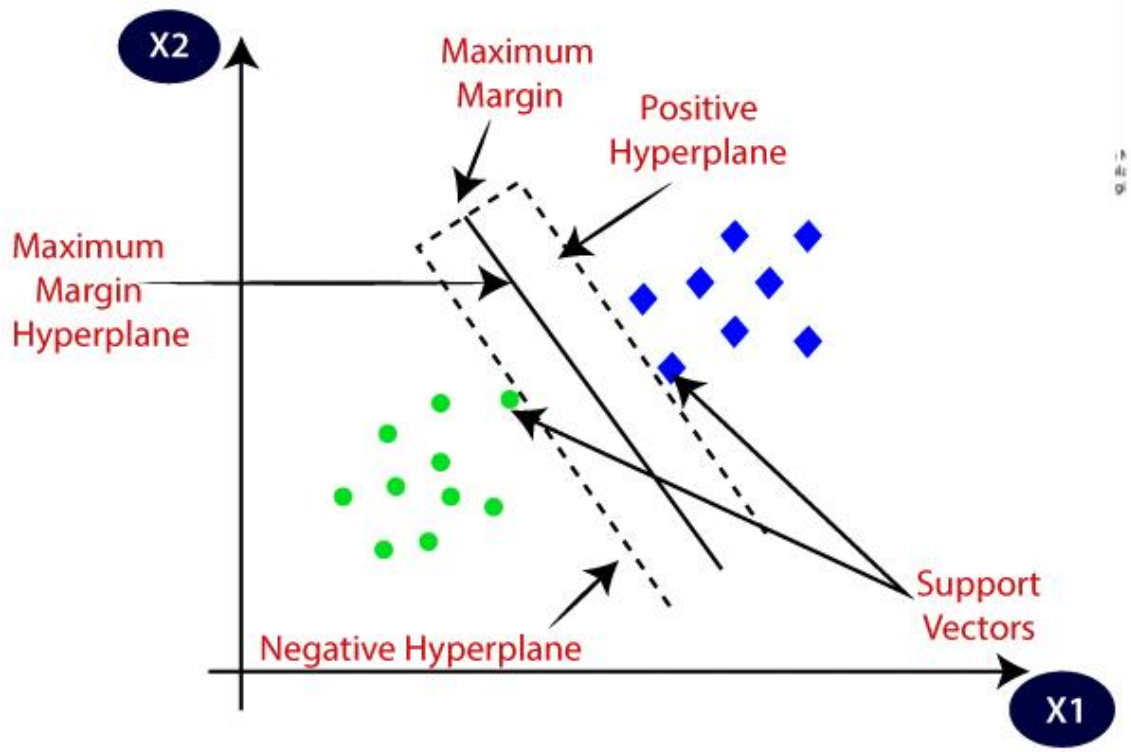
(Turing, n.d.)

Εικόνα 32 4 τύποι του μοντέλου Naive Bayes [πηγή : (Turing, χ.χ.)]

Αλγόριθμος SVM

Θεωρητικά, ο αλγόριθμος SVM, γνωστός και ως αλγόριθμος διανυσματικής μηχανής υποστήριξης, είναι γραμμικός. Αυτό που κάνει τον αλγόριθμο SVM να ξεχωρίζει σε σχέση με άλλους αλγορίθμους είναι ότι μπορεί να αντιμετωπίσει προβλήματα ταξινόμησης χρησιμοποιώντας έναν ταξινομητή SVM και προβλήματα παλινδρόμησης χρησιμοποιώντας έναν παλινδρομητή SVM. Ωστόσο, δεν πρέπει να ξεχνάμε ότι ο ταξινομητής SVM αποτελεί τη ραχοκοκαλιά της έννοιας της μηχανής διανυσμάτων υποστήριξης και, γενικά, είναι ο καταλληλότερος αλγόριθμος για την επίλυση προβλημάτων ταξινόμησης. Όντας ένας γραμμικός αλγόριθμος στον πυρήνα του μπορεί να φανταστεί κανείς σχεδόν σαν μια γραμμική ή λογιστική παλινδρόμηση. Για παράδειγμα, ένας ταξινομητής SVM δημιουργεί μια γραμμή σε έναν χώρο N διαστάσεων για να ταξινομήσει σημεία δεδομένων που ανήκουν σε δύο ξεχωριστές κλάσεις. Το SVM αναπτύχθηκε από τον Vladimir Vapnik τη δεκαετία του 1970, στο πλαίσιο ενός στοιχήματος όπου ο Vapnik προέβλεψε ότι η επίτευξη ενός ορίου απόφασης που προσπαθεί να μεγιστοποιήσει το περιθώριο μεταξύ των δύο κλάσεων θα δώσει εξαιρετικά αποτελέσματα και θα ξεπεράσει το πρόβλημα της υπερπροσαρμογής. Όλα άλλαξαν, ιδίως τη δεκαετία του '90, όταν εισήχθη η μέθοδος πυρήνα που κατέστησε δυνατή την επίλυση μη γραμμικών προβλημάτων με τη χρήση SVM. Αυτό επηρέασε σημαντικά τη σημασία και την ανάπτυξη των νευρωνικών δικτύων για ένα διάστημα, καθώς ήταν εξαιρετικά περίπλοκα. Ταυτόχρονα, το SVM ήταν πολύ πιο απλό από αυτά και εξακολουθούσε να μπορεί να επιλύει μη γραμμικά

προβλήματα ταξινόμησης με ευκολία και καλύτερη ακρίβεια. Στη σημερινή εποχή, ακόμη και με την πρόοδο της βαθιάς μάθησης και των νευρωνικών δικτύων γενικότερα, η σημασία και η εξάρτηση από το SVM δεν έχουν μειωθεί και συνεχίζει να απολαμβάνει επαίνους και συχνή χρήση σε πολυάριθμες βιομηχανίες που εμπλέκουν τη μηχανική μάθηση στη λειτουργία τους.
Ο καλύτερος τρόπος για να κατανοήσουμε τον αλγόριθμο SVM είναι να επικεντρωθείτε στον κύριο τύπο του, τον ταξινομητή SVM. Η ιδέα πίσω από τον ταξινομητή SVM είναι να καταλήξει σε μια υπερζώνη σε έναν N-διάστατο χώρο που διαχωρίζει τα σημεία δεδομένων που ανήκουν σε διαφορετικές κλάσεις. Ωστόσο, αυτό το υπερεπίπεδο επιλέγεται με βάση το περιθώριο, καθώς λαμβάνεται υπόψη το υπερεπίπεδο που παρέχει το μέγιστο περιθώριο μεταξύ των δύο κλάσεων. Αυτά τα περιθώρια υπολογίζονται χρησιμοποιώντας σημεία δεδομένων γνωστά ως διανύσματα υποστήριξης. Τα διανύσματα υποστήριξης είναι εκείνα τα σημεία δεδομένων που βρίσκονται κοντά στο υπερεπίπεδο και βοηθούν στον προσανατολισμό του.



Εικόνα 33 Διάγραμμα λειτουργίας αλγόριθμου SVM [πηγή: (Analytixlabs, n.d.)]

Πλεονεκτήματα SVM:

- Είναι ένας δυναμικός αλγόριθμος και μπορεί να επιλύσει μια σειρά προβλημάτων, συμπεριλαμβανομένων γραμμικών και μη γραμμικών προβλημάτων, προβλημάτων ταξινόμησης δυαδικών, διωνυμικών και πολλαπλών κλάσεων, καθώς και προβλημάτων παλινδρόμησης
- Ο SVM χρησιμοποιεί την έννοια των περιθωρίων και προσπαθεί να μεγιστοποιήσει τη διαφοροποίηση μεταξύ δύο κλάσεων, και μειώνει τις πιθανότητες υπερπροσαρμογής του μοντέλου, καθιστώντας το μοντέλο ιδιαίτερα σταθερό

- Λόγω της διαθεσιμότητας πυρήνων και των ίδιων των θεμελιωδών αρχών στις οποίες βασίζεται ο SVM, μπορεί εύκολα να λειτουργήσει όταν τα δεδομένα είναι σε υψηλές διαστάσεις και είναι ακριβής σε υψηλές διαστάσεις σε βαθμό που μπορεί να ανταγωνιστεί αλγορίθμους όπως ο Naïve Bayes που ειδικεύεται στην αντιμετώπιση προβλημάτων ταξινόμησης πολύ υψηλών διαστάσεων
- Ο SVM είναι γνωστός για την ταχύτητα υπολογισμού και τη διαχείριση της μνήμης του. Χρησιμοποιεί λιγότερη μνήμη, ειδικά σε σύγκριση με τους αλγορίθμους μηχανικής και βαθιάς μάθησης με τους οποίους ο SVM συχνά ανταγωνίζεται και μερικές φορές ακόμη και ξεπερνά μέχρι σήμερα.

Μειονεκτήματα SVM:

- Ενώ το SVM είναι γρήγορο και μπορεί να λειτουργήσει σε υψηλές διαστάσεις, εξακολουθεί να αποτυγχάνει μπροστά από το Naïve Bayes, παρέχοντας ταχύτερες προβλέψεις σε υψηλές διαστάσεις. Επίσης, χρειάζεται σχετικά μεγάλο χρονικό διάστημα κατά τη φάση της εκπαίδευσης
- Όπως και ορισμένοι άλλοι αλγόριθμοι μηχανικής μάθησης, οι οποίοι είναι συχνά ιδιαίτερα ευαίσθητοι σε ορισμένες από τις υπερπαραμέτρους τους, η απόδοση του SVM εξαρτάται επίσης σε μεγάλο βαθμό από τον πυρήνα που επιλέγει ο χρήστης
- Σε σύγκριση με άλλους γραμμικούς αλγορίθμους, όπως η γραμμική παλινδρόμηση, ο SVM δεν είναι ιδιαίτερα ερμηνεύσιμος, ειδικά όταν χρησιμοποιούνται πυρήνες που καθιστούν τον SVM μη γραμμικό. Έτσι, δεν είναι εύκολο να εκτιμηθεί πώς οι ανεξάρτητες μεταβλητές επηρεάζουν τη μεταβλητή-στόχο

(Analytixlabs, n.d.)

2.2.2 Εφαρμογή της εξόρυξης γνώσης

- Οικονομική ανάλυση: Ο τραπεζικός και χρηματοοικονομικός κλάδος βασίζεται σε υψηλής ποιότητας, αξιόπιστα δεδομένα. Στις αγορές δανείων, τα οικονομικά δεδομένα και τα δεδομένα χρηστών μπορούν να χρησιμοποιηθούν για διάφορους σκοπούς, όπως η πρόβλεψη των πληρωμών των δανείων και ο καθορισμός της πιστοληπτικής ικανότητας. Και οι μέθοδοι εξόρυξης δεδομένων καθιστούν τις εργασίες αυτές πιο εύχρηστες. Οι τεχνικές ταξινόμησης διευκολύνουν τον διαχωρισμό των κρίσιμων παραγόντων που επηρεάζουν τις τραπεζικές αποφάσεις των πελατών από τους άσχετους. Περαιτέρω, οι τεχνικές πολυδιάστατης ομαδοποίησης επιτρέπουν τον εντοπισμό πελατών με παρόμοια συμπεριφορά πληρωμής δανείων. Η ανάλυση και εξόρυξη δεδομένων μπορεί επίσης να βοηθήσει στον εντοπισμό ξεπλύματος χρήματος και άλλων οικονομικών εγκλημάτων.
- Βιομηχανία τηλεπικοινωνιών: Επεκτείνεται και αναπτύσσεται με ταχείς ρυθμούς, ιδίως με την έλευση του διαδικτύου. Η εξόρυξη δεδομένων μπορεί να επιτρέψει στους βασικούς παράγοντες του κλάδου να βελτιώσουν την ποιότητα των υπηρεσιών τους. Η ανάλυση προτύπων χωροχρονικών βάσεων δεδομένων μπορεί να διαδραματίσει τεράστιο ρόλο στις κινητές τηλεπικοινωνίες, στην κινητή πληροφορική, καθώς και στις υπηρεσίες ιστού και πληροφοριών. Και τεχνικές όπως η ανάλυση ακραίων τιμών μπορούν να ανιχνεύσουν δόλιους χρήστες. Επίσης, τα εργαλεία OLAP και οπτικοποίησης μπορούν να βοηθήσουν στη σύγκριση πληροφοριών, όπως η συμπεριφορά

ομάδων χρηστών, το κέρδος, η κίνηση δεδομένων, η υπερφόρτωση του συστήματος κ.ά.

- Ανίχνευση εισβολής: Η παγκόσμια συνδεσιμότητα στη σημερινή οικονομία που καθοδηγείται από την τεχνολογία έχει παρουσιάσει προκλήσεις ασφαλείας για τη διοίκηση δικτύων. Οι πόροι του δικτύου μπορούν να αντιμετωπίσουν απειλές και ενέργειες που παραβιάζουν την εμπιστευτικότητα ή την ακεραιότητά τους. Ως εκ τούτου, η ανίχνευση εισβολών έχει αναδειχθεί σε κρίσιμη πρακτική εξόρυξης δεδομένων. Περιλαμβάνει ανάλυση συσχέτισης και συσχέτισης, τεχνικές συνάθροισης, οπτικοποίηση και εργαλεία ερωτημάτων, τα οποία μπορούν να ανιχνεύσουν αποτελεσματικά τυχόν ανωμαλίες ή αποκλίσεις από την κανονική συμπεριφορά
- Λιανικό εμπόριο: Ο οργανωμένος τομέας του λιανικού εμπορίου διαθέτει σημαντικές ποσότητες δεδομένων που καλύπτουν τις πωλήσεις, το ιστορικό αγορών, την παράδοση των αγαθών, την κατανάλωση και την εξυπηρέτηση των πελατών. Οι βάσεις δεδομένων έχουν γίνει ακόμη μεγαλύτερες με την άφιξη των αγορών ηλεκτρονικού εμπορίου. Στο σύγχρονο λιανικό εμπόριο, σχεδιάζονται και κατασκευάζονται αποθήκες δεδομένων για να αξιοποιηθούν πλήρως τα οφέλη της εξόρυξης δεδομένων. Η πολυδιάστατη ανάλυση δεδομένων βοηθά στην αντιμετώπιση δεδομένων που σχετίζονται με διαφορετικούς τύπους πελατών, προϊόντων, περιοχών και χρονικών ζωνών. Οι διαδικτυακοί λιανοπωλητές μπορούν επίσης να προτείνουν προϊόντα για να αυξήσουν τα έσοδα από τις πωλήσεις και να αναλύσουν την αποτελεσματικότητα των διαφημιστικών εκστρατειών τους. Έτσι, από την παρατήρηση των αγοραστικών προτύπων έως τη βελτίωση της εξυπηρέτησης και της ικανοποίησης των πελατών, η εξόρυξη δεδομένων ανοίγει πολλές πόρτες στον τομέα αυτό
- Τριτοβάθμια Εκπαίδευση: Καθώς η ζήτηση για τριτοβάθμια εκπαίδευση αυξάνεται παγκοσμίως, τα ιδρύματα αναζητούν καινοτόμες λύσεις για να καλύψουν τις αυξανόμενες ανάγκες. Τα ιδρύματα μπορούν να χρησιμοποιήσουν την εξόρυξη δεδομένων για να προβλέψουν ποιοι φοιτητές θα εγγραφούν σε ένα συγκεκριμένο πρόγραμμα, ποιοι θα χρειαστούν πρόσθετη βοήθεια για να αποφοιτήσουν, βελτιώνοντας συνολικά τη διαχείριση των εγγραφών. Επιπλέον, η πρόβλεψη της επαγγελματικής πορείας των φοιτητών και η παρουσίαση των δεδομένων θα γίνει πιο άνετη με την αποτελεσματική ανάλυση. Με αυτόν τον τρόπο, οι τεχνικές εξόρυξης δεδομένων μπορούν να βοηθήσουν στην αποκάλυψη των κρυμμένων μοτίβων σε τεράστιες βάσεις δεδομένων στον τομέα της τριτοβάθμιας εκπαίδευσης
- Βιομηχανία ενέργειας: Τα μεγάλα δεδομένα είναι διαθέσιμα ακόμη και στον τομέα της ενέργειας σήμερα. Τα μοντέλα δέντρων αποφάσεων και η μηχανική μάθηση διανυσμάτων υποστήριξης συγκαταλέγονται στις πιο δημοφιλείς προσεγγίσεις στον κλάδο, παρέχοντας εφικτές λύσεις για τη λήψη αποφάσεων και τη διαχείριση. Επιπλέον, η εξόρυξη δεδομένων μπορεί επίσης να επιτύχει παραγωγικά οφέλη με την πρόβλεψη των εκροών ισχύος και της τιμής εκκαθάρισης της ηλεκτρικής ενέργειας
- Εξόρυξη χωρικών δεδομένων: Τα Γεωγραφικά Συστήματα Πληροφοριών (GIS) και πολλές άλλες εφαρμογές πλοήγησης χρησιμοποιούν την εξόρυξη δεδομένων για την εξασφάλιση ζωτικής σημασίας πληροφοριών και την κατανόηση των επιπτώσεών τους. Αυτή η νέα τάση περιλαμβάνει την εξόρυξη γεωγραφικών, περιβαλλοντικών και αστρονομικών δεδομένων,

συμπεριλαμβανομένων εικόνων από το διάστημα. Τυπικά, η εξόρυξη χωρικών δεδομένων μπορεί να αποκαλύψει πτυχές όπως η τοπολογία και η απόσταση

- Ανάλυση βιολογικών δεδομένων: Οι πρακτικές εξόρυξης βιολογικών δεδομένων είναι κοινές στη γονιδιωματική, την πρωτεομική και τη βιοϊατρική έρευνα. Από τον χαρακτηρισμό της συμπεριφοράς των ασθενών και την πρόβλεψη των επισκέψεων στο ιατρείο μέχρι τον εντοπισμό ιατρικών θεραπειών για τις ασθένειές τους, οι τεχνικές της επιστήμης των δεδομένων παρέχουν πολλαπλά πλεονεκτήματα. Ορισμένες από τις εφαρμογές εξόρυξης δεδομένων στον τομέα της βιοπληροφορικής είναι: Σημασιολογική ολοκλήρωση ετερογενών και κατανεμημένων βάσεων δεδομένων, Ανάλυση συσχετίσεων και διαδρομών, Χρήση εργαλείων οπτικοποίησης, Ανακάλυψη δομικών προτύπων, Ανάλυση γενετικών δικτύων και πρωτεϊνικών μονοπατιών
- Ποινικές έρευνες: Οι δραστηριότητες εξόρυξης δεδομένων χρησιμοποιούνται επίσης στην Εγκληματολογία, η οποία είναι η μελέτη των χαρακτηριστικών του εγκλήματος. Πρώτον, οι εκθέσεις εγκλημάτων που βασίζονται σε κείμενο πρέπει να μετατραπούν σε αρχεία επεξεργασίας κειμένου. Στη συνέχεια, θα πραγματοποιηθεί η διαδικασία αναγνώρισης και επεξεργασίας εγκλημάτων με την ανακάλυψη μοτίβων σε μαζικές αποθήκες δεδομένων
- Αντιμετώπιση Τρομοκρατίας: Εξελιγμένοι μαθηματικοί αλγόριθμοι μπορούν να υποδείξουν ποια μονάδα πληροφοριών θα πρέπει να πρωταγωνιστήσει στις αντιτρομοκρατικές δραστηριότητες. Η εξόρυξη δεδομένων μπορεί να βοηθήσει ακόμη και σε καθήκοντα διοίκησης της αστυνομίας, όπως ο καθορισμός του τόπου ανάπτυξης του εργατικού δυναμικού και η επισήμανση των ερευνών στις συνοριακές διαβάσεις

(Sharma, 2021)

2.2.3 Τύποι πηγών δεδομένων στην εξόρυξη δεδομένων

Flat files

- Ως επίπεδα αρχεία ορίζονται τα αρχεία δεδομένων σε μορφή κειμένου ή δυαδική μορφή με δομή που μπορεί εύκολα να εξαχθεί από αλγορίθμους εξόρυξης δεδομένων
- Τα δεδομένα που είναι αποθηκευμένα σε επίπεδα αρχεία δεν έχουν καμία σχέση ή διαδρομή μεταξύ τους, όπως αν μια σχεσιακή βάση δεδομένων είναι αποθηκευμένη σε επίπεδο αρχείο, τότε δεν θα υπάρχουν σχέσεις μεταξύ των πινάκων
- Τα επίπεδα αρχεία αναπαρίστανται με λεξικό δεδομένων (όπως αρχείο CSV)
- Εφαρμογή: Χρησιμοποιείται στην αποθήκευση δεδομένων, στη μεταφορά δεδομένων από και προς το διακομιστή κ.ά.

Σχεσιακές βάσεις δεδομένων

- Μια σχεσιακή βάση δεδομένων ορίζεται ως η συλλογή δεδομένων οργανωμένων σε πίνακες με γραμμές και στήλες
- Το φυσικό σχήμα στις σχεσιακές βάσεις δεδομένων είναι ένα σχήμα που ορίζει τη δομή των πινάκων
- Το λογικό σχήμα στις σχεσιακές βάσεις δεδομένων είναι ένα σχήμα που ορίζει τη σχέση μεταξύ των πινάκων

- Εφαρμογή: Εξόρυξη δεδομένων, μοντέλο ROLAP κ.ά.

Αποθήκη δεδομένων

- Μια αποθήκη δεδομένων ορίζεται ως η συλλογή δεδομένων που ενσωματώνονται από πολλαπλές πηγές και που θα χρησιμοποιηθούν για την υποβολή ερωτημάτων και τη λήψη αποφάσεων
- Υπάρχουν τρεις τύποι αποθηκών δεδομένων: Αποθήκη επιχειρησιακών δεδομένων (Enterprise datawarehouse), πρατήριο δεδομένων (Data Mart) και εικονική αποθήκη (Virtual Warehouse)
- Δύο προσεγγίσεις μπορούν να χρησιμοποιηθούν για την ενημέρωση των δεδομένων στο DataWarehouse: προσέγγιση με βάση το ερώτημα και προσέγγιση με βάση την ενημέρωση
- Εφαρμογή: Εξόρυξη δεδομένων, Λήψη επιχειρηματικών αποφάσεων κ.ά.

Συναλλακτικές βάσεις δεδομένων

- Οι συναλλακτικές βάσεις δεδομένων είναι μια συλλογή δεδομένων που οργανώνονται με βάση τις χρονικές σφραγίδες, την ημερομηνία κ.ά. για την αναπαράσταση των συναλλαγών στις βάσεις δεδομένων
- Αυτός ο τύπος βάσης δεδομένων έχει τη δυνατότητα να ανατρέψει ή να αναιρέσει τη λειτουργία της όταν μια συναλλαγή δεν έχει ολοκληρωθεί ή δεσμευτεί
- Εξαιρετικά ευέλικτο σύστημα όπου οι χρήστες μπορούν να τροποποιούν πληροφορίες χωρίς να αλλάζουν ευαίσθητες πληροφορίες
- Ακολουθεί την ιδιότητα ACID των DBMS
- Εφαρμογή: Τραπεζικές εργασίες, κατανεμημένα συστήματα, βάσεις δεδομένων αντικειμένων κ.ά.

Βάσεις δεδομένων πολυμέσων

- Οι βάσεις δεδομένων πολυμέσων αποτελούνται από μέσα ήχου, βίντεο, εικόνες και κείμενο
- Μπορούν να αποθηκευτούν σε αντικειμενοστραφείς βάσεις δεδομένων
- Χρησιμοποιούνται για την αποθήκευση σύνθετων πληροφοριών σε προκαθορισμένες μορφές
- Εφαρμογή: Ψηφιακές βιβλιοθήκες, βίντεο κατά παραγγελία, ειδήσεις κατά παραγγελία, μουσική βάση δεδομένων κ.ά.

Χωρική βάση δεδομένων

- Αποθήκευση γεωγραφικών πληροφοριών
- Αποθηκεύει δεδομένα με τη μορφή συντεταγμένων, τοπολογίας, γραμμών, πολυγώνων κ.ά.
- Εφαρμογή: Χάρτες, παγκόσμιος εντοπισμός θέσης κ.ά.

Βάσεις δεδομένων χρονοσειρών

- Οι βάσεις δεδομένων χρονοσειρών περιέχουν δεδομένα χρηματιστηρίου και καταγεγραμμένες δραστηριότητες χρηστών
- Χειρίζεται συστοιχία αριθμών με ευρετήριο την ώρα, την ημερομηνία κ.ά.
- Απαιτεί ανάλυση σε πραγματικό χρόνο
- Εφαρμογή: eXtremeDB, Graphite, InfluxDB κ.ά.

World Wide Web

- Ο όρος WWW αναφέρεται στον Παγκόσμιο Ιστό και είναι μια συλλογή εγγράφων και πόρων, όπως ήχος, βίντεο, κείμενο κ.ά. τα οποία αναγνωρίζονται από ομοιόμορφους εντοπιστές πόρων (URL) μέσω φυλλομετρητών ιστού, συνδέονται με σελίδες HTML και είναι προσβάσιμα μέσω του διαδικτύου
- Είναι το πιο ετερογενές αποθετήριο, καθώς συλλέγει δεδομένα από πολλούς πόρους
- Είναι δυναμικό στη φύση του, καθώς ο όγκος των δεδομένων αυξάνεται και μεταβάλλεται συνεχώς
- Εφαρμογή: Διαδικτυακές αγορές, αναζήτηση εργασίας, έρευνα, σπουδές κ.ά.

(Geeksforgeeks, 2022)

2.2 Οπτικοποίηση Δεδομένων

Από τις αρχές του 21ου αιώνα, τα ψηφιακά δεδομένα που παράγονται από τις προηγμένες τεχνολογίες αυξάνονται με γεωμετρική πρόοδο κάθε χρόνο. Τα big data είναι ένας τεράστιος όγκος συλλογής ψηφιακών δεδομένων που δεν μπορούν να υποστούν επεξεργασία μέσω των παραδοσιακών βάσεων δεδομένων και των υφιστάμενων εργαλείων. Το πρόβλημα εδώ δεν σχετίζεται μόνο με τον όγκο τους, αλλά και με τις διαφορετικές μορφές των πληροφοριών, δηλαδή τα μη δομημένα δεδομένα που παράγονται με μεγάλη ταχύτητα. Η ταχέως αυξανόμενη ροή δεδομένων που παράγονται από διάφορες πηγές, όπως τα κοινωνικά δίκτυα, οι τεχνητοί δορυφόροι της Γης, τα δίκτυα αισθητήρων κ.α. , καθιστά τη διαδικασία διαχείρισης δεδομένων, συμπεριλαμβανομένης της απόκτησης πληροφοριών και γνώσης, ακόμη πιο δύσκολη (J. Gantz, 2012). Η εφαρμογή εργαλείων οπτικοποίησης δεδομένων διαδραματίζει σημαντικό ρόλο στην εξάλειψη του προβλήματος της επεξεργασίας και ανάλυσης δεδομένων. Η οπτικοποίηση αναδύθηκε στα τέλη της δεκαετίας του 1980 ως ο κύριος τομέας της επιστήμης των πληροφοριών. Καλύπτει και ενσωματώνει τομείς όπως τα γραφικά υπολογιστών, η επεξεργασία εικόνας, η οπτικοποίηση υπολογιστών, η επεξεργασία σήματος, τα αυτοματοποιημένα συστήματα σχεδιασμού και η αλληλεπίδραση ανθρώπου-υπολογιστή. Η οπτικοποίηση παίζει επίσης ρόλο γέφυρας που συνδέει έναν υπολογιστή με ένα σύστημα ανθρώπινης όρασης. Εντοπίζει περιγραφές, δημιουργεί υποθέσεις και αναπτύσσει ιδέες από μεγάλους όγκους, συμβάλλοντας στην επιστημονική έρευνα και πρόβλεψη.
Εντοπίζει περιγραφές, δημιουργεί υποθέσεις και αναπτύσσει ιδέες από μεγάλους όγκους δεδομένων, συμβάλλοντας στην επιστημονική έρευνα και πρόβλεψη. Αν και η χρήση των τεχνολογιών υπολογιστών στην οπτικοποίηση ξεκίνησε τη δεκαετία του 1990, η οπτικοποίηση έχει μια αρχαία ιστορία, όπως οι γεωγραφικοί χάρτες, ο περιοδικός πίνακας του Μεντελέγιεφ κ.ά. (Tzu-Wei Hsu, 2004).

Η οπτικοποίηση δεδομένων είναι μια διαδικασία ερμηνείας των αποτελεσμάτων της ανάλυσης με διαφορετικούς τρόπους για την παροχή μιας πιο αποτελεσματικής διαδικασίας λήψης αποφάσεων. Ο ειδικός στην οπτικοποίηση δεδομένων Edward Tufte λέει: "Ο κόσμος είναι πολύπλοκος, δυναμικός, πολυδιάστατος, το χαρτί είναι στατικό, επίπεδο. Πώς θα αναπαραστήσουμε τον πλούσιο οπτικό κόσμο της εμπειρίας και της μέτρησης σε απλή επίπεδη επιφάνεια;" Πράγματι, στη σύγχρονη εποχή, τα δεδομένα είναι εύκολα προσβάσιμα, ωστόσο είναι δύσκολο να τα κατανοήσουμε και να τα κατανοήσουμε. Αυτός είναι ο λόγος για τον οποίο τα βελτιωμένα εργαλεία οπτικοποίησης έχουν γίνει το κύριο συστατικό της σύγχρονης κοινωνίας. Έτσι, η οπτικοποίηση των πληροφοριών μπορεί να διαδραματίσει πολύ σημαντικό ρόλο σε αυτή τη διαδικασία. Ένας συνδυασμός σωστά επιλεγμένων απεικονίσεων, λέξεων και σχημάτων μπορεί να δημιουργήσει μια πλήρη εικόνα της αντίληψης των δεδομένων. Όσον αφορά την επεξεργασία των μεγάλων δεδομένων, αυτό δεν είναι ένα απλό θέμα και απαιτεί εξαιρετικές μεθόδους και προσεγγίσεις. Η γραφική περιγραφή είναι ένας από τους πιο αποτελεσματικούς τρόπους αντικειμενικών αξιολογήσεων και λήψης της σωστής απόφασης στη λύση του προβλήματος. Ωστόσο, στην περίπτωση των μεγάλων δεδομένων, οι πιο κλασικές μέθοδοι είναι αναποτελεσματικές, ακόμη και αδύνατο να εφαρμοστούν σε συγκεκριμένα προβλήματα και ως εκ τούτου έχει αναπτυχθεί μεγάλος αριθμός μεθόδων οπτικοποίησης για την ταχεία παρουσίαση επεξεργασμένων δεδομένων. Ωστόσο, οι νέες μέθοδοι οπτικοποίησης μεγάλων δεδομένων διαμορφώνουν νέα ερευνητικά ζητήματα και λύσεις. Ταυτόχρονα, τα χαρακτηριστικά των μεγάλων όγκων δεδομένων, όπως ο όγκος, η ταχύτητα, η ποικιλία (C.L. Philip Chen, 2014), η αξία και η ειλικρίνεια, απαιτούν ευέλικτες αποφάσεις και είναι σημαντικό να μελετηθεί ο τομέας αυτός ως αντικείμενο και να διερευνηθούν τα επιστημονικά και θεωρητικά του προβλήματα.

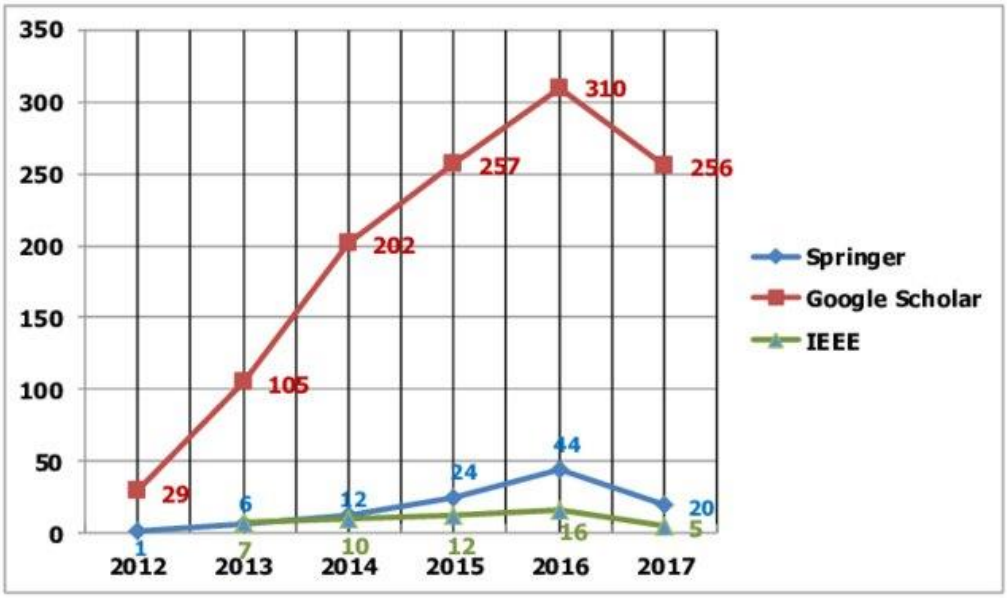
Η έννοια, η ιστορία, η ουσία και τα στάδια της οπτικοποίησης

Ο όρος οπτικοποίηση προέρχεται από τη λατινική λέξη "visualis" και αναφέρεται στη φαντασία, την παρατήρηση και την παρουσίαση των οπτικών αποτελεσμάτων της κατάλληλης παρατήρησης και ανάλυσης ψηφιακών πληροφοριών ή φυσικών γεγονότων. Ο Αμερικανός ψυχολόγος Michael Friendly δομεί την ιστορία της οπτικοποίησης δεδομένων στα ακόλουθα στάδια (Kumar, 2016) :

- μέχρι τον XVII αιώνα: πρώιμοι χάρτες και διαγράμματα
- 1600-1699: μετρήσεις και θεωρίες
- 1700-1799: νέες γραφικές μορφές
- 1800-1850: απαρχές των σύγχρονων γραφικών
- 1850-1900: χρυσή εποχή της στατιστικής
- 1900-1950: σκοτεινές εποχές
- 1950 - 1975 - αναγέννηση της οπτικοποίησης δεδομένων
- 1975-σημερινή εποχή: διαδραστική και δυναμική οπτικοποίηση δεδομένων

Σήμερα ένα ευρύ φάσμα εργαλείων ανάλυσης και οπτικοποίησης δεδομένων έχει δημιουργήσει τις προϋποθέσεις για διαδραστική και δυναμική οπτικοποίηση. Όλα αυτά οφείλονται στη διαθεσιμότητα διαδραστικών συστημάτων, στις δυνατότητες των τρισδιάστατων μοντέλων, στην αυξημένη υπολογιστική ισχύ και κυρίως, στην προσβασιμότητα των δεδομένων λόγω της διαθεσιμότητας του Διαδικτύου. Σήμερα ένα ευρύ φάσμα εργαλείων ανάλυσης και οπτικοποίησης δεδομένων έχει δημιουργήσει τις προϋποθέσεις για διαδραστική και δυναμική οπτικοποίηση. Όλα

αυτά οφείλονται στη διαθεσιμότητα διαδραστικών συστημάτων, στις δυνατότητες των τρισδιάστατων μοντέλων, στην αυξημένη υπολογιστική ισχύ και κυρίως, στην προσβασιμότητα των δεδομένων λόγω της διαθεσιμότητας του Διαδικτύου. Διεξάγονται θεμελιώδεις μελέτες στον τομέα της οπτικοποίησης μεγάλων δεδομένων σε γνωστά επιστημονικά κέντρα σε όλο τον κόσμο, όπως η Λευκή Βίβλος της Intel κ.ά (Intel IT Center, 2013).

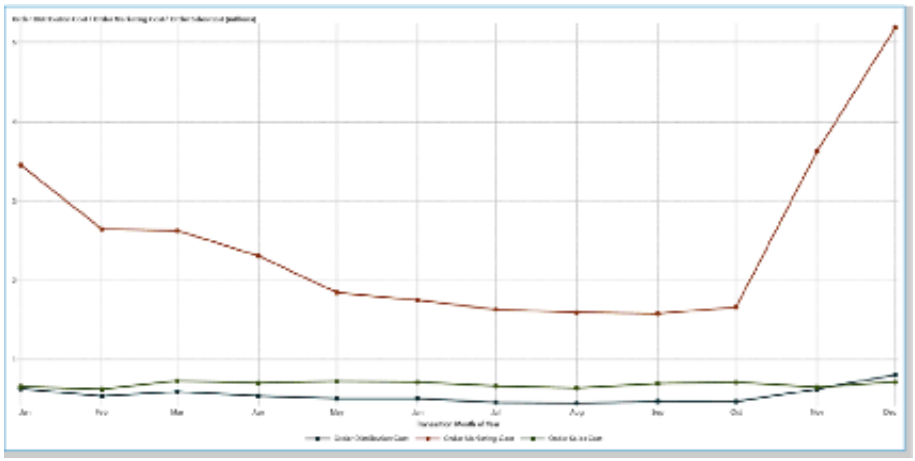


Εικόνα 34 Διανομή της έρευνας οπτικοποίησης μεγάλων δεδομένων στις βάσεις δεδομένων Springer, Google Scholar και IEEE

Παραδοσιακές μέθοδοι οπτικοποίησης δεδομένων και υφιστάμενες προσεγγίσεις

Οι πληροφορίες αποτελούν πλέον αναπόσπαστο μέρος της ανθρώπινης ζωής. Ένας μεγάλος αριθμός πληροφοριών ή δεδομένων παράγεται από διάφορες πηγές και κάθε χρόνο οι άνθρωποι πρέπει να ασχολούνται με αυτές τις πληροφορίες. Η παρουσίαση των αποτελεσμάτων της ανάλυσης δεδομένων είναι πολύ σημαντική για την καλύτερη κατανόηση. Παραδοσιακά, έχουν χρησιμοποιηθεί διάφορες μέθοδοι για την οπτικοποίηση των δεδομένων. Αυτές οι μέθοδοι οπτικοποίησης ταξινομούνται σε τρεις μεγάλες ομάδες, ως στατικές, δυναμικές και διαδραστικές. Υπάρχουν πολλές προσεγγίσεις σε αυτές τις μεθόδους και ορισμένες από αυτές μπορούν να χρησιμοποιηθούν και στις τρεις ομάδες.

- Το γραμμικό γράφημα δείχνει την αλληλεπίδραση μιας μεταβλητής με μια άλλη. Ένα τέτοιο γράφημα χρησιμοποιείται συχνά για την παρακολούθηση των αλλαγών που έχουν λάβει χώρα κατά τη διάρκεια μιας μεγάλης χρονικής περιόδου και για τη σύγκριση πολλών αντικειμένων ταυτόχρονα. Το γραμμικό γράφημα χρησιμοποιείται για την οπτική περιγραφή μιας ή περισσότερων μεταβλητών και του ρυθμού μεταβολής των πληροφοριών σχετικά με το περιεχόμενο αυτών των μεταβλητών (Muzammil Khan, 2011)

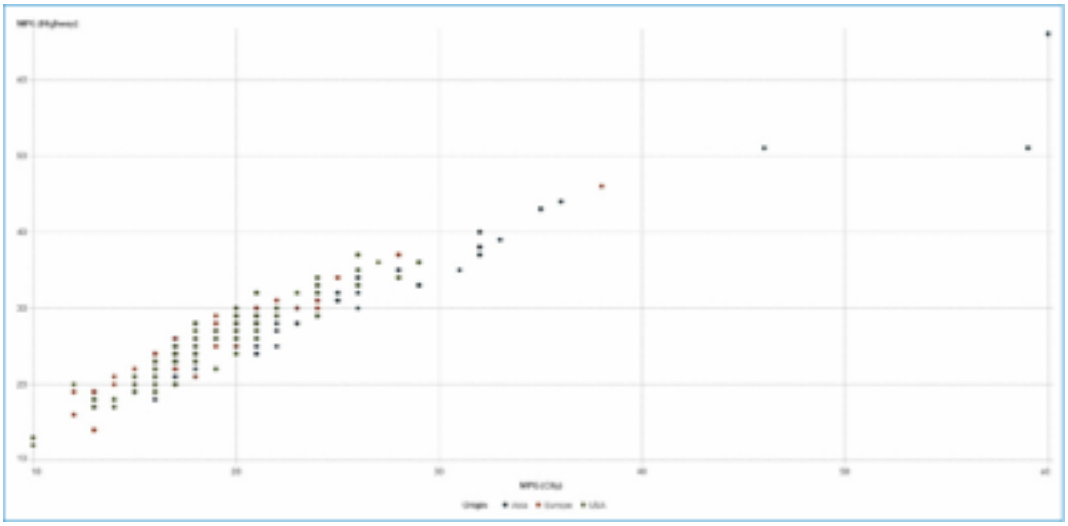


Εικόνα 35 γραμμικό γράφημα [πηγή: (SAS, n.d.)]

- Τα ραβδογράμματα χρησιμοποιούνται συχνά για τη σύγκριση ποσοτικών δεικτών δεδομένων διαφορετικών ομάδων και κατηγοριών. Τα περιεχόμενα των κατηγοριών εμφανίζονται σε ράβδους και μπορούν να είναι κατακόρυφες και οριζόντιες και το ύψος και το μήκος τους αντιστοιχούν σε ορισμένες τιμές δεδομένων. Εάν οι τιμές αυτές είναι αρκετά διαφορετικές, η διαφορά στις ράβδους θα είναι εμφανής και μπορεί να χρησιμοποιηθεί και ένα απλό ραβδόγραμμα. Όταν οι τιμές είναι πολύ κοντά η μία στην άλλη ή οι αριθμοί είναι υψηλοί και η διαφορά τους γίνεται δύσκολη. Για το σκοπό αυτό, οι ράβδοι μπορούν να παρουσιαστούν με διαφορετικά χρώματα (SAS, n.d.)

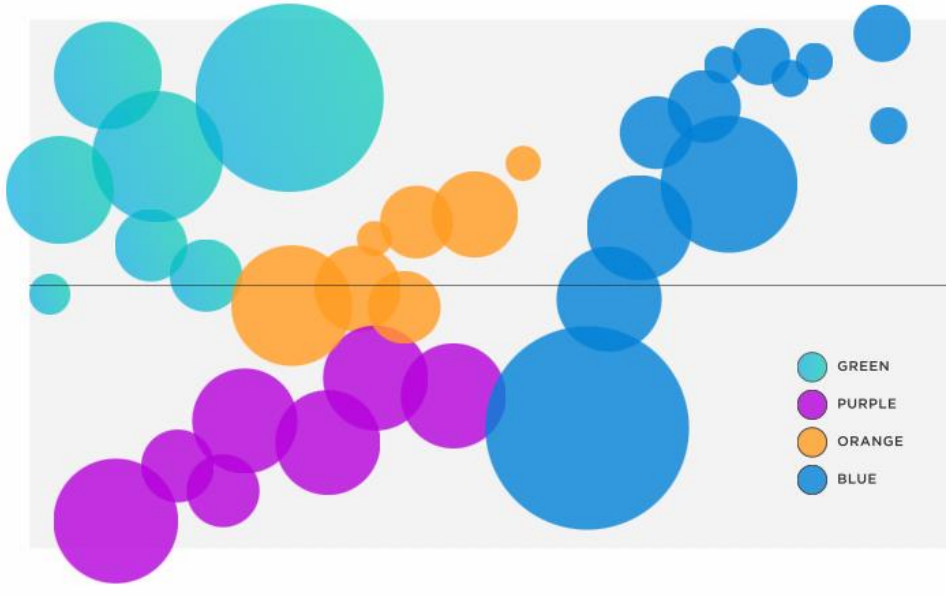
Εικόνα 36 ραβδογράμματα [πηγή: (SAS, n.d.)]

- Το διάγραμμα διασποράς χρησιμοποιεί το καρτεσιανό σύστημα συντεταγμένων για να περιγράψει το σύνολο δεδομένων κυρίως δύο μεταβλητών μαθηματικών διαγραμμάτων. Σε ορισμένες περιπτώσεις, ο αριθμός των μεταβλητών μπορεί να αυξηθεί σε 3 με τη χρήση έγχρωμων εικονιδίων. Με αυτόν τον τρόπο μπορεί να προσδιοριστεί η κατεύθυνση και η γραμμικότητα της εξάρτησης μεταξύ των μεταβλητών που εμφανίζονται στο διάγραμμα. Όσο αυξάνεται ο αριθμός των σημείων στο διάγραμμα, τόσο υψηλότερο είναι το ποσοστό συσχέτισης. Το διάγραμμα διασποράς χρησιμοποιείται κυρίως για την απεικόνιση πολυδιάστατων δεδομένων (SAS, χ.χ.)



Εικόνα 37 Διάγραμμα Διασποράς [πηγή: (SAS, n.d.)]

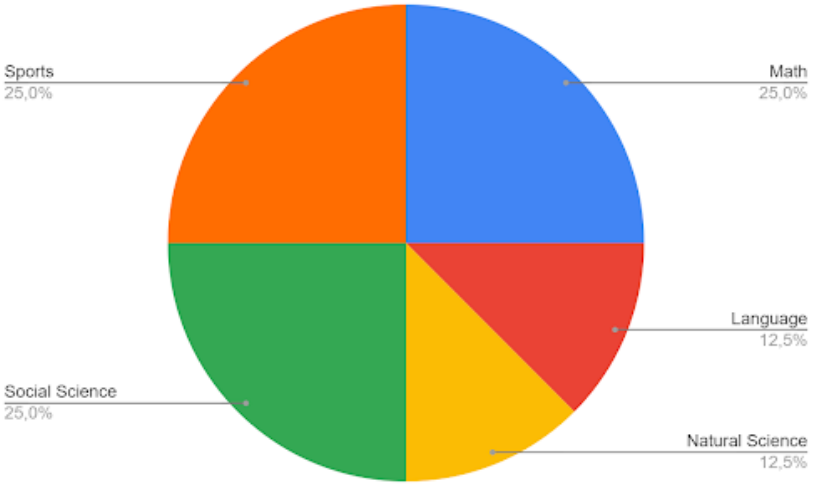
- Τα διαγράμματα φυσαλίδων είναι ένας τύπος διαγράμματος κουκκίδων, στο οποίο οι κουκκίδες παρουσιάζονται με τη μορφή φυσαλίδων νερού και διαφέρουν μεταξύ τους ως προς το μέγεθος και τη θέση τους. Εδώ, μαζί με τους άξονες X και Y, χρησιμοποιείται και ο Z. Το διάγραμμα φυσαλίδων χρησιμοποιείται συχνά για την παρουσίαση χρηματοοικονομικών δεδομένων (Muzammil Khan, 2011)
- Τα διαγράμματα φυσαλίδων είναι ένας τύπος διαγράμματος κουκκίδων, στο οποίο οι κουκκίδες παρουσιάζονται με τη μορφή φυσαλίδων νερού και διαφέρουν μεταξύ τους ως προς το μέγεθος και τη θέση τους. Εδώ, μαζί με τους άξονες X και Y, χρησιμοποιείται και ο Z και χρησιμοποιείται συχνά για την παρουσίαση χρηματοοικονομικών δεδομένων



Εικόνα 38 διαγράμματα φυσαλίδων [πηγή: (Tibco, n.d.)]

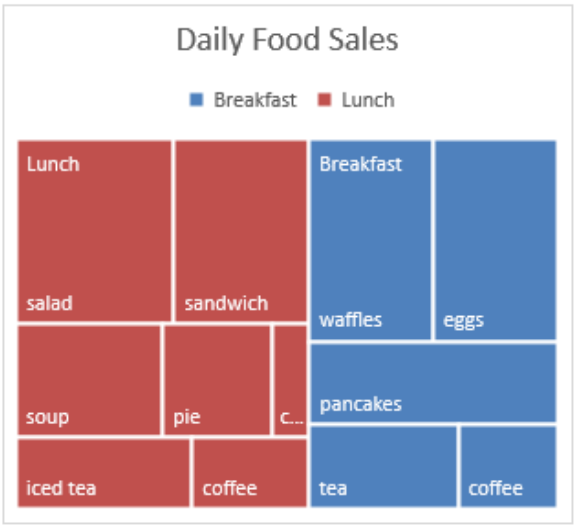
- Το διάγραμμα πίτας περιγράφει τα δεδομένα ως τμήμα κύκλου ή ως ποσοστό. Με τη βοήθειά του, τα δεδομένα παρουσιάζονται με μία μόνο σειρά και υποδεικνύει την αναλογία του μεγέθους των στοιχείων προς το άθροισμά τους. Προσφέρονται πολλές προσεγγίσεις για τη βελτίωση των κυκλικών διαγραμμάτων. Δεδομένου ότι δεν είναι τόσο εύκολη η σύγκριση των

τμημάτων του διαγράμματος με το ανθρώπινο μάτι, υπάρχουν προβλήματα με την ερμηνεία των αποτελεσμάτων (Abhishek Kaushik, 2016)



Εικόνα 39 Διάγραμμα πίτας [πηγή: (Jiménez, 2021)]

- Ο δενδρικός χάρτης εμφανίζει όλα τα δεδομένα ως στοιχεία ενός ιεραρχικού δέντρου. Παρουσιάζει ορθογώνια που σχετίζονται με τα κλαδιά του δέντρου, τα οποία διαφέρουν μεταξύ τους ως προς το χρώμα και το μέγεθος



Εικόνα 40 χάρτης δέντρων [πηγή: (Microsoft, n.d.)]

- Το Sunburst είναι σχετικά νεότερη μέθοδος και χρησιμοποιείται συχνά ως εναλλακτική λύση στους "χάρτες δέντρων". Η κύρια διαφορά μεταξύ αυτών των μεθόδων είναι ότι τα δεδομένα διαφέρουν όχι μόνο ως προς το ύψος και το πλάτος αλλά και ως προς την ακτίνα και το μήκος των κλάδων. Κατά συνέπεια, ένας τομέας με νέα δεδομένα μπορεί να αλλάξει με την αλλαγή της ακτίνας, ενώ ολόκληρο το σύστημα παραμένει αμετάβλητο. Λόγω αυτού του χαρακτηριστικού, η μέθοδος μπορεί να προσαρμοστεί ώστε να δείχνει τη δυναμική των δεδομένων με τη χρήση κινούμενων σχεδίων (Evgeniy Yur'evich Gorodov, 2013)

Εικόνα 41 Sunburst [πηγή: (Excel Dashboard School, 2022)]

Οι πιο συχνά χρησιμοποιούμενοι τύποι χαρτών είναι οι γεωγραφικοί, οι φωτογραφικοί, οι οδικοί και οι ηλεκτρονικοί θεματικοί. Προφανώς, οι γεωγραφικοί χάρτες χρησιμοποιούνται για τη σχηματική αναπαράσταση των γεωγραφικών αντικειμένων. Οι φωτογραφικοί χάρτες είναι η φωτογραφική αναπαράσταση των γεωγραφικών αντικειμένων από το δορυφόρο. Οι οδικοί χάρτες παρουσιάζουν τα σχήματα των αυτοκινητοδρόμων, των σιδηροδρόμων και άλλων οδών.
Ανάλογα με τους τομείς εφαρμογής, οι μέθοδοι απεικόνισης ομαδοποιούνται ως εξής:

- Επιστημονική απεικόνιση
- Οπτικοποίηση πληροφοριών
- Οπτικοποίηση λογισμικού

<u>Προβλήματα οπτικοποίησης μεγάλων δεδομένων</u>

Ένα από τα βασικά ζητήματα στην ανάλυση μεγάλων δεδομένων είναι η παρουσίαση των αποτελεσμάτων, δηλαδή η οπτικοποίηση. Σε γενικές γραμμές, η οπτικοποίηση δεδομένων είναι μια από τις απλούστερες και πιο τακτικές μεθόδους στην επεξεργασία και ανάλυση δεδομένων. Βοηθά στην ταχύτερη κατανόηση των παρουσιαζόμενων πληροφοριών και στη λήψη των βέλτιστων αποφάσεων μέσω της αξιολόγησης των αποτελεσμάτων. Η χρήση παραδοσιακών μεθόδων για την οπτικοποίηση δεδομένων μεγάλης κλίμακας και όγκου είναι αναποτελεσματική. Μελέτες δείχνουν ότι τα κυριότερα επιστημονικά και θεωρητικά προβλήματα της οπτικοποίησης μεγάλων δεδομένων, τα οποία έχουν χαρακτηριστικά όπως ο όγκος, η ταχύτητα και η ποικιλία, είναι τα εξής:

- Οπτικός θόρυβος. Αυτό το πρόβλημα εμφανίζεται όταν τα αντικείμενα είναι υπερβολικά συνδεδεμένα μεταξύ τους στο σύνολο δεδομένων. Ο οπτικός θόρυβος δεν αποτελεί παραμόρφωση ή βλάβη των δεδομένων, αλλά ελαχιστοποίηση ή απώλεια των εικόνων στην οθόνη. Αυτό περιπλέκει την εξαγωγή χρήσιμων πληροφοριών από ολόκληρη την εικόνα και απαιτεί πρόσθετη επεξεργασία
- Αντίληψη μεγάλης εικόνας. Το πρόβλημα μπορεί να επιλυθεί με τη διανομή των δεδομένων σε μεγαλύτερες οθόνες, ωστόσο μερικές φορές μπορεί να εμφανιστεί το πρόβλημα της αντίληψης μιας μεγάλης εικόνας. Ο ανθρώπινος

εγκέφαλος μπορεί να αντιληφθεί μια οπτική εικόνα σε ένα ορισμένο επίπεδο και μετά από ένα ορισμένο επίπεδο, ο άνθρωπος χάνει την ικανότητα αντίληψης των πληροφοριών εκτός από τα εξαιρετικά υπερφορτωμένα οπτικά δεδομένα. Οι δυνατότητες όλων των μεθόδων οπτικοποίησης περιορίζονται στις δυνατότητες των τεχνικών συσκευών που παρέχουν την παρουσίαση των δεδομένων

- Απώλεια πληροφοριών. Το πρόβλημα αυτό προκύπτει από τη λύση του προβλήματος του οπτικού θορύβου και της αντίληψης μεγάλων εικόνων. Οι προσεγγίσεις για τη λύση του προβλήματος μειώνουν τα τελικά δεδομένα, δημιουργώντας έτσι ένα νέο πρόβλημα απώλειας πληροφορίας. Δεδομένου ότι οι μέθοδοι μείωσης της οπτικής πληροφορίας βασίζονται στην εγγύτητα των αντικειμένων, τα δεδομένα αυτά συγκεντρώνονται και φιλτράρονται σύμφωνα με ένα ή περισσότερα κριτήρια. Αυτές οι προσεγγίσεις μπορούν να παραπλανήσουν τους αναλυτές, αφήνοντας απαρατήρητα τα κρυμμένα και πιο σημαντικά αντικείμενα. Επιπλέον, η συνάθροιση των δεδομένων μπορεί να απαιτεί πολύ χρόνο και υπολογιστικούς πόρους για την απόκτηση ακριβών και σχετικών πληροφοριών
- Υψηλές απαιτήσεις απόδοσης. Η γραφική ανάλυση δεν περιορίζεται μόνο στη στατική απεικόνιση των εικόνων, χρησιμοποιείται επίσης για τη δυναμική απεικόνιση, η οποία οδηγεί σε απαρατήρητο πρόβλημα στη στατική απεικόνιση. Στον συγκεκριμένο ρυθμό της οπτικοποίησης, προκύπτει η ανάγκη για υψηλές επιδόσεις επειδή η διαδικασία της ανάλυσης απαιτεί μεγάλη ποσότητα υπολογιστικών πόρων και χρόνου για το φιλτράρισμα μεγάλου αριθμού δεδομένων
- Υψηλός ρυθμός αλλαγής εικόνας. Όπως φαίνεται από το όνομα, το πρόβλημα αυτό οφείλεται στην ταχεία αλλαγή των εικόνων. Δηλαδή, κατά τη διάρκεια της παρατήρησης, ένα άτομο απλά δεν μπορεί να αντιδράσει στην ταχεία αλλαγή των δεδομένων ή της έντασής τους στην οθόνη. Η μείωση των αλλαγμένων δεδομένων δεν μπορεί να προσφέρει την επιθυμητή αποτελεσματικότητα της διαδικασίας. Ωστόσο, ο ρυθμός των ανθρώπινων αντιδράσεων προκαλεί ορισμένους περιορισμούς στη διαδικασία αυτή

(Lidong Wang, 2015)

<u>Νέες μέθοδοι και εργαλεία για την οπτικοποίηση μεγάλων δεδομένων</u>

Η οπτική αναπαράσταση μεγάλων δεδομένων στοχεύει κυρίως:

- Στον εντοπισμό της εμπιστευτικότητας και των ανωμαλιών των δεδομένων
- Την αύξηση της ευελιξίας στην αναζήτηση ορισμένων τιμών
- Σύγκριση των διαφόρων μπλοκ προκειμένου να προκύψει η σχετική διαφορά στον όγκο
- Καθιέρωση ανθρώπινων σχέσεων για τη λειτουργία σε πραγματικό χρόνο

Πολλές από τις κλασικές μεθόδους οπτικοποίησης δεν είναι σε θέση να επιλύσουν συγκεκριμένα προβλήματα μεγάλων δεδομένων, γεγονός που οδήγησε στην εμφάνιση νέων εργαλείων και τεχνικών οπτικοποίησης.

Ακολουθούν ορισμένες από τις πιο συχνά χρησιμοποιούμενες μεθόδους για την οπτικοποίηση μεγάλων δεδομένων:

- Tag Cloud: ένας όρος που περιγράφει το περιεχόμενο του εγγράφου ή του συνόλου ή μια οπτική περιγραφή του δημοφιλούς Web 2.0, η οποία αποτελείται από σύντομες φράσεις. Συνήθως, οι λέξεις-κλειδιά ή τα αναφερόμενα αντικείμενα που προέρχονται από το περιεχόμενο του εγγράφου περιγράφονται με τη χρήση τεχνικών επεξεργασίας φυσικής γλώσσας. Το μέγεθος, το χρώμα και η θέση των λέξεων διέπονται από κριτήρια σπουδαιότητας, αισθητικής και άνεσης. Η μέθοδος αυτή υιοθετεί έναν συντελεστή βάρους για κάθε στοιχείο. Όσο μεγαλύτερος είναι ο συντελεστής, τόσο μεγαλύτερο θα είναι το μέγεθος της γραμματοσειράς. Ο συντελεστής βάρους εξαρτάται από τη σημασία του στοιχείου που καθορίζεται από τον εμπειρογνώμονα, τη συχνότητα της κατάστασής του και άλλους παράγοντες (Big data: The next frontier for innovation, competition, and productivity, 2011)
- Clustergram: μια μέθοδος οπτικοποίησης που χρησιμοποιείται για να δείξει πώς τα στοιχεία των δεδομένων ενσωματώνονται στις συστάδες καθώς ο αριθμός των συστάδων αυξάνεται κατά τη διάρκεια της ανάλυσης (Ekaterina Olshannikova, 2015)
- Διαγράμματα κίνησης: επιτρέπει την αποτελεσματική εξερεύνηση και αλληλεπίδραση με μεγάλα και πολυδιάστατα δεδομένα χρησιμοποιώντας δισδιάστατα διαγράμματα φυσαλίδων. Οι φυσαλίδες (κύρια αντικείμενα αυτής της μεθόδου) μπορούν να ελέγχονται σύμφωνα με την εξεταζόμενη μεταβαλλόμενη εικόνα (Ekaterina Olshannikova, 2015)
- Dashboard: εμφανίζει αρχεία καταγραφής διαφορετικών μορφών και τα φιλτράρει με βάση επιλεγμένα εύρη δεδομένων. Το ταμπλό αποτελείται από τρία επίπεδα, δηλαδή δεδομένα (ακατέργαστα δεδομένα), ανάλυση (δεδομένα που μεταδίδονται από τους τύπους και τα επίπεδα δεδομένων σε πίνακες) και παρουσίαση (γραφικές παρουσιάσεις με βάση το επίπεδο ανάλυσης) (Ekaterina Olshannikova, 2015)
- Ροή ιστορικού: περιγράφει την εξέλιξη ενός εγγράφου που έχει τροποποιηθεί από πολλούς συγγραφείς. Επιτρέπει την παρακολούθηση του συγγραφέα που επεξεργάζεται το έγγραφο, του τι προστίθεται στο έγγραφο και του χρόνου που δαπανάται. Εδώ, ο οριζόντιος άξονας δηλώνει τον χρόνο, ενώ ο κάθετος άξονας το κείμενο που έχει εισαχθεί. Κάθε συγγραφέας έχει διαφορετικό χρωματικό κώδικα και το κάθετο μήκος της συμβολοσειράς δείχνει τον αριθμό του κειμένου που γράφτηκε από κάθε συγγραφέα (Christin Seifert, 2014)

Η οπτική αναπαράσταση της ανάλυσης μεγάλων δεδομένων είναι πολύ σημαντική για την ερμηνεία της. Όπως σημειώθηκε, η ανθρώπινη αντίληψη είναι περιορισμένη. Τα αποτελεσματικά εργαλεία οπτικοποίησης θα πρέπει να λαμβάνουν υπόψη τα χαρακτηριστικά του ανθρώπινου εγκεφάλου, όπως η κατανόηση και η αντίληψη. Οι προηγμένες μέθοδοι οπτικοποίησης δεδομένων στοχεύουν στη βελτίωση των εικόνων, των διαγραμμάτων και των κινούμενων σχεδίων. Αν και τα εργαλεία ανάκτησης μεγάλων δεδομένων είναι ικανά να επεξεργάζονται petabytes (PB) και zetabytes (ZB) δεδομένων, μερικές φορές δεν μπορούν να τα οπτικοποιήσουν. Δεδομένου ότι τα μεγάλα δεδομένα επεκτείνονται με εξαιρετικά υψηλή ταχύτητα, η έλλειψη κλιμακούμενων εργαλείων οπτικοποίησης περιπλέκει τη διαδικασία απόκτησης εμπιστευτικών πληροφοριών. Διάφορα εργαλεία οπτικοποίησης χρησιμοποιούνται για να ξεπεραστούν τα προαναφερθέντα προβλήματα:

- Το Tableau παρέχει εκτεταμένες δυνατότητες διαδραστικής απεικόνισης για απεικόνιση δεδομένων με προσανατολισμό στην επιχειρηματική ανάλυση. Είναι γρήγορο και ευέλικτο, η διεπαφή χρήστη είναι διαισθητική και διατίθενται πολλά διαγράμματα. Δεν απαιτούνται δεξιότητες κωδικοποίησης για απλούς υπολογισμούς και στατιστικά στοιχεία. Ωστόσο, η γλώσσα προγραμματισμού και το περιβάλλον είναι ενεργοποιημένα στην R μόνο για την ανάλυση μοντέλων για την προκλητική ανάλυση, την ανάλυση δεδομένων, τους στατιστικούς υπολογισμούς και την οπτικοποίηση δεδομένων, και στη συνέχεια τα αποτελέσματα μπορεί να μεταβεί στο Tableau. Το Tableau αποτελείται από τρία βασικά εργαλεία, δηλαδή το Tableau Desktop, το Tableau Server και το Tableau Public (Tableau, n.d.)
- Microsoft Power Business Intelligence (BI). Πρόκειται για μια ισχυρή υπηρεσία επιχειρησιακής ανάλυσης που βασίζεται στο νέφος και διαθέτει διαδραστικές και πλούσιες δυνατότητες οπτικοποίησης. Το Power BI αποτελείται από τρία στοιχεία: Power BI Desktop, λογισμικό ως υπηρεσία (SaaS) και εφαρμογές. Κάθε υπηρεσία είναι προσβάσιμη και ως εκ τούτου ευέλικτη και εύχρηστη. Το Power BI ενσωματώνει έως και 60 τύπους πηγών και μπορεί να ξεκινήσει την οπτικοποίηση μέσα σε λίγα λεπτά. Συνδυάζει επίσης το Microsoft Office, το SharePoint και την SQL. Αυτό το εργαλείο οπτικοποίησης δεν χρειάζεται δεξιότητες προγραμματισμού για έρευνες (Microsoft, n.d.)
- Το Plotly, επίσης γνωστό ως Plot.ly, χρησιμοποιεί τη γλώσσα προγραμματισμού Python και την εφαρμογή ιστού Django για ανάλυση και οπτικοποίηση δεδομένων. Με ορισμένες περιορισμένες επιλογές, το Plotly είναι ελεύθερα διαθέσιμο για τους χρήστες. Αυτές οι επιλογές απαιτούν επαγγελματική συνδρομή, και επίσης τα διαγράμματα και οι πίνακες οργάνων δημιουργούνται online μέσω αυτών των επιλογών. Τα διαγράμματα μπορεί να είναι στατιστικά, επιστημονικά, τρισδιάστατα, πίνακες πληροφοριών κ.ά. Επιπλέον, το Plotly χρησιμοποιεί το εργαλείο DWP (Digitizer Web Plot) για την αυτόματη ανάκτηση στατιστικών δεδομένων εικόνας (Leong, 2017)
- Το Excel 2016 δεν είναι μόνο ένας ηλεκτρονικός πίνακας για μεγάλα δεδομένα και στατιστική ανάλυση, αλλά και ένα ισχυρό εργαλείο για οπτικοποίηση. Χρησιμοποιώντας τη δυνατότητα υποβολής ερωτημάτων, το Excel ενσωματώνει HDFS, SaaS και άλλες υπηρεσίες και διαχειρίζεται ημιδομημένα δεδομένα. Σε συνδυασμό με μεθόδους οπτικοποίησης, όπως η μορφοποίηση υπό όρους και τα διαδραστικά γραφικά, το Excel 2016 είναι ένας ισχυρός ανταγωνιστής μεταξύ των εργαλείων οπτικοποίησης μεγάλων δεδομένων (Microsoft, 2016)

ΚΕΦΑΛΑΙΟ 3 : BIG DATA ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ

Η επιχειρηματικότητα είναι σημαντικό μέρος της οικονομίας μας. Βοηθά να δημιουργήσουμε νέα προϊόντα και υπηρεσίες και να κερδίσουμε χρήματα. Είναι η πρακτική να χρησιμοποιούμε τις γνώσεις και τις δεξιότητές μας για την δημιουργία κάτι καινούργιου με οικονομικό όφελος από αυτό. Ο Alfred Marshall, ο πρώτος ακαδημαϊκός που σπούδασε οικονομικά, αναγνώρισε αυτόν τον σημαντικό ρόλο της επιχειρηματικότητας. Ο Drucker, ο πατέρας του μάνατζμεντ, πίστευε επίσης στη σημασία της επιχειρηματικότητας. Η επιχειρηματικότητα είναι μια διαδικασία όπου οι άνθρωποι προσπαθούν να βρουν νέους τρόπους για να κερδίσουν χρήματα. Το

κάνουν αυτό δημιουργώντας νέες ιδέες και στη συνέχεια πουλώντας αυτές τις ιδέες σε άλλους ανθρώπους. Εάν κάποιος έχει μια καινοτόμο ιδέα, αυτή μπορεί να μετατραπεί σε επιχειρηματική ευκαιρία.
Και, σύμφωνα με τον ΟΟΣΑ, το να είσαι επιχειρηματίας σημαίνει να είσαι ο καταλύτης για την αλλαγή στην οικονομία της αγοράς. Επιχειρηματικότητα είναι όταν κάποιος αναλαμβάνει την ευθύνη για τη δική του επιτυχία, λαμβάνοντας σημαντικές αποφάσεις που επηρεάζουν την επιχείρησή του. Μπορούν να είναι από την οικονομική πλευρά (που εμπλέκονται στη διανομή, τη χρήση και τη λήψη πόρων για την επιχείρησή τους), τη διοικητική πλευρά (φροντίζοντας όλα τα διαδικαστικά να είναι εντάξει) ή και τα δύο. Οι επιχειρηματίες μπορούν να χωριστούν σε δύο ομάδες: αυτούς που αναλαμβάνουν κινδύνους και αυτούς που αποφεύγουν τους κινδύνους. Καινοτομία είναι όταν κάποιος έρχεται με μια νέα ιδέα που μπορεί να χρησιμοποιηθεί για να βγάλει χρήματα. Η επιχειρηματικότητα είναι η διαδικασία λήψης αυτών των ιδεών και μετατροπής τους σε πραγματικότητα, καθώς και της ικανότητας να βλέπεις νέους τρόπους να κάνεις πράγματα και να φέρεις αυτές τις ιδέες στην αγορά. Χρειάζεται πολλή δημιουργικότητα και καινοτομία για να είσαι επιχειρηματίας (Khan, 2020).

Συμβολή των Big Data στην επιχειρηματικότητα

Στο παρελθόν, οι επιχειρήσεις χρησιμοποιούσαν σχεσιακές βάσεις δεδομένων για την αποθήκευση και την επεξεργασία πληροφοριών. Ωστόσο, λόγω των ευκαιριών που παρέχουν το Διαδίκτυο και τα μεγάλα δεδομένα στον 21ο αιώνα, οι επιχειρήσεις χρησιμοποιούν πλέον εφαρμογές μεγάλων δεδομένων για να βελτιώσουν την αποτελεσματικότητα και την ανταγωνιστικότητά τους. Για παράδειγμα, αναλύοντας σύνολα μεγάλων δεδομένων, οι επιχειρήσεις μπορούν να προβλέψουν με μεγαλύτερη ακρίβεια τη συμπεριφορά των καταναλωτών. Μπορούν επίσης να βελτιώσουν τα σχέδια πωλήσεών τους συγκρίνοντας μαζικά δεδομένα. Επιπλέον, βελτιστοποιώντας τα βασικά προϊόντα όσον αφορά την τιμή, το κόστος και την ποιότητα, οι επιχειρήσεις μπορούν να βελτιώσουν την αποτελεσματικότητα και την ικανοποίησή τους. Τέλος, προβλέποντας με ακρίβεια τις απαιτήσεις, οι επιχειρήσεις μπορούν να εξοικονομήσουν χρόνο και χρήμα. Τα μεγάλα δεδομένα είναι ένας τρόπος συλλογής δεδομένων που είναι πολύ μεγάλα για να τα χειρίζονται κανονικά συστήματα υπολογιστών. Μπορεί να χρησιμοποιηθεί για τη βελτίωση των επιχειρηματικών λειτουργιών, βοηθώντας στη διαχείριση του αποθέματος, στο συντονισμό με τους προμηθευτές και στη λήψη καλύτερων αποφάσεων σχετικά με τα οικονομικά.

Big Data και καινοτομία

Πρόσφατη έρευνα έδειξε ότι τα Big Data θα μπορούσαν να φέρουν διαφορετικά είδη καινοτομιών, όπως βελτιώσεις διαδικασιών και προϊόντων, από μια αρχιτεκτονική σε μια αρθρωτή βάση. Ένα παράδειγμα είναι η χρήση δεδομένων για την αύξηση του επιπέδου εξατομίκευσης στις υπηρεσίες (Tempini, 2017). Ωστόσο, οι ερευνητές αρχίζουν επίσης να εξετάζουν πώς οι οργανωτικοί παράγοντες μπορούν να βοηθήσουν στη διευκόλυνση αυτού του τύπου καινοτομίας, όπως η πελατοκεντρικότητα, η προσανατολισμένη στα δεδομένα διαχείριση και η εφαρμογή ευέλικτων πρακτικών. Από τη μία πλευρά, τα Big Data μπορούν να βοηθήσουν την εταιρεία να επεκτείνει τις δυνατότητές της με έναν νέο τρόπο, πέρα από τα όρια αυτού που μπορεί να κάνει η εταιρεία αυτή τη στιγμή. Αυτό μπορεί να οδηγήσει σε νέες επιχειρηματικές ευκαιρίες ή ακόμα και σε ένα επιχειρηματικό μοντέλο

καινοτομίας που επεκτείνει τον τομέα λειτουργίας της εταιρείας (Luigi Mario De Luca, 2020). Υπάρχει ακόμη πολλή έρευνα που πρέπει να γίνει για το πώς τα μεγάλα δεδομένα μπορούν να ενεργοποιήσουν τους μηχανισμούς που δημιουργούν αξία. Μέχρι στιγμής, όλες οι ερευνητικές μελέτες έχουν επικεντρωθεί στο πώς να λειτουργήσουν τα μεγάλα δεδομένα για τις εταιρείες. Ο νέος μηχανισμός πίστωσης αξίας χρησιμοποιεί διαφορετικές στρατηγικές για την αποτύπωση αξίας από μεγάλα δεδομένα. Για παράδειγμα, μια στρατηγική είναι η βελτίωση των υφιστάμενων μηχανισμών όπως η διαφήμιση. Ένα άλλο είναι να κατανοήσετε τους τρέχοντες πελάτες με έναν βαθύτερο τρόπο ή να βρείτε ένα νέο είδος πελάτη που μπορεί να δει μια ευρύτερη αξία σε αυτά τα δεδομένα. Αυτό το μοντέλο βασίζεται στη θεωρία της αμφίδρομης αγοράς. Αυτό σημαίνει ότι υπάρχουν δύο (ή περισσότερα) διαφορετικές ομάδες πελατών που συνδέονται μεταξύ τους μέσω μιας πλατφόρμας. Πρόσφατα, οι ερευνητές άρχισαν να βλέπουν τους διαφορετικούς τρόπους με τους οποίους τα Big Data μπορούν να βοηθήσουν στην προώθηση της καινοτομίας.Μπορεί να γίνει περισσότερη έρευνα σχετικά με το πώς τα Big Data μπορούν να συμβάλουν στην ενθάρρυνση της καινοτομίας, καθώς η τεχνολογία έχει διευκολύνει τη συλλογή συγκεκριμένων δεδομένων γρήγορα και με χαμηλότερο κόστος. Επιπλέον, τα Big Data μπορούν να βοηθήσουν τις εταιρείες να εκμεταλλευτούν την αξία τους (καινοτομία από δεδομένα), καθώς και να ενεργοποιήσουν έναν συγκεκριμένο τύπο καινοτομίας προϊόντος που θα μπορεί να συλλέγει και να αναλύει δεδομένα (καινοτομία ως δεδομένα) (AmirGandomi, 2015).

3.1 Η επιχειρηματικότητα στην 4η Βιομηχανική Επανάσταση

Στην τέταρτη βιομηχανική επανάσταση (4IR), η STARA (smart technology, artificial intelligence, robotics, and algorithms) προβλέπεται να αντικαταστήσει το ένα τρίτο των θέσεων εργασίας που υπάρχουν σήμερα. Σχεδόν διπλάσιες από τις σημερινές εργασίες θα διεκπεραιώνονται από ρομπότ. Προβλέπεται ότι μέχρι το 2025, 85 εκατομμύρια θέσεις εργασίας μπορεί να εκτοπιστούν από την αλλαγή στον καταμερισμό εργασίας μεταξύ ανθρώπων και μηχανών, ενώ 97 εκατομμύρια νέοι ρόλοι μπορεί να προκύψουν που θα είναι πιο προσαρμοσμένοι στον νέο καταμερισμό εργασίας μεταξύ ανθρώπων, μηχανών και αλγορίθμων. Οι βιομηχανικοί ψυχολόγοι διαδραματίζουν ολοένα και πιο σημαντικό ρόλο στο χώρο εργασίας λόγω αυτών των τάσεων από την άποψη της στρατηγικής νοημοσύνης. Στόχος του παρόντος άρθρου είναι να παρουσιάσει μια κριτική επισκόπηση των βιομηχανικών ψυχολόγων στους μελλοντικούς χώρους εργασίας στο πλαίσιο της 4ης βιομηχανικής επανάστασης. Ένα μοντέλο ικανοτήτων τίθεται για τους βιομηχανικούς ψυχολόγους προκειμένου να επιτελέσουν ρόλο στρατηγικής νοημοσύνης στους οργανισμούς (Oosthuizen, 2022). Η επιτάχυνση της καινοτομίας και η ταχύτητα της αναστάτωσης είναι δύσκολο να κατανοηθούν ή να προβλεφθούν και ότι αυτοί οι παράγοντες αποτελούν πηγή συνεχών εκπλήξεων,και σε όλους τους κλάδους υπάρχουν σαφείς ενδείξεις ότι οι τεχνολογίες που στηρίζουν την τέταρτη βιομηχανική επανάσταση έχουν σημαντικό αντίκτυπο στις επιχειρήσεις. Στο κομμάτι της τροφοδοσίας παρατηρείται η εισαγωγή νέων τεχνολογιών που δημιουργούν εντελώς νέους τρόπους εξυπηρέτησης των υφιστάμενων αναγκών και διαταράσσουν σημαντικά τις υφιστάμενες αλυσίδες αξίας του κλάδου. Η διαταραχή προέρχεται επίσης από ευέλικτους, καινοτόμους ανταγωνιστές οι οποίοι, χάρη στην πρόσβαση σε παγκόσμιες ψηφιακές πλατφόρμες για έρευνα, ανάπτυξη, μάρκετινγκ, πωλήσεις και διανομή, μπορούν να εκτοπίσουν τους καθιερωμένους κατεστημένους ανταγωνιστές ταχύτερα από ποτέ, βελτιώνοντας την ποιότητα, την ταχύτητα ή την τιμή στην οποία παρέχεται η αξία. Αλλαγές

συμβαίνουν και στην πλευρά της ζήτησης, καθώς η αυξανόμενη διαφάνεια, η δέσμευση των καταναλωτών και τα νέα πρότυπα συμπεριφοράς των καταναλωτών (που βασίζονται όλο και περισσότερο στην πρόσβαση σε δίκτυα κινητής τηλεφωνίας και δεδομένα) αναγκάζουν τις εταιρείες να προσαρμόσουν τον τρόπο με τον οποίο σχεδιάζουν, προωθούν και παρέχουν προϊόντα και υπηρεσίες. Μια βασική τάση είναι η ανάπτυξη τεχνολογικά υποστηριζόμενων πλατφορμών που συνδυάζουν τόσο τη ζήτηση όσο και την προσφορά για να διαταράξουν τις υπάρχουσες δομές του κλάδου, όπως αυτές που βλέπουμε στο πλαίσιο της οικονομίας του "διαμοιρασμού" ή της οικονομίας "κατά παραγγελία". Αυτές οι τεχνολογικές πλατφόρμες, οι οποίες καθίστανται εύχρηστες με τη χρήση smartphone, συγκεντρώνουν ανθρώπους, περιουσιακά στοιχεία και δεδομένα, δημιουργώντας έτσι εντελώς νέους τρόπους κατανάλωσης αγαθών και υπηρεσιών. Επιπλέον, μειώνουν τα εμπόδια για τις επιχειρήσεις και τα άτομα να δημιουργήσουν πλούτο, αλλάζοντας το προσωπικό και επαγγελματικό περιβάλλον των εργαζομένων. Συνολικά, υπάρχουν τέσσερις κύριες επιπτώσεις που έχει η τέταρτη βιομηχανική επανάσταση στις επιχειρήσεις:

- στις προσδοκίες των πελατών
- στη βελτίωση των προϊόντων
- στη συνεργατική καινοτομία
- στις οργανωτικές μορφές

Είτε πρόκειται για καταναλωτές, είτε για επιχειρήσεις, οι πελάτες βρίσκονται ολοένα και περισσότερο στο επίκεντρο της οικονομίας, η οποία αφορά τη βελτίωση του τρόπου εξυπηρέτησης των πελατών. Τα φυσικά προϊόντα και οι υπηρεσίες, εξάλλου, μπορούν πλέον να ενισχυθούν με ψηφιακές δυνατότητες που αυξάνουν την αξία τους. Οι νέες τεχνολογίες καθιστούν τα περιουσιακά στοιχεία πιο ανθεκτικά και ανθεκτικά, ενώ τα δεδομένα και οι αναλύσεις μετασχηματίζουν τον τρόπο συντήρησής τους. Εν τω μεταξύ, ένας κόσμος εμπειριών πελατών, υπηρεσιών βασισμένων σε δεδομένα και απόδοσης περιουσιακών στοιχείων μέσω αναλύσεων. Η εμφάνιση παγκόσμιων πλατφορμών και άλλων νέων επιχειρηματικών μοντέλων σημαίνει ότι το ταλέντο, η κουλτούρα και οι οργανωτικές μορφές θα πρέπει να επανεξεταστούν. Συνολικά, η αδυσώπητη μετάβαση από την απλή ψηφιοποίηση (η Τρίτη Βιομηχανική Επανάσταση) στην καινοτομία που βασίζεται σε συνδυασμούς τεχνολογιών (η Τέταρτη Βιομηχανική Επανάσταση) αναγκάζει τις εταιρείες να επανεξετάσουν τον τρόπο με τον οποίο δραστηριοποιούνται. Η ουσία, ωστόσο, είναι η ίδια: οι ηγέτες των επιχειρήσεων και τα ανώτερα στελέχη πρέπει να κατανοήσουν το μεταβαλλόμενο περιβάλλον τους, να αμφισβητήσουν τις παραδοχές των επιχειρησιακών ομάδων τους και να καινοτομούν αδιάκοπα και συνεχώς (Schwab, 2016).

3.2 BIG DATA σε επιμέρους κλάδους

Logistics

Μεγάλες εταιρείες στον τομέα των logistics χρησιμοποιούν λογισμικό για τη συλλογή δεδομένων σχετικά με τις δραστηριότητές τους. Αυτό τους βοηθά να μάθουν πώς να κάνουν τις διαδικασίες τους πιο αποτελεσματικές. Σήμερα, οι πάροχοι υπηρεσιών logistics πρέπει να διαχειρίζονται πολλά προϊόντα και δεδομένα. Ορισμένα από αυτά τα δεδομένα προέρχονται από την αποστολή εκατομμυρίων προϊόντων καθημερινά. Αυτά τα δεδομένα περιλαμβάνουν στοιχεία όπως η ημερομηνία, η ώρα και ο τόπος κάθε αποστολής, το βάρος και οι διαστάσεις κάθε προϊόντος και πληροφορίες οδηγού

και οχήματος για κάθε αποστολή. Οι εταιρείες χρησιμοποιούν τεχνολογίες μεγάλων δεδομένων και ανάλυσης για να βελτιώσουν την επιχειρηματική τους αποτελεσματικότητα και την εμπειρία των πελατών τους. Οι αλυσίδες εφοδιασμού θα είχαν πολλά οφέλη εάν είχαν πολλούς παρόχους logistics που θα μπορούσαν να μοιράζονται συνεχώς πληροφορίες. Αυτό θα διευκόλυνε τη δημιουργία νέων συνεργασιών και νέων υπηρεσιών που μπορούν να βοηθήσουν τις επιχειρήσεις να προβλέψουν τη ζήτηση της αγοράς.

Λιανικό Εμπόριο

Το λιανικό εμπόριο είναι ένας τομέας της αγοράς όπου οι επιχειρήσεις μπορούν να χρησιμοποιήσουν μεγάλα δεδομένα και επιχειρηματικές αναλύσεις για να βελτιώσουν το εισόδημά τους. Αυτό μπορεί να περιλαμβάνει πράγματα όπως η παροχή καλύτερης εξυπηρέτησης πελατών, η στόχευση διαφημίσεων και η παρακολούθηση παραγγελιών. Για να είναι μια επιχείρηση κερδοφόρα, θα πρέπει να συγχωνεύσει πολλά διαφορετικά μέρη της εφοδιαστικής αλυσίδας, τους πελάτες της, καθώς και τις υπηρεσίες και τη διαφήμισή της. Οι μεγάλες επιχειρήσεις συλλέγουν πολλά δεδομένα από διαφορετικά μέρη των δραστηριοτήτων τους. Αυτό μπορεί να περιλαμβάνει πράγματα όπως ο αριθμός των συναλλαγών που πραγματοποιούν οι πελάτες, το ποσό του αποθέματος που διαθέτουν, η τοποθεσία και οι κινήσεις των προϊόντων, ακόμη και τα οικονομικά δεδομένα της εταιρείας. Το ενοποιημένο πληροφοριακό σύστημα ERP συλλέγει δεδομένα από διαφορετικές πηγές για να βοηθήσει τις επιχειρήσεις να λειτουργούν πιο αποτελεσματικά. Βοηθά τις εταιρείες να παρακολουθούν το απόθεμα, τους προμηθευτές και τους πελάτες τους και μπορούν ακόμη και να μάθουν από λάθη του παρελθόντος. Π.χ. το κέντρο δεδομένων Wal-Mart χρησιμοποιεί μηχανική εκμάθηση για τη συλλογή δεδομένων από διαφορετικά μέρη της εταιρείας ταυτόχρονα. Με αυτόν τον τρόπο, η Wal-Mart μπορεί να διαχειρίζεται τα αποθέματα της και την αλυσίδα εφοδιασμού πιο αποτελεσματικά (C. L. Philip ChenChun, 2014).

Big Data στο Ηλεκτρονικό Εμπόριο

Το Big Data Business Analytics είναι ένας τρόπος χρήσης δεδομένων για τη βελτίωση των επιχειρηματικών λειτουργιών. Το ηλεκτρονικό εμπόριο (eCommerce) είναι ένας κλάδος που είναι ιδιαίτερα καλός στη χρήση μεγάλων δεδομένων, επειδή περιλαμβάνει τη συλλογή δεδομένων από πολλές διαφορετικές πηγές (όπως κλικ σε έναν ιστότοπο, likes στα μέσα κοινωνικής δικτύωσης κ.λπ.). Αυτό καθιστά τα μεγάλα δεδομένα πολύτιμο πλεονέκτημα για τις επιχειρήσεις ηλεκτρονικού εμπορίου, οι οποίες μπορούν να τα χρησιμοποιήσουν για να βελτιώσουν τις προσπάθειες μάρκετινγκ και πωλήσεων, καθώς και τα κέρδη τους. Υπάρχουν διάφοροι τρόποι που τα big data βοηθούν στον τουρισμό:

- Μια από τις πιο αποτελεσματικές χρήσεις των μεγάλων δεδομένων στον ταξιδιωτικό κλάδο συνδέεται με τη διαχείριση των εσόδων. Για να μεγιστοποιήσουν τα οικονομικά αποτελέσματα, τα ξενοδοχεία και άλλες τουριστικές επιχειρήσεις πρέπει να είναι σε θέση να πωλούν το σωστό προϊόν, στον σωστό πελάτη, τη σωστή στιγμή, στη σωστή τιμή, μέσω του σωστού καναλιού, και τα μεγάλα δεδομένα μπορούν να είναι ανεκτίμητα για αυτό. Ειδικότερα, τα εσωτερικά δεδομένα, όπως τα προηγούμενα ποσοστά πληρότητας, τα έσοδα δωματίων και οι τρέχουσες κρατήσεις, μπορούν να συνδυαστούν με εξωτερικά δεδομένα, όπως πληροφορίες σχετικά με τοπικές

εκδηλώσεις, πτήσεις και σχολικές διακοπές, προκειμένου να προβλεφθεί με μεγαλύτερη ακρίβεια και να προβλεφθεί η ζήτηση. Ως αποτέλεσμα αυτού, τα ξενοδοχεία είναι στη συνέχεια σε θέση να διαχειρίζονται καλύτερα τις τιμές τους, αυξάνοντάς τες σε περιόδους υψηλής ζήτησης, προκειμένου να μεγιστοποιήσουν τα έσοδα που παράγονται

- Μια άλλη σημαντική χρήση των μεγάλων δεδομένων στην τουριστική βιομηχανία είναι η διαχείριση της φήμης. Στην εποχή του διαδικτύου, οι πελάτες μπορούν να αφήνουν κριτικές σε ένα ευρύ φάσμα διαφορετικών πλατφορμών, συμπεριλαμβανομένων των ιστότοπων κοινωνικής δικτύωσης, των μηχανών αναζήτησης και των ειδικών ιστότοπων κριτικών, μοιράζοντας τις απόψεις και τις εμπειρίες τους. Επιπλέον, οι πελάτες ελέγχουν όλο και περισσότερο αυτές τις κριτικές και συγκρίνουν διάφορα ξενοδοχεία πριν κάνουν κράτηση. Τα δεδομένα αυτά, σε συνδυασμό με τα σχόλια που αποκτώνται εσωτερικά, μπορούν να χρησιμοποιηθούν για τον εντοπισμό των σημαντικότερων δυνατών και αδύνατων σημείων και των σημείων όπου οι πελάτες εντυπωσιάζονται ή απογοητεύονται. Μόλις συγκεντρωθούν αυτές οι πληροφορίες, τα ξενοδοχεία μπορούν να τις χρησιμοποιήσουν για να ενημερώσουν τις εκπαιδευτικές τους προσπάθειες, προκειμένου να προβούν σε βελτιώσεις και να διασφαλίσουν ότι οι μελλοντικές κριτικές θα είναι θετικές
- Στον ταξιδιωτικό κλάδο, το μάρκετινγκ μπορεί να είναι δύσκολο να γίνει σωστά, επειδή οι δυνητικοί πελάτες είναι τόσο διαφορετικοί ως προς το ποιοι είναι, από πού προέρχονται και τι αναζητούν. Ωστόσο, τα μεγάλα δεδομένα μπορούν να βοηθήσουν τις τουριστικές εταιρείες να υιοθετήσουν μια πιο στρατηγική προσέγγιση στις προσπάθειες μάρκετινγκ, στοχεύοντας στους σωστούς ανθρώπους με τον σωστό τρόπο. Πιο συγκεκριμένα, τα μεγάλα δεδομένα μπορούν να βοηθήσουν τις επιχειρήσεις να εντοπίσουν τις κύριες τάσεις που υπάρχουν μεταξύ των πελατών τους, πού υπάρχουν οι ομοιότητες και ποιες είναι οι καλύτερες ευκαιρίες μάρκετινγκ. Μπορούν επίσης να βοηθήσουν τις επιχειρήσεις να κατανοήσουν πού βρίσκονται αυτοί οι άνθρωποι και πότε το μάρκετινγκ είναι πιο σχετικό με αυτούς. Αυτό μπορεί να επιτρέψει την αποστολή μηνυμάτων μάρκετινγκ, με βάση τον χρόνο, την τοποθεσία και άλλα δεδομένα, επιτρέποντας την παροχή πιο στοχευμένου διαφημιστικού περιεχομένου
- Τα ξενοδοχεία και οι άλλες επιχειρήσεις του ταξιδιωτικού και τουριστικού κλάδου έχουν ένα ευρύ φάσμα αλληλεπιδράσεων με τους πελάτες και κάθε μία από αυτές τις αλληλεπιδράσεις μπορεί να παρέχει πολύτιμα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν για τη βελτίωση της συνολικής εμπειρίας του πελάτη. Τα δεδομένα αυτά μπορεί να περιλαμβάνουν τα πάντα, από συζητήσεις στα μέσα κοινωνικής δικτύωσης και διαδικτυακές κριτικές μέχρι δεδομένα χρήσης υπηρεσιών. Με αποτελεσματική χρήση, οι πληροφορίες αυτές μπορούν να αποκαλύψουν ποιες υπηρεσίες χρησιμοποιούν οι πελάτες περισσότερο, ποιες δεν χρησιμοποιούν καθόλου και ποιες είναι πιο πιθανό να ζητήσουν ή να μιλήσουν γι' αυτές. Μέσω αυτών των δεδομένων, οι εταιρείες μπορούν να λαμβάνουν πιο τεκμηριωμένες, βασισμένες στα δεδομένα αποφάσεις σχετικά με τις υπηρεσίες που παρέχουν σήμερα, τις υπηρεσίες που δεν χρειάζεται πλέον να παρέχουν, τις υπηρεσίες που θέλουν να εισαγάγουν και τη νέα τεχνολογία στην οποία επιλέγουν να επενδύσουν
- Τέλος, όσοι δραστηριοποιούνται στον κλάδο των ταξιδιών και του τουρισμού μπορούν επίσης να χρησιμοποιήσουν τα μεγάλα δεδομένα για να

συγκεντρώσουν και να αναλύσουν πληροφορίες σχετικά με τους κύριους ανταγωνιστές τους, προκειμένου να κατανοήσουν καλύτερα τι προσφέρουν στους πελάτες τους άλλα ξενοδοχεία ή επιχειρήσεις. Και πάλι, τα δεδομένα αυτά μπορούν να αποκτηθούν από διάφορες πηγές, καθώς δεν λείπουν τα μέρη όπου οι πελάτες πηγαίνουν για να μοιραστούν τις απόψεις τους για τα ξενοδοχεία και τις ταξιδιωτικές εταιρείες, ιδίως στο διαδίκτυο (Revfine, n.d.)

Big Data στην Φαρμακοβιομηχανία

Με την πάροδο των ετών, η ζήτηση για δεδομένα έχει αυξηθεί εκθετικά και η ταχεία ενσωμάτωση θεωρείται πλέον επιχειρηματική απαίτηση. Αυτό ισχύει ιδιαίτερα για τις φαρμακευτικές εταιρείες, καθώς ανέκαθεν βασίζονταν σε εμπειρικά δεδομένα για τον εντοπισμό μοτίβων, τον έλεγχο θεωριών και την κατανόηση της αποτελεσματικότητας των θεραπειών. Με τις ελπίδες του κόσμου να εναποτίθενται στη φαρμακοβιομηχανία περισσότερο από ποτέ άλλοτε στην εποχή της πανδημίας COVID, η ανάλυση μεγάλων δεδομένων έπαιξε καθοριστικό ρόλο στην ανάπτυξη φαρμάκων και εμβολίων. Παραδοσιακά, οι ερευνητές εφάρμοζαν μια επαναληπτική διαδικασία φυσικής δοκιμής διαφόρων ενώσεων για την ανακάλυψη νέων φαρμάκων. Αυτό απαιτεί τεράστιο χρόνο και πόρους, ενώ το κόστος ανάπτυξης αυτών των φαρμάκων μπορεί επίσης να γίνει όλο και πιο ακριβό. Όμως, με τη βοήθεια της ανάλυσης δεδομένων, οι ερευνητές μπορούν να αξιοποιήσουν την προγνωστική μοντελοποίηση για την ανακάλυψη φαρμάκων. Η προγνωστική μοντελοποίηση επιτρέπει στους ερευνητές να προβλέπουν τις αλληλεπιδράσεις, την τοξικότητα και την αναστολή των φαρμάκων και έτσι επιταχύνει την όλη διαδικασία. Έτσι, η ανάλυση μεγάλων δεδομένων στη φαρμακευτική βιομηχανία βοηθά στην ανακάλυψη φαρμάκων. Οι κλινικές δοκιμές είναι ζωτικής σημασίας στον κόσμο των φαρμακευτικών και βιοεπιστημών, καθώς χρησιμοποιούνται για να ελεγχθεί κατά πόσον μια συγκεκριμένη θεραπεία είναι αποτελεσματική και ασφαλής για τα ανθρώπινα υποκείμενα. Επιπλέον, οι κλινικές δοκιμές είναι δαπανηρές, χρονοβόρες και αρκετές κλινικές δοκιμές αποτυγχάνουν, καθώς η πρόσληψη του κατάλληλου ασθενούς για τη δοκιμή είναι αρκετά δύσκολη. Με τη βοήθεια των μεγάλων δεδομένων, οι φαρμακευτικές εταιρείες μπορούν να στρατολογήσουν τους κατάλληλους ασθενείς για κλινικές δοκιμές, χρησιμοποιώντας δεδομένα όπως γενετικές πληροφορίες, χαρακτηριστικά προσωπικότητας και την κατάσταση της νόσου, τα οποία με τη σειρά τους θα αυξήσουν το ποσοστό επιτυχίας του φαρμάκου. Αυτό επιτρέπει επίσης την ιατρική ακριβείας, όπου η διάγνωση και η θεραπεία των διαταραχών πραγματοποιούνται με τη χρήση σχετικών δεδομένων σχετικά με τη γενετική σύνθεση ενός ασθενούς, τα πρότυπα συμπεριφοράς κ.ά. Με αυτή την προσέγγιση, οι φαρμακευτικές εταιρείες μπορούν να αναπτύξουν εξατομικευμένα φάρμακα που είναι κατάλληλα για τα γονίδια και τον τρέχοντα τρόπο ζωής ενός μεμονωμένου ασθενούς. Μεγάλη ανάλυση δεδομένων στην έρευνα και την ανάπτυξη Με τις γνώσεις που αποκτώνται από ιστορικές και σε πραγματικό χρόνο πηγές δεδομένων, όπως τα μέσα κοινωνικής δικτύωσης, οι αισθητήρες IoT, τα αρχεία καταγραφής και τα δεδομένα ασθενών, οι φαρμακευτικές εταιρείες μπορούν να δημιουργήσουν κατατοπιστικές αναλύσεις. Αυτό είναι ένα από τα μεγαλύτερα οφέλη της ανάλυσης μεγάλων δεδομένων στη φαρμακευτική βιομηχανία. Με τη συλλογή μεγάλου όγκου δεδομένων που παράγονται στα διάφορα στάδια της αλυσίδας αξίας, από την ανακάλυψη φαρμάκων έως την πραγματική χρήση, μπορούν να χρησιμοποιήσουν την ανάλυση μεγάλων δεδομένων και να αποκτήσουν χρήσιμες πληροφορίες που είναι επωφελείς για την έρευνα και την ανάπτυξη. Η εφαρμογή της

ανάλυσης μεγάλων δεδομένων στη φαρμακευτική βιομηχανία δεν σταματά μόνο στα φάρμακα και τις δοκιμές. Με τον αυξανόμενο ανταγωνισμό στον κόσμο των βιοεπιστημών, η μεγάλη φαρμακευτική βιομηχανία γίνεται όλο και πιο έξυπνη όσον αφορά την ανάλυση και την προώθηση της αποτελεσματικότητας των λειτουργιών πωλήσεων και μάρκετινγκ. Αναλύοντας τις πληροφορίες από τα μέσα κοινωνικής δικτύωσης, τα δημογραφικά στοιχεία, τα ηλεκτρονικά ιατρικά αρχεία και άλλες πηγές δεδομένων, οι φαρμακευτικές εταιρείες μπορούν να εντοπίσουν και να αξιοποιήσουν υποεξυπηρετούμενες και νέες αγορές. Επιπλέον, μπορούν επίσης να αναλύουν την αποτελεσματικότητα των προσπαθειών πωλήσεων και να λαμβάνουν σημαντικές αποφάσεις στις στρατηγικές μάρκετινγκ και πωλήσεων.
Η έλευση της ανάλυσης μεγάλων δεδομένων στη φαρμακευτική βιομηχανία υπήρξε μόνο επαναστατική. Δεν υπάρχει καμία αμφιβολία ότι οι φαρμακευτικές εταιρείες μπορούν να κάνουν τις θεραπείες πιο αποτελεσματικές με τη βοήθεια των μεγάλων δεδομένων. Μόνο οι φαρμακευτικές εταιρείες που θα συνεχίσουν τις επενδύσεις τους στην ανάλυση μεγάλων δεδομένων θα προσπαθήσουν να προχωρήσουν με καινοτόμες εφαρμογές. Έτσι, η ανάπτυξη ενός στρατηγικού σχεδίου για την ενσωμάτωση της ανάλυσης μεγάλων δεδομένων με την υποδομή και τα συστήματα της επιχείρησης είναι ζωτικής σημασίας (Intone, 2022).

Big Data στον Τραπεζικό Κλάδο

Ο τραπεζικός τομέας είναι ο κινητήριος μοχλός που τροφοδοτεί τις οικονομίες, τα έθνη και τους οργανισμούς. Παράγει επίσης τεράστιες ποσότητες δεδομένων κάθε δευτερόλεπτο. Κάθε συναλλαγή αφήνει ένα ίχνος και παράγει δεδομένα που προηγουμένως θεωρούνταν στατικά και χρήσιμα μόνο στους ελεγκτές για τους σκοπούς της λογιστικής και του ελέγχου. Ωστόσο, καθώς οι τεχνολογίες μεγάλων δεδομένων σε άλλους τομείς, όπως η υγειονομική περίθαλψη, άρχισαν να δείχνουν τις πραγματικές τους δυνατότητες, αρχίσαμε να ενσωματώνουμε παλαιωμένα δεδομένα σε αυτά τα συστήματα και αρχίσαμε να βλέπουμε πραγματικά τις δυνατότητες οικονομικών πληροφοριών που θα μπορούσαν να χρησιμοποιηθούν για διάφορους σκοπούς. Κατά συνέπεια, τα μεγάλα δεδομένα στον τραπεζικό τομέα έχουν ανεκμετάλλευτες δυνατότητες. Πίσω στο 2008, οι τεχνολογίες Big Data και Business Intelligence βοήθησαν σε αυτή την προσπάθεια και επέτρεψαν στις τράπεζες και τα χρηματοπιστωτικά ιδρύματα να αμφισβητήσουν το status quo, δίνοντας το έναυσμα για την εμφάνιση των Big Data στον τραπεζικό τομέα. Οι τράπεζες χρησιμοποιούν τεχνολογίες Big Data και BI, όπως το Hadoop και τα RDBMS, σε όλες τις διαδικασίες τους, αλλάζοντας το πρόσωπο του τραπεζικού τομέα προς το καλύτερο. Τα Μεγάλα Δεδομένα έχουν συμβάλει στη διαμόρφωση οργανισμών και ιδρυμάτων σε όλο τον κόσμο, από την ψηφιοποίηση όλων των τραπεζικών διαδικασιών έως τη μετατροπή των αναπτυσσόμενων οικονομιών από τις συναλλαγές με μετρητά σε ψηφιακές συναλλαγές.

- Οι πελάτες λαμβάνουν εξατομικευμένες τραπεζικές λύσεις: Τα μεγάλα δεδομένα, όταν συνδυάζονται με αποτελεσματικά εργαλεία και τεχνολογίες, μπορούν να παρέχουν στις τράπεζες καλύτερη κατανόηση των μεμονωμένων πελατών με βάση τις εισροές που λαμβάνουν. Αυτό περιλαμβάνει τις επενδυτικές τους συνήθειες, τις αγοραστικές τους συνήθειες, τα επενδυτικά τους κίνητρα και το προσωπικό ή οικονομικό τους υπόβαθρο. Για παράδειγμα, μπορούν να προβλέψουν και να αποτρέψουν την απομάκρυνση, έχοντας ένα πλήρες προφίλ και δεδομένα πελατών. Βρίσκουν τον καλύτερο τρόπο για την

επίλυση τυχόν υφιστάμενων προβλημάτων. Τα μεγάλα δεδομένα χρησιμοποιούνται από τον τραπεζικό κλάδο για να γνωρίσουν τους πελάτες τους. Ως αποτέλεσμα, δημιουργούν προϊόντα, υπηρεσίες και άλλες προσφορές με βάση τα υπάρχοντα προφίλ πελατών που είναι προσαρμοσμένα στις συγκεκριμένες ανάγκες τους

- Τμηματοποίηση των πελατών: Η τμηματοποίηση των πελατών επιτρέπει στις τράπεζες να στοχεύουν καλύτερα τους πελάτες τους με τις καταλληλότερες εκστρατείες μάρκετινγκ. Οι εκστρατείες αυτές προσαρμόζονται στη συνέχεια ώστε να ανταποκρίνονται στις ανάγκες τους με πιο ουσιαστικό τρόπο. Οι τράπεζες θα αποκτήσουν πολύτιμες γνώσεις σχετικά με τη συμπεριφορά των χρηστών συνδυάζοντας τη μηχανική μάθηση και την τεχνητή νοημοσύνη με τα μεγάλα δεδομένα. Αυτό τους επιτρέπει επίσης να βελτιστοποιήσουν ανάλογα την εμπειρία των πελατών τους. Επιπλέον, οι τράπεζες θα είναι σε θέση να κατηγοριοποιούν τους πελάτες τους με βάση διάφορες παραμέτρους, όπως οι προτιμώμενες δαπάνες πιστωτικών καρτών ή ακόμη και η καθαρή περιουσία, έχοντας τη δυνατότητα να παρακολουθούν και να εντοπίζουν κάθε συναλλαγή του πελάτη
- Αποτελεσματική ανάλυση της ανατροφοδότησης πελατών: Μέσω της ανατροφοδότησης, τα εργαλεία Big Data μπορούν να παρέχουν στις τράπεζες ερωτήσεις, σχόλια και ανησυχίες των πελατών. Αυτή η ανατροφοδότηση τις βοηθά να ανταποκρίνονται εγκαίρως. Οι πελάτες θα παραμείνουν πιστοί σε μια εταιρεία εάν πιστεύουν ότι οι τράπεζές τους εκτιμούν τα σχόλιά τους και επικοινωνούν μαζί τους άμεσα
- Ανίχνευση και πρόληψη της απάτης: Μία από τις πιο δύσκολες προκλήσεις που αντιμετωπίζει σήμερα ο τραπεζικός κλάδος είναι η ανίχνευση της απάτης και η πρόληψη αμφισβητήσιμων συναλλαγών. Τα μεγάλα δεδομένα στον τραπεζικό τομέα τους δίνουν τη δυνατότητα να διασφαλίσουν ότι δεν πραγματοποιούνται ανεπίσημες συναλλαγές. Θα διασφαλίσει επίσης τη συνολική ασφάλεια και προστασία του τραπεζικού κλάδου. Επιπλέον, οι τράπεζες μπορούν να χρησιμοποιήσουν τα μεγάλα δεδομένα για να αποτρέψουν την απάτη και να κάνουν τους πελάτες να αισθάνονται πιο ασφαλείς, παρακολουθώντας τα μοτίβα δαπανών των πελατών και εντοπίζοντας ασυνήθιστη συμπεριφορά

(Mathur, 2022)

3.3 Μελέτες Περίπτωσης

Alumil

Η εταιρεία Alumil ιδρύθηκε το 1988 και δραστηριοποιείται στην παραγωγή προφίλ αλουμινίου. Τα επόμενα έτη η εταιρεία αναπτύσσεται, και συγκεκριμένα το 1990 ξεκινάει η λειτουργία της πρώτης γραμμής διέλασης αλουμινίου 1.600 μετρικών τόνων. Το 1993 ιδρύεται η θυγατρική της εταιρεία Alusys, με σκοπό την πώληση εξαρτημάτων στη Νότιο Ελλάδα και την τεχνική υποστήριξη των πελατών της. Μέχρι σήμερα έχει ιδρυθεί ειδικό τμήμα έρευνας και ανάπτυξης με σκοπό τη σχεδίαση νέων προϊόντων, τη δημιουργία τεχνικών καταλόγων, καθώς και τη μελέτη κατασκευών από αλουμίνιο.
Ένας από τους κύριους τρόπους που μια εταιρεία σαν την Alumil μπορεί να κάνει χρήση των big dataείναι για την βελτίωση των λειτουργιών της.

Τα προϊόντα της είναι :

- Πόρτες και παράθυρα
- Πόρτες εισόδου
- Συμπληρωματικά συστήματα κουφωμάτων
- Υαλοπετάσματα
- Συστήματα Σκίασης
- Κάγκελα και περιφράξεις
- Προσόψεις Κτηρίων
- Δομικά Υλικά
- Διαχωριστικά
- Αίθρια

Επίσης τα συστήματα της είναι πιστοποιημένα σε:

- Θερμομόνωση
- Αεροστεγάνωση/Αεροδιαπερατότητα
- Αντοχή σε Ανεμοπίεση
- Υδατοστεγανότητα
- Αντίσταση κατά της Διάρρηξης

Βάση όλων των άνωθι, καθώς και των διαδικασιών της σαν εταιρεία μπορεί να γίνει χρήση των big data. Συγκεκριμένα τα προϊόντα της απευθύνονται σε ιδιώτες, αρχιτέκτονες και κατασκευαστές. Έτσι μπορεί η εταιρεία να αποκτήσει βαθύτερη κατανόηση των περιοχών τυχόν αναποτελεσματικότητας, ώστε να βελτιωθεί η παραγωγή και το τελικό προϊόν. Ακολουθούν ορισμένα παραδείγματα χρήσης των big dataστην περίπτωση της εταιρείας Alumil:

Βελτιστοποίηση της αλυσίδας εφοδιασμού

Αξιοποιώντας τα μεγάλα δεδομένα, η Alumil μπορεί να βελτιώσει την αλυσίδα εφοδιασμού της και να αποκτήσει ανταγωνιστικό πλεονέκτημα στην αγορά.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της εφοδιαστικής αλυσίδας στην Alumil είναι η χρήση της παρακολούθησης σε πραγματικό χρόνο. Με τη χρήση αισθητήρων και άλλων τεχνολογιών, η Alumil μπορεί να παρακολουθεί την αλυσίδα εφοδιασμού της σε πραγματικό χρόνο και να εντοπίζει γρήγορα τυχόν προβλήματα ή σημεία συμφόρησης. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτίωση των διαδικασιών και την αύξηση της αποδοτικότητας. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένας συγκεκριμένος προμηθευτής παραδίδει συστηματικά με καθυστέρηση, η Alumil μπορεί να λάβει μέτρα για να βελτιώσει την απόδοση του προμηθευτή ή να βρει έναν νέο προμηθευτή.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της αλυσίδας εφοδιασμού στην Alumil είναι μέσω της χρήσης προγνωστικών αναλύσεων. Με την ανάλυση ιστορικών δεδομένων και τη χρήση αλγορίθμων μηχανικής μάθησης, η Alumil μπορεί να κάνει προβλέψεις σχετικά με τη μελλοντική απόδοση της αλυσίδας εφοδιασμού και να λάβει προληπτικά μέτρα για την πρόληψη των προβλημάτων πριν αυτά εμφανιστούν. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι μια συγκεκριμένη πρώτη ύλη είναι πιθανό να είναι σε έλλειψη

στο εγγύς μέλλον, η Alumil μπορεί να λάβει μέτρα για την εξασφάλιση εναλλακτικών πηγών ή την αύξηση των επιπέδων αποθεμάτων.
Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για την πρόβλεψη της ζήτησης. Αναλύοντας δεδομένα από διάφορες πηγές, όπως παραγγελίες πελατών και τάσεις της αγοράς, η Alumil μπορεί να κάνει ακριβείς προβλέψεις για τη μελλοντική ζήτηση των προϊόντων της. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των επιπέδων αποθεμάτων και τη μείωση του κινδύνου έλλειψης αποθεμάτων ή υπεραποθεμάτων.
Επιπλέον, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση των προμηθευτών. Με την ανάλυση δεδομένων από τους προμηθευτές, η Alumil μπορεί να εντοπίσει περιοχές για βελτίωση στην αλυσίδα εφοδιασμού της και να διασφαλίσει ότι οι προμηθευτές της πληρούν τα πρότυπα ποιότητας και παράδοσης. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένας συγκεκριμένος προμηθευτής παραδίδει σταθερά υποβαθμισμένα προϊόντα, η Alumil μπορεί να λάβει μέτρα για τη βελτίωση των διαδικασιών του προμηθευτή ή να βρει έναν νέο προμηθευτή. Αυτό μπορεί να συμβάλει στη διασφάλιση ότι τα προϊόντα της Alumil είναι υψηλής ποιότητας και ανταποκρίνονται στις προσδοκίες των πελατών.
Τέλος, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της εφοδιαστικής αλυσίδας. Με την ανάλυση δεδομένων από διάφορες πηγές, όπως τα χρονοδιαγράμματα αποστολής και οι χρόνοι παράδοσης, η Alumil μπορεί να βελτιστοποιήσει τις διαδικασίες και να μειώσει το κόστος μεταφοράς. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι μια συγκεκριμένη διαδρομή μεταφοράς προκαλεί συνεχώς καθυστερήσεις, η Alumil μπορεί να λάβει μέτρα για την εξεύρεση μιας ταχύτερης ή πιο αποδοτικής εναλλακτικής λύσης.

Ανάλυση της συμπεριφοράς των πελατών

Αξιοποιώντας τα μεγάλα δεδομένα, η Alumil μπορεί να αποκτήσει πολύτιμες γνώσεις σχετικά με τη συμπεριφορά και τις προτιμήσεις των πελατών της, οι οποίες μπορούν να βοηθήσουν την εταιρεία να λάβει τεκμηριωμένες επιχειρηματικές αποφάσεις.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση της συμπεριφοράς των πελατών στην Alumil είναι μέσω της χρήσης δεδομένων πελατών. Συλλέγοντας και αναλύοντας δεδομένα από διάφορες πηγές, όπως συναλλαγές πελατών, επισκέψεις στον ιστότοπο και ανατροφοδότηση πελατών, η Alumil μπορεί να αποκτήσει μια ολοκληρωμένη κατανόηση της συμπεριφοράς των πελατών της. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για την τμηματοποίηση της πελατειακής βάσης, την κατανόηση των αναγκών και των προτιμήσεών τους και τη βελτίωση της εμπειρίας των πελατών τους.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση της συμπεριφοράς των πελατών της Alumil είναι μέσω της χρήσης δεδομένων από τα μέσα κοινωνικής δικτύωσης. Με την ανάλυση δεδομένων από πλατφόρμες μέσων κοινωνικής δικτύωσης, όπως το Facebook, το Twitter και το Instagram, η Alumil μπορεί να κατανοήσει τι λένε οι πελάτες της για τα προϊόντα και τις υπηρεσίες της, καθώς και τις απόψεις και τις προτιμήσεις τους. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτίωση των στρατηγικών μάρκετινγκ και επικοινωνίας της και για την παροχή πιο εξατομικευμένων εμπειριών για τους πελάτες της.
Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για την ανάλυση του συναισθήματος των πελατών. Αναλύοντας τα σχόλια, τις κριτικές και τα σχόλια των

πελατών, η Alumil μπορεί να κατανοήσει πώς αισθάνονται οι πελάτες της για τα προϊόντα και τις υπηρεσίες της. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τον εντοπισμό τομέων προς βελτίωση, όπως τα χαρακτηριστικά των προϊόντων ή η εξυπηρέτηση των πελατών, και να γίνουν αλλαγές που θα βελτιώσουν την εμπειρία των πελατών.
Επιπλέον, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για εξατομίκευση. Με την ανάλυση των δεδομένων πελατών, η Alumil μπορεί να κατανοήσει τις ανάγκες και τις προτιμήσεις των πελατών της και να δημιουργήσει εξατομικευμένες εμπειρίες για κάθε πελάτη. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένας συγκεκριμένος πελάτης έχει προτίμηση σε ένα συγκεκριμένο προϊόν ή υπηρεσία, η Alumil μπορεί να συστήσει παρόμοια προϊόντα ή υπηρεσίες που μπορεί να ενδιαφέρουν τον πελάτη. Αυτό μπορεί να συμβάλει στην αύξηση της ικανοποίησης και της αφοσίωσης των πελατών.
Τέλος, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για προγνωστική ανάλυση. Με την ανάλυση των δεδομένων των πελατών, η Alumil μπορεί να κάνει προβλέψεις σχετικά με τη μελλοντική συμπεριφορά των πελατών, όπως ποια προϊόντα και υπηρεσίες είναι πιθανό να αγοράσουν στο μέλλον.

Βελτιστοποίηση της διαδικασίας παραγωγής

Αξιοποιώντας τα μεγάλα δεδομένα, η Alumil μπορεί να αποκτήσει πολύτιμες πληροφορίες για τις παραγωγικές της διαδικασίες, να εντοπίσει τομείς προς βελτίωση και να λάβει αποφάσεις βάσει δεδομένων που θα αυξήσουν την αποδοτικότητα και την παραγωγικότητα.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της παραγωγικής διαδικασίας στην Alumil είναι μέσω της χρήσης αλγορίθμων μηχανικής μάθησης. Αναλύοντας δεδομένα από τις μηχανές παραγωγής της, όπως λειτουργικά δεδομένα, κατανάλωση ενέργειας και παραγωγή, η Alumil μπορεί να εντοπίσει μοτίβα και τάσεις που υποδεικνύουν αναποτελεσματικότητα στις παραγωγικές της διαδικασίες. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για την ανάπτυξη προγνωστικών μοντέλων που μπορούν να προβλέψουν και να αποτρέψουν τις διακοπές της παραγωγής, να αυξήσουν την αποδοτικότητα της παραγωγής και να μειώσουν το κόστος.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της παραγωγικής διαδικασίας στην Alumil είναι μέσω της χρήσης της προγνωστικής συντήρησης. Αναλύοντας δεδομένα από τα μηχανήματα παραγωγής της, η Alumil μπορεί να προβλέψει πότε τα μηχανήματα είναι πιθανό να παρουσιάσουν βλάβη και να λάβει προληπτικά μέτρα για να αποφύγει δαπανηρές διακοπές λειτουργίας. Αυτό μπορεί να συμβάλει στη μείωση του κόστους συντήρησης και στην αύξηση της συνολικής αποδοτικότητας της παραγωγής.
Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για τη βελτιστοποίηση των διαδικασιών. Αναλύοντας δεδομένα από τις διαδικασίες παραγωγής της, η Alumil μπορεί να εντοπίσει τομείς για βελτίωση, όπως η αύξηση της παραγωγικής απόδοσης και η βελτίωση της ποιότητας των προϊόντων. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των παραγωγικών διαδικασιών της, την αύξηση της αποδοτικότητας και τη μείωση του κόστους.

Πωλήσεις και μάρκετινγκ

Στο σημερινό ταχέως μεταβαλλόμενο επιχειρηματικό περιβάλλον, οι εταιρείες πρέπει να είναι σε θέση να προσαρμόζονται και να εξελίσσονται γρήγορα προκειμένου να παραμείνουν ανταγωνιστικές. Ένας από τους βασικούς τρόπους για να το επιτύχουν αυτό είναι η αξιοποίηση των μεγάλων δεδομένων για την προώθηση των προσπαθειών πωλήσεων και μάρκετινγκ. Αναλύοντας δεδομένα σχετικά με τη συμπεριφορά των πελατών, τις τάσεις της αγοράς και τις επιδόσεις των πωλήσεων, η Alumil μπορεί να αποκτήσει πολύτιμες πληροφορίες που μπορούν να ενημερώσουν τις στρατηγικές πωλήσεων και μάρκετινγκ της.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τις πωλήσεις και το μάρκετινγκ στην Alumil είναι μέσω της ανάλυσης της συμπεριφοράς των πελατών. Με την ανάλυση δεδομένων από τις αλληλεπιδράσεις των πελατών, όπως οι ηλεκτρονικές αγορές, η επισκεψιμότητα του ιστότοπου και τα σχόλια των πελατών, η Alumil μπορεί να αποκτήσει βαθύτερη κατανόηση των αναγκών και των προτιμήσεων των πελατών της. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για την ενημέρωση των προσπαθειών πωλήσεων και μάρκετινγκ, όπως η ανάπτυξη στοχευμένων εκστρατειών μάρκετινγκ και η προσφορά εξατομικευμένων προϊόντων και υπηρεσιών.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τις πωλήσεις και το μάρκετινγκ της Alumil είναι μέσω της ανάλυσης των τάσεων της αγοράς. Με την ανάλυση δεδομένων σχετικά με τις τάσεις της αγοράς, όπως οι επιδόσεις των ανταγωνιστών, οι τάσεις του κλάδου και οι προτιμήσεις των πελατών, η Alumil μπορεί να παραμείνει μπροστά από τις εξελίξεις και να αναπτύξει στρατηγικές πωλήσεων και μάρκετινγκ που ανταποκρίνονται στις μεταβαλλόμενες ανάγκες των πελατών της. Οι πληροφορίες αυτές μπορούν επίσης να βοηθήσουν την Alumil να εντοπίσει νέες ευκαιρίες στην αγορά και να παραμείνει μπροστά από τους ανταγωνιστές της.
Επιπλέον, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση της απόδοσης των πωλήσεων. Με την ανάλυση δεδομένων σχετικά με τις επιδόσεις των πωλήσεων, όπως τα στοιχεία πωλήσεων, τα σχόλια των πελατών και τις τάσεις της αγοράς, η Alumil μπορεί να εντοπίσει τομείς για βελτίωση των προσπαθειών πωλήσεων και μάρκετινγκ. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των στρατηγικών πωλήσεων και μάρκετινγκ, όπως η προσαρμογή της τιμολόγησης, η βελτίωση των προσφορών προϊόντων και η αύξηση της δέσμευσης των πελατών.
Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για την τμηματοποίηση των πελατών. Αναλύοντας δεδομένα σχετικά με τη συμπεριφορά των πελατών, η Alumil μπορεί να τμηματοποιήσει την πελατειακή της βάση σε διαφορετικές ομάδες με βάση παράγοντες όπως δημογραφικές πληροφορίες, αγοραστική συμπεριφορά και προτιμήσεις. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για την ανάπτυξη στοχευμένων εκστρατειών μάρκετινγκ και την προσφορά εξατομικευμένων προϊόντων και υπηρεσιών σε κάθε τμήμα πελατών.
Τέλος, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων βάση των μεγάλων δεδομένων. Αναλύοντας δεδομένα σχετικά με τη συμπεριφορά των πελατών, τις τάσεις της αγοράς και την απόδοση των πωλήσεων, η Alumil μπορεί να λάβει αποφάσεις βάσει δεδομένων που θα βελτιώσουν τις προσπάθειες πωλήσεων και μάρκετινγκ. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι μια συγκεκριμένη εκστρατεία μάρκετινγκ δεν έχει απήχηση στους πελάτες, η Alumil μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες για να προβεί σε αλλαγές που θα βελτιώσουν

τις προσπάθειες μάρκετινγκ και θα αυξήσουν τη δέσμευση των πελατών της. Με τη χρήση της ανάλυσης της συμπεριφοράς των πελατών, της ανάλυσης των τάσεων της αγοράς, της ανάλυσης της απόδοσης των πωλήσεων, της τμηματοποίησης των πελατών και της λήψης αποφάσεων βάση των μεγάλων δεδομένων, η Alumil μπορεί να αποκτήσει πολύτιμες πληροφορίες για τους πελάτες της και την αγορά, να ενημερώσει τις στρατηγικές πωλήσεων και μάρκετινγκ και να βελτιώσει τη συνολική απόδοση των πωλήσεών της.

Ποιοτικός έλεγχος

Ο ποιοτικός έλεγχος αποτελεί κρίσιμο στοιχείο κάθε παραγωγικής διαδικασίας, και η Alumil δεν αποτελεί εξαίρεση. Η διασφάλιση ότι τα προϊόντα πληρούν τις απαιτήσεις και τα πρότυπα των πελατών είναι απαραίτητη για τη διατήρηση της θετικής φήμης, την οικοδόμηση της εμπιστοσύνης των πελατών και την προώθηση της μακροπρόθεσμης επιτυχίας. Τα μεγάλα δεδομένα μπορούν να διαδραματίσουν καθοριστικό ρόλο στη βελτιστοποίηση του ποιοτικού ελέγχου στην Alumil, παρέχοντας πολύτιμες πληροφορίες για τη διαδικασία παραγωγής.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τον ποιοτικό έλεγχο στην Alumil είναι μέσω της παρακολούθησης σε πραγματικό χρόνο. Με τη συλλογή δεδομένων από διάφορες πηγές παραγωγής, όπως αισθητήρες, κάμερες και άλλα συστήματα παρακολούθησης, η Alumil μπορεί να αποκτήσει μια εικόνα της παραγωγικής διαδικασίας σε πραγματικό χρόνο. Αυτές οι πληροφορίες μπορούν στη συνέχεια να αναλυθούν για τον έγκαιρο εντοπισμό πιθανών προβλημάτων ποιότητας, όπως δυσλειτουργίες του εξοπλισμού ή αποκλίσεις της διαδικασίας, επιτρέποντας στην Alumil να λάβει διορθωτικά μέτρα πριν τα προϊόντα φτάσουν στον πελάτη.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τον έλεγχο της ποιότητας στην Alumil είναι μέσω της προληπτικής συντήρησης. Με την ανάλυση δεδομένων σχετικά με την απόδοση και τη χρήση του εξοπλισμού, η Alumil μπορεί να εντοπίσει πιθανά προβλήματα πριν αυτά εμφανιστούν. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για τον προγραμματισμό της προληπτικής συντήρησης και την ελαχιστοποίηση του κινδύνου βλάβης του εξοπλισμού κατά τη διάρκεια της παραγωγής, η οποία μπορεί να οδηγήσει σε προβλήματα ποιότητας και διακοπή της παραγωγής.
Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για τη βελτιστοποίηση της διαδικασίας. Με την ανάλυση δεδομένων σχετικά με τις διαδικασίες παραγωγής, η Alumil μπορεί να εντοπίσει περιοχές προς βελτίωση και να προβεί σε αλλαγές που θα βελτιώσουν την ποιότητα και την αποδοτικότητα. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένα συγκεκριμένο στάδιο παραγωγής διαρκεί περισσότερο από το αναμενόμενο, η Alumil μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες για να προβεί σε αλλαγές που θα μειώσουν τον απαιτούμενο χρόνο για το συγκεκριμένο στάδιο, γεγονός που μπορεί να οδηγήσει σε βελτιωμένη ποιότητα και υψηλότερη αποδοτικότητα της παραγωγής.
Επιπλέον, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση των βαθύτερων αιτιών. Όταν εμφανίζονται προβλήματα ποιότητας, η Alumil μπορεί να χρησιμοποιήσει τα δεδομένα για να εντοπίσει τη βασική αιτία του προβλήματος και να προβεί σε αλλαγές για να αποτρέψει την επανάληψή του στο μέλλον. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένα συγκεκριμένο στάδιο παραγωγής οδηγεί σε υψηλό αριθμό ελαττωμάτων, η Alumil μπορεί να χρησιμοποιήσει αυτές τις

πληροφορίες για να προβεί σε αλλαγές που θα εξαλείψουν τη βασική αιτία του προβλήματος και θα βελτιώσουν την ποιότητα του προϊόντος.
Τέλος, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για συνεχή βελτίωση. Συλλέγοντας και αναλύοντας δεδομένα σχετικά με τις διαδικασίες παραγωγής και την ποιότητα, η Alumil μπορεί να κάνει συνεχείς βελτιώσεις που θα οδηγήσουν σε καλύτερα προϊόντα και αυξημένη ικανοποίηση των πελατών. Οι πληροφορίες αυτές μπορούν επίσης να χρησιμοποιηθούν για τον εντοπισμό τομέων για βελτίωση των διαδικασιών, όπως η μείωση της σπατάλης, η αύξηση της αποδοτικότητας και η μείωση του κόστους, γεγονός που μπορεί να οδηγήσει σε μια πιο βιώσιμη και κερδοφόρα επιχείρηση.

EFOOD.GR

Το efood είναι υπηρεσία delivery στην Ελλάδα μέσω της οποίας μπορεί κανείς να παραγγείλει από 20.000 καταστήματα σε 100 πόλεις της Ελλάδας. Από φαγητό και καφέ, μέχρι τα καθημερινά σου ψώνια και τα απαραίτητα από μανάβικα, ιχθυοπωλεία, minimarket, κρεοπωλεία, ζαχαροπλαστεία, κάβες, bar και φούρνους. Στο σημερινό ταχέως εξελισσόμενο και διαρκώς μεταβαλλόμενο επιχειρηματικό τοπίο, οι εταιρείες αναζητούν συνεχώς νέους και καινοτόμους τρόπους συλλογής, ανάλυσης και αξιοποίησης των δεδομένων. Το eFood.gr είναι μια εξέχουσα υπηρεσία παράδοσης φαγητού στην Ελλάδα, η οποία παρέχει στους πελάτες της ένα ευρύ φάσμα επιλογών φαγητού από τοπικά εστιατόρια. Προκειμένου να παραμείνει μπροστά από τον ανταγωνισμό και να παρέχει την καλύτερη εμπειρία στον πελάτη, το eFood.gr αξιοποιεί τα μεγάλα δεδομένα για να κατανοήσει τη συμπεριφορά των πελατών.

Ανάλυση συμπεριφοράς πελάτη

Πρώτα απ' όλα, τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη συλλογή τεράστιου όγκου πληροφοριών σχετικά με τη συμπεριφορά των πελατών. Οι πληροφορίες αυτές μπορούν να ληφθούν από διάφορες πηγές, συμπεριλαμβανομένων των αναλύσεων ιστοτόπων, των ανατροφοδοτήσεων και των ερευνών πελατών, των μέσων κοινωνικής δικτύωσης και των δεδομένων συναλλαγών. Με αυτές τις πληροφορίες, η eFood.gr μπορεί να αποκτήσει μια ολοκληρωμένη κατανόηση των προτιμήσεων των πελατών, των αγοραστικών συνηθειών και των παραγόντων που επηρεάζουν τη διαδικασία λήψης αποφάσεων. Οι πληροφορίες αυτές είναι ζωτικής σημασίας για την ανάπτυξη στοχευμένων στρατηγικών μάρκετινγκ και τη δημιουργία εξατομικευμένων εμπειριών για τους πελάτες.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση της συμπεριφοράς των πελατών στην eFood.gr είναι μέσω της χρήσης προγνωστικών αναλύσεων. Η προγνωστική ανάλυση είναι μια διαδικασία που χρησιμοποιεί ιστορικά δεδομένα και αλγόριθμους μηχανικής μάθησης για να κάνει προβλέψεις για μελλοντικά γεγονότα. Για παράδειγμα, το eFood.gr μπορεί να χρησιμοποιήσει την προγνωστική ανάλυση για να εντοπίσει μοτίβα στη συμπεριφορά των πελατών και να προβλέψει τι είναι πιθανό να παραγγείλουν στο μέλλον. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για την παροχή εξατομικευμένων συστάσεων στους πελάτες, βελτιώνοντας τη συνολική εμπειρία του πελάτη και αυξάνοντας την πιθανότητα επανάληψης της επιχείρησης.

Βελτιστοποίηση παράδοσης

Το eFood.gr αξιοποιεί τα μεγάλα δεδομένα για τη βελτιστοποίηση της διαδικασίας παράδοσης. Μπορούν να χρησιμοποιηθούν για την παρακολούθηση των οχημάτων παράδοσης σε πραγματικό χρόνο. Με τη βοήθεια του εντοπισμού GPS και άλλων τεχνολογιών, η eFood.gr μπορεί να παρακολουθεί τη θέση και την κίνηση των οχημάτων παράδοσης σε πραγματικό χρόνο. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της διαδρομής παράδοσης, μειώνοντας το χρόνο ταξιδιού και αυξάνοντας την αποτελεσματικότητα της διαδικασίας παράδοσης. Επιπλέον, η παρακολούθηση σε πραγματικό χρόνο μπορεί επίσης να βοηθήσει την eFood.gr να ανταποκριθεί γρήγορα σε τυχόν απροσδόκητα προβλήματα, όπως καθυστερήσεις στην κυκλοφορία, διασφαλίζοντας ότι οι παραδόσεις εξακολουθούν να πραγματοποιούνται εγκαίρως. Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της παράδοσης στο eFood.gr είναι μέσω της χρήσης προγνωστικών αναλύσεων. Για παράδειγμα, το eFood.gr μπορεί να χρησιμοποιήσει την προγνωστική ανάλυση για να προβλέψει τους χρόνους παράδοσης με βάση δεδομένα του παρελθόντος και μοτίβα κίνησης. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των διαδρομών παράδοσης και να διασφαλιστεί ότι οι παραδόσεις γίνονται στην ώρα τους. Επιπλέον, η προγνωστική ανάλυση μπορεί επίσης να χρησιμοποιηθεί για την πρόβλεψη πιθανών προβλημάτων που μπορεί να προκύψουν κατά τη διαδικασία παράδοσης, επιτρέποντας στην eFood.gr να τα επιλύσει προληπτικά πριν γίνουν πρόβλημα. Με τη βοήθεια εντοπισμού GPS και άλλων τεχνολογιών, η eFood.gr μπορεί να παρακολουθεί τη συμπεριφορά των πελατών σε πραγματικό χρόνο και να ανταποκρίνεται ανάλογα. Για παράδειγμα, εάν ένας πελάτης αντιμετωπίζει καθυστέρηση στην παράδοση, το eFood.gr μπορεί να χρησιμοποιήσει την παρακολούθηση σε πραγματικό χρόνο για να επιλύσει γρήγορα το ζήτημα και να διασφαλίσει ότι ο πελάτης θα παραλάβει την παραγγελία του στην ώρα της. Αυτή η ικανότητα γρήγορης ανταπόκρισης στις ανάγκες των πελατών μπορεί να βελτιώσει σημαντικά την ικανοποίηση και την αφοσίωση των πελατών.
Η eFood.gr μπορεί να συγκεντρώσει δεδομένα σχετικά με τα οχήματα παράδοσης, τους οδηγούς και τις διαδρομές και να χρησιμοποιήσει αυτές τις πληροφορίες για τη βελτιστοποίηση της κατανομής των πόρων. Για παράδειγμα, η eFood.gr μπορεί να χρησιμοποιήσει δεδομένα σχετικά με τις διαδρομές παράδοσης και τη χρήση των οχημάτων για να προσδιορίσει ποια οχήματα υπολειτουργούν και να τα ανακατευθύνει σε περιοχές με μεγαλύτερη ζήτηση για υπηρεσίες παράδοσης. Αυτό μπορεί να οδηγήσει σε αποδοτικότερη χρήση των πόρων και βελτιωμένους χρόνους παράδοσης.

Βελτιστοποίηση μενού

Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση του μενού στο eFood.gr είναι μέσω της χρήσης της ανάλυσης δεδομένων πελατών. Αναλύοντας τη συμπεριφορά και τις προτιμήσεις των πελατών, το eFood.gr μπορεί να προσδιορίσει ποια στοιχεία του μενού είναι πιο δημοφιλή και ποια δεν έχουν καλή απόδοση. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των προσφορών του μενού και να διασφαλιστεί ότι οι πελάτες έχουν τις καλύτερες δυνατές επιλογές. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι ένα συγκεκριμένο στοιχείο του μενού δεν είναι δημοφιλές, το eFood.gr μπορεί να το αφαιρέσει από το μενού και να το

αντικαταστήσει με μια πιο δημοφιλή επιλογή. Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση του μενού στο eFood.gr είναι μέσω της χρήσης προγνωστικών αναλύσεων. Για παράδειγμα, το eFood.gr μπορεί να χρησιμοποιήσει την προγνωστική ανάλυση για να προβλέψει ποια στοιχεία του μενού θα είναι πιο δημοφιλή στο μέλλον με βάση τη συμπεριφορά και τις προτιμήσεις των πελατών στο παρελθόν. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τη βελτιστοποίηση των προσφορών του μενού και να διασφαλιστεί ότι οι πελάτες θα έχουν τις καλύτερες δυνατές επιλογές.

Πωλήσεις και μάρκετινγκ

Προκειμένου να βελτιώσει τις προσπάθειες πωλήσεων και μάρκετινγκ, η eFood.gr αξιοποιεί τα μεγάλα δεδομένα για να αποκτήσει πολύτιμες πληροφορίες σχετικά με τη συμπεριφορά και τις προτιμήσεις των πελατών της.
Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση πωλήσεων και μάρκετινγκ στην eFood.gr είναι μέσω της χρήσης της ανάλυσης δεδομένων πελατών. Αναλύοντας τη συμπεριφορά και τις προτιμήσεις των πελατών, η eFood.gr μπορεί να αποκτήσει πολύτιμες πληροφορίες σχετικά με το τι θέλουν οι πελάτες και πώς αλληλεπιδρούν με τις προσφορές της εταιρείας. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για την ανάπτυξη στοχευμένων εκστρατειών μάρκετινγκ και στρατηγικών πωλήσεων που είναι προσαρμοσμένες στις συγκεκριμένες ανάγκες και προτιμήσεις των πελατών της. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι μια συγκεκριμένη ομάδα πελατών είναι πιο πιθανό να αγοράσει ένα συγκεκριμένο στοιχείο του μενού, η eFood.gr μπορεί να στοχεύσει σε αυτούς τους πελάτες με εκστρατείες μάρκετινγκ που περιλαμβάνουν αυτό το στοιχείο του μενού.
Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση πωλήσεων και μάρκετινγκ στην eFood.gr είναι μέσω της χρήσης προγνωστικών αναλύσεων. Για παράδειγμα, το eFood.gr μπορεί να χρησιμοποιήσει την προγνωστική ανάλυση για να προβλέψει ποια στοιχεία του μενού θα είναι πιο δημοφιλή στο μέλλον με βάση τη συμπεριφορά και τις προτιμήσεις των πελατών του παρελθόντος. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για την ανάπτυξη στοχευμένων εκστρατειών μάρκετινγκ και στρατηγικών πωλήσεων που είναι προσαρμοσμένες στις συγκεκριμένες ανάγκες και προτιμήσεις των πελατών της. Τα μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για την παρακολούθηση της συμπεριφοράς των πελατών σε πραγματικό χρόνο. Για παράδειγμα, εάν ένας πελάτης αναζητά ένα συγκεκριμένο στοιχείο του μενού αλλά δεν μπορεί να το βρει, το eFood.gr μπορεί να χρησιμοποιήσει την παρακολούθηση σε πραγματικό χρόνο για να προσθέσει γρήγορα το στοιχείο στο μενού και να παρέχει στον πελάτη τις επιλογές που αναζητά. Αυτή η ικανότητα γρήγορης ανταπόκρισης στις ανάγκες των πελατών μπορεί να βελτιώσει σημαντικά την ικανοποίηση των πελατών και να αυξήσει τις πωλήσεις.

Ποιοτικός έλεγχος

Στον κλάδο της παράδοσης τροφίμων, η διατήρηση υψηλών επιπέδων ποιότητας τροφίμων είναι υψίστης σημασίας. Οι πελάτες αναμένουν τα τρόφιμα που παραλαμβάνουν να είναι υψηλής ποιότητας και ασφαλή για κατανάλωση, και οι εταιρείες πρέπει να λαμβάνουν όλα τα απαραίτητα μέτρα για να διασφαλίσουν ότι αυτό συμβαίνει. Το eFood.gr είναι μια κορυφαία υπηρεσία παράδοσης φαγητού στην

Ελλάδα που παρέχει στους πελάτες της ένα ευρύ φάσμα επιλογών φαγητού από τοπικά εστιατόρια. Προκειμένου να διατηρήσει τη φήμη του για υψηλής ποιότητας τρόφιμα και να διασφαλίσει την ικανοποίηση των πελατών του, το eFood.gr αξιοποιεί τα μεγάλα δεδομένα για την ανάλυση του ποιοτικού ελέγχου. Ένας από τους βασικούς τρόπους με τους οποίους τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση του ποιοτικού ελέγχου στο eFood.gr είναι μέσω της χρήσης των ανατροφοδοτήσεων των πελατών. Συλλέγοντας και αναλύοντας τα σχόλια των πελατών, η eFood.gr μπορεί να αποκτήσει πολύτιμες πληροφορίες σχετικά με την ποιότητα των τροφίμων της και το επίπεδο ικανοποίησης των πελατών της. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για τον εντοπισμό τομέων προς βελτίωση και την πραγματοποίηση αλλαγών στις διαδικασίες προμήθειας, παρασκευής και παράδοσης τροφίμων της εταιρείας. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι οι πελάτες είναι συστηματικά δυσαρεστημένοι με την ποιότητα ενός συγκεκριμένου στοιχείου του μενού, η eFood.gr μπορεί να λάβει μέτρα για τη βελτίωση της ποιότητας του συγκεκριμένου στοιχείου ή να το αφαιρέσει εντελώς από το μενού. Ένας άλλος τρόπος με τον οποίο τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάλυση του ποιοτικού ελέγχου στο eFood.gr είναι μέσω της χρήσης δεδομένων για την ασφάλεια των τροφίμων. Με τη βοήθεια αισθητήρων και άλλων τεχνολογιών, το eFood.gr μπορεί να παρακολουθεί τη θερμοκρασία και την ασφάλεια των τροφίμων κατά τη μεταφορά, την αποθήκευση και την προετοιμασία. Αυτές οι πληροφορίες μπορούν στη συνέχεια να χρησιμοποιηθούν για τον εντοπισμό και την πρόληψη περιστατικών ασφάλειας τροφίμων, όπως η μόλυνση ή η αλλοίωση των τροφίμων. Αυτό μπορεί να βοηθήσει να διασφαλιστεί ότι τα τρόφιμα που παραδίδονται στους πελάτες είναι ασφαλή για κατανάλωση και υψηλής ποιότητας.
Αναλύοντας δεδομένα, το eFood.gr μπορεί να παρακολουθεί την προετοιμασία των τροφίμων σε πραγματικό χρόνο και να διασφαλίζει ότι πληρούν τα πρότυπα ποιότητας. Για παράδειγμα, εάν τα δεδομένα δείχνουν ότι μια συγκεκριμένη κουζίνα χρησιμοποιεί λανθασμένες θερμοκρασίες μαγειρέματος, το eFood.gr μπορεί να παρέμβει και να διορθώσει το ζήτημα σε πραγματικό χρόνο, αποτρέποντας την παράδοση υποβαθμισμένων τροφίμων. Αυτή η δυνατότητα παρακολούθησης της προετοιμασίας των τροφίμων σε πραγματικό χρόνο μπορεί να βελτιώσει σημαντικά την ποιότητα των τροφίμων και την ικανοποίηση των πελατών.

Walmart

Η Walmart είναι ο μεγαλύτερος λιανοπωλητής στον κόσμο και η μεγαλύτερη εταιρεία παγκοσμίως βάσει εσόδων, με πάνω από δύο εκατομμύρια υπαλλήλους και 20.000 καταστήματα σε 28 χώρες.
Με επιχειρήσεις αυτής της κλίμακας δεν αποτελεί έκπληξη το γεγονός ότι έχουν δει εδώ και καιρό την αξία της ανάλυσης δεδομένων. Το 2004, όταν ο τυφώνας Sandy έπληξε τις ΗΠΑ, διαπίστωσαν ότι απροσδόκητες ιδέες μπορούσαν να έρθουν στο φως όταν τα δεδομένα μελετούνταν ως σύνολο και όχι ως μεμονωμένα σύνολα. Επιχειρώντας να προβλέψει τη ζήτηση για προμήθειες έκτακτης ανάγκης ενόψει του επερχόμενου τυφώνα Σάντι, η CIO Linda Dillman κατέληξε σε μερικά εκπληκτικά στατιστικά στοιχεία. Εκτός από τους φακούς και τον εξοπλισμό έκτακτης ανάγκης, η αναμενόμενη κακοκαιρία είχε οδηγήσει σε έξαρση των πωλήσεων ταρτών με φράουλες σε διάφορες άλλες τοποθεσίες. Πρόσθετες προμήθειες από αυτές στάλθηκαν σε καταστήματα στο πέρασμα του τυφώνα Φράνσις το 2012 και πουλήθηκαν εξαιρετικά καλά.

Από τότε η Walmart έχει αναπτύξει σημαντικά το τμήμα Big Data και analytics, παραμένοντας συνεχώς στην αιχμή του δόρατος. Το 2015, η εταιρεία ανακοίνωσε ότι βρισκόταν στη διαδικασία δημιουργίας του μεγαλύτερου ιδιωτικού νέφους δεδομένων στον κόσμο, ώστε να καταστεί δυνατή η επεξεργασία 2,5 petabytes πληροφοριών κάθε ώρα (ProjectPro, 2022).

CERN

Το CERN είναι ο διεθνής οργανισμός επιστημονικής έρευνας που λειτουργεί τον Μεγάλο Επιταχυντή Αδρονίων (LHC), το μεγαλύτερο και πιο προηγμένο πείραμα φυσικής της ανθρωπότητας. Οι επιταχυντές, που βρίσκονται μέσα σε 17 μίλια σηράγγων θαμμένων 600 πόδια κάτω από την επιφάνεια της Ελβετίας και της Γαλλίας, έχουν ως στόχο να προσομοιώσουν τις συνθήκες στο σύμπαν χιλιοστά του δευτερολέπτου μετά τη Μεγάλη Έκρηξη. Αυτό επιτρέπει στους φυσικούς να αναζητήσουν ασύλληπτα θεωρητικά σωματίδια, όπως το μποζόνιο Higgs, τα οποία θα μπορούσαν να μας δώσουν μια άνευ προηγουμένου εικόνα της σύνθεσης του σύμπαντος. Τα έργα του CERN, όπως ο LHC, δεν θα ήταν εφικτά αν δεν υπήρχε το Διαδίκτυο και τα Μεγάλα Δεδομένα. Το Διαδίκτυο δημιουργήθηκε αρχικά στο CERN τη δεκαετία του 1990 και ο Tim Berners-Lee, ο άνθρωπος που συχνά αναφέρεται ως ο "πατέρας του Διαδικτύου", ανέπτυξε το πρωτόκολλο υπερκειμένου που συγκρατεί τον Παγκόσμιο Ιστό ενώ εργαζόταν στο CERN. Ο αρχικός σκοπός του ήταν να διευκολύνει την επικοινωνία μεταξύ ερευνητών σε όλο τον κόσμο. Μόνο ο LHC παράγει περίπου 30 petabytes πληροφοριών ετησίως, 15 τρισεκατομμύρια σελίδες τυπωμένου κειμένου. Οι συγκρούσεις που παρακολουθούνται στον LHC συμβαίνουν πολύ γρήγορα και τα υποατομικά "συντρίμμια" που προκύπτουν και περιέχουν τα ασύλληπτα, περιζήτητα σωματίδια υπάρχουν μόνο για μερικά εκατομμυριοστά του δευτερολέπτου πριν διασπαστούν. Οι ακριβείς συνθήκες που προκαλούν την απελευθέρωση των σωματιδίων που αναζητά το CERN συμβαίνουν μόνο υπό πολύ ακριβείς συνθήκες, με αποτέλεσμα πολλές εκατοντάδες εκατομμύρια συγκρούσεις να πρέπει να παρακολουθούνται και να καταγράφονται κάθε δευτερόλεπτο με την ελπίδα ότι οι αισθητήρες θα τις εντοπίσουν. Οι αισθητήρες του LHC καταγράφουν εκατοντάδες εκατομμύρια συγκρούσεις μεταξύ σωματιδίων, ορισμένα από τα οποία επιτυγχάνουν ταχύτητες μόλις ένα κλάσμα κάτω από την ταχύτητα του φωτός καθώς επιταχύνονται γύρω από τον επιταχυντή. Αυτό δημιουργεί τεράστιο όγκο δεδομένων και απαιτεί πολύ ευαίσθητο και ακριβή εξοπλισμό για τη μέτρηση και την καταγραφή των αποτελεσμάτων. Ο LHC χρησιμοποιείται σε τέσσερα κύρια πειράματα, στα οποία συμμετέχουν περίπου 8000 αναλυτές από όλο τον κόσμο. Χρησιμοποιούν τα δεδομένα για να αναζητήσουν ασύλληπτα θεωρητικά σωματίδια και να διερευνήσουν τις απαντήσεις σε ερωτήματα που αφορούν την αντιύλη, τη σκοτεινή ύλη και τις επιπλέον διαστάσεις στο χώρο και το χρόνο. Τα δεδομένα συλλέγονται από αισθητήρες στο εσωτερικό του επιταχυντή που παρακολουθούν εκατοντάδες εκατομμύρια συγκρούσεις σωματιδίων κάθε δευτερόλεπτο. Οι αισθητήρες συλλαμβάνουν φως, οπότε είναι ουσιαστικά κάμερες, με ανάλυση 100 megapixel, ικανές να καταγράφουν εικόνες σε απίστευτα υψηλές ταχύτητες. Τα δεδομένα αυτά αναλύονται στη συνέχεια από αλγορίθμους που είναι συντονισμένοι ώστε να εντοπίζουν τις αποκαλυπτικές ενεργειακές υπογραφές που αφήνει πίσω της η εμφάνιση και η εξαφάνιση των εξωτικών σωματιδίων που αναζητά το CERN. Οι αλγόριθμοι συγκρίνουν τις εικόνες που προκύπτουν με θεωρητικά δεδομένα που εξηγούν πώς πιστεύουμε ότι θα ενεργήσουν τα σωματίδια-στόχοι, όπως το μποζόνιο

Higgs. Εάν τα αποτελέσματα ταιριάζουν, αυτό αποτελεί απόδειξη ότι οι αισθητήρες έχουν βρει τα σωματίδια-στόχους (CERN, 2013).

ROLLS - ROYCE

Η Rolls-Royce κατασκευάζει τεράστιους κινητήρες που χρησιμοποιούνται από 500 αεροπορικές εταιρείες και περισσότερες από 150 ένοπλες δυνάμεις. Αυτοί οι κινητήρες παράγουν τεράστιες ποσότητες ενέργειας και δεν αποτελεί έκπληξη το γεγονός ότι μια εταιρεία που έχει συνηθίσει να ασχολείται με μεγάλους αριθμούς έχει αγκαλιάσει ολόψυχα τα Big Data. Πρόκειται για έναν κλάδο εξαιρετικά υψηλής τεχνολογίας, όπου οι αποτυχίες και τα λάθη μπορούν να κοστίσουν δισεκατομμύρια, καθώς και ανθρώπινες ζωές. Επομένως, είναι ζωτικής σημασίας η εταιρεία να είναι σε θέση να παρακολουθεί την υγεία των προϊόντων της για να εντοπίζει πιθανά προβλήματα πριν αυτά εμφανιστούν. Τα δεδομένα που συλλέγει η Rolls-Royce τη βοηθούν να σχεδιάζει πιο ανθεκτικά προϊόντα, να συντηρεί αποτελεσματικά τα προϊόντα και να παρέχει καλύτερες υπηρεσίες στους πελάτες της.
Η Rolls-Royce χρησιμοποιεί τις διαδικασίες Big Data σε τρεις βασικούς τομείς των δραστηριοτήτων της: το σχεδιασμό, την κατασκευή και την υποστήριξη μετά την πώληση. Ας δούμε με τη σειρά κάθε τομέα. Ο Paul Stein, επικεφαλής επιστημονικός υπεύθυνος της εταιρείας, λέει: "Διαθέτουμε τεράστιες συστάδες υπολογιστών υψηλής ισχύος που χρησιμοποιούνται στη διαδικασία σχεδιασμού. Παράγουμε δεκάδες terabytes δεδομένων σε κάθε προσομοίωση ενός από τους κινητήρες μας. Στη συνέχεια πρέπει να χρησιμοποιήσουμε κάποιες αρκετά εξελιγμένες τεχνικές υπολογιστών για να εξετάσουμε αυτό το τεράστιο σύνολο δεδομένων και να απεικονίσουμε αν το συγκεκριμένο προϊόν που έχουμε σχεδιάσει είναι καλό ή κακό. Η οπτικοποίηση των μεγάλων δεδομένων είναι εξίσου σημαντική με τις τεχνικές που χρησιμοποιούμε για τον χειρισμό τους". Στην πραγματικότητα, ελπίζουν τελικά να είναι σε θέση να οπτικοποιήσουν τα προϊόντα τους σε λειτουργία σε όλες τις πιθανές ακραίες συμπεριφορές στις οποίες χρησιμοποιούνται. Τα συστήματα παραγωγής της εταιρείας δικτυώνονται όλο και περισσότερο και επικοινωνούν μεταξύ τους στην προσπάθεια για ένα δικτυωμένο βιομηχανικό περιβάλλον, το Διαδίκτυο των Πραγμάτων (IoT). Όσον αφορά την υποστήριξη μετά την πώληση, οι κινητήρες και τα συστήματα πρόωσης της Rolls-Royce είναι όλα εξοπλισμένα με εκατοντάδες αισθητήρες που καταγράφουν κάθε μικροσκοπική λεπτομέρεια σχετικά με τη λειτουργία τους και αναφέρουν τυχόν αλλαγές στα δεδομένα σε πραγματικό χρόνο στους μηχανικούς, οι οποίοι στη συνέχεια αποφασίζουν τον καλύτερο τρόπο δράσης. Η Rolls-Royce διαθέτει λειτουργικά κέντρα σέρβις σε όλο τον κόσμο, στα οποία εξειδικευμένοι μηχανικοί αναλύουν τα δεδομένα που ανατροφοδοτούνται από τους κινητήρες τους. Μπορούν να συνδυάσουν τα δεδομένα από τους κινητήρες τους για να αναδείξουν παράγοντες και συνθήκες υπό τις οποίες οι κινητήρες μπορεί να χρειάζονται συντήρηση. Σε ορισμένες περιπτώσεις, οι άνθρωποι θα παρέμβουν στη συνέχεια για να αποφύγουν ή να μετριάσουν οτιδήποτε είναι πιθανό να προκαλέσει πρόβλημα. Όλο και περισσότερο, η Rolls-Royce αναμένει ότι οι υπολογιστές θα πραγματοποιούν οι ίδιοι την παρέμβαση. Με τους κινητήρες της πολιτικής αεροπορίας τόσο αξιόπιστους όσο είναι, η έμφαση μετατοπίζεται στη διατήρηση της μέγιστης δυνατής απόδοσής τους, εξοικονομώντας καύσιμα στις αεροπορικές εταιρείες και τηρώντας τα προγράμματά τους. Οι αναλύσεις μεγάλων δεδομένων βοηθούν τη Rolls-Royce να εντοπίζει τις ενέργειες συντήρησης ημέρες ή εβδομάδες νωρίτερα, έτσι ώστε οι αεροπορικές εταιρείες να μπορούν να προγραμματίζουν τις εργασίες χωρίς οι επιβάτες να υφίστανται καμία διακοπή. Για να υποστηριχθεί αυτό,

τα αναλυτικά συστήματα επί των κινητήρων επεξεργάζονται μεγάλο όγκο δεδομένων που παράγονται σε κάθε πτήση και μεταδίδουν μόνο τα σχετικά στοιχεία στο έδαφος για περαιτέρω ανάλυση. Μόλις φτάσουν στην πύλη, όλα τα δεδομένα της πτήσης είναι διαθέσιμα στους μηχανικούς για να τα εξετάσουν και να εντοπίσουν τα μικρά περιθώρια βελτίωσης των επιδόσεων. Ο τεράστιος αριθμός παραγόντων που λαμβάνονται υπόψη σημαίνει ότι όταν κάτι πάει στραβά, όλα όσα συνέβαλαν μπορούν να εντοπιστούν και το σύστημα μπορεί να μάθει να προβλέπει πότε και πού είναι πιθανό να επαναληφθεί το πρόβλημα. Ολοκληρώνοντας τον κύκλο, οι πληροφορίες αυτές επιστρέφουν στη διαδικασία σχεδιασμού (Rao, 2021).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα μεγάλα δεδομένα είναι ένας όρος που εμφανίστηκε την τελευταία δεκαετία για να περιγράψει τον τεράστιο όγκο δεδομένων που είναι πλέον διαθέσιμος στη σύγχρονη κοινωνία. Η εργασία αυτή παρείχε έναν ορισμό των μεγάλων δεδομένων, διερεύνησε τη σύντομη ιστορία τους και τόνισε τη σημασία τους στη σύγχρονη κοινωνία.
Στην εισαγωγή συζητήθηκε η εκθετική αύξηση των δεδομένων τα τελευταία χρόνια και η εμφάνιση νέων τεχνολογιών που επιτρέπουν τη συλλογή και την ανάλυση τεράστιων ποσοτήτων δεδομένων. Αποδείχθηκε ότι τα μεγάλα δεδομένα έχουν γίνει κρίσιμο μέρος πολλών κλάδων, όπως η οικονομία, η υγειονομική περίθαλψη και η κυβέρνηση, και έχουν μεταμορφώσει τον τρόπο με τον οποίο προσεγγίζουμε τα προβλήματα και λαμβάνουμε αποφάσεις.Προσφέρουν τεράστιες δυνατότητες για τη βελτίωση της λήψης αποφάσεων και την αποκάλυψη νέων γνώσεων. Έχουν επιτρέψει νέες μορφές έρευνας και ανάλυσης και έχουν φέρει επανάσταση στον τρόπο με τον οποίο συλλέγουμε, αποθηκεύουμε, επεξεργαζόμαστε και αναλύουμε πληροφορίες.
Τα μεγάλα δεδομένα έχουν αναδειχθεί σε κρίσιμο συστατικό της σύγχρονης κοινωνίας και η σημασία τους θα συνεχίσει να αυξάνεται τα επόμενα χρόνια. Καθώς αναπτύσσονται νέες τεχνολογίες και τεχνικές για την ανάλυση δεδομένων, οι δυνατότητες των μεγάλων δεδομένων να μετασχηματίσουν τους κλάδους και να προωθήσουν την καινοτομία είναι τεράστιες. Είναι σημαντικό τα άτομα και οι οργανισμοί να ενημερώνονται και να προσαρμόζονται στις νέες εξελίξεις στον τομέα για να αξιοποιήσουν πλήρως τη δύναμη των μεγάλων δεδομένων.
Τα τρία χαρακτηριστικά των μεγάλων δεδομένων, όγκος, ταχύτητα και ποικιλία, είναι θεμελιώδη για την κατανόηση των προκλήσεων και των ευκαιριών που παρουσιάζονται από τις τεράστιες ποσότητες δεδομένων που είναι διαθέσιμες στον σημερινό κόσμο.
Το πρώτο χαρακτηριστικό, ο όγκος, αναφέρεται στον τεράστιο όγκο δεδομένων που παράγονται και συλλέγονται. Αποδείχθηκε ότι ο όγκος αυτός αυξάνεται εκθετικά και ότι οι παραδοσιακές μέθοδοι αποθήκευσης και ανάλυσης δεδομένων είναι ανεπαρκείς για την αντιμετώπισή του. Αυτό δημιουργεί σημαντικές προκλήσεις, αλλά προσφέρει επίσης τεράστιες δυνατότητες για νέες γνώσεις και ανακαλύψεις.
Το δεύτερο χαρακτηριστικό, η ταχύτητα, αναφέρεται στην ταχύτητα με την οποία παράγονται τα δεδομένα και πρέπει να υποβληθούν σε επεξεργασία. Τονίστηκε ότι αυτό αποκτά ολοένα και μεγαλύτερη σημασία, καθώς περισσότερα δεδομένα

παράγονται σε πραγματικό χρόνο και απαιτούν άμεση προσοχή. Το γεγονός αυτό έχει οδηγήσει στη δημιουργία νέων εργαλείων και τεχνολογιών για την ανάλυση δεδομένων σε πραγματικό χρόνο.

Το τρίτο χαρακτηριστικό, η ποικιλία, αναφέρεται στους διαφορετικούς τύπους δεδομένων που παράγονται, από δομημένα και μη δομημένα δεδομένα έως κείμενο, εικόνες και βίντεο. Η ποικιλία παρουσιάζει σημαντικές προκλήσεις για την παραδοσιακή ανάλυση δεδομένων, αλλά προσφέρει επίσης ευκαιρίες για νέες μορφές ανάλυσης και ανακάλυψης.

Τα μεγάλα δεδομένα χρησιμοποιούνται σε διάφορους τομείς, από τις επιχειρήσεις και τα χρηματοοικονομικά μέχρι την υγειονομική περίθαλψη και την ιατρική, την κυβέρνηση και τη δημόσια πολιτική, τις μεταφορές και την εφοδιαστική, τα μέσα κοινωνικής δικτύωσης και την ψυχαγωγία.

Στις επιχειρήσεις και τα χρηματοοικονομικά, τα μεγάλα δεδομένα χρησιμοποιούνται για την απόκτηση γνώσεων σχετικά με τη συμπεριφορά των καταναλωτών, τη βελτίωση της διαχείρισης της αλυσίδας εφοδιασμού και την ανάπτυξη νέων προϊόντων και υπηρεσιών. Στην υγειονομική περίθαλψη και την ιατρική, τα μεγάλα δεδομένα χρησιμοποιούνται για τη βελτίωση των αποτελεσμάτων των ασθενών, την ανάπτυξη νέων θεραπειών και την ενίσχυση των προσπαθειών για τη δημόσια υγεία. Στην κυβέρνηση και τη δημόσια πολιτική, τα μεγάλα δεδομένα χρησιμοποιούνται για την ενημέρωση των πολιτικών αποφάσεων, την παρακολούθηση των κοινωνικών και οικονομικών τάσεων και τη βελτίωση των δημόσιων υπηρεσιών. Στις μεταφορές και τα logistics, τα μεγάλα δεδομένα χρησιμοποιούνται για τη βελτιστοποίηση των δικτύων logistics, τη μείωση της συμφόρησης και τη βελτίωση της ασφάλειας. Στα μέσα κοινωνικής δικτύωσης και την ψυχαγωγία, τα μεγάλα δεδομένα χρησιμοποιούνται για την εξατομίκευση του περιεχομένου και τη βελτίωση της δέσμευσης των χρηστών.

Οι πιθανές εφαρμογές των μεγάλων δεδομένων είναι τεράστιες και συνεχίζουν να εξελίσσονται, και είναι σαφές ότι ο αντίκτυπος των μεγάλων δεδομένων θα συνεχίσει να αυξάνεται τα επόμενα χρόνια. Ωστόσο, είναι σημαντικό να αναγνωρίσουμε ότι η χρήση των μεγάλων δεδομένων δημιουργεί επίσης σημαντικές ηθικές και κοινωνικές προκλήσεις, όπως ανησυχίες για την προστασία της ιδιωτικής ζωής και το ενδεχόμενο αλγοριθμικής προκατάληψης. Ως εκ τούτου, είναι σημαντικό να συνεχίσουμε να διερευνούμε τα δυνητικά οφέλη των μεγάλων δεδομένων, ενώ παράλληλα να εργαζόμαστε για τον μετριασμό των πιθανών κινδύνων και μειονεκτημάτων τους.

Η ποιότητα και η διαχείριση των δεδομένων αποτελούν βασικές προκλήσεις στην εποχή των μεγάλων δεδομένων, καθώς ο μεγάλος όγκος και η ποικιλία των δεδομένων μπορεί να καταστήσει δύσκολη τη διασφάλιση της ακρίβειας, της συνέπειας και της πληρότητας. Οι ανησυχίες για την προστασία της ιδιωτικής ζωής και την ασφάλεια αποτελούν επίσης σημαντική πρόκληση, καθώς η χρήση μεγάλων δεδομένων μπορεί να εγείρει ανησυχίες σχετικά με παραβιάσεις δεδομένων, κλοπή ταυτότητας και μη εξουσιοδοτημένη πρόσβαση σε προσωπικές πληροφορίες.
Τα νομικά και ηθικά ζητήματα αποτελούν επίσης σημαντικό πρόβλημα, καθώς η χρήση μεγάλων δεδομένων μπορεί να εγείρει ερωτήματα σχετικά με την κατάλληλη χρήση προσωπικών πληροφοριών, τις πιθανές διακρίσεις και προκαταλήψεις και την ευθύνη των οργανισμών για την προστασία της ιδιωτικής ζωής και των δικαιωμάτων των ατόμων. Επιπλέον, οι τεχνολογικοί περιορισμοί και οι απαιτήσεις υποδομής μπορεί να αποτελέσουν πρόκληση για τους οργανισμούς, καθώς η επεξεργαστική ισχύς και η χωρητικότητα αποθήκευσης που απαιτούνται για την ανάλυση μεγάλων δεδομένων μπορεί να είναι δαπανηρή και απαιτητική σε πόρους.
Παρά τις προκλήσεις αυτές, είναι σαφές ότι τα μεγάλα δεδομένα έχουν τη δυνατότητα να φέρουν επανάσταση στον τρόπο με τον οποίο εργαζόμαστε, ζούμε και αλληλεπιδρούμε μεταξύ μας. Ως εκ τούτου, είναι σημαντικό να συνεχίσουμε να διερευνούμε τρόπους για να ξεπεράσουμε αυτές τις προκλήσεις και τους περιορισμούς, όπως επενδύοντας σε συστήματα ποιότητας και διαχείρισης δεδομένων, εφαρμόζοντας ισχυρά πρωτόκολλα προστασίας της ιδιωτικής ζωής και ασφάλειας και αντιμετωπίζοντας νομικές και ηθικές ανησυχίες. Με τον τρόπο αυτό, μπορούμε να ξεκλειδώσουμε το πλήρες δυναμικό των μεγάλων δεδομένων και να αξιοποιήσουμε τη δύναμή τους για την προώθηση της καινοτομίας και της προόδου.
Συμπερασματικά, τα μεγάλα δεδομένα έχουν φέρει επανάσταση στον τρόπο με τον οποίο συλλέγουμε, αποθηκεύουμε, επεξεργαζόμαστε και αναλύουμε τεράστιες ποσότητες πληροφοριών. Όπως φαίνεται σε αυτή την εργασία, τα μεγάλα δεδομένα έχουν αποκτήσει ολοένα και μεγαλύτερη σημασία στη σύγχρονη κοινωνία, με ένα ευρύ φάσμα εφαρμογών στις επιχειρήσεις, την υγειονομική περίθαλψη, την κυβέρνηση, τις μεταφορές και την ψυχαγωγία.
Τα χαρακτηριστικά των μεγάλων δεδομένων, συμπεριλαμβανομένου του όγκου, της ταχύτητας και της ποικιλίας, έχουν παράσχει στους οργανισμούς νέες ευκαιρίες για την απόκτηση γνώσεων και τη βελτίωση της λήψης αποφάσεων. Επιπλέον, η ανάπτυξη νέων εργαλείων και τεχνολογιών για την ανάλυση μεγάλων δεδομένων,

όπως το Hadoop, το ApacheSpark και οι αλγόριθμοι μηχανικής μάθησης, έχουν διευκολύνει τους οργανισμούς να χειρίζονται και να εξάγουν αξία από τεράστιες ποσότητες δεδομένων.

Ωστόσο, τα μεγάλα δεδομένα παρουσιάζουν επίσης σημαντικές προκλήσεις και περιορισμούς. Σε αυτές περιλαμβάνονται ζητήματα που σχετίζονται με την ποιότητα και τη διαχείριση των δεδομένων, ζητήματα προστασίας της ιδιωτικής ζωής και ασφάλειας, νομικά και ηθικά ζητήματα, καθώς και τεχνολογικοί περιορισμοί και απαιτήσεις υποδομής.

Παρά τις προκλήσεις αυτές, τα δυνητικά οφέλη των μεγάλων δεδομένων δεν μπορούν να αγνοηθούν. Είναι ζωτικής σημασίας οι οργανισμοί να αντιμετωπίζουν τις προκλήσεις και τους περιορισμούς που σχετίζονται με τα μεγάλα δεδομένα με υπεύθυνο και ηθικό τρόπο, προκειμένου να διασφαλιστεί ότι η χρήση αυτής της τεχνολογίας είναι διαφανής, ασφαλής και επωφελής για όλους.

Ατενίζοντας το μέλλον, είναι σαφές ότι τα μεγάλα δεδομένα θα συνεχίσουν να διαδραματίζουν ολοένα και πιο σημαντικό ρόλο στον τρόπο με τον οποίο ζούμε και εργαζόμαστε. Καθώς όλο και περισσότεροι οργανισμοί αγκαλιάζουν τα μεγάλα δεδομένα και επενδύουν σε νέες τεχνολογίες και υποδομές, είναι πιθανό να δούμε ακόμη περισσότερες εφαρμογές και οφέλη αυτής της τεχνολογίας να αναδύονται. Ωστόσο, είναι σημαντικό να παραμείνουμε σε εγρήγορση σχετικά με τις προκλήσεις και τους περιορισμούς των μεγάλων δεδομένων και να συνεργαστούμε για την εξεύρεση λύσεων που μεγιστοποιούν τις δυνατότητές τους και ταυτόχρονα ελαχιστοποιούν τους κινδύνους τους.

Βιβλιογραφία

IBM. (χ.χ.). *What is predictive analytics?* Ανάκτηση από www.ibm.com/: https://www.ibm.com/topics/predictive-analytics

Abhishek Kaushik, S. N. (2016, 02). An Anatomy of Data Visualization. *IJCSNS International Journal of Computer Science and Network Security, 16*(2).

Accern. (2022, 09 08). *Structured vs Unstructured Data vs Semi-Structured Data (Differences)*. Ανάκτηση από Accern: https://accern.com/blog/structured-vs-semi-structured-vs-unstructured-data/

Adeel Shiraz Hashmi, T. A. (2016). Big Data Mining: Tools & Algorithms. *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, (σσ. 36-40).

Alan Nugent, F. H. (2013). *Big Data For Dummies.* John Wiley & Sons.

Al-Barhamtoshy, H. M. (2014). A Data Analytic Framework for Unstructured Text. *Life Science Journal 2014*, 339-347.

Alexandros Labrinidis, H. J. (2012). Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment, 5*(12), 2032-2033.

Allouche, G. (2014, 12 15). *Natural Language Processing and Big Data: A Powerful Combination*. Ανάκτηση από Data Versity: https://www.dataversity.net/natural-language-processing-big-data-powerful-combination/#

allthingsdistributed. (2012, 01 18). *Amazon DynamoDB – a Fast and Scalable NoSQL Database Service Designed for Internet Scale Applications*. Ανάκτηση από www.allthingsdistributed.com: https://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html

AMAZON. (χ.χ.). *What is Apache HBase?* Ανάκτηση από www.aws.amazon.com: https://aws.amazon.com/big-data/what-is-hbase/

AmirGandomi, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), σσ. 137-144.

Analytixlabs. (χ.χ.). *Introduction to SVM – Support Vector Machine Algorithm of Machine Learning*. Ανάκτηση από www.analytixlabs.co.in/: https://www.analytixlabs.co.in/blog/introduction-support-vector-machine-algorithm/#What_Is_a_Support_Vector_Machine

Apache. (χ.χ.). *What is Apache Cassandra?* Ανάκτηση από cassandra.apache.org: https://cassandra.apache.org/_/cassandra-basics.html

Apple. (χ.χ.). *Τι είναι ένα API;*. Ανάκτηση από Support Apple: https://support.apple.com/el-gr/guide/shortcuts-mac/apd2e30c9d45/mac

Avast. (χ.χ.). *What Is Metadata: Definition and Meaning*. Ανάκτηση από Avast Academy: https://www.avast.com/c-what-is-metadata

awario. (χ.χ.). *All social media analytics on one dashboard*. Ανάκτηση από www.awario.com: https://awario.com/social-listening-dashboards/

Axians. (χ.χ.). *The development of video analytics for the smart city*. Ανάκτηση από www.axians.com/: https://www.axians.com/news/development-video-analytics-smart-city/

Big data: The next frontier for innovation, competition, and productivity. (2011, 05). *McKinsey Global Institute*.

bigdatapath. (χ.χ.). *Big Data Path* . Ανάκτηση από bigdatapath.wordpress.com/: https://bigdatapath.wordpress.com/2019/05/27/what-is-nosql-nosql-features-types-what-is-advantages/

briefcam. (χ.χ.). *VIDEO ANALYTICS FOR HEALTHCARE*. Ανάκτηση από www.briefcam.com/: https://www.briefcam.com/who-we-serve/healthcare-hospitals/

Britannica, T. E. (2022, 09 09). *TCP-IP*. Ανάκτηση από www.britannica.com/: https://www.britannica.com/technology/TCP-IP

Burns, E. (χ.χ.). *Definition machine learning*. Ανάκτηση από TechTarget: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

C. L. Philip ChenChun, C.-Y. Z. (2014, 08). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, σσ. 314-347.

C.L. Philip Chen, C.-Y. Z. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 314-347.

CERN. (2013, 09 18). *Attacking CERN's big data problem*. Ανάκτηση από www.old.gigaom.com/: https://old.gigaom.com/2013/09/18/attacking-cerns-big-data-problem/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+OmMalik+%28GigaOM%3A+Tech%29

CFI Team. (2022, 10 06). *Bar Charts (Data Visualization and Technical Analysis)*. Ανάκτηση από www.corporatefinanceinstitute.com: https://corporatefinanceinstitute.com/resources/business-intelligence/bar-charts-data-visualization-and-technical-analysis/

Christin Seifert, V. S. (2014, 01). Visual Analysis and Knowledge Discovery for Text.

Data Flair. (χ.χ.). *Data Mining Algorithms – 13 Algorithms Used in Data Mining*. Ανάκτηση από data-flair.training/: https://data-flair.training/blogs/data-mining-algorithms/

Data Mining vs Big Data . (χ.χ.). Ανάκτηση από www.javatpoint.com: https://www.javatpoint.com/data-mining-vs-big-data

dbpedia. (χ.χ.). *About: InfiniteGraph*. Ανάκτηση από www.dbpedia.org: https://dbpedia.org/page/InfiniteGraph

Deepashree Karanjkar, K. B. (2019, 12 02). NOSQL OVER RDBMS IN IMAGE STORING USING MONGODB. *NCRD's Technical Review*, 4. Ανάκτηση από www.guru99.com: http://ncrdsims.edu.in/site/views/pdfs/technical%20review%202019/23.%20NOSQL-OVER-RDBMS-IN-IMAGE-STORING-USING-MONGODB-Deepashree-K-Kanchan-B-Prof-Mrunali-M.pdf

Dialani, P. (2020, 10 29). *The Future of Data Revolution will be Unstructured Data*. Ανάκτηση από Analytics Insight.: https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/

Ekaterina Olshannikova, A. O. (2015, 06). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data, 2*(1).

Evgeniy Yur'evich Gorodov, V. V. (2013, 01). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering, 4*, σσ. 1-7.

Excel Dashboard School. (2022, 04 14). *Sunburst Chart*. Ανάκτηση από www.exceldashboardschool.com: https://exceldashboardschool.com/creating-sunburst-chart/

Feldman R., S. J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridgε: Cambridge University Press.

Financesonline.com. (2022, 11 04). Ανάκτηση από Finances Online: https://financesonline.com/how-much-data-is-created-every-day/
Fowler, M. (2012, 01 09). *NosqlDefinition*. Ανάκτηση από www.martinfowler.com: https://martinfowler.com/bliki/NosqlDefinition.html
Geeksforgeeks. (2022, 02 23). *How Neural Networks Can Be Used For Data Mining?* Ανάκτηση από www.geeksforgeeks.org: https://www.geeksforgeeks.org/how-neural-networks-can-be-used-for-data-mining/
Geeksforgeeks. (2022, 07 08). *Types of Sources of Data in Data Mining*. Ανάκτηση από www.geeksforgeeks.org/: https://www.geeksforgeeks.org/types-of-sources-of-data-in-data-mining/?ref=rp
Gonçalves, A. (2017). *SOCIAL MEDIA ANALYTICS STRATEGY - Using Data to Optimize Business Performance.* Las Vegas,: Apress.
google. (χ.χ.). *What is a Data Lake?* Ανάκτηση από google cloud: https://cloud.google.com/learn/what-is-a-data-lake
GreatLearning. (χ.χ.). *www.mygreatlearning.com*. Ανάκτηση από Features of NoSQL: https://www.mygreatlearning.com/no-sql/tutorials/features-of-nosql
Greenplum Database. (χ.χ.). *The World's First Open Source Massively Parallel Data Warehouse*. Ανάκτηση από www.greenplum.org/: https://greenplum.org/769-2/
guru99. (2022, 10 08). *NoSQL Tutorial: What is, Types of NoSQL Databases & Example*. Ανάκτηση από guru99: https://www.guru99.com/nosql-tutorial.html
Halper, F. (2018, 09 25). *3 Use Cases for Unstructured Data*. Ανάκτηση από www.tdwi.org: https://tdwi.org/articles/2018/09/25/data-all-3-use-cases-unstructured-data.aspx
Harrison, G. (2012). *Next Generation Databases NoSQL, NewSQL, and Big Data.* New York: apress.
https://database.guide. (2016, 06 20). *What does ACID mean in Database Systems?* Ανάκτηση από database.guide: https://database.guide/what-is-acid-in-databases/
HyperTable Inc. (χ.χ.). *OVERVIEW* . Ανάκτηση από hypertable.com: https://hypertable.com/documentation/
IBM. (2021, 03 01). *Classification*. Ανάκτηση από www.ibm.com/: https://www.ibm.com/docs/en/db2/10.1.0?topic=algorithms-classification
IBM Cloud Education. (2021, 06 29). *Structured vs. Unstructured Data: What's the Difference?* Ανάκτηση από IBM: https://www.ibm.com/cloud/blog/structured-vs-unstructured-data
IBM. (χ.χ.). *What is a chatbot?* Ανάκτηση από IBM: https://www.ibm.com/topics/chatbots
IBM. (χ.χ.). *What is Apache Hadoop?* Ανάκτηση από IBM: https://www.ibm.com/analytics/hadoop
ibm. (χ.χ.). *What is HDFS?* Ανάκτηση από IBM: https://www.ibm.com/topics/hdfs
ibm. (χ.χ.). *What is MapReduce?* Ανάκτηση από IBM: https://www.ibm.com/topics/mapreduce
IBM. (χ.χ.). *What is the k-nearest neighbors algorithm?* Ανάκτηση από www.ibm.com: https://www.ibm.com/topics/knn
Intel IT Center. (2013). *Big Data Visualization: Turning Big Data Into Big Insights*. Ανάκτηση από www.intel.com: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-visualization-turning-big-data-into-big-insights.pdf

Intone. (2022, 05 13). *Big data analytics in the pharmaceutical industry How is Big data analytics revolutionizing the pharma industry?* Ανάκτηση από www.intone.com: https://intone.com/big-data-analytics-in-pharmaceutical-industry/

J. Gantz, D. R. (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.

J. Wang, J. L. (2017). The Application of Data Mining Technology to Big Data. *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC*, *2*, σσ. 284-288.

Jiawei Han, M. K. (2011). *Data Mining: Concepts and Techniques.* Morgan Kaufmann; 3rd edition (.

Jiménez, G. F. (2021, 08 18). *Data Visualization: Pie Chart*. Ανάκτηση από www.datasketch.co: https://www.datasketch.co/blog/data-visualization-pie-chart/

Kalpesh Adhatrao, A. G. (2013, 09). Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 3*(5).

Khalid Adam Ismail Hammad, M. A. (2015). Big Data Analysis and Storage. *Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering.* Orlando, Florida, US.

Khan, M. L. (2020). Big Data and Entrepreneurship. *Handbook of Media Management and Business*.

Kumar, S. (2016). A Review of Recent Trends and Issues in Visualization. *International Journal on Computer Science and Engineering (IJCSE)*, 41.

Leong, N. (2017, 12 09). *Python For Data Science — A Guide to Data Visualization with Plotly*. Ανάκτηση από www.towardsdatascience.com/: https://towardsdatascience.com/python-for-data-science-a-guide-to-data-visualization-with-plotly-969a59997d0c

Lidong Wang, G. W. (2015, 07 22). Big Data and Visualization: Methods, Challenges and Technology Progress.

Longzhi Yang, J. L. (2019). Towards Big data Governance in Cybersecurity. *Data-Enabled Discovery and Applications*, 3.

Luigi Mario De Luca, D. H. (2020, 08). How and when do big data investments pay off? The role of marketing affordances and service innovation. *Journal of the Academy of Marketing Science , 49*(1).

Marr, B. (2016, 09 12). *Big Data: The Mind-Blowing Potential Of Voice Analytics*. Ανάκτηση από www.linkedin.com: https://www.linkedin.com/pulse/big-data-mind-blowing-potential-voice-analytics-bernard-marr

Master’s in Data Science. (χ.χ.). *Introduction to Machine Learning Algorithms*. Ανάκτηση από www.mastersindatascience.org: https://www.mastersindatascience.org/learning/machine-learning-algorithms/

Mathur, V. (2022, 08 05). *Big Data In Banking Industry: Benefits, Uses and Challenges*. Ανάκτηση από www.analyticssteps.com: https://www.analyticssteps.com/blogs/big-data-banking-industry-benefits-uses-and-challenges

Merlin Packiam, S. J. (2015). An empirical study on text analytics in big data. *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).*

Microsoft. (2016, 06 23). *Excel and big data*. Ανάκτηση από www.microsoft.com: https://www.microsoft.com/en-us/microsoft-365/blog/2016/06/23/excel-and-big-data/
Microsoft. (2022, 08 11). *Introduction to Azure Blob Storage*. Ανάκτηση από learn.microsoft.com: https://learn.microsoft.com/en-us/azure/storage/blobs/storage-blobs-introduction
microsoft. (2022, 11 19). *Transactions (Transact-SQL)*. Ανάκτηση από learn.microsoft.com: https://learn.microsoft.com/en-us/sql/t-sql/language-elements/transactions-transact-sql?view=sql-server-ver16
Microsoft. (χ.χ.). *Create a treemap chart in Office*. Ανάκτηση από www.support.microsoft.com: https://support.microsoft.com/en-us/office/create-a-treemap-chart-in-office-dfe86d28-a610-4ef5-9b30-362d5c624b68
Microsoft. (χ.χ.). *What is Power BI?* Ανάκτηση από www.microsoft.com: https://powerbi.microsoft.com/en-us/what-is-power-bi/
MongoDB. (χ.χ.). *What is NoSQL?* Ανάκτηση από MongoDB: https://www.mongodb.com/nosql-explained
Motashim Rasool, W. K. (2015, 04 15). Big Data: Study in Structured and Unstructured Data. *HCTL Open International Journal of Technology Innovations and Research (IJTIR)*.
Muzammil Khan, S. S. (2011, 12). Data and Information Visualization Methods, and Interactive Mechanisms: A Survey. *International Journal of Computer Applications, 34*(1).
Naveen, J. (2017, 11 26). *Top 5 sources of big data*. Ανάκτηση από Allerin: https://www.allerin.com/blog/top-5-sources-of-big-data
Nealon, C. (2018, 04 02). *App for early autism detection launched on World Autism Awareness Day, April 2*. Ανάκτηση από www.buffalo.edu/: https://www.buffalo.edu/news/releases/2018/04/001.html
NoSQL. (χ.χ.). *Your Ultimate Guide to the Non-Relational Universe!* Ανάκτηση από NoSQL: http://nosql-database.org
Oosthuizen, R. M. (2022, 07 06). The Fourth Industrial Revolution – Smart Technology, Artificial Intelligence, Robotics and Algorithms: Industrial Psychologists in Future Workplaces. *Frontiers in Artificial Intelligence* .
oreilly. (2004, 06). *Mastering Oracle SQL, 2nd Edition by Sanjay Mishra, Alan Beaulieu*. Ανάκτηση από www.oreilly.com/: https://www.oreilly.com/library/view/mastering-oracle-sql/0596006322/ch03.html
OrientDB LTD. (χ.χ.). *OrientDB*. Ανάκτηση από https://orientdb.org/: https://dbdb.io/db/orientdb
Patil, H. A. (2010, 01 01). “Cry Baby”: Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant’s Cry. *Advances in Speech Recognition*, σσ. 323–348.
Pedamkar, P. (χ.χ.). *Statistical Analysis Types*. Ανάκτηση από www.educba.com/: https://www.educba.com/statistical-analysis-types/
Petar Ristoski, H. P. (2016). Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey. *Journal of Web Semantics, 36*, σσ. 1-22.
Project Pro. (2022, 09 14). *Healthcare Big Data Projects, Applications and Examples*. Ανάκτηση από Project Pro: https://www.projectpro.io/article/5-healthcare-applications-of-hadoop-and-big-data/85

ProjectPro. (2022, 09 22). *How Big Data Analysis helped increase Walmarts Sales turnover?* Ανάκτηση από www.projectpro.io/: https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

Rao, P. (2021, 05 06). *How Rolls Royce uses Big Data?* Ανάκτηση από www.poonamrao.medium.com: https://poonamrao.medium.com/rolls-royce-big-data-use-case-6a090db2aa

Revfine. (χ.χ.). *5 Ways Big Data Can Benefit the Travel Industry*. Ανάκτηση από www.revfine.com: https://www.revfine.com/big-data-travel-industry/

Rob Kitchin, G. M. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society January*, 1-10.

Rouse, M. (2015, 06 06). *PP database (massively parallel processing database)*. Ανάκτηση από www.techtarget.com: http://searchdatamanagement.techtarget.com/.

S. Bansal, D. R. (2014, 01). Transitioning From Relational Database to Big Data. *Engineering, International Journal of Advanced Research in Computer Science and Software*, *4*(1), σ. 628.

SAS. (χ.χ.). *Data Visualization Techniques*. Ανάκτηση από www.sas.com/: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-visualization-techniques-106006.pdf

SAS. (χ.χ.). *SAS*. Ανάκτηση από Big Data: What it is and why it matters: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Scheyer, M. (2021, 05 20). *Real-time video analytics and the road to Vision Zero*. Ανάκτηση από www.cities-today.com: https://cities-today.com/industry/real-time-video-analytics-and-the-road-to-vision-zero/

Schwab, K. (2016, 14 01). *The Fourth Industrial Revolution: what it means, how to respond*. Ανάκτηση από www.weforum.org: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Scylladb. (χ.χ.). *Introduction to DynamoDB*. Ανάκτηση από www.scylladb.com: https://www.scylladb.com/learn/dynamodb/introduction-to-dynamodb/

Sharma, R. (2021, 01 8). *12 Most Useful Data Mining Applications of 2023*. Ανάκτηση από www.upgrad.com/: https://www.upgrad.com/blog/12-most-useful-data-mining-applications/

Shukla, A. (2014, 07 29). *Why traditional database systems fail to support "big data"*. Ανάκτηση από yahoo finance: https://finance.yahoo.com/news/why-traditional-database-systems-fail-210014958.html

Sinha, T. (2021, 04 07). *Cloud Data Lake vs. Data Warehouse vs. Data Mart*. Ανάκτηση από IBM: https://www.ibm.com/cloud/blog/cloud-data-lake-vs-data-warehouse-vs-data-mart

Smallcombe, M. (2022, 01 03). *Structured vs Unstructured Data: 5 Key Differences*. Ανάκτηση από intergate: https://www.integrate.io/blog/structured-vs-unstructured-data-key-differences/#four

SmartDraw. (χ.χ.). *Line Graph*. Ανάκτηση από www.smartdraw.com: https://www.smartdraw.com/line-graph/

Suharjito, S. (2018, 01). Implementation of Database Massively Parallel Processing System to Build Scalability on Process Data Warehouse. *Procedia Computer Science ·*, 69.

Tableau. (χ.χ.). *Why choose Tableau?* Ανάκτηση από www.tableau.com: https://www.tableau.com/why-tableau

Tempini, N. (2017). Till data do us part: Understanding data-based value creation in data-intensive infrastructures. *Information and Organization*, 191-210.
Tibco. (χ.χ.). *What is a Bubble Chart?* Ανάκτηση από www.tibco.com: https://www.tibco.com/reference-center/what-is-a-bubble-chart
Turing. (χ.χ.). *An Introduction to Naive Bayes Algorithm for Beginners*. Ανάκτηση από www.turing.com: https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners#probability,-bayes-theory,-and-conditional-probability
tutorialspoint. (χ.χ.). *Neo4j - Quick Guide*. Ανάκτηση από www.tutorialspoint.com: https://www.tutorialspoint.com/neo4j/neo4j_quick_guide.htm
Tzu-Wei Hsu, L. I. (2004). MonkEllipse: Visualizing the History of Information Visualization. *IEEE Symposium on Information Visualization*.
V, K. (2021, 04 14). *Artificial Neural Network, Its inspiration and the Working Mechanism*. Ανάκτηση από www.analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2021/04/artificial-neural-network-its-inspiration-and-the-working-mechanism/
V. Gnanaprakash, N. K. (2021). Automatic number plate recognition using deep learning. *IOP Conference Series: Materials Science and Engineering, 1084*.
w3schools. (χ.χ.). *HTML Unicode (UTF-8) Reference*. Ανάκτηση από www.w3schools.com: https://www.w3schools.com/charsets/ref_html_utf8.asp
Wei-SenChen, Y.-K. (2009, 03). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications, 36*(2), σσ. 4075-4086.
www.geeksforgeeks.org. (2021, 11 20). *What is Semi-structured data?* Ανάκτηση από geeksforgeeks: https://www.geeksforgeeks.org/what-is-semi-structured-data/
www.google.com. (χ.χ.). *Cloud Bigtable*. Ανάκτηση από cloud.google.com: https://cloud.google.com/bigtable
www.javatpoint.com. (χ.χ.). *Database Schema*. Ανάκτηση από javatpoint: https://www.javatpoint.com/database-schema
www.json.org. (χ.χ.). *Introducing JSON*. Ανάκτηση από JSON: https://www.json.org/json-en.html
www.mathworks.com. (χ.χ.). *What Is Deep Learning?* Ανάκτηση από Mathworks: https://www.mathworks.com/discovery/deep-learning.html
www.tutorialspoint.com. (χ.χ.). *SQL - Overview*. Ανάκτηση από tutorialspoint: https://www.tutorialspoint.com/sql/sql-overview.htm