Πολυτεχνείο Κρήτης

Τμήμα Μηχανικών Παραγωγής και Διοίκησης
Μεταπτυχιακό Πρόγραμμα Σπουδών
Επιχειρησιακή Έρευνα

Επιβλέπων: Καθηγητής Ματσατσίνης Νικόλαος

Τίτλος
Αναλυτική Μεγάλων Όγκων Δεδομένων: Επισκόπηση και Οδηγός Εφαρμογής στις Επιχειρήσεις
Big Data Analytics: Overview and Business Implementation Guide

Μεταπτυχιακή Διατριβή

Ειρήνη Αναστασιάδου

Α.Μ 2016019011

Χανιά, Οκτώβριος 2019

Ευχαριστίες

Με το πέρας της μεταπτυχιακής μου διατριβής θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή κ. Νικόλαο Ματσατσίνη αρχικά για την εμπιστοσύνη που μου έδειξε με την ανάθεση της διπλωματικής εργασίας, αλλά κυρίως για την πολύτιμη καθοδήγηση, επιμονή και υποστήριξη του, η οποία ήταν καθοριστικής σημασίας για την έκβαση της.

Επιπλέον, αισθάνομαι την ανάγκη να ευχαριστήσω την οικογένεια μου και όλους τους δικούς μου ανθρώπους που με τη στήριξη και την υπομονή τους με βοήθησαν για την υλοποίηση του πρώτου στόχου μου στην ακαδημαϊκή μου πορεία.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Περίληψη

Η έκρηξη του όγκου των δεδομένων που παρατηρείται τα τελευταία χρόνια οδηγεί στην ανάπτυξη νέων μεθόδων Ανάλυση Μεγάλων Όγκων Δεδομένων με στόχο την ανακάλυψη γνώσης και την αξιοποίησή της με στόχο την βελτίωση της λειτουργίας των επιχειρήσεων. Πιο συγκεκριμένα ο όρος «Μεγάλα Δεδομένα» χαρακτηρίζεται από τεράστιο όγκο, ποικιλία και συνεχή συσσώρευση νέων δεδομένων. Πλέον τα δεδομένα προέρχονται από πολλές πηγές και παρέχονται σε διαφορετικές μορφές. Συνεπώς αναδύεται η ανάγκη για διαχείριση του όγκου δεδομένων και όχι απλά ο υπολογισμός αυτού.

Στα πλαίσια της μεταπτυχιακής μου διατριβής αρχικά γίνεται καταγραφή του θεωρητικού πλαισίου στα πρώτα τρία κεφάλαια. Στο πρώτο κεφάλαιο δίνεται ο ορισμός των Μεγάλων Δεδομένων τα χαρακτηριστικά τους, η ιστορική αναδρομή και η βιβλιογραφική επισκόπηση. Επίσης θεωρήθηκε σημαντικό να γίνει αποσαφήνιση όρων που συχνά μπερδεύονται μεταξύ τους, όπως η Ανάλυση Δεδομένων και η Αναλυτική. Στο δεύτερο κεφάλαιο αναλύονται τα χαρακτηριστικά των βάσεων δεδομένων και στο τρίτο κεφάλαιο η Αρχιτεκτονική των Big Data. Σε αυτό το κεφάλαιο αναφέρονται οι τρόποι αποθήκευσης δεδομένων και η πλατφόρμα Hadoop.

Στο τέταρτο κεφάλαιο της παρούσας διατριβής θα παρουσιαστούν οι τεχνικές ανάλυσης και στο επόμενο κεφάλαιο οι διαδικασίες διακυβέρνησης και διαχείρισης δεδομένων. Στα κεφάλαια έξι και εφτά παρουσιάζονται τα μεγάλα δεδομένα ως μέσο στρατηγικής και ανταγωνιστικότητας και οι ρόλοι των ομάδων που συμμετέχουν στο έργο αντίστοιχα. Στο όγδοο κεφάλαιο γίνεται λόγος για το πως η Τεχνητή Νοημοσύνη μπορεί να εφαρμοστεί σε έναν οργανισμό και να επιφέρει Βαθιά Εκμάθηση. Τέλος παρουσιάζεται μια μελέτη περίπτωσης.

1 Εισαγωγή

1.1 Ορισμός

Αν και ο όρος "μεγάλα δεδομένα" έχει γίνει δημοφιλής, δεν υπάρχει γενική συναίνεση για το τι πραγματικά σημαίνει. Συχνά, πολλοί επαγγελματίες αναλυτές δεδομένων θα υπονοούσαν τη διαδικασία εξόρυξης, μετασχηματισμού, και φόρτωσης (Extraction Transformation Load) για μεγάλα σύνολα δεδομένων με το χαρακτηρισμό Μεγάλα Δεδομένα. Μια δημοφιλής περιγραφή των Big Data βασίζεται σε τρία βασικά χαρακτηριστικά των δεδομένων: όγκος, ταχύτητα και ποικιλία (Volume, Velocity, Variety ή 3Vs). Παρ 'όλα αυτά, δεν συλλαμβάνει με ακρίβεια όλες τις πτυχές του Big Data. Για να δοθεί μια ολοκληρωμένη έννοια των Μεγάλων Δεδομένων, θα διερευνηθεί αυτός ο όρος από ιστορική άποψη και θα εξεταστεί η εξέλιξη του από παλαιότερα χρόνια σε σχέση με σημερινές ερμηνείες (Buyya, Calheiros and Dastjerdi 2016).

Ιστορικά, ο όρος Big Data είναι αρκετά ασαφής και κακώς καθορισμένος. Δεν είναι ακριβής όρος και δεν έχει ιδιαίτερη σημασία εκτός από την έννοια του μεγέθους της. Η λέξη "μεγάλη" είναι πολύ γενική. το ερώτημα πόσο "μεγάλο" είναι μεγάλο και πόσο "μικρό" είναι μικρό είναι σε σχέση με το χρόνο, το διάστημα και την περίσταση. Από μια εξελικτική προοπτική, το μέγεθος του όρου "Big Data" είναι πάντα εξελισσόμενο. Αν χρησιμοποιούμε την τρέχουσα παγκόσμια χωρητικότητα κίνησης του Internet ως αναφορά μέτρησης, η σημασία του όγκου Big Data θα ήταν μεταξύ του εύρους terabyte (TB ή 10^{12} ή 2^{40}) και zettabyte (ZB ή 10^{21} ή 2^{70}). Βάσει των ιστορικών δεδομένων ο ρυθμός αύξησης της κυκλοφορίας, έχει εισέλθει στην εποχή του ZB το 2015, σύμφωνα με τη Cisco. Με σκοπό να γίνει κατανοητή η σημασία του αντίκτυπου του όγκου δεδομένων, παρουσιάζεται το μέσο μέγεθος των διαφόρων αρχείων δεδομένων που εμφανίζονται στον Πίνακας 1 (Buyya, Calheiros and Dastjerdi 2016).

Πίνακας 1: Ορισμοί Μέγεθος διαφορετικών αρχείων δεδομένων (Buyya, Calheiros and Dastjerdi 2016)

Μέσο	*Μέσος όρος μεγέθους αρχείων δεδομένων*	*Σημειώσεις (2014)*
Ιστοσελίδα	1,6-2 MB	Συνήθως 100 αντικείμενα

eBook	1-5 MB	200-350 σελίδες
Τραγούδια	3,5-5,8 MB	Συνήθως 1,9 MB/λεπτό
Ταινίες	100-120 GB	60 καρέ το λεπτό

Ο κύριος στόχος αυτού του κεφαλαίου είναι να παράσχει μια ιστορική άποψη για τα Big Data και να υποστηρίξει ότι δεν είναι μόνο 3Vs, αλλά μάλλον 3^2Vs ή 9Vs. Αυτά τα πρόσθετα χαρακτηριστικά μεγάλων δεδομένων αντικατοπτρίζουν το πραγματικό κίνητρο πίσω από την Αναλυτική Μεγάλων Δεδομένων (Big Data Analytics, BDA). Θεωρείται ότι αυτές οι διευρυμένες λειτουργίες διευκρινίζουν ορισμένα βασικά ερωτήματα σχετικά με την ουσία της BDA: ποια προβλήματα μπορεί να αντιμετωπίσει το Big Data και ποια προβλήματα δεν πρέπει να συγχέονται ως BDA. Αυτά τα θέματα καλύπτονται στο κεφάλαιο μέσω της ανάλυσης των ιστορικών εξελίξεων με τις σχετικές τεχνολογίες που υποστηρίζουν την επεξεργασία μεγάλων δεδομένων. (Buyya, Calheiros and Dastjerdi 2016):

1.2 Αξία των Big Data

Ο πρωταρχικός λόγος για τον οποίο τα Big Data αναπτύχθηκαν γρήγορα τα τελευταία χρόνια οφείλεται στο γεγονός ότι παρέχουν μακροπρόθεσμη επιχειρηματική αξία. Η αξία συλλαμβάνεται και από την άποψη της άμεσης κοινωνικού ή χρηματικού κέρδους και με τη μορφή ενός στρατηγικού ανταγωνιστικού πλεονεκτήματος. Λόγω του ευρέος φάσματος εφαρμογών της, τα Big Data αγκαλιάζονται από όλους τους τύπους βιομηχανιών, που κυμαίνονται από την υγειονομική περίθαλψη, τη χρηματοδότηση και την ασφάλιση, στον ακαδημαϊκό και μη κερδοσκοπικό τομέα (Big Data Framework 2018).

Υπάρχουν διάφοροι τρόποι με τους οποίους η αξία μπορεί να ληφθεί μέσω των μεγάλων δεδομένων και ο τρόπος που λειτουργούν οι επιχειρήσεις μπορεί να διευκολύνει την ανάπτυξη ή να την κάνουν πιο αποτελεσματική. Καθένα από αυτά οδηγεί στον ψηφιακό μετασχηματισμό των οργανισμών και έχουν μακροπρόθεσμο αντίκτυπο στον τρόπο με τον οποίο θα δραστηριοποιηθούν

οι επιχειρήσεις, τον τρόπο που σχεδιάζονται, οργανώνονται και να διαχειρίζονται (Big Data Framework 2018).

Οι επιχειρήσεις μπορούν να λάβουν αξία από τα μεγάλα δεδομένα με έναν από τους ακόλουθους πέντε τρόπους (Big Data Framework 2018):

1) Δημιουργία διαφάνειας. Χρησιμοποιώντας τα δεδομένα ενός οργανισμού για τον προσδιορισμό των μελλοντικών αποφάσεων κάνει μια οργάνωση ολοένα και πιο διαφανή και σπάει τα φράγματα μεταξύ στα διαφορετικά τμήματα. Τα μεγάλα δεδομένα αναλύονται σε διαφορετικά όρια και μπορούν να αναγνωρίσουν διάφορες ανεπάρκειες. Στους κατασκευαστικούς οργανισμούς, για παράδειγμα, τα Big Data μπορούν να εντοπίζουν ευκαιρίες βελτίωσης σε όλο το τμήμα Ε & Α, τη μηχανική και την παραγωγή με σκοπό την ταχύτερη διάθεση νέων προϊόντων στην αγορά (Big Data Framework 2018).
2) Ανακάλυψη με γνώμονα τα δεδομένα. Καθώς οι επιχειρήσεις δημιουργούν και αποθηκεύουν όλο και περισσότερες συναλλαγές δεδομένων σε ψηφιακές μορφές, θα είναι διαθέσιμα περισσότερα δεδομένα απόδοσης. Τα Μεγάλα δεδομένα μπορούν να παρέχουν τεράστιες νέες ιδέες που ίσως δεν είχαν εντοπιστεί προηγουμένως από την εύρεση μοτίβων ή τάσεων στα σύνολα δεδομένων. Στον ασφαλιστικό κλάδο, για παράδειγμα, τα Big Data μπορεί να βοηθήσουν να προσδιοριστούν τα επικερδή προϊόντα και να παρέχουν βελτιωμένους τρόπους υπολογισμού των ασφάλιστρων (Big Data Framework 2018).
3) Τμηματοποίηση και προσαρμογή. Η ανάλυση της Μεγάλης Κλίμακας Δεδομένων παρέχει μια βελτιωμένη ευκαιρία να προσαρμόσουν τις προσφορές της αγοράς προϊόντων σε συγκεκριμένα τμήματα πελατών προκειμένου να αυξηθούν τα έσοδα. Τα δεδομένα σχετικά με τη συμπεριφορά των χρηστών ή των πελατών καθιστούν δυνατή τη δημιουργία διαφορετικών προφίλ πελατών τα οποία μπορούν να στοχεύσουν ανάλογα. Συνδεδεμένοι λιανοπωλητές, για παράδειγμα, μπορούν να προσαρμόσουν την προσφορά προϊόντων στις ιστοσελίδες τους ώστε να ταιριάζουν με τον τρέχοντα πελάτη και να αυξήσουν τα ποσοστά μεταστροφής τους (Big Data Framework 2018).
4) Η δύναμη της αυτοματοποίησης. Οι υποκείμενοι αλγόριθμοι που αναλύουν σύνολα μεγάλων δεδομένων μπορούν να χρησιμοποιηθούν για να αντικαταστήσουν χειρωνακτικών αποφάσεων και αποφάσεις που απαιτούν πολλούς υπολογισμούς. Ο

αυτοματισμός μπορεί να βελτιστοποιήσει τις επιχειρηματικές διαδικασίες και να βελτιώσει την ακρίβεια ή τους χρόνους απόκρισης. Οι έμποροι λιανικής πώλησης, για παράδειγμα, μπορούν να εκμεταλλευτούν τους μεγάλους αλγόριθμους δεδομένων για τις αποφάσεις αγοράς ή να καθορίσουν τι απόθεμα θα προσφέρει ένα βέλτιστο ποσοστό απόδοσης (Big Data Framework 2018).

5) Καινοτομία και νέα προϊόντα. Τα Μεγάλα δεδομένα μπορούν να εντοπίσουν πρότυπα που προσδιορίζουν την ανάγκη για νέα προϊόντα ή να αυξήσουν το σχεδιασμό των σημερινών προϊόντων ή υπηρεσιών. Με την ανάλυση δεδομένων που έχουν αγοράσει, οι οργανισμοί μπορούν να εντοπίσουν τη ζήτηση για προϊόντα που η ίδια αγνοεί. Τα πανεπιστήμια ή τα κολέγια, για παράδειγμα, θα μπορούσαν να μελετήσουν την επισκεψιμότητα τους στον ιστότοπο και τους όγκους αναζήτησης για να προβλέψουν την ταξινόμηση και την κατανομή των μαθημάτων διδασκαλίας αναλόγως (Big Data Framework 2018).

Ποιο αναλυτικά η αξία των μεγάλων δεδομένων αφορά και άλλους τομείς όπως φαίνεται παρακάτω:

1.2.1 Σημασία στην εθνική ανάπτυξη

Την υφιστάμενη περίοδο ο κόσμος έχει εισέλθει στην εποχή της πληροφορίας. Η εκτεταμένη χρήση του διαδικτύου, του διαδικτύου των πραγμάτων (IoT), της υπολογιστικής Cloud και άλλων αναδυόμενων τεχνολογιών πληροφορικής έχει αυξήσει τα παραγόμενα δεδομένα με ένα πρωτοφανή βαθμό καθιστώντας όλο και πιο σύνθετες τις δομές και τους τύπους των δεδομένων. Μια ανάλυση σε βάθος για την αξιοποίηση των μεγάλων δεδομένων θα διαδραματίσει σημαντικό ρόλο στην προώθηση της διατήρησης της οικονομικής ανάπτυξης χωρών και την ενίσχυση ανταγωνιστικότητας των επιχειρήσεων (Jin, et al. 2015).

Στο μέλλον τα μεγάλα δεδομένα θα γίνουν ένα νέο σημείο οικονομικής ανάπτυξης. Με τα μεγάλα δεδομένα οι εταιρείες θα αναβαθμιστούν και θα μετατρέψουν τη λειτουργία του τρόπου Ανάλυσης ως Υπηρεσία (AaaS) αλλάζοντας έτσι το περιβάλλον της Τεχνολογίας Πληροφοριών και άλλων βιομηχανιών. Με αυτή την έννοια, παγκόσμιου κολοσσοί όπως οι IBM, Google, Microsoft και Oracle έχουν ήδη ξεκινήσει τον σχεδιασμό τεχνικής ανάπτυξης στην εποχή των μεγάλων δεδομένων (Jin, et al. 2015).

Σε εθνικό επίπεδο, η ικανότητα απόκτησης, επεξεργασίας και χρήσης τεράστιου όγκου δεδομένων θα αποτελέσει ένα ορόσημο δύναμης των κρατών. Τα δεδομένα της ύπατης εξουσίας μιας χώρας στον ψηφιακό κόσμο θα λαμβάνουν μια ακόμα θέση κυριαρχίας εκτός από την γη, τη θάλασσα και τον αέρα (Jin, et al. 2015).

Στην Κίνα, μια κυβερνητική αναφορά εισηγείται ότι ο κυβερνοχώρος καθώς και η θάλασσα και το διάστημα αποτελούν βασικά κλειδιά του πυρήνα των εθνικών συμφερόντων. Η καθυστέρηση με την έννοια της έλλειψης ενέργειας και πρωτοβουλίας στης αναλυτική των μεγάλων δεδομένων δεν καταλήγει μόνο σε απώλεια στρατηγικού πλεονεκτήματος αλλά και στην ανάπτυξη τρωτών σημείων στην ασφάλεια του διαδικτύου. Με αυτή την έννοια, η πρωτοβουλία "Big Data Research and Development Initiative" που ανακοινώθηκε 29 Μαρτίου του 2012, δεν είναι μόνο ένα στρατηγικό σχέδιο που προωθεί τις ΗΠΑ (Kalil 2012) σε τομείς υψηλής τεχνολογίας, αλλά και ένα σχέδιο για την προστασία της εθνικής ασφάλειας και προώθησης της κοινωνικοοικονομικής ανάπτυξης (Jin, et al. 2015).

Σε γενικές γραμμές οι δυτικές χώρες εκπροσωπούμενες από τις ΗΠΑ κινούνται στο πλαίσιο της εθνικής τους ατζέντας προς έναν εκσυγχρονισμό της εθνικής τους δύναμης μέσω της έρευνας μεγάλων δεδομένων και εφαρμογών. Αναμένεται ότι οι μελλοντικές οικονομικές και πολιτικές εντατικές προσπάθειες για επικράτηση μεταξύ χωρών θα βασίζεται στην εκμετάλλευση του δυναμικού μεγάλων δεδομένων, μεταξύ των παραδοσιακών προσανατολισμών. Εν ολίγοις, η έρευνα και οι εφαρμογές της αναλυτικής μεγάλων δεδομένων έχουν στρατηγική σημασία και βαρύτητα για την ενίσχυση της ανταγωνιστικότητας οποιασδήποτε χώρας (Jin, et al. 2015).

1.2.2 Σημασία στον τομέας της βιομηχανίας

Οι σύγχρονες επιχειρήσεις έρχονται αντιμέτωπες με νέες προκλήσεις, η ψηφιοποίηση και η μηχανογράφηση περιλαμβάνονται στο πλαίσιο των Μεγάλων δεδομένων. Η έρευνα για κοινά προβλήματα καθώς και η ανακάλυψη βασικών τεχνολογιών θα επιτρέψει στις βιομηχανίες να δαμάσουν την πολυπλοκότητα που προκαλείται από τη διασύνδεση δεδομένων και να αντιμετωπίσουν την αβεβαιότητα που προκαλείται από τον πλεονασμό ή έλλειψη δεδομένων. Τα στελέχη μιας εταιρίας ελπίζουν να βγάζουν στην επιφάνεια δεδομένα που σχετίζονται με τη ζήτηση, τη γνώση ακόμα και τη νοημοσύνη ώστε να εκμεταλλεύονται πλήρως την αξία των

μεγάλων δεδομένων. Αυτό αναπόφευκτα σημαίνει ότι τα δεδομένα δεν αποτελούν πλέον παραπροϊόν του βιομηχανικού τομέα, αλλά έχει γίνει κλειδί σύνδεσης όλων των τομέων. Με αυτή την έννοια, η μελέτη κοινών προβλημάτων και οι βασικές τεχνολογίες των μεγάλων δεδομένων θα αποτελέσουν το επίκεντρο της νέας γενιάς Πληροφορικής και των εφαρμογών του. Τα Μεγάλα δεδομένα δεν θα αποτελούν μόνο την κινητήρια δύναμη για υψηλή ανάπτυξη και ανταγωνιστικότητα, αλλά και για την ανάδειξη νέων εργαλείων για τις βιομηχανίες για τη βελτίωση της ανταγωνιστικότητας τους (Jin, et al. 2015).

Για παράδειγμα τα τελευταία χρόνια το υπολογιστικό νέφος (cloud computing) εξελίχθηκε αρχικά από μια ασαφή έννοια σε μια ώριμη και σύγχρονη τεχνολογία. Πολλές μεγάλες εταιρίες όπως οι Google, Microsoft, Amazon, Facebook, Alibaba, Baidu, Tencent και άλλοι γίγαντες της Πληροφορικής εργάζονται σε τεχνολογίες cloud computing και υπηρεσίες υπολογιστικής cloud-based. Τα Μεγάλα δεδομένα και το υπολογιστικό νέφος θεωρούνται δύο πλευρές ενός νομίσματος διότι τα μεγάλα δεδομένα είναι μια εφαρμογή του υπολογιστικού νέφους μάλιστα με μεγάλη επιτυχία και την ίδια στιγμή το υπολογιστικό νέφος παρέχει την υποδομή πληροφορικής για τα Μεγάλα δεδομένα. Οι στενοί σύνδεσμοι μεταξύ αυτών των δύο τεχνολογιών αναμένεται να αλλάξει το οικοσύστημα του Διαδικτύου και μάλιστα να επηρεάσει τον κλάδο της πληροφορίας (Jin, et al. 2015).

1.2.3 Σημασία στον τομέα της επιστημονικής έρευνας

Τα Μεγάλα δεδομένα έχουν ωθήσει την επιστημονική κοινότητα να επανεξετάσει τη μεθοδολογία της επιστημονικής έρευνας και πυροδότησε μια επανάσταση στην επιστημονική σκέψη και στις μεθόδους της. Είναι γνωστό ότι η επιστημονική μέθοδος βασίζεται στο πείραμα. Αργότερα αναδύθηκε η θεωρητική επιστήμη που χαρακτηρίστηκε από τη μελέτη νόμων και θεωρημάτων. Ωστόσο, επειδή η θεωρητική ανάλυση είναι πολύ περίπλοκη και δεν είναι εφικτή για την επίλυση πρακτικών προβλημάτων, η επιστημονική κοινότητα άρχισε να αναζητά μεθόδους προσομοίωσης που οδήγησαν στην υπολογιστική επιστήμη.

Η εμφάνιση των Μεγάλων δεδομένων δημιούργησε ένα νέο υπόδειγμα έρευνας, όπου με τα Μεγάλα δεδομένα οι ερευνητές θα πρέπει να βρούνε ή να προβούν στην εξόρυξη της επιθυμητής πληροφορίας, γνώσης και νοημοσύνης. Δεν χρειάζεται πλέον οι ερευνητές να έχουν άμεση

πρόσβαση στα αντικείμενα που εξετάζει. Το 2007 ο βραβευμένος με βραβείο Turing Jim Gray παρουσίασε στην ομιλία του το τέταρτο υπόδειγμα στην οποία και ξεχώρισε την επιστήμη έντασης δεδομένων από την υπολογιστική επιστήμη. Σήμερα με την εξερεύνηση των δεδομένων ενοποιείται η θεωρία με το πείραμα και την προσομοίωση. Σύμφωνα με τον Gray (2007) τα βήματα ενοποίησης είναι τα εξής:

- Σύλληψη δεδομένων από όργανα μετρήσεων ή από παραγωγή τους από προσομοίωση
- Επεξεργασία από κάποιο λογισμικό
- Αποθήκευση πληροφορίας/ γνώσης
- Ανάλυση βάσης δεδομένων/ αρχείων με τη χρήση διαχείρισης δεδομένων και στατιστικής

Ο Gray θεωρούσε ότι το τέταρτο υπόδειγμα ίσως είναι ο μόνος συστηματικός τρόπος για την επίλυση κάποιων από των πιο δύσκολων παγκοσμίων προκλήσεων.

1.2.4 Σημασία στην αναδυόμενη διεπιστημονική έρευνα

Μια αναδυόμενη διεπιστημονική αρχή, που ονομάζεται επιστήμη δεδομένων αποκτά σταδιακά ισχύ. Στοχεύει στην εξαγωγή γνώσης από δεδομένα απαιτεί τη χρήση των μεγάλων δεδομένων ως ερευνητικό αντικείμενο. Η επιστήμη δεδομένων εκτείνεται σε πολλούς κλάδους συμπεριλαμβανομένης της επιστήμης της πληροφορικής, των μαθηματικών, τις κοινωνικές επιστήμες, την επιστήμη των δικτύων, την ψυχολογία και τα οικονομικά. Χρησιμοποιεί διάφορες τεχνικές και θεωρίες από πολλά πεδία συμπεριλαμβανομένης της επεξεργασίας σημάτων, τη θεωρία πιθανοτήτων, μηχανική μάθηση, στατιστική, προγραμματισμός υπολογιστών, μηχανική δεδομένων, αναγνώριση προτύπων, οπτικοποίηση, ασαφή μοντέλα, την αποθήκευση δεδομένων και τον υπολογισμό υψηλής απόδοσης (Jin, et al. 2015).

Η ζήτηση για επιστήμονες δεδομένων (και παρόμοιοι τίτλοι εργασίας) έχει αυξηθεί σημαντικά και πολλοί άνθρωποι συμμετέχουν ενεργά στον τομέα της επιστήμης των δεδομένων (Big Data Framework 2018). Ως αποτέλεσμα, η γνώση και η εκπαίδευση σχετικά με την επιστήμη των δεδομένων έχει πολύ πιο επαγγελματική. Ενώ οι στατιστικές και ανάλυση δεδομένων ως επί το

πλείστων παρέμεινε προηγουμένως ένα ακαδημαϊκό πεδίο, γρήγορα γίνεται ένα δημοφιλές θέμα μεταξύ των της ακαδημαϊκής κοινότητας και του ενεργού πληθυσμού (Big Data Framework 2018).

Έχουν δημιουργηθεί πολλά ερευνητικά κέντρα/ινστιτούτα για τα μεγάλα δεδομένα τα τελευταία χρόνια σε διάφορα πανεπιστήμια σε όλο το κόσμο (όπως το University of California στο Berkeley, Columbia University, New York University, Tsinghua University, Eindhoven University of Technology, και Chinese University του Hong Kong). Πολλά πανεπιστήμια και ερευνητικά ιδρύματα έχουν ακόμη δημιουργήσει προπτυχιακά προγράμματα και / ή μεταπτυχιακά μαθήματα σχετικά με την ανάλυση δεδομένων, συμπεριλαμβανομένων των επιστημόνων δεδομένων και των μηχανικών δεδομένων (Jin, et al. 2015). Στην Ελλάδα λειτουργούν το 2018 τρία μεταπτυχιακά προγράμματα

- "Data Science" στο Διεθνές Πανεπιστήμιο στην Θεσσαλονίκη
- Advanced Software Engineering – Data Analytics στο Business College στην Αθήνα
- Business Analytics στο Οικονομικό Πανεπιστήμιο Αθηνών στην Αθήνα

1.2.5 Σημασία για την ενίσχυση αντίληψης του παρόντος

Τα μεγάλα δεδομένα ειδικά τα μεγάλα δικτυακά δεδομένα, περιέχουν ένα πλούτο από κοινωνικές πληροφορίες και έτσι μπορούν να θεωρηθούν ως ένα χαρτογραφημένο δίκτυο μιας κοινωνίας. Για το σκοπό αυτό, αναλύοντας μεγάλα δεδομένα και συνοψίζοντας περαιτέρω για την εύρεση ενδείξεων και νόμων που εμπεριέχει εμμέσως μπορεί να μας βοηθήσει καλύτερα να αντιληφθούμε την παρούσα κατάσταση.

Για παράδειγμα, αναπτύχθηκαν δύο παραδείγματα δεικτών ενδιαφέροντος στην Κίνα όπου και γίνεται μεγάλη χρήση των δεδομένων που είναι διαθέσιμα στο Διαδίκτυο. Από το 2007, το Κέντρο Έρευνας και Αξιολόγησης της Κίνας, συνδέεται με το Πανεπιστήμιο Renmin της Κίνας, έχει εκδώσει ετήσια "Ανάπτυξη της Κίνας Ευρετήριο ". Ο κατάλογος, με τέσσερις μεμονωμένους δείκτες για την υγεία, την εκπαίδευση, το βιοτικό επίπεδο και το κοινωνικό περιβάλλον, σκοπεύει να μετρήσει το status quo και να αποσαφηνίσει τα προβλήματα της ανάπτυξης της Κίνας. Παρέχει μια επιστημονική βάση για μια εύλογη μέτρηση της συνολικής ανάπτυξης της Κίνας. Ως μια άλλη προσπάθεια, από το 2010, το πρακτορείο ειδήσεων Xinhua, μαζί με το Dow Jones Newswires,

δημοσιεύουν δύο φορές το χρόνο "Διεθνή χρηματοοικονομικά κέντρα Xinhua-Dow Jones Δείκτης ανάπτυξης". Συγκρίνοντας και αναλύοντας διάφορους υποκειμενικούς και αντικειμενικούς δείκτες και συνδυάζοντας ποιοτικά και ποσοτικά κριτήρια με ποσοτική ανάλυση, ο δείκτης αυτός αποκαλύπτει την τρέχουσα εξέλιξη, το καθεστώς και τους νόμους των διεθνών χρηματοπιστωτικών κέντρων (Jin, et al. 2015).

Οι πληροφορίες εξόρυξης που περιέχονται σε μεγάλα δεδομένα μπορούν επίσης να βοηθήσουν τους ανθρώπους να λαμβάνουν καλύτερες αποφάσεις. Για παράδειγμα, στην προεδρική εκλογή των Ηνωμένων Πολιτειών τον Νοέμβριο του 2012, του Μπαράκ Ομπάμα η ομάδα καμπάνιας βοήθησε τον Ομπάμα αναλύοντας μεγάλα δεδομένα για να νικήσει τον Romney και να επανεκλεγεί. Τους δεκαοκτώ μήνες πριν την Ημέρα Εκλογών, η ομάδα ανάλυσης δεδομένων του Ομπάμα δημιούργησε τεράστια δεδομένα επεξεργασίας. Μέσω συλλογής και ανάλυσης δεδομένων σε πραγματικό χρόνο, όχι μόνο θα μπορούσε να πει στην ομάδα εκστρατείας πώς να βρει ψηφοφόρους και για να πάρουν την προσοχή τους, αλλά ανέλυσε επίσης την τάση των ψηφοφόρων να ψηφίσουν. Κάθε βράδυ, η ομάδα ανάλυσης δεδομένων διεξήγαγε προσομοίωση για την εκλογή και παρουσίαζε τα αποτελέσματα της προσομοίωσης την επόμενη μέρα για να βοηθήσουν στην κατανόηση της πιθανότητας να κερδίσει ο Ομπάμα σε κάποιες περιοχές, βάσει των οποίων η ομάδα μπορεί να διαθέσει τους πόρους με μεγαλύτερη ακρίβεια. Αργότερα τα γεγονότα έδειξαν ότι η ομάδα ανάλυσης δεδομένων έπαιξε ένα κρίσιμο ρόλο στην επανεκλογή του Ομπάμα, πολύ πέρα από τη φαντασία των ανθρώπων (Issenberg 2012) .

Η ανάλυση και η εξόρυξη μεγάλων δεδομένων μπορεί επίσης να διασφαλίσει αποτελεσματικά τη δημόσια ασφάλεια και την καταπολέμηση των εγκλημάτων. Για παράδειγμα, το 2012 η μεγάλη ανάλυση δεδομένων διαδραμάτισε σημαντικό ρόλο στην αποκάλυψη της ποινικής υπόθεσης του Zhou Kehua, ενός διαβόητου σειριακού δολοφόνου και ληστή στην Κίνα, ο οποίος πέθανε σε μια ανταλλαγή πυροβολισμών με την αστυνομία. Από τη σειρά των ένοπλων ληστειών και των ανθρωποκτονιών όπου ο Zhou ήταν ύποπτος, η αστυνομία διεξήγαγε μια ολοκληρωμένη εξέταση τεράστιας ποικιλίας των δεδομένων βίντεο και έλαβε με επιτυχία μια βιντεοκασέτα όπου ο ύποπτος αγόρασε πρωινό χωρίς καμουφλάζ. Μετά από αυτό, παρακολούθησαν τον Zhou στο Internet cafe όπου επισκέπτονταν τακτικά και με επιτυχία απέκτησαν δύο σαφείς κούπες από τον ύποπτο όταν είχαν πρόσβαση στο Διαδίκτυο. Σύμφωνα με τις προτιμήσεις του σε ιστοσελίδες που

σχετίζονται με το Sichuan και το Chongqing της Κίνα, η αστυνομία διαπίστωσε ότι ο ύποπτος ήταν από την περιοχή με διάλεκτο Sichuan. Βάσει της συνοπτικής ανάλυσης από διάφορες πληροφορίες που προέρχονται από μεγάλα δεδομένα, η αστυνομία συγκέντρωσε τα χαρακτηριστικά και τις ενέργειες του ύποπτου κατά τη διάπραξη των εγκλημάτων. Η ανάλυση αυτή διαδραμάτισε καθοριστικό ρόλο στην αστυνομία να αναπτύξει τις δυνάμεις της και τελικά συλλάβει τον Zhou (Wikipedia n.d.).

1. Σημασία στη βοήθεια των ανθρώπων να προβλέψουν καλύτερα το μέλλον

Μέσω της αποτελεσματικής προσαρμογής και της ακριβούς ανάλυσης σχετικά με την πολυπαραγοντική χρήση ετερογενών και μεγάλων δεδομένων, μπορεί να επιτευχθούν καλύτερες προβλέψεις για τις μελλοντικές τάσεις των γεγονότων. Είναι δυνατή η ανάλυση μεγάλων δεδομένων ακόμη και για την προωθήσει βιώσιμων εξελίξεων της κοινωνίας και της οικονομίας και για να δημιουργούν περαιτέρω νέες βιομηχανίες που σχετίζονται με τις υπηρεσίες δεδομένων (Jin et al. 2015).

Η ικανότητα των μεγάλων δεδομένων δικτύου έχει αναπτυχθεί πολύ και εφαρμόζεται αποτελεσματικά στον τομέα της ασφάλειας και του στρατού. Για παράδειγμα, ήδη από το 2010, οι Ηνωμένες Πολιτείες έδωσαν στη δημοσιότητα μια έκθεση με τίτλο «Κινεζική αποθήκευση πυρηνικών κεφαλών και συστημάτων χειρισμού» (Stokes 2010), που ισχυρίστηκε ότι οι ΗΠΑ βρήκαν πυρηνικές βάσεις στην Κίνα σε περιοχές όπως Shaanxi, Jiangxi, και Sichuan. Η έκθεση ακόμα παρουσίασε τα ονόματα των πόλεων και των κομητειών όπου βρίσκονταν οι πυρηνικές βάσεις. Αυτές οι αναφορές προκάλεσαν αίσθηση σε παγκόσμιο επίπεδο. Μέσα από αυτή την έκθεση, το Ινστιτούτο Έργων του 2049 των ΗΠΑ έλαβε την προσοχή του κοινού. Ιδρύθεν στην Ουάσιγκτον, DC, το 2008, το ινστιτούτο αυτό χρησιμοποιεί δημόσια διαθέσιμα δεδομένα και έγγραφα (όπως περιοδικά και έγγραφα συνεδρίου) για την ανάλυση και πρόβλεψη θεμάτων ασφάλειας στην Κίνα που σχετίζονται με το στρατό και την οικονομία της. Συμπλήρωσαν την αναφορά μέσω κάθετων αναζητήσεων και συστηματική ανάλυση μεγάλων δεδομένων. Τον Μάρτιο του 2013, το Ινστιτούτο δημοσίευσε επίσης μια έκθεση έρευνας σχετικά με τα Ανώνυμα Αέρια Οχήματα της Κίνας (UAV) στην οποία υπήρχε εκτενής ανάλυση της έρευνας, της ανάπτυξης, του εξοπλισμού και της λειτουργίας ανάπτυξη UAV στην Κίνα. Υπολόγισαν επίσης

ότι το μελλοντικό UAV της Κίνας θα είναι σε θέση, να εντοπίσει και να στοχεύσει Αμερικάνικα αεροπλανοφόρα (Easton and Hsiao 2013).

Η ανάλυση πρόβλεψης βασισμένη σε μεγάλα δεδομένα εφαρμόστηκε για να θίξει κοινωνικά θέματα, συμπεριλαμβανομένης της δημόσιας υγείας και της οικονομικής ανάπτυξης. Οι Ginsberg, et al. το 2009 διαπίστωσαν ότι, αν ο όγκος των αιτημάτων που υποβλήθηκαν στην Google με λέξεις-κλειδιά όπως "σύμπτωμα γρίπης" και "θεραπεία γρίπης" έχει αυξήθεί σε μια περιοχή, μετά από μερικές εβδομάδες, ο αριθμός των ασθενών με γρίπη στους χώρους έκτακτης ανάγκης των νοσοκομείων θα αυξηθεί ανάλογα. Με αυτή την ανακάλυψη, θα είναι σε θέση να προβλέψουν με καθυστέρηση μόνο μιας μέρας αντίμετρα εκ των προτέρων. Όσον αφορά την οικονομική ανάπτυξη, τα Ηνωμένα Έθνη ξεκίνησαν πρόσφατα ένα νέο έργο, αποκαλούμενο Global Pulse[5] το οποίο αναμένει να χρησιμοποιήσει μεγάλα δεδομένα για την προώθηση της ανάπτυξης της παγκόσμιας οικονομίας. Τα Ηνωμένα Έθνη θα διεξαγάγουν τη λεγόμενη συναισθηματική ανάλυση, η οποία χρησιμοποιεί τη λογισμικό επεξεργασίας του φυσικού λόγου για την ανάλυση μηνυμάτων κειμένου στα μέσα κοινωνικής δικτύωσης προκειμένου να προβλεφθούν κοινωνικά ζητήματα όπως το ποσοστό ανεργίας, περικοπές δαπανών και εκδηλώσεις νόσων σε μια δεδομένη περιοχή. Συνολικά σκοπός είναι να χρησιμοποιηθούν ψηφιακά σήματα έγκαιρης προειδοποίησης για να καθοδηγήσουν προγράμματα βοήθειας προκειμένου να αποφευχθεί η επανεμφάνιση φτώχειας σε μια περιοχή.

1.3 Κίνητρα για τη χρήση Μεγάλων Δεδομένων

Το Big Data προέκυψε την τελευταία δεκαετία από τον συνδυασμό επιχειρησιακών αναγκών και τεχνολογικές καινοτομίες. Ορισμένες εταιρείες που διαθέτουν μεγάλα δεδομένα στον πυρήνα της στρατηγικής τους έχουν γίνει πολύ επιτυχημένες στις αρχές του 21ου αιώνα. Διάσημα παραδείγματα περιλαμβάνουν την Apple, Amazon, Facebook και Netflix. (Big Data Framework 2018)

[5] https://www.unglobalpulse.org/projects/BigDataforDevelopment (15/11/2018)

Ένας αριθμός επιχειρηματικών οδηγών βρίσκονται στο επίκεντρο αυτής της επιτυχίας και εξηγούν γιατί το Big Data γρήγορα αυξήθηκε για να γίνει ένα από τα πιο πολυπόθητα θέματα στον κλάδο. Έξι κύριοι λόγοι μπορούν να εντοπιστούν:

1) Η ψηφιοποίηση της κοινωνίας.
2) Η μείωση του τεχνολογικού κόστους.
3) Συνδεσιμότητα μέσω του cloud computing;
4) Αυξημένη γνώση της επιστήμης των δεδομένων.
5) εφαρμογές κοινωνικών μέσων μαζικής ενημέρωσης;
6) Το επερχόμενο Internet-of-Things (IoT).

Σε αυτή την ενότητα, θα διερευνηθεί μια επισκόπηση υψηλού επιπέδου για κάθε έναν από αυτούς τους επιχειρηματικούς οδηγούς. Καθένα από τα παραπάνω αυξάνει το ανταγωνιστικό πλεονέκτημα των επιχειρήσεων δημιουργώντας νέες ροές εισοδήματος από την μείωση του λειτουργικό κόστος (Big Data Framework 2018).

1.3.1 Η ψηφιοποίηση της κοινωνίας

Τα μεγάλα δεδομένα σε μεγάλο βαθμό οδηγούνται από τον καταναλωτή και προσανατολίζονται στον καταναλωτή. Τα περισσότερα από τα δεδομένα στον κόσμο παράγονται από τους καταναλωτές, οι οποίοι είναι πάντα «online». Οι περισσότεροι άνθρωποι δαπανούν τώρα 4-6 ώρες κάθε ημέρα και παράγουν δεδομένα μέσω διάφορων συσκευών και (κοινωνικών) εφαρμογών. Με κάθε κλικ, φωτογραφία ή μήνυμα, δημιουργούνται νέα δεδομένα σε μια βάση δεδομένων κάπου σε όλο τον κόσμο. Επειδή όλοι έχουν τώρα ένα smartphone στην τσέπη τους, η δημιουργία δεδομένων συνοψίζεται σε ακατανόητα ποσά. Ορισμένες μελέτες εκτιμούν ότι το 60% των δεδομένων δημιουργήθηκε εντός των δύο τελευταίων χρόνων, γεγονός που αποτελεί καλή ένδειξη του ρυθμού με τον οποίο η κοινωνία έχει ψηφιοποιηθεί (Big Data Framework 2018)

1.3.2 Η μείωση του τεχνολογικού κόστους

Τεχνολογία που σχετίζεται με τη συλλογή και την επεξεργασία πολλών δεδομένων (υψηλής ποικιλίας) έχουν γίνει ολοένα και πιο προσιτά. Το κόστος αποθήκευσης δεδομένων και επεξεργαστών συνεχίζει να μειώνεται, επιτρέποντας στις μικρές επιχειρήσεις και τα άτομα να

εμπλακούν με τα Μεγάλα Δεδομένα. Για τη χωρητικότητα αποθήκευσης, ο συχνά αναφερόμενος νόμος του Moore εξακολουθεί να θεωρεί ότι η πυκνότητα αποθήκευσης (και ως εκ τούτου ικανότητα) διπλασιάζεται κάθε δύο χρόνια (Big Data Framework 2018).

Εκτός από την πτώση των δαπανών αποθήκευσης, ένας δεύτερος βασικός παράγοντας που συμβάλλει στην προσιτότητα των Big Data ήταν η ανάπτυξη ανοικτών λογισμικών. Το πιο δημοφιλές πλαίσιο λογισμικού (σήμερα θεωρείται το πρότυπο για τα Μεγάλα Δεδομένα) είναι Apache Hadoop για κατανεμημένη αποθήκευση και επεξεργασία. Λόγω της υψηλής διαθεσιμότητας αυτών των πλαισίων λογισμικού σε ανοιχτό επίπεδο, έχει γίνει όλο και πιο φθηνή η έναρξη των έργων Big Data σε οργανισμούς[1].

1.3.3 Συνδεσιμότητα μέσω του υπολογιστικού νέφους

Περιβάλλοντα υπολογιστών Cloud (όπου τα δεδομένα αποθηκεύονται εξ αποστάσεως σε κατανεμημένα συστήματα αποθήκευσης) έχουν καταστήσει δυνατή την ταχεία κλιμάκωση ή μείωση της υποδομής πληροφορικής και τη διευκόλυνση ενός μοντέλου pay -as-you-go. Αυτό σημαίνει ότι οι οργανισμοί που θέλουν να επεξεργαστούν τεράστιες ποσότητες δεδομένων (και συνεπώς έχουν μεγάλες απαιτήσεις αποθήκευσης και επεξεργασίας) δεν χρειάζεται να επενδύσουν σε μεγάλες ποσότητες υποδομής πληροφορικής. Αντί για αυτό, μπορούν να χορηγήσουν άδεια για την ικανότητα αποθήκευσης και επεξεργασίας κα χρειάζονται και πληρώνουν μόνο για τα ποσά που πραγματικά χρησιμοποίησαν. Ως αποτέλεσμα, τα περισσότερα από αυτές τις λύσεις μεγάλων δεδομένων αξιοποιούν τις δυνατότητες του cloud computing για την παροχή λύσεων στις επιχειρήσεις (Big Data Framework 2018).

1.3.4 Εφαρμογή κοινωνικών μέσων δικτύωσης

Όλοι κατανοούν τον αντίκτυπο που έχουν τα κοινωνικά μέσα στην καθημερινή ζωή. Ωστόσο, στη μελέτη της Μεγάλης Κλίμακας Δεδομένων, τα κοινωνικά μέσα ενημέρωσης διαδραματίζουν πρωταρχικό ρόλο. Όχι μόνο εξαιτίας της καθαρότητας των δεδομένων που παράγεται καθημερινά μέσω πλατφορμών όπως το Twitter, το Facebook, LinkedIn και Instagram, αλλά και επειδή τα κοινωνικά μέσα παρέχουν δεδομένα της ανθρώπινης συμπεριφοράς σε πραγματικό χρόνο. Τα δεδομένα των κοινωνικών μέσων παρέχουν πληροφορίες για τις συμπεριφορές, τις προτιμήσεις και τις απόψεις του κοινού σε μια κλίμακα που δεν ήταν ποτέ γνωστή πριν. Λόγω αυτού, είναι

εξαιρετικά πολύτιμες πληροφορίες για όποιον είναι σε θέση να αντλήσει νόημα από αυτές τις μεγάλες ποσότητες δεδομένων. Τα δεδομένα των κοινωνικών μέσων μπορούν να χρησιμοποιούνται για τον εντοπισμό των προτιμήσεων των πελατών για την ανάπτυξη προϊόντων, που στοχεύουν νέους πελάτες μελλοντικές αγορές, ή ακόμα και να στοχεύσουν πιθανούς ψηφοφόρους στις εκλογές. Τα δεδομένα των κοινωνικών μέσων ενδέχεται να είναι ένας από τους σημαντικότερους επιχειρηματικές κίνητρα των Big Data (Big Data Framework 2018).

1.3.5 Το επερχόμενο δίκτυο Internet of Things

Το Διαδίκτυο των πραγμάτων (IoT) είναι το δίκτυο των φυσικών συσκευών, οχημάτων, οικιακών συσκευών και αντικειμένων ενσωματωμένα με ηλεκτρονικά, λογισμικά, αισθητήρες, ενεργοποιητές και συνδέονται με το διαδίκτυο η οποία επιτρέπει σε αυτά τα αντικείμενα να συνδέονται και να ανταλλάσσουν δεδομένα. Πλέον οι κατασκευαστές καταναλωτικών αγαθών ξεκινούν να συμπεριλαμβάνουν αισθητήρες σε οικιακές συσκευές. Ενώ το μέσο νοικοκυριό το 2010 διέθετε περίπου 10 συσκευές που συνδέονται με το διαδίκτυο ο αριθμός αναμένεται να αυξηθεί στα 50 ανά νοικοκυριό έως το 2020. Παραδείγματα αυτών των συσκευών είναι θερμοστάτες, ανιχνευτές καπνού, τηλεοράσεις, ηχοσυστήματα και ακόμη και έξυπνα ψυγεία (Big Data Framework 2018)

Κάθε μία από αυτές τις συνδεδεμένες συσκευές παράγει δεδομένα που ανταλλάσσονται μέσω του Διαδικτύου και τα οποία μπορεί να αναλυθούν για να ανακτήσει την αξία. Παρόμοια με τα κοινωνικά μέσα ενημέρωσης, τα δεδομένα που παράγονται μέσω συσκευών IoT είναι τεράστιες από την άποψη της ποσότητας και μπορούν να παρέχουν πληροφορίες για τη συμπεριφορά των καταναλωτών. Ως εκ τούτου, είναι εξαιρετικά πολύτιμη (Big Data Framework 2018).

1.4 Ιστορική Αναδρομή

Για να καταγράψουμε την ουσία των Big Data, παρέχουμε την προέλευση και την ιστορία της BDA και στη συνέχεια προτείνουμε ακριβή ορισμό της Αναλυτικής Μεγάλων Δεδομένων. Έχουν διεξαχθεί αρκετές μελέτες σχετικά με τις ιστορικές απόψεις και τις εξελίξεις στην περιοχή της BDA.

1.4.1 Σημαντικές χρονικές στιγμές

Το μεγαλύτερο χρονικό διάστημα της ιστορικής ανασκόπησης για το Big Data ανήκει στην περιγραφή του Bernard Marr (Marr 2015). Αυτός ανίχνευσε την προέλευση των Μεγάλων Δεδομένων στα 18.000 π.Χ. με τα ραβδιά συλλογής. Το Ishango Bone ανακαλύφθηκε το 1960 σε αυτό που είναι τώρα η Ουγκάντα και πιστεύεται ότι είναι ένα από τα πρώτα στοιχεία τεκμηρίωσης της προϊστορικής αποθήκευσης δεδομένων. Οι παλαιολιθικοί άνδρες θα σηματοδοτούσαν εγκοπές σε ραβδιά ή οστά, για να παρακολουθούν την εμπορική δραστηριότητα ή τις προμήθειες. Θα συγκρίνουν τα ραβδιά και τις εγκοπές για να πραγματοποιήσουν στοιχειώδεις υπολογισμούς, επιτρέποντάς τους να κάνουν προβλέψεις όπως το πόσο καιρό θα παρέμεναν τα τρόφιμά τους.

Ο Gil Press[i] παρέχει μια σύντομη ιστορία των Big Data ξεκινώντας από το 1944, η οποία βασίστηκε στο έργο του Rider. Καλύπτει 68 χρόνια ιστορίας της εξέλιξης των Big Data μεταξύ 1944 και 2012 και απεικονίζει 32 γεγονότα που σχετίζονται με τα μεγάλα δεδομένα. Όπως αναφέρει ο Press στο άρθρο του, η λεπτή γραμμή μεταξύ της ανάπτυξης δεδομένων και των μεγάλων δεδομένων έχουν μπερδευτεί. Πολύ συχνά, ο ρυθμός αύξησης των δεδομένων αναφέρεται ως "έκρηξη πληροφοριών". αν και τα "δεδομένα" και "πληροφορίες" χρησιμοποιούνται συχνά εναλλακτικά, οι δύο όροι έχουν διαφορετικές σημασίες. Η μελέτη του Press είναι αρκετά ολοκληρωμένη και καλύπτει εμφανίσεις της BDA μέχρι τον Δεκέμβριο του 2013. Από τότε, έχουν υπάρξει πολλά σημαντικά γεγονότα Big Data. Παρ 'όλα αυτά, η ανασκόπηση κάλυψε τόσο τα γεγονότα Big Data όσο και τα επιστημονικά δεδομένα. Σε αυτό το βαθμό, θα μπορούσε ο όρος "επιστήμη των δεδομένων" να θεωρηθεί ως συμπληρωματική σημασία για την BDA.

Ο Frank Ohlhorst το 2012 καθιέρωσε την προέλευση του όρου Big Data το *1880* όταν πραγματοποιήθηκε η δέκατη απογραφή των ΗΠΑ. Αυτή η απογραφή χρειάστηκε οκτώ χρόνια για να καταγραφεί και εκτιμήθηκε ότι η απογραφή του 1890 θα χρειαζόταν περισσότερα από 10 χρόνια χρησιμοποιώντας τις τότε διαθέσιμες μεθόδους (Press 2013).

Η εισροή των δεδομένων απογραφής οδήγησε το *1881* στην εφεύρεση της μηχανής χαρτογράφησης Hollerith (κάρτες διάτρησης), η οποία "εξημέρωσε" τα μεγάλα δεδομένα και τους άφησε να ολοκληρώσουν την εργασία σε περίπου ένα χρόνο. Μετέτρεψε τον Hollerith σε έναν

επιχειρηματία και η εταιρεία του τελικά έγινε μέρος αυτού που γνωρίζουμε ως IBM (Winshuttle 1996).

Η υπερφόρτωση των πληροφοριών συνεχίστηκε με την έκρηξη στον πληθυσμό των ΗΠΑ το *1932*, την έκδοση αριθμών κοινωνικής ασφάλισης και τη γενική ανάπτυξη της γνώσης (έρευνα) που απαιτούσε πιο λεπτομερή και οργανωμένη τήρηση αρχείων.

Το *1940* οι βιβλιοθήκες, η αρχική πηγή της οργάνωσης και αποθήκευσης δεδομένων, έπρεπε να προσαρμόσουν τις μεθόδους αποθήκευσης τους για να ανταποκριθούν στην ταχέως αυξανόμενη ζήτηση νέων εκδόσεων και έρευνας (Winshuttle 1996).

Οι μελετητές άρχισαν να αναφέρονται σε αυτήν την απίστευτη επέκταση των πληροφοριών ως "έκρηξη πληροφοριών". Πρώτα αναφέρεται από το Σύνταγμα Lawton (εφημερίδα) το *1941*, ο όρος επεκτάθηκε σε ένα άρθρο New Statesman τον Μάρτιο του 1964, το οποίο αναφερόταν στη δυσκολία διαχείρισης της ποσότητας των διαθέσιμων πληροφοριών (Winshuttle 1996).

Η πρώτη προειδοποίηση για την ανάπτυξη της γνώσης ως έρχεται πρόβλημα αποθήκευσης και ανάκτηση ήρθε το *1944*, όταν ο Fremont Rider, ένας βιβλιοθηκονόμος του Πανεπιστημίου Wesleyan, έκρινε ότι οι Αμερικανικές Βιβλιοθήκες Πανεπιστημίου διπλασιάζονται σε μέγεθος κάθε δεκαέξι χρόνια. Δεδομένου αυτού του ρυθμού ανάπτυξης, ο Rider εκτιμά ότι η βιβλιοθήκη Yale το 2040 θα έχει "περίπου 200.000.000 τόμους, οι οποίοι θα καταλαμβάνουν περισσότερα από 6.000 μίλια ράφια απαιτώντας ένα προσωπικό πάνω από έξι χιλιάδες άτομα" (Winshuttle 1996).

Ο Claude Shannon δημοσίευσε το *1948* μια Μαθηματική Θεωρία Επικοινωνίας η οποία καθιέρωσε ένα πλαίσιο για τον προσδιορισμό των ελάχιστων απαιτήσεων δεδομένων για τη μετάδοση πληροφοριών σε θορυβώδη κανάλια. Αυτό ήταν ένα έργο ορόσημο που επέτρεψε μεγάλο μέρος της σημερινής υποδομής. Χωρίς αυτή την κατανόηση, τα δεδομένα θα ήταν "μεγαλύτερα" από ό,τι σήμερα. Συνεχίστηκε με το "Some Factors Affecting Telegraph Speed" της Nyquist, το οποίο επέτρεψε να δοκιμάσουμε αναλογικά σήματα και να τα εκπροσωπήσουμε ψηφιακά, το οποίο είναι το θεμέλιο της σύγχρονης επεξεργασίας δεδομένων (Winshuttle 1996).

Η έννοια της εικονικής μνήμης αναπτύχθηκε από τον Γερμανό φυσικό Fritz-Rudolf Güntsch (*1956*) ως ιδέα που επεξεργάστηκε την πεπερασμένη αποθήκευση ως άπειρη. Η αποθήκευση, η οποία διαχειρίζεται ολοκληρωμένο υλικό και λογισμικό για την απόκρυψη των λεπτομερειών από τον χρήστη, μας επέτρεψε να επεξεργαζόμαστε δεδομένα χωρίς τους περιορισμούς της μνήμης υλικού, οι οποίοι προηγουμένως ανάγκασαν το πρόβλημα να χωριστεί (καθιστώντας την λύση αντανακλαστική της αρχιτεκτονικής του υλικού, μια πολύ αφύσικη πράξη).

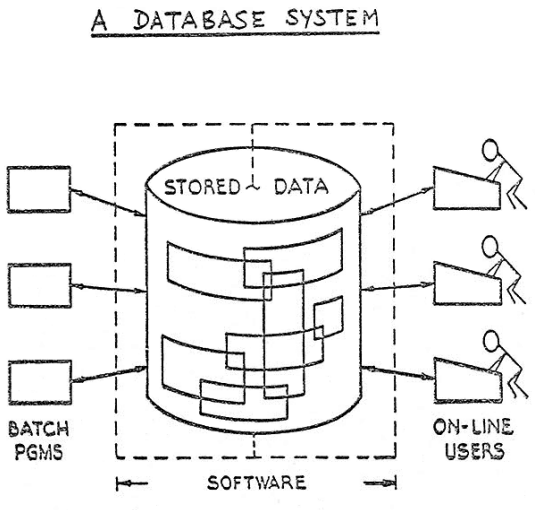
Ο επιστήμονας πληροφοριών, Derek Price το *1961*, γενίκευσε τα ευρήματα του Rider για να συμπεριλάβει το σύνολο της επιστημονικής γνώσης. Η επιστημονική επανάσταση, όπως την ονόμασε, ήταν υπεύθυνη για την ταχεία επικοινωνία νέων ιδεών ως επιστημονικές πληροφορίες. Αυτή η ταχεία ανάπτυξη είχε τη μορφή νέων περιοδικών που διπλασιάζονται κάθε 15 χρόνια (Winshuttle 1996).

Οι επιστήμονες έχουν επιδιώξει την αναγνώριση της φωνής σχεδόν για όσο διάστημα χτίζουν υπολογιστές. Το *1962*, ο William C. Dersch της IBM παρουσίασε την Μηχανή Shoebox στην Παγκόσμια Έκθεση. Αυτή ήταν η πρώτη μηχανή για την κατανόηση 16 λέξεων και δέκα ψηφίων σε προφορικά αγγλικά χρησιμοποιώντας τα τότε διαθέσιμα δεδομένα και την ικανότητα να τα επεξεργάζεται αποτελεσματικά. Ωστόσο, παρέμεινε ένας μακρύς δρόμος για τη μετατροπή αυτής της καινοτομίας αναγνώρισης ομιλίας σε εμπορικά εφικτά προϊόντα. Αυτό το ταξίδι θα απαιτούσε άλματα στην ισχύ επεξεργασίας και μειωμένο κόστος υπολογιστών. Τα περισσότερα δεδομένα θα συμβάλουν επίσης τελικά στην προώθηση της κατάρτισης των συστημάτων αναγνώρισης ομιλίας (Winshuttle 1996).

Στις αρχές της δεκαετίας του 1960 (*1963*), η Price παρατήρησε ότι το τεράστιο ποσό της επιστημονικής έρευνας ήταν υπερβολικά μεγάλο για τους ανθρώπους να πληροφορηθούν Τα αφηρημένα περιοδικά, τα οποία δημιουργήθηκαν στα τέλη της δεκαετίας του 1800 ως τρόπος διαχείρισης της αυξανόμενης βάσης γνώσεων, αναπτύσσονταν επίσης στην ίδια τροχιά (πολλαπλασιάζοντας τον παράγοντα δέκα κάθε μισό αιώνα) και είχαν ήδη φθάσει σε ένα "κρίσιμο μέγεθος". Δεν ήταν πλέον μια λύση αποθήκευσης ή οργάνωσης για πληροφορίες (Winshuttle 1996).

Όχι μόνο υπήρξε η άνοδος των πληροφοριών στον τομέα της επιστήμης, αλλά και στον επιχειρηματικό τομέα. Λόγω της εισροής πληροφοριών στη δεκαετία του 1960 (*1966*), οι περισσότεροι οργανισμοί άρχισαν να σχεδιάζουν, να αναπτύσσουν και να εφαρμόζουν κεντρικά υπολογιστικά συστήματα που τους επέτρεπαν να αυτοματοποιήσουν τα συστήματά τους απογραφής τους (Winshuttle 1996).

Το *1970*, ο Edgar F. Codd, μαθητής της Οξφόρδης που εργάζεται στο εργαστήριο IBM Research Lab, δημοσίευσε ένα έγγραφο που δείχνει πώς μπορούν να αποκτήσουν πρόσβαση οι πληροφορίες που αποθηκεύονται σε μεγάλες βάσεις δεδομένων χωρίς να γνωρίζουν πώς έχουν διαρθρωθεί οι πληροφορίες ή που διέμεναν στη βάση δεδομένων. Μέχρι τότε, η ανάκτηση πληροφοριών απαιτούσε σχετικά εξελιγμένες γνώσεις υπολογιστών ή ακόμα και τις υπηρεσίες ειδικών - μια χρονοβόρα και δαπανηρή εργασία. Σήμερα, οι περισσότερες συναλλαγές δεδομένων ρουτίνας-πρόσβαση σε τραπεζικούς λογαριασμούς, χρησιμοποιώντας πιστωτικές κάρτες, μετοχές διαπραγμάτευσης, κράτηση ταξιδιού, αγοράζοντας τα πράγματα σε απευθείας σύνδεση - όλοι χρησιμοποιούν δομές βασισμένες στη θεωρία σχεσιακών βάσεων δεδομένων. Στην Εικόνα 1 παρουσιάζεται ένα σύστημα σχεσιακής βάσης δεδομένων που παρουσίασε ο Codd (Winshuttle 1996).



Εικόνα 1: Σχεδιάγραμμα του Codd για τη σχεσιακή βάση δεδομένων (Winshuttle, 1996)

Η Απογραφή Πληροφοριών Ροής, που διεξήχθη από το Υπουργείο Ταχυδρομείων και Τηλεπικοινωνιών στην Ιαπωνία, άρχισε να παρακολουθεί τον όγκο των πληροφοριών που κυκλοφορούν στη χώρα αυτή για το *1975*. Με τον αριθμό των λέξεων που χρησίμευαν ως μονάδα μέτρησης σε όλα τα μέσα ενημέρωσης, η μελέτη διαπίστωσε ότι η παροχή πληροφοριών υπερέβησαν σε μεγάλο βαθμό την κατανάλωση πληροφοριών και η ζήτηση για απλή επικοινωνία είχε σταματήσει. Μάλλον, η ζήτηση για αμφίδρομη, πιο εξατομικευμένη επικοινωνία ήταν σε άνοδο, καλύπτοντας τις ανάγκες των ατόμων (Winshuttle 1996).

Στα μέσα της δεκαετίας του 1970 (*1976*), τα συστήματα προγραμματισμού υλικών απαιτήσεων (MRP) σχεδιάστηκαν ως εργαλείο για να βοηθήσουν τις κατασκευαστικές εταιρείες να οργανώσουν και να προγραμματίσουν τις πληροφορίες τους. Γύρω από αυτό το διάστημα, οι υπολογιστές μόλις αρχίζουν να κερδίζουν τη δημοτικότητα με τις επιχειρήσεις. Αυτές οι αλλαγές σηματοδότησαν τη στροφή προς τις επιχειρηματικές διαδικασίες και τις λογιστικές δυνατότητες, καθώς σχηματίστηκαν εταιρείες όπως οι Oracle, JD Edwards και SAP. Ήταν η Oracle που παρουσίασε τελικά την αρχική εμπορικά δομημένη γλώσσα προγραμματισμού βάσεων δεδομένων (SQL) (Winshuttle 1996).

Καθώς οι πληροφορίες άρχισαν να αναπτύσσονται γρηγορότερα, μειώθηκαν οι συμπυκνωμένες επιλογές αποθήκευσης και οργάνωσης δεδομένων. Στη συζήτησή του τον Απρίλιο του *1980*, "Πού πηγαίνουμε από εδώ;" ο Tjomsland δήλωσε: "Όσοι συνδέονται με συσκευές αποθήκευσης έχουν συνειδητοποιήσει εδώ και καιρό ότι ο Πρώτος Νόμος της Πάρκινσον μπορεί να παραφράζεται για να περιγράψει τη βιομηχανία μας:" Τα δεδομένα διευρύνθηκαν για να γεμίσουν τον διαθέσιμο χώρο ". Πιστεύω ότι διατηρούνται μεγάλα ποσά δεδομένων, επειδή οι χρήστες δεν έχουν κανένα τρόπο να εντοπίζουν παρωχημένα δεδομένα. οι κυρώσεις για την αποθήκευση παρωχημένων δεδομένων είναι λιγότερο εμφανείς από ό, τι οι κυρώσεις για την απόρριψη δυνητικά χρήσιμων δεδομένων (Winshuttle 1996). "

Καθώς η τεχνολογία συνέχισε να προχωράει, κάθε βιομηχανία άρχισε να επωφελείται από νέους τρόπους οργάνωσης, αποθήκευσης και παραγωγής δεδομένων. Οι εταιρείες άρχισαν να χρησιμοποιούν τα δεδομένα για να παρέχουν απαντήσεις για καλύτερες επιχειρηματικές αποφάσεις. Στην παρακολούθηση της ροής των πληροφοριών, που δημοσιεύτηκε στο περιοδικό

Science, ο συντάκτης Ithiel de Sola Pool, εξέτασε την ανάπτυξη πληροφοριών σε 17 μεγάλα μέσα επικοινωνίας από το 1960 έως το 1977. Ο ίδιος αναγνώρισε τη μαζική ανάπτυξη πληροφοριών στην επέκταση της ραδιοτηλεοπτικής βιομηχανίας (*1983*) (Winshuttle 1996).

Μετά την άνοδο των συστημάτων MRP, ο Σχεδιασμός Πόρων Παραγωγής (MRP II) εισήχθη στη δεκαετία του 1980 (*1985*) με έμφαση στη βελτιστοποίηση των διαδικασιών παραγωγής με το συγχρονισμό των υλικών με τις απαιτήσεις παραγωγής. Το MRP II περιελάμβανε τομείς όπως η διαχείριση του καταστήματος και της διανομής, η διαχείριση έργων, η χρηματοδότηση, ο ανθρώπινος πόρος και η μηχανική. Δεν ήταν πολύ καιρό μετά την υιοθέτηση αυτής της τεχνολογίας που άλλες βιομηχανίες άρχισαν να παρατηρούν και τελικά να υιοθετούν την ίδια την τεχνολογία ERP (π.χ. κυβερνητικές υπηρεσίες και οργανισμούς στον τομέα των υπηρεσιών) (Winshuttle 1996).

Το *1985*, οι Barry Devlin και Paul Murphy καθόρισαν μια αρχιτεκτονική για την αναφορά και την ανάλυση των επιχειρήσεων στην IBM (Devlin & Murphy, IBM Systems Journal 1988), η οποία έγινε η βάση της αποθήκευσης δεδομένων. Η καρδιά αυτής της αρχιτεκτονικής και η αποθήκευση δεδομένων γενικά είναι η ανάγκη για υψηλής ποιότητας και συνεπή αποθήκευση ιστορικά πλήρων και ακριβών δεδομένων (Winshuttle 1996).

Στο άρθρο του ο Hal Becker (*1986*), "Μπορούν οι χρήστες να απορροφήσουν δεδομένα με τα σημερινά ποσοστά; Αύριο;", δήλωσε ότι" η πυκνότητα αποκωδικοποίησης που επιτεύχθηκε από τον Gutenberg ήταν περίπου 500 σύμβολα (χαρακτήρες) ανά κυβική ίντσα - 500 φορές την πυκνότητα [4.000 π.Χ. Σουμέρι] δισκίων αργίλου. Μέχρι το έτος 2000, η μνήμη τυχαίας προσπέλασης ημιαγωγών θα πρέπει να αποθηκεύει $1.25x10^{11}$ bytes ανά κυβική ίντσα.

Στα τέλη της δεκαετίας του '80 (*1988*) έως τις αρχές της δεκαετίας του '90 παρατηρήθηκε αύξηση των συστημάτων ERP (Enterprise Resource Planning) καθώς έγιναν πιο εξελιγμένά και απέκτησαν την ικανότητα συντονισμού και ενσωμάτωσης σε ολόκληρες εταιρείες. Τα τεχνολογικά θεμέλια των συστημάτων MRP, MRP II και ERP άρχισαν να ενσωματώνουν τομείς στους τομείς της μεταποίησης, της διανομής, της λογιστικής, της χρηματοδότησης, της διαχείρισης ανθρώπινων πόρων, της διαχείρισης έργων, της διαχείρισης αποθεμάτων, των υπηρεσιών και της συντήρησης, σε όλη την εταιρεία (Winshuttle 1996).

Το *1989*, ο Howard Dresner επεκτάθηκε στον δημοφιλές πλέον όρο "Business Intelligence (BI)", που σχεδιάστηκε αρχικά από τον Hans Peter Luhn το 1958. Ο Dresner χαρακτήρισε τον όρο ως "έννοιες και μεθόδους για τη βελτίωση της λήψης επιχειρηματικών αποφάσεων χρησιμοποιώντας συστήματα στήριξης βάσει πραγματικών δεδομένων" . Λίγο αργότερα, ανταποκρινόμενοι στην ανάγκη βελτίωσης του BI, άρχισαν να εμφανίζονται εταιρείες όπως τα Business Objects, Actuate, Crystal Reports και MicroStrategy, προσφέροντας την παρουσίαση και την ανάλυση των δεδομένων της εταιρείας (Winshuttle 1996).

Το *1992*, η Crystal Reports δημιούργησε την πρώτη απλή αναφορά βάσης δεδομένων χρησιμοποιώντας τα Windows. Αυτές οι αναφορές επέτρεψαν στις επιχειρήσεις να δημιουργήσουν μια ενιαία αναφορά από μια ποικιλία πηγών δεδομένων με ελάχιστο γραπτό κώδικα. Αυτό με τη σειρά του βοήθησε να ελαττωθεί η πίεση του κορεσμένου από δεδομένα τοπίου και επέτρεψε στις επιχειρήσεις να χρησιμοποιούν επιχειρηματική ευφυΐα με προσιτό τρόπο. Η δεκαετία του 1990 (συγκεκριμένα από το *1995*) ήταν μια εποχή εκρηκτικής ανάπτυξης για την τεχνολογία και τα δεδομένα Business Intelligence (BI) άρχισαν να συσσωρεύονται με τη μορφή εγγράφων του Microsoft Excel (Winshuttle 1996).

Η εισροή πληροφοριών οδήγησε σε μια νέα πρόκληση για τη διαχείριση των δεδομένων και στην αύξηση του κόστους δημοσίευσης και αποθήκευσης όλων. Δεδομένου ότι τα δεδομένα έγιναν πιο δύσκολα να διατηρηθούν, για να προσφέρουν περισσότερη λειτουργικότητα, η ψηφιακή αποθήκευση έγινε γρήγορα πιο αποδοτική για την αποθήκευση δεδομένων από ό, τι το χαρτί, όπου και αρχίζει να εμφανίζεται η πλατφόρμα BI. Οι R.J.T. Οι Morris και B.J. Truskowski διερεύνησαν την αποθήκευση δεδομένων στο άρθρο τους The Evolution of Storage Systems, που δημοσιεύθηκε στο IBM Systems Journal (*1996*) (Winshuttle 1996).

Η έκρηξη δεδομένων έφερε επίσης περισσότερες προκλήσεις στους προμηθευτές ERP. Η ανάγκη επανασχεδιασμού των προϊόντων ERP, συμπεριλαμβανομένου του διαχωρισμού του εμποδίου της ιδιοκτησίας και της προσαρμογής, υποχρέωνε τους πωλητές να αγκαλιάσουν τη συνεργατική επιχείρηση μέσω του ενδοδικτύου με απρόσκοπτο τρόπο (*1996*) (Winshuttle 1996).

Μια άλλη ιστορική ανασκόπηση συνέβαλε το Visualizing.org, το οποίο επικεντρώθηκε στο χρονοδιάγραμμα του τρόπου με τον οποίο εφαρμόζεται η Αναλυτική Μεγάλων Δεδομένων. Η

ιστορική περιγραφή καθορίζεται κυρίως από γεγονότα που σχετίζονται με τα μεγάλα δεδομένα και προωθούνται από πολλές εταιρείες του Διαδικτύου και της πληροφορικής, όπως η Google, το YouTube, το Yahoo, το Facebook, το Twitter και την Apple. Τόνισε επίσης τη σημαντική επίδραση της Hadoop στην ιστορία της BDA. Τόνισε κυρίως τον σημαντικό ρόλο της Hadoop στο BDA. Με βάση αυτές τις μελέτες, παρουσιάζεται η ιστορία των Big Data, της Hadoop, και το οικοσύστημα τους στην Εικόνα 2 (Buyya, Calheiros and Dastjerdi 2016).

1.4.2 Ανάλυση κατά περιόδους

Big Data 1.0

Η ανάλυση δεδομένων, η αναλυτική δεδομένων και τα μεγάλα δεδομένα προέρχονται από τον μακροχρόνιο τομέα της διαχείρισης βάσης δεδομένων διαχείριση. Βασίζεται σε μεγάλο βαθμό στις τεχνικές αποθήκευσης, εξαγωγής και βελτιστοποίησης που είναι συνηθισμένες σε δεδομένα που αποθηκεύονται σε συστήματα διαχείρισης σχεσιακής βάσης δεδομένων (RDBMS). Η διαχείριση βάσεων δεδομένων και η αποθήκευση δεδομένων θεωρούνται τα βασικά συστατικά της μεγάλης κλίμακας δεδομένων στη φάση 1.0. Παρέχει το θεμέλιο της σύγχρονης ανάλυσης δεδομένων όπως το γνωρίζουμε σήμερα, χρησιμοποιώντας γνωστές τεχνικές όπως ερωτήματα βάσης δεδομένων, online αναλυτική επεξεργασία και πρότυπά εργαλεία αναφοράς. Τα χαρακτηριστικά της Φάσης 1.0 περιγράφονται στον Πίνακας 2 (Big Data Framework 2018).

Εικόνα 2: Ιστορική αναδρομή μεγάλων δεδομένων (Buyya, Calheiros and Dastjerdi 2016)

Big Data 2.0

Από τις αρχές του 2000, το Διαδίκτυο και ο Παγκόσμιος Ιστός άρχισαν να προσφέρουν μοναδικές συλλογές δεδομένων και ευκαιρίες ανάλυσης δεδομένων. Με την επέκταση της επισκεψιμότητας στον ιστό και των ηλεκτρονικών καταστημάτων, οι εταιρείες όπως οι Yahoo, Amazon και το eBay άρχισαν να αναλύουν τη συμπεριφορά των πελατών, αναλύοντας τα κλικ, τα δεδομένα τοποθεσίας για τα συγκεκριμένα IP και τα αρχεία καταγραφής αναζήτησης. Αυτό άνοιξε έναν εντελώς νέο κόσμο δυνατοτήτων (Big Data Framework 2018).

Σε σχέση με την ανάλυση δεδομένων, την αναλυτική δεδομένων, και τις Μεγάλης Κλίμακας δεδομένα, η επισκεψιμότητα Ιστού βασισμένη σε HTTP εισήγαγε μια τεράστια αύξηση σε ημιδομημένα και αδόμητα δεδομένα. Εκτός από το πρότυπο δομημένου τύπου δεδομένων, οι οργανισμοί χρειάζονται τώρα να βρουν νέες προσεγγίσεις και λύσεις αποθήκευσης για να αντιμετωπιστούν αυτοί οι νέοι τύποι δεδομένων προκειμένου να αναλυθούν αποτελεσματικά. Η άφιξη και η ανάπτυξη των δεδομένων των κοινωνικών μέσων ενημέρωσης επιδείνωσε σημαντικά την ανάγκη για εργαλεία, τεχνολογίες και αναλυτικές τεχνικές που ήταν σε θέση να εξάγουν σημαντικές πληροφορίες από αυτά τα αδόμητα δεδομένα (Big Data Framework 2018).

Big Data 3.0

Παρόλο που το μη δομημένο περιεχόμενο, που βασίζεται στο διαδίκτυο εξακολουθεί να αποτελεί την κύρια εστίαση για πολλές οργανώσεις στην ανάλυση δεδομένων, την αναλυτική δεδομένων και στα μεγάλα δεδομένα, οι τρέχουσες δυνατότητες ανάκτησης πολύτιμων πληροφοριών αναδύονται από τις κινητές συσκευές (Big Data Framework 2018).

Οι κινητές συσκευές δίνουν όχι μόνο τη δυνατότητα να αναλύουν δεδομένα συμπεριφοράς (όπως κλικ και ερωτήματα αναζήτησης), αλλά επίσης παρέχουν τη δυνατότητα αποθήκευσης και ανάλυσης δεδομένων βάσει τοποθεσίας (GPS data). Με την πρόοδο αυτών των κινητών συσκευών, είναι δυνατό να παρακολουθείτε η κίνηση, να αναλύετε η φυσική συμπεριφορά και ακόμη και τα δεδομένα που σχετίζονται με την υγεία (αριθμός βημάτων που παίρνετε ανά ημέρα). Αυτά τα δεδομένα προσφέρουν ένα εντελώς νέο φάσμα ευκαιριών, από τη μεταφορά, το σχεδιασμό της πόλης και την υγεία (Big Data Framework 2018).

Ταυτόχρονα, η άνοδος των συσκευών που βασίζονται σε αισθητήρες, αυξάνουν τα δεδομένα όπως ποτέ άλλοτε. Φημισμένα στοιχεία όπως το «Internet of Things» (IoT), εκατομμύρια τηλεοράσεις, θερμοστάτες, αξεσουάρ και ακόμη και τα ψυγεία δημιουργούν τώρα zettabytes δεδομένων κάθε ημέρα. Και ο αγώνας για την απόσπαση σημαντικών και πολύτιμων πληροφοριών από αυτά τα νέα δεδομένα πηγών έχει μόλις αρχίσει (Big Data Framework 2018).

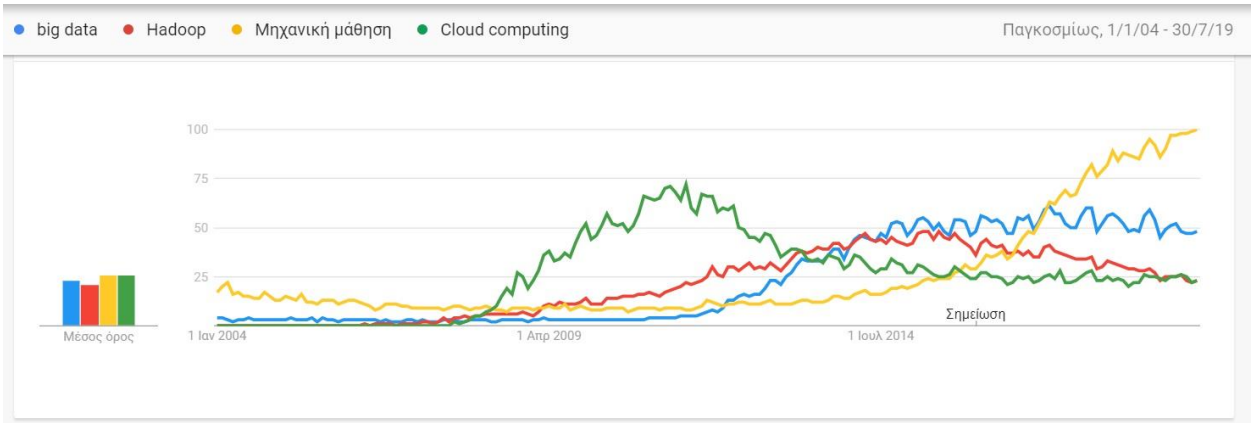
Μια περίληψη των τριών φάσεων στα Big Data παρατίθεται στον Πίνακας 2.

Πίνακας 2: Οι τρείς βασικές φάσεις της ιστορίας των Μεγάλων Δεδομένων

BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

Πίνακας 2: Οι τρείς βασικές φάσεις της ιστορίας των Μεγάλων Δεδομένων (Big Data Framework 2018)

Στην παρούσα μεταπτυχιακή διατριβή παρουσιάζουμε τη συσχέτιση των πιο σημαντικών θεμάτων με τη βοήθεια του εργαλείου Google Trends. Στην παρακάτω Εικόνα 3 παρουσιάζεται στον άξονα x χρονολογία από τον Ιανουάριο του 2004 έως 30 Ιουλίου 2019 και στον άξονα y το ενδιαφέρον των επισκεπτών όσων αφορά τους όρους "Big Data, Hadoop, Machine learning, Cloud Computing" σε κλίμακα από το ένα έως το 100.



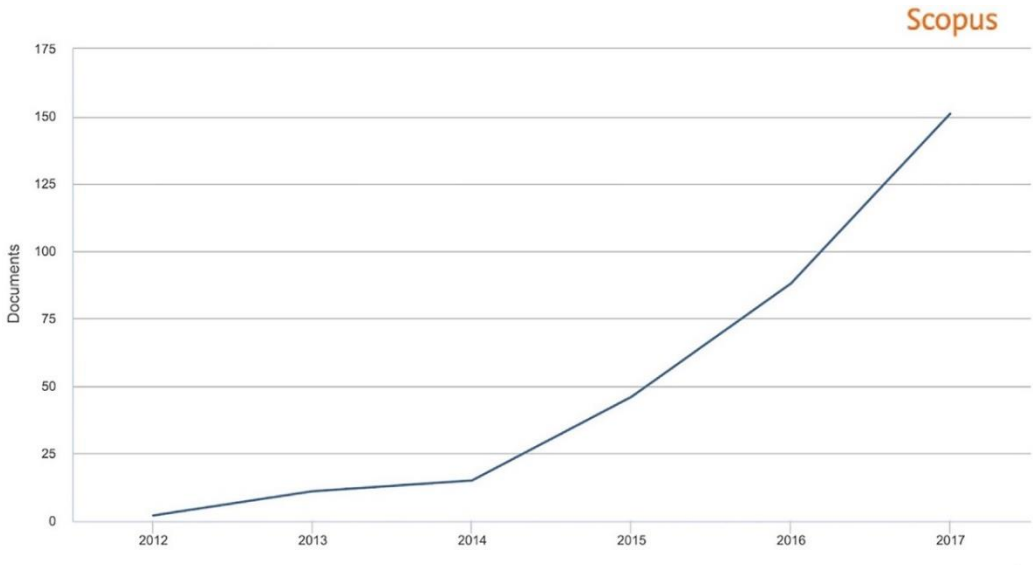
Εικόνα 3: Big Data Trend μέχρι το 2019

1.5 Βιβλιογραφική Επισκόπηση

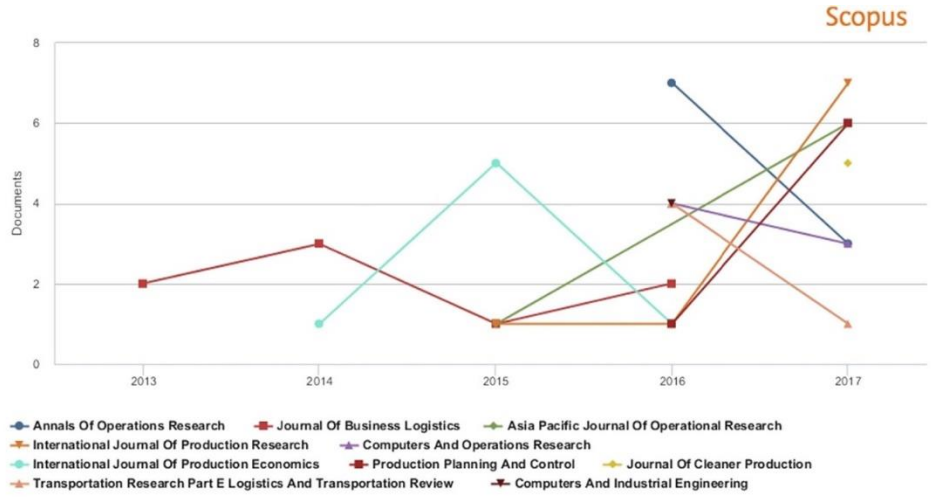
Για την βιβλιογραφική επισκόπηση της Μεγάλης κλίμακας δεδομένων χρησιμοποιήθηκε η έρευνα των (Govindan, et al. 2018), όπου εκπόνησαν ευρεία αναζήτηση για να τεκμηριωθεί ο αριθμός

των δημοσιεύσεων που δημοσιεύθηκαν στην περιοχή της «Αναλυτική Μεγάλης κλίμακας δεδομένων». Η περίοδος έρευνας που χρησιμοποιήθηκε ήταν από το 2012 έως τις 14 Μαρτίου 2018. Μια επισκόπηση του αριθμού των άρθρων παρουσιάζεται στην Εικόνα 4. Αφού εξετάστηκε ειδική αναζήτηση του περιοδικού ,Εικόνα 6 δείχνει σαφώς ότι το «Annals of Operations Research» ήταν το κορυφαίο περιοδικό όσον αφορά τον αριθμό των εγγράφων από το Scopus (Govindan, et al. 2018).

Προκειμένου να βελτιωθεί η αναζήτηση για τον προσδιορισμό του τύπου εγγράφου, εξετάστηκαν περαιτέρω οι βάσεις δεδομένων Scopus. Σε αυτήν την αναζήτηση, τα έγγραφα του συνεδρίου, τα κεφάλαια βιβλίων, βιβλία και οι σύντομες έρευνες αποκλείστηκαν. Μετά την εξαίρεση αυτών των στοιχείων, εξετάστηκαν συνολικά 313. Οι τύποι εγγράφων που εξετάζονται σε αυτή την ενότητα περιλάμβαναν άρθρα, άρθρα στον Τύπο και κριτικές. Ο αριθμός των μεγάλων εγγράφων ανάλυσης δεδομένων που συνεισέφεραν ανάλογα με τη χώρα προέλευσης αναλύθηκε επίσης. Οι δέκα πρώτες χώρες παρουσιάζονται στην Εικόνα 7. Οι Ηνωμένες Πολιτείες είναι πρώτες με 94 έγγραφα, ακολουθούμενα από την Κίνα με 88 έγγραφα, το Ηνωμένο Βασίλειο με 36 έγγραφα, Ινδία με δεκαπέντε έγγραφα και Γερμανία με δεκατέσσερα έγγραφα. Επιπλέον, η θεματική ταξινόμηση των δημοσιευμένων παρουσιάζεται στην Εικόνα 8 (Govindan, et al. 2018).

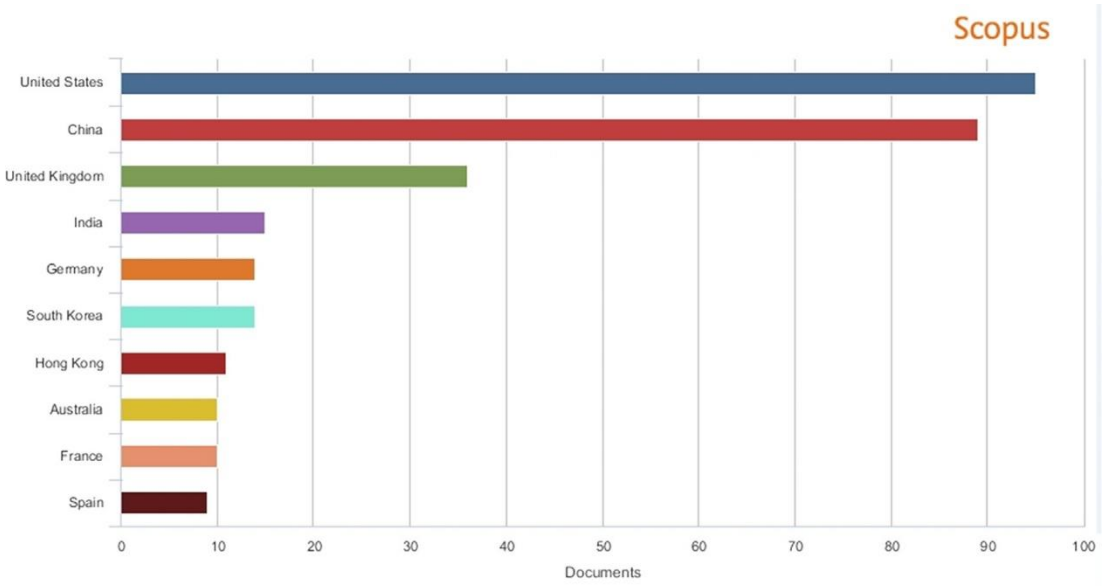


Εικόνα 4:Αριθμός άρθρων για την Αναλυτική Μεγάλης κλίμακας δεδομένων στο Scopus (Govindan, et al. 2018)

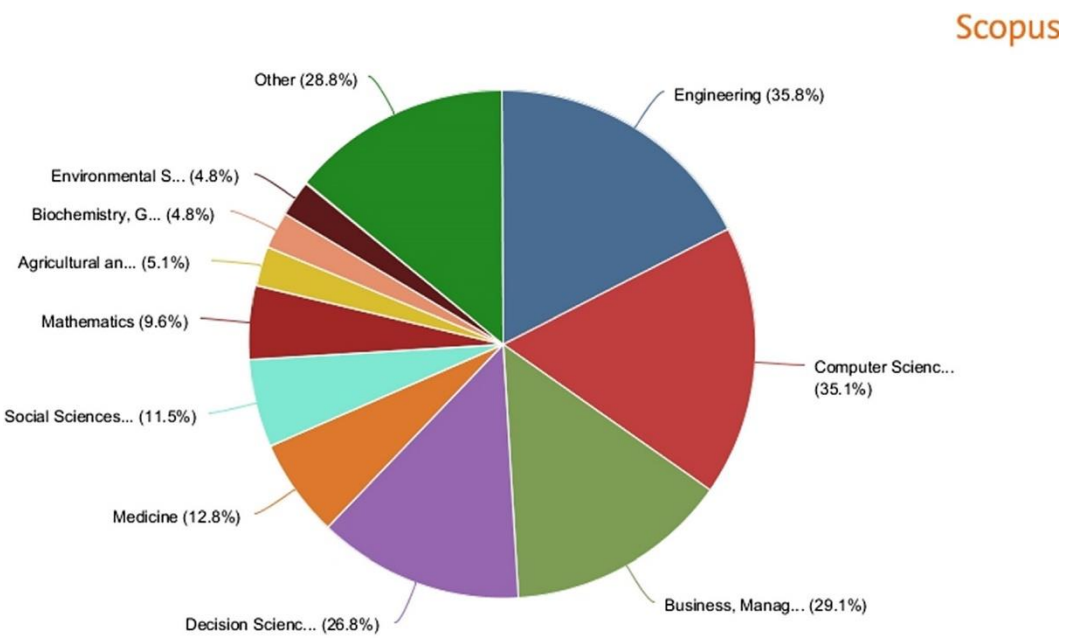


Εικόνα 5: Μερίδιο κορυφαίων διεθνών περιοδικών με τις υψηλότερες συνειδφορές στη δημοσίευση μεγάλων θεμάτων ανάλυσης δεδομένων (Govindan, et al. 2018)

Οι πρώτες θέσεις μοιράζονται μεταξύ της Μηχανικής (35,8%) και της Επιστήμης των Υπολογιστών (35,1%). Στη συνέχεια ακολουθεί η Επιχειρηματική Διοίκηση με ποσοστό 29,1% και η τομέας Λήψης Αποφάσεων με ποσοστό 26,8%. Η Ιατρική, Κοινωνικές επιστήμες, τα Μαθηματικά η Αγροτική ανάπτυξη, η Βιοχημεία και η Γενετική και το Περιβάλλον με ποσοστά 12,8%, 11,5%, 9,6%, 5,1%, 4,8% και 4,8% αντίστοιχα. Τέλος υπάρχει και ένα μεγάλο ποσοστό 28,8% που περιλαμβάνει διάφορους τομείς (Govindan, et al. 2018).



Εικόνα 7:Μερίδιο προέλευσης χώρας δημοσιεύσεων θεμάτων Μεγάλης κλίμακας δεδομένων (Govindan, et al. 2018)



Εικόνα 8:Ταξινόμηση θεμάτων των μεγάλων κλίμακας δεδομένων σύμφωνα με το Scopus. (Govindan, et al. 2018)

1.6 Χαρακτηριστικά Μεγάλης Κλίμακας Δεδομένων

Μπορεί να γίνει εύκολα κατανοητό ότι, ούτε ο χθεσινός όγκος δεδομένων (απόλυτο μέγεθος) ούτε ο σημερινός όγκος δεδομένων μπορεί να οριστεί ως "μεγάλος". Επιπλέον, το σημερινό "μεγάλο" μπορεί να γίνει αύριο "μικρό". Για να αποσαφηνιστεί με ακρίβεια ο όρος Big Data, θα γίνει διερεύνηση ενός ορισμού που βασίζεται στον συνδυασμό των προσεγγίσεων των (Tweed 2008) και (Copi, Cohen and McMahon 2014)

Με βάση την προσέγγιση ορισμού των, θα αποσαφηνιστεί πρώτα ο ιστορικός ορισμός από μια εξελικτική σκοπιά (λεξικό νόημα). Στη συνέχεια, επεκτείνεται ο όρος από 3Vs σε 9Vs ή 3^2Vs με βάση το κίνητρο (καθοριστική έννοια), που θα προσθέσουν περισσότερες ιδιότητες για τον όρο.

Gartner: - ορισμός 3Vs

Από το 1997, έχουν προστεθεί πολλά χαρακτηριστικά στα Μεγάλα Δεδομένα. Μεταξύ αυτών των χαρακτηριστικών, τρία από αυτά είναι τα πιο δημοφιλή και έχουν ευρέως αναφερθεί και

υιοθετηθεί. Η πρώτη είναι η λεγόμενη ερμηνεία του Gartner ή 3Vs. Η προέλευση αυτού του όρου μπορεί να ανιχνευθεί από τον Φεβρουάριο του 2001, από τον (Laney 2001), όπου εξηγεί ότι λόγω της αύξησης των δραστηριοτήτων ηλεκτρονικού εμπορίου, τα δεδομένα έχουν αυξηθεί κατά τρεις διαστάσεις, και συγκεκριμένα:

- Volume, που σημαίνει την εισερχόμενη ροή δεδομένων και τον σωρευτικό όγκο δεδομένων (Όγκος)
- Velocity, η οποία αντιπροσωπεύει το ρυθμό των δεδομένων που αναπτύσσονται για την υποστήριξη της αλληλεπίδρασης και δημιουργούνται από τις αλληλεπιδράσεις (Ταχύτητα)
- Variety, που υποδηλώνει την ποικιλία ασυμβίβαστων και ασυνεπών μορφών δεδομένων και δομών δεδομένων (Ποικιλία)

Σύμφωνα με την ιστορία του χρονοδιαγράμματος των μεγάλων δεδομένων [30], ο ορισμός των 3V's του Douglas Laney (Laney 2001) θεωρούνται ευρέως ως τα "κοινά" χαρακτηριστικά των Big Data, αλλά σταμάτησε να αποδίδει αυτά τα χαρακτηριστικά με τον όρο "μεγάλα δεδομένα".

IBM- ορισμός 4Vs

Η IBM πρόσθεσε ένα άλλο χαρακτηριστικό ή το "V" για τον όρο "Veracity" στην κορυφή των 3Vs σημείωσης, η οποία είναι γνωστή ως τα 4V της Big Data. Ορίζει κάθε "V" ως εξής

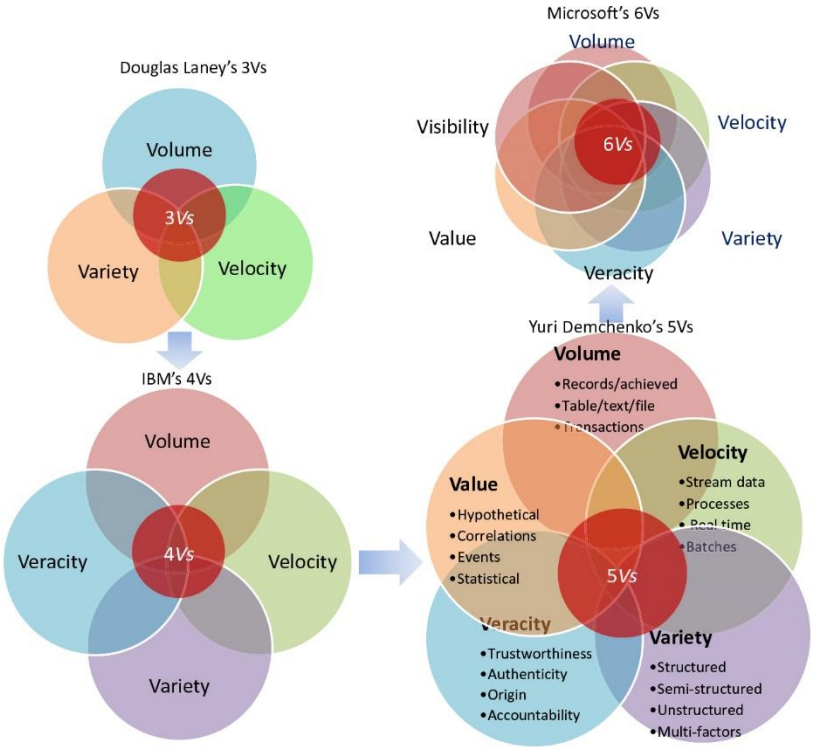
- Ο όγκος αντιπροσωπεύει την κλίμακα των δεδομένων (Volume)
- Η ταχύτητα υποδηλώνει την ανάλυση δεδομένων ροής (Velocity)
- Η ποικιλία δηλώνει διαφορετικές μορφές δεδομένων (Variety)
- Η ακρίβεια συνεπάγεται την αβεβαιότητα των δεδομένων (Veracity)

Οι (Zikopoulos, et al. 2012) εξήγησαν τον λόγο πίσω από την πρόσθετη διάσταση "V" ή την αξιοπιστία, η οποία είναι η απάντηση στα προβλήματα ποιότητας που οι πελάτες της IBM αντιμετώπισαν ξεκινώντας δραστηριότητες με τα μεγάλα δεδομένα.

Microsoft- 6Vs ορισμός

Για τη μεγιστοποίηση της επιχειρηματικής αξίας, η Microsoft επέκτεινε τα χαρακτηριστικά των 3Vs του Douglas Laney σε 6Vs, η οποία πρόσθεσε την μεταβλητότητα, την αλήθεια και την ορατότητα Εικόνα 9 (Buyya, Calheiros and Dastjerdi 2016).

- Ο όγκος αντιπροσωπεύει την κλίμακα δεδομένων (Volume)
- Η ταχύτητα υποδηλώνει την ανάλυση δεδομένων ροής (Velocity)
- Η ποικιλία δηλώνει διαφορετικές μορφές δεδομένων (Variety)
- Η εγκυρότητα επικεντρώνεται στην αξιοπιστία των πηγών δεδομένων (Veracity)
- Η μεταβλητότητα αφορά στην πολυπλοκότητα του συνόλου δεδομένων (Variability). Σε σύγκριση με την "Ποικιλία" (Variety) (ή τη μορφή διαφορετικών δεδομένων), αυτό σημαίνει τον αριθμό των μεταβλητών στα σύνολα δεδομένων.
- Η ορατότητα υπογραμμίζει ότι πρέπει να έχετε μια πλήρη εικόνα των δεδομένων για να παρθεί μια σωστή απόφαση (Visibility)



Εικόνα 9: Ορισμός Μεγάλων Δεδομένων από 3Vs, 4Vs, 5Vs και 6Vs (Buyya, Calheiros and Dastjerdi 2016)

Περισσότερα Vs για τα μεγάλα δεδομένα

Όλοι αυτοί οι ορισμοί, όπως οι 3Vs, 4Vs, 5Vs ή ακόμα και οι 11 Vs, προσπαθούν κυρίως να αρθρώσουν μια πτυχή των δεδομένων. Οι περισσότεροι από αυτούς είναι ορισμοί που βασίζονται σε δεδομένα, αλλά δεν καταφέρνουν να εκφράζουν σαφώς τα Big Data σε σχέση με την ουσία της BDA. Ο παρουσιάζει 7 τύπους ορισμών των μεγάλων δεδομένων. Κάθε τύπος προσπαθεί να περιγράψει μια ξεχωριστή οπτική των μεγάλων δεδομένων (Elliott 2013).

1.7 Αποσαφήνιση όρων

Στον επιχειρηματικό και επιστημονικό τομέα, οι περισσότεροι οργανισμοί διαφοροποιούνται μεταξύ τεσσάρων διαφορετικών μορφών αναγνώρισης προτύπων σε σύνολα δεδομένων. Παρόλο που και οι τέσσερις ορισμοί είναι στενά συνδεδεμένοι, υπάρχουν ορισμένες λεπτές διαφορές μεταξύ τους που έχουν αντίκτυπο για το σχεδιασμό λύσεων Big Data (Big Data Framework 2018).

1.7.1 Ανάλυση δεδομένων

Η ανάλυση δεδομένων είναι μια διαδικασία επιθεώρησης, καθαρισμού, μετασχηματισμού και μοντελοποίησης δεδομένων με στόχο την ανακάλυψη χρήσιμων πληροφοριών, υποδεικνύοντας συμπεράσματα και υποστηρίζοντας τη λήψη αποφάσεων. Η ανάλυση δεδομένων έχει πολλαπλές όψεις και προσεγγίσεις, έχοντας μια ποικιλία από ονόματα, σε διαφορετικούς τομείς των επιχειρήσεων, της επιστήμης και της κοινωνικής επιστήμης (Big Data Framework 2018).

Η ανάλυση των δεδομένων - με την κυριολεκτική έννοια - υπήρξε εδώ και αιώνες. Ο πρωταρχικός σκοπός της ανάλυσης δεδομένων είναι η αναθεώρηση των υφιστάμενων δεδομένων προκειμένου να σχηματιστούν πρότυπα που έχουν συμβεί στο παρελθόν. Ως εκ τούτου, αναφέρεται επίσης συχνά ως περιγραφική ανάλυση δεδομένων. Ένα παράδειγμα ανάλυσης δεδομένων θα ήταν να αναθεωρηθεί το μοτίβο πωλήσεων διαφορετικών καταστημάτων όπως διαμορφώνεται μέσα στο έτος (Big Data Framework 2018).

1.7.2 Αναλυτική

Η Αναλυτική δεδομένων αφορά στην ανακάλυψη, την ερμηνεία και η επικοινωνία σημαντικών προτύπων από τα δεδομένα. Ιδιαίτερα πολύτιμες σε περιοχές πλούσιες με καταγεγραμμένες πληροφορίες, τα αναλυτικά στοιχεία βασίζονται στην ταυτόχρονη εφαρμογή στατιστικών, τον

προγραμματισμό υπολογιστών και την επιχειρησιακή έρευνα για την ποσοτικοποίηση της απόδοσης (Big Data Framework 2018).

Η Αναλυτική περιλαμβάνει ένα αυξανόμενο πεδίο των δυνατοτήτων της επιστήμης των δεδομένων, συμπεριλαμβανομένων των στατιστικών στοιχείων, τα μαθηματικά, μηχανική μάθηση, πρόβλεψη μοντελοποίησης, εξόρυξη δεδομένων, το γνωστικό υπολογιστή και την τεχνητή νοημοσύνη (Big Data Framework 2018).

Υπάρχουν τέσσερις κατηγορίες αναλυτικών στοιχείων που πρέπει να λάβουν υπόψη οι οργανώσεις (Big Data Framework 2018):

1) Περιγραφικές αναλύσεις: Περιγραφικές αναλύσεις ή εξόρυξη δεδομένων βρίσκονται στη βάση της αλυσίδας αξίας μεγάλων δεδομένων, αλλά μπορούν να είναι πολύτιμες για την αποκάλυψη μοτίβων που προσφέρουν γνώση. Ένα απλό παράδειγμα περιγραφικών αναλυτικών στοιχείων είναι η επανεξέταση του αριθμού των ανθρώπων που επισκέφθηκε τον ιστότοπο της εταιρείας τους τελευταίους μήνες. Τα περιγραφικά στοιχεία ανάλυσης μπορούν να είναι χρήσιμα στον κύκλο των πωλήσεων, για παράδειγμα, προκειμένου να εντοπίζονται οι εποχιακές τάσεις και να προσαρμόζονται οι αποφάσεις της αγοράς (Big Data Framework 2018).
2) Διαγνωστικές αναλύσεις: Οι διαγνωστικές αναλύσεις χρησιμοποιούνται για ανακάλυψη ή για τον προσδιορισμό της αιτίας που συνέβη κάτι. Σε μια καμπάνια μάρκετινγκ κοινωνικών μέσων, για παράδειγμα, διαγνωστικά αναλυτικά στοιχεία μπορούν να χρησιμοποιηθούν για να καθορίσουν γιατί ορισμένες διαφημίσεις είχαν ως αποτέλεσμα την αύξηση ποσοστών μεταστροφής εν δυνάμει πελατών. Οι διαγνωστικές αναλύσεις παρέχουν πολύτιμες πληροφορίες για τους οργανισμούς, γιατί τους βοηθά να καταλάβουν ποιες αποφάσεις επηρεάζουν την απόδοση της εταιρείας (Big Data Framework 2018).
3) Προγνωστικά αναλυτικά στοιχεία: Οι προβλέψεις των αναλυτικών στοιχείων χρησιμοποιούν Μεγάλα δεδομένα για να προσδιορίσουν τα προηγούμενα πρότυπα πρόβλεψης του μέλλοντος. Από τις τάσεις ή τα πρότυπα στα υπάρχοντα σύνολα δεδομένων, οι προβλέψιμοι αλγόριθμοι υπολογίζουν την πιθανότητα να συμβεί κάποιο συμβάν. Για παράδειγμα, ορισμένες εταιρείες χρησιμοποιώντας προγνωστικές αναλύσεις

για τη βαθμολόγηση των πωλήσεων, υποδεικνύοντας ποιες εισερχόμενες πωλήσεις οδηγούν να έχουν την υψηλότερη πιθανότητα μετατροπής σε πραγματικό πελάτη. Προβλεπόμενη σωστά τα αναλυτικά στοιχεία μπορούν να χρησιμοποιηθούν για την υποστήριξη πωλήσεων, μάρκετινγκ ή για άλλους τύπους σύνθετων προβλέψεων (Big Data Framework 2018).

4) Προειδοποιητικές αναλύσεις: Οι αναλυτικές αναλύσεις είναι το τελευταίο και πιο πολύτιμο επίπεδο. Ενώ τα αναλυτικά στοιχεία των Big Data γενικεύουν ένα θέμα, τα αναλυτικά στοιχεία δίνουν μια εστίαση τύπου λέιζερ για να απαντηθούν συγκεκριμένες ερωτήσεις. Για παράδειγμα, στον τομέα της υγειονομικής περίθαλψης, μπορείτε να διαχειριστείτε καλύτερα τον πληθυσμό των ασθενών χρησιμοποιώντας συνταγογραφικές αναλύσεις για τη μέτρηση του αριθμού των ασθενών που είναι παθολογικά παχύσαρκοι προσθέτοντας φίλτρα για παράγοντες όπως ο διαβήτης και τα επίπεδα LDL χοληστερόλης για να προσδιοριστεί πού να εστιάσει η θεραπεία. Το ίδιο πρότυπο μπορεί να εφαρμοστεί σε σχεδόν οποιαδήποτε βιομηχανία ή πρόβλημα (Big Data Framework 2018).

Ενώ η ανάλυση των δεδομένων στοχεύει στη στήριξη της λήψης αποφάσεων με την ανασκόπηση προηγούμενων δεδομένων (δηλ. Περιγραφικών ή διαγνωστικών αναλύσεων), η ανάλυση αφορά κυρίως τα δεδομένα Big Data για τη βελτιστοποίηση του μέλλοντος (δηλ. προγνωστικά ή προειδοποιητικά αναλυτικά στοιχεία). Για το σκοπό αυτό, τα αναλυτικά στοιχεία κάνουν χρήση (πολύπλοκων) αλγορίθμων για να βρουν μοτίβα στα δεδομένα για να παρέχουν συμβουλές σχετικά με την καλύτερη δυνατή πορεία δράσης για έναν οργανισμό (δηλ. συστάσεις). Ένα δημοφιλές και ευρέως χρησιμοποιούμενο εργαλείο ανάλυσης είναι το Google Analytics που μπορεί να χρησιμοποιήσουν οι οργανισμοί για την πρόβλεψη της κυκλοφορίας ιστότοπων και τη βελτιστοποίηση των διαφημίσεων στο διαδίκτυο (Big Data Framework 2018).

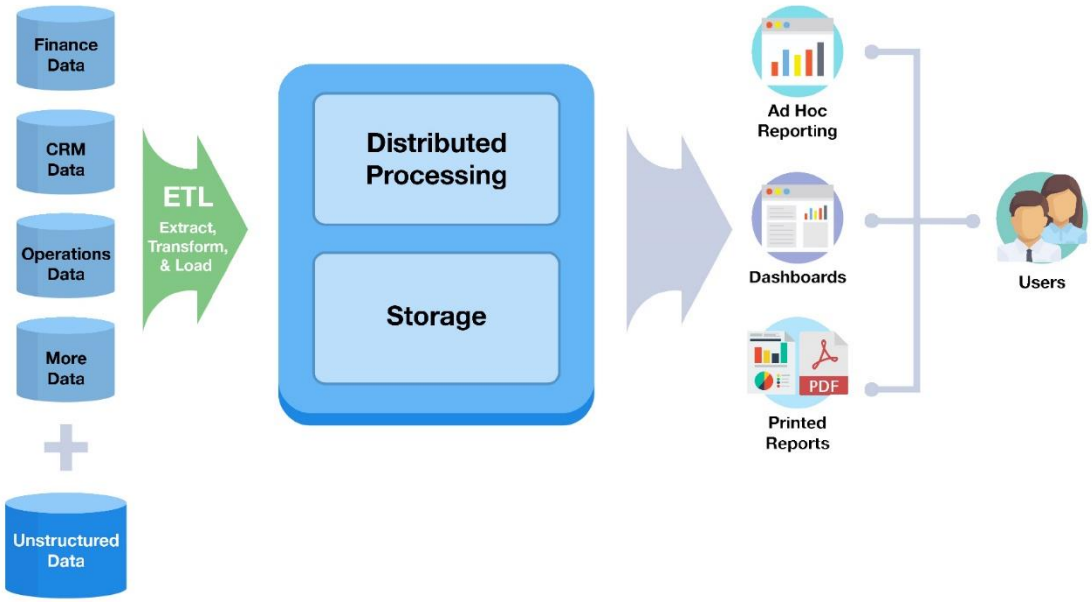
1.7.3 Επιχειρηματική Νοημοσύνη

Η Business Intelligence (BI) περιλαμβάνει τις στρατηγικές και τις τεχνολογίες που χρησιμοποιούνται από τις επιχειρήσεις για την ανάλυση δεδομένων των επιχειρηματικών πληροφοριών. Η Επιχειρηματική Νοημοσύνη χρησιμοποιεί τόσο ανάλυση δεδομένων όσο και έντεκα τεχνικών ανάλυσης για την ενοποίηση και συνοπτική πληροφόρηση που είναι ιδιαίτερα χρήσιμη σε ένα επιχειρηματικό πλαίσιο (Big Data Framework 2018).

Η βασική πρόκληση της Επιχειρηματικής Νοημοσύνης είναι να παγιώσει τα διαφορετικά επιχειρηματικά συστήματα πληροφοριών και τις πηγές δεδομένων σε μια ενιαία ολοκληρωμένη αποθήκη δεδομένων στην οποία αναλύσεις και αναλυτική δεδομένων μπορεί να διεξαχθεί. Μια αποθήκη δεδομένων είναι μια (μεγάλη) κεντρική βάση δεδομένων σε έναν οργανισμό που συνδυάζει μια ποικιλία διαφορετικών βάσεων δεδομένων από διαφορετικές πηγές. Ένα παράδειγμα επιχειρηματικής ευφυΐας θα ήταν να οικοδομήσουμε ένα ταμπλό διαχείρισης που απεικονίζει βασικούς δείκτες απόδοσης (Key Performance Indicator, KPI) σε διαφορετικές χώρες στον κόσμο (Big Data Framework 2018).

1.7.4 Μεγάλης Κλίμακας Δεδομένα και Αναλυτική

Τα Big Data χαρακτηρίζονται από τέσσερα βασικά χαρακτηριστικά - τα τέσσερα V's. Τα μεγάλα δεδομένα χρησιμοποιούν τόσο τεχνικές ανάλυσης δεδομένων όσο και τεχνικές αναλυτικής και συχνά χτίζονται από τα δεδομένα σε αποθήκες δεδομένων επιχείρησης (όπως χρησιμοποιούνται στη BI). Ως εκ τούτου, μπορεί να θεωρηθεί ως το 'Επόμενο βήμα' στην εξέλιξη της Business Intelligence (Big Data Framework 2018).



Εικόνα 10: Big Data περιλαμβάνει μη δομημένα δεδομένα και απαιτεί κατανεμημένη αποθήκευση/ επεξεργασία (Big Data Framework 2018)

Τα Big Data, ωστόσο, απαιτούν μια διαφορετική προσέγγιση από την Επιχειρηματική Ευφυΐα για μια τους παρακάτω λόγους.

- Τα δεδομένα που αναλύονται σε περιβάλλοντα μεγάλων δεδομένων είναι μεγαλύτερα από τα πιο παραδοσιακά BI και συνεπώς οι λύσεις που μπορούν να αντιμετωπίσουν απαιτούν διακριτή και κατανεμημένη αποθήκευση και λύσεις επεξεργασίας (Big Data Framework 2018).
- Τα μεγάλα δεδομένα χαρακτηρίζονται από την ποικιλία των πηγών δεδομένων και περιλαμβάνουν αδόμητα και ή ημιδομημένα δεδομένα. Οι λύσεις Big Data χρειάζονται, για παράδειγμα, να είναι σε θέση να επεξεργαστούν εικόνες αρχείων ήχου. Η διαφορά μεταξύ Big Data και Business Intelligence απεικονίζεται στην Εικόνα 10 (Big Data Framework 2018).

1.7.4.1 Περιγραφική, προγνωστική και κανονιστική Αναλυτική

Η εφαρμογή των αναλυτικών στοιχείων μπορεί να χωριστεί σε τρεις κύριες κατηγορίες, συγκεκριμένα περιγραφικές, προγνωστικές και κανονιστικές αναλύσεις. Περιγραφικές αναλύσεις περιλαμβάνουν τη χρήση προηγμένων τεχνικών για τον εντοπισμό σχετικών δεδομένων και τον εντοπισμό αξιοσημείωτων προτύπων για να περιγράψουμε καλύτερα και να κατανοήσουμε τι συμβαίνει στο σύνολο δεδομένων. Η εξόρυξη δεδομένων, η υπολογιστική διαδικασία της ανακάλυψης προτύπων σε μεγάλα σύνολα δεδομένων, της μηχανικής μάθησης, στατιστικής και συστημάτων βάσεων δεδομένων, περιλαμβάνονται σε αυτήν την κατηγορία. Το περιγραφικό μοντέλο περιγράφει έτσι τι συνέβη, αλλά μια περιγραφή από μόνη της δεν αρκεί ποτέ για τους λήπτες απόφασης. Τα περιγραφικά μοντέλα μπορούν να δώσουν μια σαφή εξήγηση για τον τρόπο που εκτελέστηκαν κάποιες διεργασίες και γιατί συνέβησαν ορισμένα γεγονότα. Οι εταιρείες μπορούν να παρακολουθήσουν παλαιότερες ενέργειες και να δούνε τι συνέβη, ίσως ακόμη και να εντοπίσουν την αιτία, αλλά αυτό που είναι σημαντικό είναι το μέλλον και πώς θα πρέπει να συμπεριφέρονται στο μέλλον. Αυτό απαιτεί μοντέλα πρόβλεψης (Coninck 2017).

Πρόβλεψη των αναλύσεων που συχνά θεωρείται ως ένα υποσύνολο της επιστήμης των δεδομένων επισημοποιούν τον τρόπο αυτο-οργάνωσης ενός προγνωστικού μοντέλου OR μέσω Big Data. Η χρήση δεδομένων, στατιστικών αλγορίθμων και διάφορων τεχνικών εκμάθησης μηχανών για τον

εντοπισμό της πιθανότητας μελλοντικών αποτελεσμάτων βάσει ιστορικών δεδομένων Το μοντέλο προβλέπει επομένως τι είναι πιθανό να συμβεί, με βάση τα διαθέσιμα δεδομένα. Ως εκ τούτου, ένα πλούσιο και εκτεταμένο σύνολο δεδομένων είναι το κλειδί. Η ποσότητα των διαθέσιμων δεδομένων δεν είναι το πρόβλημα, αλλά ο πλούτος των δεδομένων είναι συχνά αμφισβητήσιμος. Αυτό είναι σίγουρα απαραίτητο όταν οι άνθρωποι θέλουν να εκτελέσουν αναλυτικές αναλύσεις. Όταν η εφαρμογή των μαθηματικών και υπολογιστικών αλγορίθμων επιτρέπει στους υπεύθυνους λήψης αποφάσεων για να δούνε όχι μόνο το μέλλον των δικών τους διαδικασιών και ευκαιριών, αλλά παρουσιάζει ακόμη και την καλύτερη πορεία δράσης που πρέπει να ακολουθήσουν για να επωφεληθούν από την προοπτική, με βάση τα δεδομένα. Οι απαιτήσεις για ακριβή και αξιόπιστα αποτελέσματα αναλυτικών αναλύσεων είναι τα υβριδικά δεδομένα, οι ολοκληρωμένες προβλέψεις και συνταγές, λαμβάνοντας υπόψη παρενέργειες, προσαρμοστικούς αλγόριθμους και έναν σαφή μηχανισμό ανάδρασης Το τελικό παράδειγμα ενός κανονιστικού μοντέλου είναι το σύστημα υποστήριξης αποφάσεων (Coninck 2017).

Ο σχετικά νέος τομέας της ανάλυσης ωριμάζει και στρέφεται από μια πρωταρχική εστίαση στις στατιστικές και στα οικονομετρικά μοντέλα περιγραφικής και προγνωστικής ανάλυσης σε κανονιστικά αναλυτικά στοιχεία, με επίκεντρο την Επιχειρησιακή Έρευνα και Διοίκηση τα μοντέλα βελτιστοποίησης και συστήματα υποστήριξης αποφάσεων (Coninck 2017).

1.7.5 Αναλυτική και Επιχειρησιακή Έρευνα

Οι τομείς της Αναλυτικής δεδομένων και της Επιχειρησιακής Έρευνας έχουν πολλά βασικά στοιχεία που έχουν κοινό. Το γεγονός ότι και οι δύο οι τομείς λειτουργούν με ποσοτικά και συχνά μαθηματικά μοντέλα για την επίλυση πραγματικών προβλημάτων ίσως είναι η πιο ορατή ομοιότητα. Άτομα που εργάζονται τακτικά σε ένα από τα δύο πεδία έχουν το ίδιο υπόβαθρο (εφαρμοσμένα μαθηματικά, βιομηχανική μηχανική, επιστήμη υπολογιστών) ή τα ίδια συμφέροντα, οπότε αυτή η αλληλεπίδραση ήταν μία από τις αιτίες της αυξανόμενης κοινής προσέγγισης. Πιο συχνά από ό, τι όχι, οι απαιτούμενες δεξιότητες για τους αιτούντες και στους δύο επιχειρηματικούς τομείς είναι αρκετά ισοδύναμες.

Η Αναλυτική αποτελεί μια διαδικασία end-to-end, οποία περιλαμβάνει 1) την εξόρυξη δεδομένων 2) δημιουργία μιας πρότυπης λύσης 3) απόφαση 4) εφαρμογή εκτέλεση και 5) παραγωγή αξίας.

Αντίθετα η Επιχειρησιακή Έρευνα φαίνεται τις περισσότερες φορές ως ένα εργαλείο (toolbox), το οποίο χρησιμοποιείται για την επίλυση προβλημάτων, ενώ η αναλυτική σαν μια ολοκληρωμένη διαδικασία απαραίτητη για τη λήψη αποφάσεων Παρόλα αυτά η ΕΕ είναι μέρος της κανονιστικής αναλυτικής (prescriptive analytics) με δεδομένους περιορισμούς και σε δεδομένες καταστάσεις προβάλλεται η καλύτερη πιθανή απόφαση. Τελικά η ΕΕ βρίσκεται στην κορυφή των αναλυτικών δυνατοτήτων και της ενδεχόμενης αξίας (Robinson 2014).

1.8 Πακέτα Ανάλυσης Big Data

Η εργασία με δεδομένα και η διεξαγωγή αναλύσεων και αναλυτικής δεδομένων απαιτεί εξειδικευμένες γνώσεις. Στους περισσότερους οργανισμούς, ενδιαφέρονται μόνο για την εύρεση των αποτελεσμάτων και λύσεων σε ορισμένα ερωτήματα– απαιτούν δηλαδή ένα προϊόν δεδομένων. Ένα προϊόν δεδομένων (data product) είναι μια εφαρμογή που εκτελεί λειτουργίες ανάλυσης δεδομένων ή αναλυτικής σε μια συγκεκριμένη είσοδο και κυρίως έχουν μια εύκολη στην κατανοητή διεπαφή χρήστη. Οι χρήστες των προϊόντων δεδομένων δεν χρειάζεται να κατανοήσουν όλους τους υποκείμενους αλγόριθμους, αλλά είναι σε θέση να εκτελέσουν μόνο ορισμένα ερωτήματα για να βρουν συγκεκριμένες απαντήσεις. Επομένως, τα προϊόντα δεδομένων μπορούν να θεωρηθούν ένας από τους βασικούς στόχους των Big Data (Big Data Framework 2018).

Εξαιτίας του αυξανόμενου ενδιαφέροντος για τα BD και της αυξημένης χρήσης του σε οργανώσεις επιχειρήσεων, πολλά προϊόντα δεδομένων έχουν αναπτυχθεί. Οι λύσεις Μεγάλων δεδομένων είναι ένας γρήγορος τρόπος για να ξεκινήσουν οι επιχειρήσεις την αξιοποίηση του δυναμικού της ανάλυσης Big Data, επειδή οι επιχειρήσεις δεν χρειάζεται να αναπτύξουν όλα τα απαιτούμενα προϊόντα δεδομένων εσωτερικά. Το μειονέκτημα των (εμπορικών) λύσεων Big Data είναι ότι συχνά είναι δαπανηρές και είναι δύσκολο να αλλάξουν οποιοσδήποτε από τους υποκείμενους αλγορίθμους της λύσης Big Data (Big Data Framework 2018).

Υπάρχουν πολλές λύσεις Big Data διαθέσιμες στην αγορά και σχεδόν κάθε μεγάλη επιχείρηση IT (Google, Amazon, Microsoft, SAP, κ.λπ.) προσφέρει πλέον μία ή περισσότερες λύσεις Big Data. Επιπλέον, οι νεοσύστατες εταιρείες διαδραματίζουν πολύ σημαντικό ρόλο στην ανάπτυξη λύσεων

Big Data επειδή καταλήγουν σε νέα και καινοτόμα προϊόντα δεδομένων (Big Data Framework 2018).

Ως προσέγγιση με βάση τα δεδομένα, η BI & A (Business Intelligence και Analytics) έχει τις ρίζες της στο μακρόχρονο τομέα διαχείρισης βάσεων δεδομένων. Εξαρτάται σε μεγάλο βαθμό από διάφορες τεχνολογίες συλλογής, εξόρυξης και ανάλυσης δεδομένων (Chaudhuri et al., 2011, Watson et al Wixom 2008). Οι τεχνολογίες και εφαρμογές BI & A που υιοθετούνταν το 2012 στη βιομηχανία, μπορεί να θεωρηθούν BI & A 1.0, όπου τα δεδομένα είναι κατά βάση διαρθρωμένα, συλλέγονται από εταιρείες μέσω διαφόρων συστημάτων παλαιού τύπου, και συχνά αποθηκεύονται σε εμπορικά συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) (Buyya, Calheiros και Dastjerdi 2016)

Εικόνα 11: Πακέτα Big Data (Big Data Framework 2018)

Οι αναλυτικές τεχνικές που χρησιμοποιούνται συνήθως σε αυτά τα συστήματα, που διαδόθηκαν στη δεκαετία του 1990, στηρίζονται κυρίως σε μεθόδους στατιστικής που αναπτύχθηκαν στη δεκαετία του 1970 και τις τεχνικές εξόρυξης δεδομένων που αναπτύχθηκε στη δεκαετία του 1980.

Η διαχείριση δεδομένων και αποθήκευση θεωρείται το θεμέλιο του BI & A 1.0. Σχεδιασμός των mart δεδομένων και εργαλείων για την εξόρυξη, τον μετασχηματισμό και το φορτίο (ETL) είναι απαραίτητα για τη μετατροπή και την ενσωμάτωση δεδομένων για συγκεκριμένες επιχειρήσεις. Απόδοση βάσης δεδομένων, διαδικτυακή αναλυτική επεξεργασία (OLAP) και χρήση εργαλείων

αναφοράς, που βασίζονται σε διαισθητικά, αλλά απλά γραφικά ώστε να διερευνήσουν σημαντικά χαρακτηριστικά δεδομένων. Η διαχείριση των επιδόσεων των επιχειρήσεων πραγματοποιείται με τη βοήθεια scorecards και dashboards που αναλύουν και να απεικονίζουν μια ποικιλία μετρήσεων απόδοσης. Εκτός από αυτές τις καθιερωμένες λειτουργίες αναφοράς επιχειρήσεων, η στατιστική ανάλυση και οι τεχνικές εξόρυξης δεδομένων υιοθετήθηκαν για την ανάλυση της σύνδεσης, τον κατακερματισμό, ταξινόμηση και ανάλυση παλινδρόμησης, ανίχνευσης ανωμαλία και προγνωστικής μοντελοποίησης σε διάφορες επιχειρηματικές εφαρμογές.

1.9 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (Artificial Intelligence) είναι η νοημοσύνη που εμφανίζεται από τις μηχανές, σε αντίθεση με τη φυσική νοημοσύνη (Natural Intelligence) που εμφανίζεται από ανθρώπους και άλλα ζώα. Ο τομέας της Τεχνητής Νοημοσύνης πρωτοεμφανίστηκε από επιστήμονες υπολογιστών στο Dartmouth το 1956 και έχει δει μια εκρηκτική ανάπτυξη, ιδιαίτερα από το 2015 (Big Data Framework 2018).

Ενώ η Τεχνητή Νοημοσύνη μπορεί να θεωρηθεί ως πλήρης τομέας επιστήμης από μόνη της, είναι συνυφασμένη με τα μεγάλα δεδομένα, επειδή ο όγκος και η ποικιλία των πηγών δεδομένων είναι συχνά μαζικός (από άποψη όγκου) και ποικίλος (από άποψη αισθητήρων). Επιπλέον, πολλοί από τους αλγόριθμους στατιστικής και μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση μεγάλων συνόλων δεδομένων είναι παρόμοιοι με αυτούς που χρησιμοποιούνται στην Τεχνητή Νοημοσύνη (Big Data Framework 2018).

Ο τομέας γνώσης της Τεχνητής Νοημοσύνης εξελίχθηκε με την πάροδο των χρόνων για να συμπεριλάβει τους Αλγόριθμους Μηχανικής Μάθησης και τέλος στην έννοια του Deep Learning, που οδηγεί η σημερινή έκρηξη του AI. Κατά τη διάρκεια της εξέλιξης της Τεχνητής Νοημοσύνης, έχουν γίνει οι βασικοί αλγόριθμοι πιο σύνθετοι Εκτός από τις τεχνικές προκλήσεις και την πολυπλοκότητά της, η Τεχνητή νοημοσύνη εγείρει επίσης πολλά κοινωνιολογικά και δεοντολογικά ζητήματα που κάνουν το θέμα ακόμη πιο πολύπλοκο (Big Data Framework 2018).

Ένα δημοφιλές παράδειγμα της εφαρμογής του AI είναι αυτο-οδήγηση αυτοκινήτων. Ο τελικός στόχος της αυτόνομης οδήγησης είναι τα αυτοκίνητα να πρέπει να μιμούνται τις ίδιες συμπεριφορές με τους «φυσικούς» ανθρώπους ενώ οδηγούν (ή κατά προτίμηση ακόμα καλύτερη συμπεριφορά χωρίς ατυχήματα). Τα δεδομένα εισόδου πρέπει να προέρχονται από διαφορετικούς αισθητήρες (μεγάλη ποικιλία) και να επεξεργάζονται χιλιάδες σήματα κάθε δευτερόλεπτο (υψηλή ταχύτητα και υψηλή ένταση) καθώς αλλάζουν οι συνθήκες κυκλοφορίας (Big Data Framework 2018).

1.10 **Υπολογιστικό** Νέφος

Το Υπολογιστικό Νέφος διαδραματίζει κρίσιμο ρόλο στη διαδικασία της BDA, καθώς προσφέρει πρόσβαση προσανατολισμένο στις συνδρομές σε υπολογιστική υποδομή, στα δεδομένα και τις υπηρεσίες εφαρμογών. Ο αρχικός στόχος της BDA ήταν η μόχλευση των αγαθών hardware για την κατασκευή συμπλεγμάτων υπολογιστών και την εξάπλωση της υπολογιστικής ικανότητας για ανίχνευση ιστού και φόρτων εργασίας του συστήματος ευρετηρίου. Λόγω του τεράστιου όγκου του συνόλου δεδομένων, ψάχνοντας για χαμηλότερο κόστος και σφάλμα η υπολογιστική ικανότητα είναι ένας σημαντικός παράγοντας για την εφαρμογή της BDA. Από την άλλη πλευρά, η εφαρμογή του CC υποστηρίχθηκε με τρία μοντέλα υπηρεσιών, τέσσερα μοντέλα ανάπτυξης και πέντε χαρακτηριστικά, που είναι ο αποκαλούμενος ορισμός 3S-4D-5C (Buyya, Calheiros and Dastjerdi 2016).

- Προσανατολισμός υπηρεσιών ή μοντέλων εξυπηρέτησης 3S (SaaS, PaaS και IaaS)
- Προσαρμοσμένη παράδοση ή μοντέλα ανάπτυξης 4D (ιδιωτικό, δημόσιο, κοινοτικό και υβριδικό cloud)
- Κοινόχρηστη υποδομή ή χαρακτηριστικά 5C (κατ 'απαίτηση, ευρεία πρόσβαση δικτύου, πηγή πόρων, γρήγορη ελαστικότητα και μετρημένη εξυπηρέτηση)

Αυτό σημαίνει ότι η φύση των χαρακτηριστικών του cloud το καθιστά την πιο προσιτή υποδομή για πολλές μικρές έως μεσαίες εταιρείες για να μπορέσουν να εφαρμόσουν το BDA.

Το cloud δεν μας επιτρέπει μόνο την εύκολη διαδικασία scale out, δηλαδή τη σύνδεση μηχανημάτων χαμηλότερης απόδοσης για να κάνουν συλλογικά την εργασία μιας προηγμένης

μηχανής. Όταν συζητείται η BDA, αρκετά συχνά η μόνη εστίαση είναι πως γίνεται η διαδικασία scale-out. Αν και ο συνολικός όγκος δεδομένων μπορεί να τείνει να αυξάνεται, ο ημερήσιος όγκος για κάθε μια μεμονωμένη περίπτωση θα μπορούσε να είναι μέτρια και κυμαινόμενη, ή απαιτήσεις επεξεργασίας μεγάλων δεδομένων που απαιτούνται για την Επιχειρηματική Ευφυία μπορεί να διαφέρει από καιρό σε καιρό. Εάν μπορούμε να εκμεταλλευτούμε την ελαστική φύση του cloud, μπορούμε να σώσουμε ένα σημαντικό ποσό του κόστους που οφείλεται στα οφέλη απόσβεσης που παρέχονται από τα συστήματα cloud. Η ελαστική φύση του cloud μπορεί να μειώσει το συνολικό κόστος του υπολογισμού για διαφορετικούς τύπους φόρτων εργασίας μεγάλων δεδομένων, όπως π.χ. παρτίδα, μικρο-παρτίδα, διαδραστική, σε πραγματικό χρόνο και κοντά σε πραγματικό χρόνο (Buyya, Calheiros and Dastjerdi 2016).

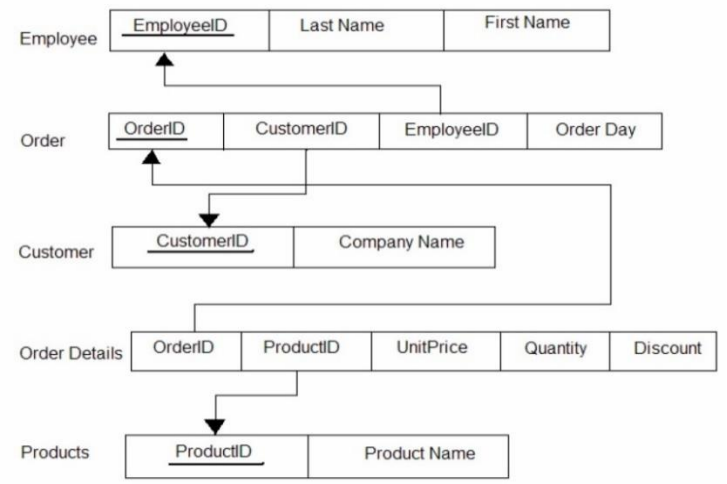
Λαμβάνοντας το Yahoo σαν παράδειγμα ταξινομώντας ένα TB δεδομένα, χρειάστηκαν 3,5 λεπτά για να ολοκληρωθούν 910 κόμβοι το 2008, αλλά χρειάστηκαν μόνο 62 δευτερόλεπτα πάνω από 1460 κόμβους το 2009. Για να γίνει scale out της υπολογιστικής χωρητικότητας θα έχει τεράστια διαφορά ανεξάρτητα από τη βελτίωση του κάθε κόμβου λόγω τεχνολογικής ανάπτυξης. Αυτό σημαίνει ότι η υποδομή του cloud παρέχει υπολογιστική ευελιξία αν το φόρτο εργασίας Big Data ή οι επιχειρηματικές απαιτήσεις το χρειάζονται. Για παράδειγμα, το Amazon Web Service (AWS) προσφέρει στιγμιότυπα στιγμιότυπων σε ένα κλάσμα της κανονικής τιμής. Αν ο φόρτος εργασίας απαιτεί μόνο τη λειτουργία παρτίδας, μπορούμε να αξιοποιήσουμε τα spot της AWS για την αύξηση της υπολογιστικής ικανότητας και την ολοκλήρωση της εργασίας σε πολύ μικρότερο χρονικό διάστημα (Buyya, Calheiros and Dastjerdi 2016).

Μια δημοφιλής και ανοιχτή πλατφόρμα που αναπτύσσεται ευρέως σε μια υποδομή cloud είναι ο Hadoop, του οποίου η εφαρμογή εμπνέεται από το Google MapReduce και το GFS (Buyya, Calheiros and Dastjerdi 2016).

2 Βάσεις Δεδομένων

Με την εξέλιξη της τεχνολογίας και τον αυξημένο όγκο δεδομένων που ρέουν μέσα και έξω από τους οργανισμούς. Αναδύεται το ζήτημα του τρόπου αποθήκευσης, οργάνωσης και διαχείρισης των δεδομένων. Δεδομένου ότι τα δεδομένα που δημιουργούνται αυξάνονται σε όγκο και ποικιλομορφία η αποτελεσματική διαχείριση δεδομένων γίνεται όλο και πιο σημαντική για να αποφευχθούν ζητήματα της κλίμακας και της πολυπλοκότητας. Συγκεκριμένα, με δεδομένα που ενημερώνονται συχνά, από διαφορετικούς ανθρώπους προτείνετε η χρήση συστημάτων διαχείρισης δεδομένων, "Database Management Systems" (DBMS) αντί να διατηρηθούν τα δεδομένα σε μεμονωμένα αρχεία ή σε στατιστικά πακέτα όπως το SAS, SPSS, Stata, και R (Foster, Ghani and Jarmin 2017)

Ένα Σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) είναι μια συλλογή προγραμμάτων που επιτρέπουν στους χρηστές να δημιουργούν και να συντηρούν μια βάση δεδομένων. Τα ΣΔΒΔ συχνά απεικονίζουν τα δεδομένα σε μια δομή τύπου πίνακα. Αυτή είναι η αφορμή για την εισαγωγή του σχεσιακού μοντέλου (relational model), όπου τα δεδομένα απεικονίζονται να αποτελούνται από σχέσεις (Εικόνα 12). Η πρόσβαση σε μια βάση δεδομένων συνήθως επιτυγχάνεται μέσω μιας γλώσσας ερωτήσεων (query language). Η πιο διαδεδομένη, που χρησιμοποιείται από τα περισσότερα ΣΔΒΔ, είναι η SQL[12].



[12] https://bit.ly/2YQ9GDQ (7/8/2019)

Εικόνα 12: Παράδειγμα του σχεσιακού μοντέλου (https://bit.ly/2YQ9GDQ)

Στη συνέχεια θεωρείται σημαντικό να ειπωθεί πότε και για ποιο λόγο θα χρησιμοποιείται ένα εργαλείο ΣΔΒΔ. Για αυτό το λόγο παρουσιάζονται τρία παραδείγματα σετ δεδομένων (Foster, Ghani and Jarmin 2017).

1. 10.000 αρχεία που περιγράφουν ερευνητικές επιχορηγήσεις, καθένα από τα οποία προσδιορίζει τον κύριο ερευνητή, το θεσμό, τον ερευνητικός τομέα, τον τίτλο της πρότασης, την ημερομηνία χορήγησης και το ποσό χρηματοδότησης με διαχωρισμένη με κόμμα τιμή.
2. 10 εκατομμύρια εγγραφές σε διάφορες μορφές από τις υπηρεσίες χρηματοδότησης, τα API ιστού και τις θεσμικές πηγές που περιγράφουν τους ανθρώπους, τις επιχορηγήσεις, τους οργανισμούς χρηματοδότησης και τα διπλώματα ευρεσιτεχνίας.
3. 10 δισεκατομμύρια μηνύματα Twitter και συναφή μεταδεδομένα-γύρω από 10 terabytes (1013 bytes) συνολικά και αυξάνονται σε ένα terabyte το μήνα.

Στην περίπτωση του συνόλου δεδομένων 1 (10.000 εγγραφές που περιγράφουν την έρευνα), μπορεί να είναι εφικτό να παραμείνουν τα δεδομένα στον αρχικό τους φάκελο, να χρησιμοποιηθούν υπολογιστικά φύλλα, πίνακες περιστροφής ή να γραφτούν προγράμματα σε γλώσσα σεναρίων, όπως η Python ή η R για να θέσει ερωτήσεις από αυτά τα αρχεία. Για παράδειγμα, κάποιος εξοικειωμένος με τέτοιες γλώσσες μπορεί να δημιουργήσει γρήγορα ένα σενάριο για την εξαγωγή από το σύνολο δεδομένων 1 όλων των επιδοτήσεων που χορηγούνται σε έναν ερευνητή, να υπολογίσει το μέσο μέγεθος επιχορήγησης και να υπολογίσει τις επιχορηγήσεις που πραγματοποιήθηκαν στο έτος σε διάφορους τομείς. Πιο αναλυτικά τα "σενάρια" ("scripts") είναι διακριτά από τον βασικό κώδικα της εφαρμογής, καθώς γράφονται συνήθως σε διαφορετική γλώσσα και συχνά δημιουργούνται ή τροποποιούνται από τον τελικό χρήστη (Foster et. al.,2017). Ωστόσο, η προσέγγιση αυτή παρουσιάζει επίσης μειονεκτήματα. Τα σενάρια παρέχουν έμφυτο έλεγχο στη δομή του αρχείου. Αυτό σημαίνει ότι αν εισάγονται νέα δεδομένα σε διαφορετική μορφή, πρέπει αυτά να είναι ενημερωμένα. Επίσης τα σενάρια δεν μπορούν ακόμα να «τρέξουν» το πρόσφατα αποκτηθέν αρχείο. Ένα ακόμα μειονέκτημα είναι ότι τα σενάρια μπορούν επίσης εύκολα να γίνουν αδικαιολόγητα αργά όσο αυξάνεται ο όγκος δεδομένων. Ένα σενάριο Python ή R δεν θα χρειαστεί πολύ για να αναζητήσει μια λίστα με 1.000 υποτροφίες για να βρεθούν εκείνες

που αφορούν ένα συγκεκριμένο ίδρυμα. Στην περίπτωση βέβαια που οι πληροφορίες αφορούν 1 εκατομμύριο επιχορηγήσεις, και για κάθε επιχορήγηση πρέπει να αναζητήσετε μια λίστα 100.000 ερευνητών και για κάθε ερευνητή, θέλετε να αναζητήσετε μια λίστα με 10 εκατομμύρια έγγραφα για να διαπιστώσετε εάν ο εν λόγω ερευνητής αναγράφεται ως συγγραφέας κάθε χαρτιού. Τώρα πρέπει να γίνουν 1.000.000 × 100.000 × 10.000.000 = 10^{18} συγκρίσεις. Το απλό σενάριο μπορεί τώρα να τρέξει για ώρες ή ακόμα και ημέρες. Η διαδικασία μπορεί να επιταχύνετε με τη διαδικασία αναζήτησης δημιουργώντας δείκτες, έτσι, για παράδειγμα, όταν χορηγηθεί επιχορήγηση, μπορείτε να βρείτε τους συνδεδεμένους ερευνητές σε συνεχή χρόνο και όχι σε χρόνο ανάλογο με τον αριθμό των ερευνητών. Ωστόσο, η κατασκευή τέτοιων δεικτών είναι χρονοβόρες και επιρρεπείς σε

Πίνακας 3: Πότε χρησιμοποιούνται οι διαφορετικές τεχνολογίες διαχείρισης και ανάλυσης δεδομένων (Foster, Ghani and Jarmin 2017)

σφάλματα (Foster, Ghani and Jarmin 2017).

Text files, spreadsheets, and scripting language
• Your data are small • Your analysis is simple • You do not expect to repeat analyses over time
Statistical packages
• Your data are modest in size • Your analysis maps well to your chosen statistical package
Relational database
• Your data are structured • Your data are large • You will be analyzing changed versions of your data over time • You want to share your data and analyses with others
NoSQL database
• Your data are unstructured • Your data are extremely large

Για τους λόγους αυτούς, η χρήση αποκλειστικά γλωσσών ενεργειών σεναρίων για ανάλυση δεδομένων σπανίως συνιστάται. Επίσης είναι ανέφικτο οι υπολογισμοί ανάλυσης να μπορούν να εκτελεστούν σε συστήματα βάσεων δεδομένων για αυτό το λόγο μια γλώσσα προγραμματισμού θα είναι συχνά απαραίτητη. Αλλά πολλοί οι υπολογισμοί πρόσβασης και χειραγώγησης των δεδομένων αντιμετωπίζονται καλύτερα σε μια βάση δεδομένων (Foster, Ghani and Jarmin 2017).

Οι ερευνητές στις κοινωνικές επιστήμες χρησιμοποιούν συχνά στατιστικά πακέτα όπως R, SAS, SPSS και Stata για. Επειδή αυτά τα συστήματα ενσωματώνουν κάποια ακατέργαστη διαχείριση δεδομένων, στατιστική ανάλυση και τις δυνατότητες γραφικών σε ένα ενιαίο πακέτο, ένας

ερευνητής μπορεί συχνά να πραγματοποιήσει ένα έργο ανάλυσης δεδομένων μέτριου μεγέθους μέσα στο ίδιο περιβάλλον (Foster, Ghani and Jarmin 2017). Για αυτό το λόγο γίνεται συνοπτική παρουσίαση των πακέτων που χρησιμοποιούνται ανάλογα με τον όγκο και το είδος των δεδομένων στον Πίνακας 3.

2.1 Διαφορετικοί τύποι δεδομένων

Στην επιστήμη των υπολογιστών, μια δομή δεδομένων είναι ένας τρόπος για την οργάνωση και την αποθήκευση δεδομένων σε έναν υπολογιστή ώστε να μπορεί να έχει πρόσβαση και να τροποποιείται αποτελεσματικά. Πιο συγκεκριμένα, μια δομή δεδομένων είναι μια συλλογή τιμών δεδομένων, οι σχέσεις μεταξύ τους και οι λειτουργίες ή οι πράξεις που μπορούν να εφαρμοστούν στα δεδομένα.

Για την ανάλυση των δεδομένων, είναι σημαντικό να αναφερθούν ότι υπάρχουν τρεις συνήθεις τύποι (Big Data Framework 2018):

Δομημένα δεδομένα: Τα δομημένα δεδομένα είναι δεδομένα που τηρούν ένα προκαθορισμένο μοντέλο δεδομένων και είναι επομένως εύκολο να αναλυθούν. Τα δομημένα δεδομένα συμμορφώνονται με μια μορφή πίνακα με σχέση μεταξύ των διαφόρων σειρών και στηλών. Κοινά παραδείγματα δομημένων δεδομένων είναι αρχεία Excel ή βάσεις δεδομένων SQL. Καθένα από αυτά έχει δομημένες σειρές και στήλες που μπορούν να ταξινομηθούν.

Μη δομημένα δεδομένα: Τα μη δομημένα δεδομένα είναι πληροφορίες που είτε δεν έχουν προκαθορισμένο μοντέλο δεδομένων είτε δεν είναι οργανωμένα με προκαθορισμένο τρόπο. Μη δομημένες πληροφορίες συνήθως μεγάλα κείμενα, τα οποία ενδέχεται να περιέχουν δεδομένα όπως ημερομηνίες, αριθμούς και γεγονότα. Αυτό έχει ως αποτέλεσμα παρατυπίες και ασάφειες που δυσκολεύουν τη χρήση παραδοσιακών προγραμμάτων σε σύγκριση με τα δεδομένα που είναι αποθηκευμένα σε δομές βάσεων δεδομένων. Τα συνηθισμένα παραδείγματα μη δομημένων δεδομένων περιλαμβάνουν αρχεία ήχου, βίντεο ή No-SQL βάσεων δεδομένων.

Ημι-δομημένα δεδομένα: Τα ημι-δομημένα δεδομένα είναι μια μορφή δομημένων δεδομένων που δεν προσαρμόζονται σύμφωνα με την τυπική δομή των μοντέλων δεδομένων που σχετίζονται με

σχεσιακές βάσεις δεδομένων ή με άλλες μορφές πίνακες δεδομένων, αλλά παρόλα αυτά να περιέχουν ετικέτες ή άλλους δείκτες για να διαχωριστούν τα σημασιολογικά στοιχεία και να επιβάλλουν ιεραρχίες αρχείων και πεδίων εντός των δεδομένων. Επομένως, είναι επίσης γνωστή ως αυτο-περιγραφόμενη δομή. Παραδείγματα ημι-δομημένων δεδομένων περιλαμβάνουν JSON και XML είναι μορφές ημι-δομημένων δεδομένων.

Οι περισσότερες «παραδοσιακές» τεχνικές ανάλυσης και ανάλυσης δεδομένων (συμπεριλαμβανομένων των περισσότερων επιχειρησιακών ευφυών λύσεων) έχουν την ικανότητα να επεξεργάζονται δομημένα δεδομένα. Η επεξεργασία μη δομημένων ή ημιδομημένων δεδομένων είναι ωστόσο πολύ πιο πολύπλοκη και απαιτεί ξεχωριστές λύσεις για ανάλυση. Μία τελευταία κατηγορία δεδομένων είναι τα μεταδεδομένα. Από τεχνική άποψη, δεν αποτελούν ξεχωριστή δομή δεδομένων, αλλά είναι ένα από τα πιο σημαντικά στοιχεία για ανάλυση μεγάλων δεδομένων. Τα μεταδεδομένα είναι δεδομένα σχετικά με τα δεδομένα. Παρέχει πρόσθετες πληροφορίες σχετικά με ένα συγκεκριμένο σύνολο δεδομένων. Σε ένα σύνολο φωτογραφιών, για παράδειγμα, τα μεταδεδομένα θα μπορούσαν να περιγράψουν πότε και πού πραγματοποιήθηκαν φωτογραφίες. Τα μεταδεδομένα παρέχουν έπειτα πεδία για ημερομηνίες και τοποθεσίες που τα ίδια, μπορούν να θεωρηθούν δομημένα δεδομένα. Για το λόγο αυτό, τα μεταδεδομένα χρησιμοποιούνται συχνά σαν λύσεις Big Data για αρχική ανάλυση (Big Data Framework 2018).

Εικόνα 13: Τέσσερα είδη δεδομένων (Big Data Framework 2018)

2.2 Τύποι Βάσεων Δεδομένων

Βάση δεδομένων είναι μια δομημένη συλλογή δεδομένων σχετικά με οντότητες και τις σχέσεις τους. Υποδείγματα αντικειμένων πραγματικού κόσμου-και δύο οντοτήτων είναι (π.χ., υποτροφίες, ερευνητές, πανεπιστήμια) και τις σχέσεις (π.χ., "Steven Weinberg "εργάζεται στο" Πανεπιστήμιο του Τέξας στο Ώστιν ") - και αποδίδει δομή με τρόπους που επιτρέπουν σε αυτές

τις οντότητες και τις σχέσεις να είναι ερωτήματα για ανάλυση. Ένα σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) είναι ένα σετ λογισμικού σχεδιασμένο για την ασφαλή αποθήκευση και αποτελεσματική διαχείριση βάσεων δεδομένων, και ακόμα για να βοηθήσει με τη συντήρηση και την ανακάλυψη των σχέσεων που αντιπροσωπεύει αυτή η βάση δεδομένων. Γενικά, ένα ΣΔΒΔ περιλαμβάνει τρία βασικά στοιχεία, όπως φαίνεται στον Πίνακας 4 το μοντέλο δεδομένων του (το οποίο καθορίζει τον τρόπο με τον οποίο τα δεδομένα αντιπροσωπεύονται), τη γλώσσα του ερωτήματος (που καθορίζει τον τρόπο με τον οποίο ο χρήστης αλληλεπιδρά με τα δεδομένα) και υποστήριξη για συναλλαγές και ανάκτηση συνθηκών (για να εξασφαλιστεί η αξιόπιστη εκτέλεση παρά τις αποτυχίες του συστήματος) (Foster, Ghani and Jarmin 2017)

Πίνακας 4: Κύρια σημεία των ΣΔΒΔ (Foster, Ghani and Jarmin 2017)

	Data model	**Query language**	**Transactions, crash recovery**
User-facing	For example: relational, semi-structured	For example: SQL (for relational), XPath (for semi-structured)	Transactions
Internal	Mapping data to storage systems; creating and maintaining indices	Query optimization and evaluation; consistency	Locking, concurrency control, recovery

Τα σχεσιακά DBMSs είναι τα πιο ευρέως χρησιμοποιούμενα και ώριμα συστήματα, και θα είναι αποτελεί τη βέλτιστη λύση αναλύσεων. Επιτρέπουν την αποτελεσματική αποθήκευση, οργάνωση και ανάλυση μεγάλων ποσοτήτων δεδομένων πίνακα: δεδομένα οργανωμένα σε πίνακα όπου οι σειρές αντιπροσωπεύουν οντότητες (π.χ. επιχορηγήσεις έρευνας) και οι στήλες αντιπροσωπεύουν χαρακτηριστικά αυτών των οντοτήτων (π.χ. κύριος ερευνητής, επίπεδο χρηματοδότησης). Στη συνέχεια, η γλώσσα SQL, Structured Query Language μπορεί να χρησιμοποιηθεί για να εκτελέσει

Type	**Examples**	**Advantages**	**Disadvantages**	**Uses**
Relational database	MySQL, PostgreSQL, Oracle, SQL	Consistency (ACID)	Fixed schema; typically harder to scale	Transactional systems: order processing, retail, hospitals, etc.

Πίνακας 5: Τύποι βάσεων δεδομένων (πρώτη σειρά) και τύποι βάσεων NoSQL (υπόλοιπες σειρές) (Foster, Ghani and Jarmin 2017)

		throughput	higher-level queries	
Column store	Cassandra, HBase	Same as key-value; distributed; better compression at column level	Not immediately consistent; using all columns is inefficient	Large-scale analysis
Document store	CouchDB, MongoDB	Index entire document (JSON)	Not immediately consistent; no higher-level queries	Web applications
Graph database	Neo4j, InfiniteGraph	Graph queries are fast	Difficult to do non-graph analysis	Recommendation systems, networks, routing

ένα ευρύ φάσμα αναλύσεων, οι οποίες εκτελούνται με υψηλή απόδοση λόγω εξελιγμένων τεχνικών ευρετηρίασης και προγραμματισμού επερωτήσεων (Πίνακας 5).

Τα εναλλακτικά συστήματα DBMS του NoSQL υποκινήθηκαν συνήθως από μια επιθυμία για την κλιμάκωση των ποσοτήτων δεδομένων και / ή του αριθμού των χρηστών που μπορούν να υποστηρίζονται και / ή να αντιμετωπίζουν μη δομημένα δεδομένα που δεν αντιπροσωπεύεται εύκολα σε πίνακα. Για παράδειγμα, ένα σύστημα βασικής τιμής (key value store) μπορεί να οργανώσει μεγάλο αριθμό αρχείων, καθένα από τα οποία συνδέει ένα αυθαίρετο κλειδί με αυθαίρετη τιμή. Αυτές οι αποθήκες δεδομένων και ειδικότερα παραλλαγές που ονομάζονται καταστήματα εγγράφων που επιτρέπουν την αναζήτηση κειμένου στις αποθηκευμένες τιμές, χρησιμοποιούνται ευρέως για την οργάνωση και τη διεκπεραίωση των δισεκατομμυρίων αρχείων που μπορούν να ληφθούν από τις μηχανές αναζήτησης (Foster, Ghani and Jarmin 2017).

Σχεσιακές (SQL) και μη σχεσιακές βάσεις δεδομένων (NoSQL) μπορούν επίσης να χρησιμοποιηθούν μαζί. Μια κοινή λύση αποθήκευσης είναι η πρώτη φόρτωση όλων των δεδομένων σε μια μεγάλη βάση δεδομένων NoSQL. Αυτή η προσέγγιση κάνει όλα διαθέσιμα δεδομένα μέσω μιας κοινής (αν και περιορισμένης) διεπαφής ερωτήματος. Ο ερευνητής μπορεί στη συνέχεια να εξάγει από αυτή τη βάση δεδομένων τα συγκεκριμένα στοιχεία που παρουσιάζουν ενδιαφέρον, φορτώνοντας αυτά τα στοιχεία σε ένα σχεσιακό ΣΔΒΔ, (π.χ. βάση δεδομένων γραφημάτων), ή ένα στατιστικό πακέτο για λεπτομερέστερη ανάλυση. Ως μέρος της διαδικασίας φόρτωσης δεδομένων από τη βάση δεδομένων NoSQL σε μια σχεσιακή βάση δεδομένων, ο ερευνητής θα ορίσει αναγκαστικά τα σχήματα, τις σχέσεις μεταξύ οντοτήτων και ούτω καθεξής. Τα αποτελέσματα της ανάλυσης μπορούν να αποθηκευτούν σε μια σχεσιακή βάση δεδομένων ή πίσω στην αποθήκη της NoSQL (Foster, Ghani and Jarmin 2017).

2.2.1 Σχεσιακό Σύστημα Διαχείρισης Βάσεων Δεδομένων

Τα σχεσιακά DBMSs εφαρμόζουν το μοντέλο σχεσιακών δεδομένων, στο οποίο τα δεδομένα αντιπροσωπεύονται ως σύνολα αρχείων που οργανώνονται σε πίνακες. Η κύρια έννοια που αποτελεί τη βάση του σχεσιακού μοντέλου δεδομένων είναι ένας πίνακας (επίσης αναφέρεται ως σχέση): ένα σύνολο γραμμών (επίσης γνωστό ως πλειάδες, εγγραφές ή παρατηρήσεις), το καθένα

με τις ίδιες στήλες (επίσης που αναφέρονται ως πεδία, ιδιότητες ή μεταβλητές). Μια βάση δεδομένων αποτελείται από πολλαπλούς πίνακες (Foster, Ghani and Jarmin 2017).

Η χρήση του σχεσιακού μοντέλου δεδομένων προβλέπει φυσική ανεξαρτησία: ένας δεδομένος πίνακας μπορεί να αποθηκευτεί με πολλούς διαφορετικούς τρόπους. Ερωτήματα SQL γράφονται σύμφωνα με τη λογική αναπαράσταση των πινάκων (δηλ. τον ορισμό του σχήματος τους). Κατά συνέπεια, ακόμη και αν η δομή των δεδομένων αλλάξει (π.χ., χρησιμοποιείται διαφορετική διάταξη) για να αποθηκευτούν τα δεδομένα στο δίσκο ή δημιουργήθηκε ένας νέος δείκτης για να επιταχυνθεί πρόσβαση σε ορισμένα ερωτήματα), τα ερωτήματα δεν χρειάζεται να αλλάξουν. Ακόμα ένα πλεονέκτημα του σχεσιακού μοντέλου δεδομένων είναι ότι, αφού ένας πίνακας είναι μια δομή, με τη μαθηματική έννοια, απλές και διαισθητικές λειτουργίες (π.χ., ένωση, τομή) μπορεί να χρησιμοποιηθούν για χειρισμό των δεδομένων. Με αυτό τον τρόπο μπορεί εύκολα, να καθοριστεί η διασταύρωση από δύο σχέσεις. Η βάση δεδομένων περαιτέρω διασφαλίζει ότι τα δεδομένα συμμορφώνονται με το μοντέλο (π.χ. τύποι δεδομένων, κλειδί μοναδικότητας, σχέσεις οντοτήτων), παρέχοντας ουσιαστικά την ποιότητα του πυρήνα ασφάλειας (Foster, Ghani and Jarmin 2017).

2.2.2 Δομημένη Γλώσσα ερωτημάτων SQL

Η SQL (structured query language) αποτελεί σήμερα την πιο διαδεδομένη γλώσσα διαχείρισης σχεσιακών βάσεων δεδομένων. Η SQL παρέχει δυνατότητες για (Foster, Ghani and Jarmin 2017).

- τον ορισμό, τη διαγραφή και τη μεταβολή πινάκων και κλειδιών,
- τη σύνταξη ερωτήσεων (queries),
- την εισαγωγή, διαγραφή και μεταβολή στοιχείων,
- τον ορισμό όψεων (views) πάνω στα δεδομένα,
- τον ορισμό δικαιωμάτων πρόσβασης,
- τον έλεγχο της ακεραιότητας των στοιχείων,
- τον έλεγχο συναλλαγών (transaction)

2.2.3 Μη Δομημένη Γλώσσα Ερωτημάτων NoSQL

Ενώ τα σχεσιακά ΣΔΒΔ κυριαρχούσαν στον κόσμο της βάσης δεδομένων για πολλές δεκαετίες, άλλες τεχνολογίες βάσεων δεδομένων υπάρχουν και πράγματι έχουν γίνει δημοφιλείς για διάφορες κατηγορίες εφαρμογών τα τελευταία χρόνια. Αυτές οι εναλλακτικές τεχνολογίες ήταν συνήθως λόγω της επιθυμίας κλιμάκωσης των ποσοτήτων δεδομένων και / ή αριθμού των χρηστών που μπορούν να υποστηριχθούν ή / και να υποστηρίξουν εξειδικευμένους τύπους δεδομένων (π.χ. μη δομημένα δεδομένα, γραφήματα) (Foster, Ghani and Jarmin 2017).

Πίνακας 6: Βασικές διαφορές μεταξύ SQL και NoSQL (Foote 2016).

SQL	
Πλεονεκτήματα	Μειονεκτήματα
Λειτουργούν με δομημένα δεδομένα	Επεκτείνονται μόνο κατακόρυφα
Υποστηρίζουν συναλλαγές ACID και την ένωση οντοτήτων (join)	Η κανονικοποίηση των δεδομένων και τα πολλά joins επηρεάζουν την ταχύτητα
Ενσωματωμένη ακεραιότητα δεδομένων (data integrity) και μεγάλη διαλειτουργικότητα με πολλά συστήματα	Πρόβλημα με την επεξεργασία ημιδομημένων δεδομένων
Δυνατότητα περιορισμών (constraints)	
Απλή γλώσσα ανάκτησης δεδομένων	
NoSQL	
Πλεονεκτήματα	Μειονεκτήματα
Επεκτείνονται οριζόντια και επεξεργάζονται δομημένα και ημιδομημένα δεδομένα.	Περιορισμένη λειτουργία joins πινάκων
Λειτουργούν χωρίς την ύπαρξη μοντέλου οντοτήτων – συσχετίσεων	Τα δεδομένα δεν είναι κανονικοποιημένα

Είναι συνεχώς διαθέσιμες	Δεν έχουν ενσωματωμένη ακεραιότητα δεδομένων (data integrity)
Πολλές βάσεις NoSQL είναι ανοιχτού κώδικα	Λιγότερο συνεπείς, δεν εφαρμόζουν τέλεια το
Πολλές βάσεις υποστηρίζουν συναλλαγές ACID	

2.2.3.1 Θεώρημα CAP

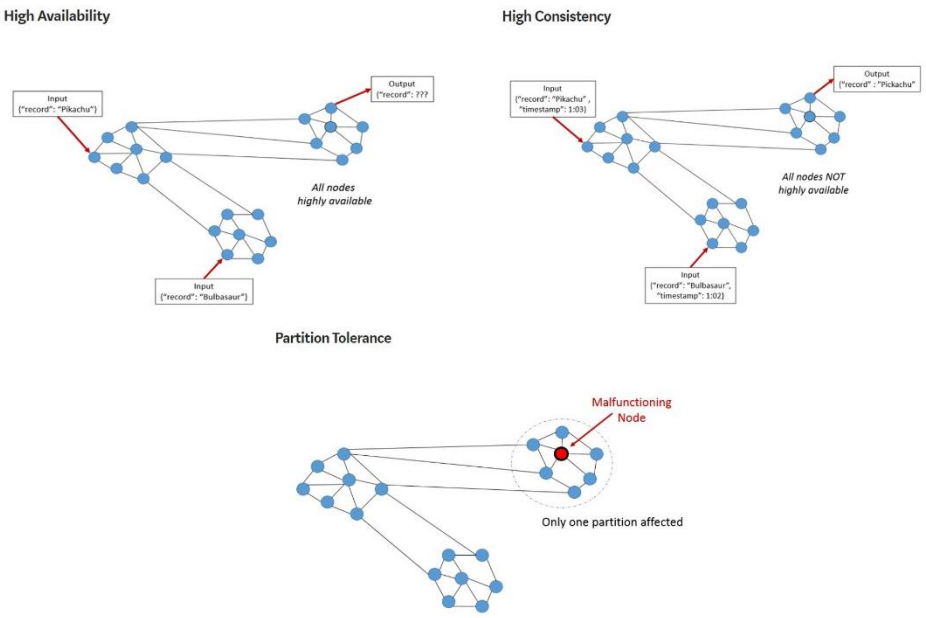
Για πολλά χρόνια, οι μεγάλοι πωλητές σχεσιακών βάσεων δεδομένων όπως η Oracle, IBM, Sybase, και σε μικρότερο βαθμό η Microsoft αποτέλεσαν το στυλοβάτη αποθήκευσης δεδομένων. Κατά τη διάρκεια της έκρηξης του Διαδικτύου, οι νεοσύστατες επιχειρήσεις αναζητούν εναλλακτικές λύσεις χαμηλού κόστους σε εμπορικά σχεσιακά ΣΔΒΔ MySQL και PostgreSQL. Ωστόσο, αυτά τα συστήματα αποδείχθηκαν ανεπαρκή για μεγάλες τοποθεσίες, καθώς δεν μπορούσαν να αντεπεξέλθουν σε μεγάλες αιχμές κυκλοφορίας, για παράδειγμα, όταν πολλοί πελάτες ξαφνικά ήθελαν να παραγγείλουν το ίδιο στοιχείο, δηλαδή δεν υπήρχε κλιμάκωση (Foster, Ghani and Jarmin 2017)

Το θεώρημα Brewer (προς τιμή του επιστήμονα υπολογιστών Eric Brewer) ή αλλιώς CAP theorem (Consistency - Availability – Partition Tolerance), υποδεικνύει ότι τα κατανεμημένα συστήματα αποθήκευσης μπορούν να ικανοποιούν ταυτόχρονα μόνο 2 από τα 3 χαρακτηριστικά τα οποία είναι: συνέπεια, διαθεσιμότητα, ανοχή κατάτμησης.

Στην Εικόνα 14 φαίνεται ποια χαρακτηριστικά ικανοποιούν οι πιο δημοφιλείς βάσεις δεδομένων. Για παράδειγμα, ο SQL Server της Microsoft ικανοποιεί την συνέπεια και διαθεσιμότητα αλλά όχι τον διαχωρισμό των δεδομένων (Nazrul 2018). Η συνέπεια δείχνει ότι όλοι οι υπολογιστές βλέπουν τα ίδια δεδομένα την ίδια ώρα. Στο παρελθόν, όταν έπρεπε να επεκταθούν οι αποθηκευτικοί και επεξεργαστικοί πόροι ενός συστήματος, η επέκταση γινόταν κατακόρυφα. Ωστόσο, με την πρόοδο της παράλληλης επεξεργασίας και των κατανεμημένων συστημάτων, η επέκταση πλέον γίνεται οριζόντια, προσθέτοντας δηλαδή περισσότερα υπολογιστικά συστήματα ώστε να εργάζονται παράλληλα.

Με βάση αυτή την λογική αναπτύσσει τα λογισμικά της η κοινότητα του Apache όπως το Spark και το Hadoop. Το θεώρημα Brewer είναι πολύ σημαντικό στην επιλογή του κατάλληλου συστήματος, καθώς η κάθε βάση δεδομένων ικανοποιεί διαφορετικά χαρακτηριστικά.

• *Συνέπεια (Consistency)*: Ο όρος αυτός δηλώνει ότι όλοι οι κόμβοι βλέπουν ταυτόχρονα τα ίδια δεδομένα. Με απλά λόγια, η εκτέλεση μιας εντολής ανάκτησης δεδομένων θα επιστρέψει την τιμή της πιο πρόσφατης εγγραφής, με αποτέλεσμα όλοι οι κόμβοι να επιστρέψουν τα ίδια αποτελέσματα. Ένα σύστημα έχει συνέπεια εάν μια συναλλαγή ξεκινά με το σύστημα σε συνεπή κατάσταση και τελειώνει με το σύστημα σε συνεπή κατάσταση. Σε αυτό το μοντέλο, ένα σύστημα μπορεί να μετατοπιστεί σε μια ασυνεπή κατάσταση κατά τη διάρκεια μιας συναλλαγής, αλλά εάν υπάρξει κάποιο σφάλμα σε οποιοδήποτε στάδιο της διαδικασίας, ολόκληρη η συναλλαγή ακυρώνεται (Nazrul 2018)



Εικόνα 14:Απεικόνιση χαρακτηριστικών CAP (Nazrul 2018)

• *Διαθεσιμότητα (Availability):* Αυτός ο όρος δηλώνει ότι κάθε ερώτημα λαμβάνει απάντηση στην επιτυχία / αποτυχία. Η επίτευξη διαθεσιμότητας σε ένα κατανεμημένο σύστημα απαιτεί το

σύστημα να παραμείνει λειτουργικό 100% του χρόνου. Κάθε πελάτης λαμβάνει μια απάντηση, ανεξάρτητα από την κατάσταση κάθε μεμονωμένου κόμβου στο σύστημα. Αυτή η μέτρηση είναι τετριμμένη για μέτρηση: είτε μπορούν να υποβληθούν εντολές ανάγνωσης / εγγραφής, είτε δεν μπορούν. Ως εκ τούτου, οι βάσεις δεδομένων είναι ανεξάρτητες από το χρόνο, καθώς οι κόμβοι πρέπει να είναι διαθέσιμοι σε απευθείας σύνδεση ανά πάσα στιγμή. Η υψηλή διαθεσιμότητα δεν είναι εφικτή κατά την ανάλυση δεδομένων ροής με υψηλή συχνότητα (Nazrul 2018).

Ανοχή κατάτμησης (Partition Tolerance): Ο όρος αυτός δηλώνει ότι το σύστημα συνεχίζει να λειτουργεί, παρά τον αριθμό μηνυμάτων που καθυστερούν από το δίκτυο μεταξύ των κόμβων. Ένα σύστημα που είναι ανεκτικό σε κατάτμηση μπορεί να υποστηρίξει οποιαδήποτε αποτυχία κόμβου που δεν έχει ως αποτέλεσμα την αποτυχία ολόκληρου του δικτύου. Τα αρχεία δεδομένων αντιγράφονται επαρκώς σε άλλους κόμβους ώστε τα αρχεία να μην χαθούν σε περίπτωση αποτυχίας ενός κόμβου. Στα σύγχρονα κατανεμημένα συστήματα, η ανοχή κατάτμησης δεν είναι επιλογή, είναι αναγκαιότητα. Ως εκ τούτου, πρέπει το σύστημα να ισορροπεί μεταξύ συνέπειας και διαθεσιμότητας (Nazrul 2018)

Τα κατανεμημένα συστήματα μας επιτρέπουν να επιτύχουμε ένα υψηλό επίπεδο υπολογιστικής ισχύος και διαθεσιμότητας το οποίο δεν ήταν διαθέσιμο στο παρελθόν.

Τα συστήματα αυτά έχουν υψηλότερες επιδόσεις, χαμηλότερη καθυστέρηση και είναι διαθέσιμα κάθε στιγμή. Ένα τεράστιο πλεονέκτημά τους είναι ότι τα κατανεμημένα συστήματα λειτουργούν με υπολογιστές προσιτού κόστους (commodity hardware). Ωστόσο, τα κατανεμημένα συστήματα είναι πιο πολύπλοκα από τα μεμονωμένα υπολογιστικά συστήματα. Η κατανόηση της πολυπλοκότητας τους και η πραγματοποίηση των κατάλληλων επιλογών όσον αφορά την παραδοχή του CAP θεωρήματος, είναι το τίμημα της επέκτασης σε οριζόντια κλίμακα (Nazrul 2018)

3

3 Αρχιτεκτονική Μεγάλων Δεδομένων

Αυτό το κεφάλαιο παρέχει μια επισκόπηση θεμελιωδών και βασικών θεματικών πεδίων που σχετίζονται με την Αρχιτεκτονική Μεγάλων Δεδομένων. Στο κεφάλαιο θα γίνει περιγραφή της επισκόπησης NIST (National Institute of Standards and Technology) Big Data Reference Architecture (NBDRA), και στη συνέχεια θα αναφερθούν τα βασικά στοιχεία της κατανεμημένης αποθήκευσης / επεξεργασίας. Το κεφάλαιο θα ολοκληρωθεί με μια επισκόπηση του πλαισίου λογισμικού ανοιχτού κώδικα Hadoop.

Όλοι όσοι μελετούν σήμερα τον τομέα των Big Data θα πρέπει να έχουν μια βασική κατανόηση του πώς Μεγάλα περιβάλλοντα δεδομένων σχεδιάζονται και λειτουργούν σε περιβάλλοντα επιχειρήσεων και τον τρόπο με τον οποίο τα δεδομένα ρέουν μέσα στα διαφορετικά επίπεδα ενός οργανισμού. Κατανόηση των βασικών στοιχείων της αρχιτεκτονικής Big Data θα βοηθήσει τους μηχανικούς συστημάτων, τους επιστήμονες δεδομένων, τους προγραμματιστές λογισμικού, τους αρχιτέκτονες δεδομένων, και τους υπεύθυνους λήψης αποφάσεων για να κατανοήσουν πώς τα στοιχεία των μεγάλων δεδομένων ταιριάζουν μαζί, και να αναπτύξουν ή να αντλήσουν λύσεις Big Data.

3.1 Η αναφορική Αρχιτεκτονική NIST

Προκειμένου να επωφεληθεί μια επιχείρηση από τις δυνατότητες των Big Data, είναι απαραίτητο να υπάρχει η τεχνολογία για την ανάλυση τεράστιων ποσοτήτων δεδομένων. Δεδομένου ότι το Big Data είναι μια εξέλιξη από «παραδοσιακά» δεδομένα, οι τεχνολογίες Big Data θα πρέπει να ταιριάζουν στο υπάρχον επιχειρηματικό περιβάλλον πληροφορικής. Για το λόγο αυτό, είναι χρήσιμο να έχουμε μια κοινή δομή που να εξηγεί το πώς τα Μεγάλα Δεδομένα συμπληρώνουν και διαφέρουν από τα υπάρχοντα αναλυτικά στοιχεία, την Business Intelligence, και τα συστήματα των βάσεων δεδομένων. Αυτή η κοινή δομή ονομάζεται αρχιτεκτονική αναφοράς (Reference Architecture) (Chang Wo 2015).

Μια αρχιτεκτονική αναφοράς είναι ένα έγγραφο ή ένα σύνολο εγγράφων στα οποία ο διαχειριστής του έργου ή άλλο ενδιαφερόμενο μέρος μπορεί να αναφερθεί για βέλτιστες πρακτικές. Στο πλαίσιο της πληροφορικής, η αναφορική αρχιτεκτονική μπορεί να χρησιμοποιηθεί για την επιλογή της

καλύτερης μεθόδου παράδοσης για συγκεκριμένες τεχνολογίες και έγγραφα όπως υλικό, λογισμικό, διαδικασίες, προδιαγραφές και διαμορφώσεις, όπως λογικά στοιχεία και αλληλεξαρτήσεις. Συνοπτικά, μια αρχιτεκτονική αναφοράς μπορεί να γίνει ως πόρος που καταγράφει τις εμπειρίες που αποκτήθηκαν από έργα του παρελθόντος (Chang Wo 2015).

Ο στόχος μιας αρχιτεκτονικής αναφοράς είναι να δημιουργήσει ένα ανοιχτό πρότυπο, ένα το οποίο μπορούν να χρησιμοποιήσουν προς όφελός τους. Το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας (NIST) - ένας από τους κορυφαίους οργανισμούς στην ανάπτυξη προτύπων−έχει αναπτύξει μια τέτοια αρχιτεκτονική αναφοράς: η Αρχιτεκτονική Αναφοράς Μεγάλων Δεδομένων NIST (Chang Wo 2015).

Τα πλεονεκτήματα της χρήσης μιας αρχικής αρχιτεκτονικής αναφοράς δεδομένων "ανοιχτού κώδικα" περιλαμβάνουν:

• Παρέχει μια κοινή γλώσσα για τους διάφορους ενδιαφερόμενους.

• Ενθαρρύνει την τήρηση κοινών προτύπων, προδιαγραφών και προτύπων.

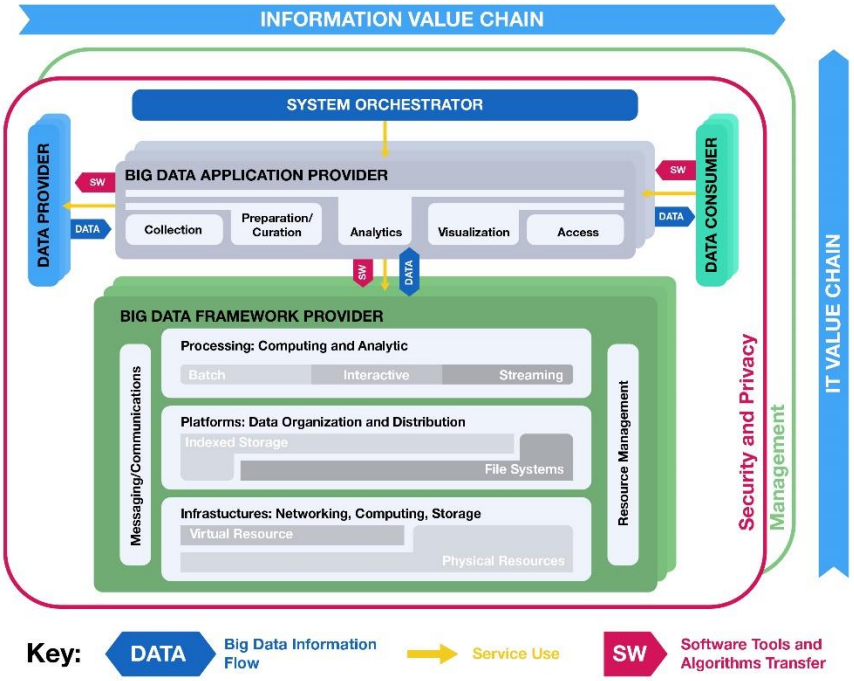
• Παρέχει συνεπείς μεθόδους για την εφαρμογή της τεχνολογίας για την επίλυση παρόμοιων προβλημάτων.

• Εικονογραφεί και βελτιώνει την κατανόηση των διαφόρων στοιχείων Big Data, των διαδικασιών και των συστημάτων, στο πλαίσιο μιας τεχνογνωσίας ενός εννοιολογικού μοντέλου.

• Διευκολύνει την ανάλυση των υποψήφιων προτύπων διαλειτουργικότητας, φορητότητας, επαναχρησιμοποίησης, και την επεκτασιμότητα.

Η Αρχιτεκτονική Αναφοράς Μεγάλων Δεδομένων NIST είναι ουδέτερη από το είδος του πωλητή και μπορεί να χρησιμοποιηθεί από κάθε οργανισμό που στοχεύει στην ανάπτυξη μιας αρχιτεκτονικής Big Data. Η αναφορά NIST, παρουσιάζεται στην Εικόνα 15 και αντιπροσωπεύει ένα σύστημα μεγάλων δεδομένων που αποτελείται από πέντε λογικά λειτουργικά στοιχεία ή ρόλους που συνδέονται με διεπαφές διαλειτουργικότητας (δηλ. υπηρεσίες). Δύο πεδία περιβάλλουν τα στοιχεία, αντιπροσωπεύοντας τον συνυφασμένο χαρακτήρα της διαχείρισης και

της ασφάλειας και την ιδιωτικότητα με τα πέντε στοιχεία. Στις επόμενες παραγράφους, κάθε μία θα εξεταστεί λεπτομερέστερα, μαζί με ορισμένα παραδείγματα.



Εικόνα 15: Αρχιτεκτονική Αναφοράς NIST (Chang Wo 2015)

Η Αρχιτεκτονική Αναφοράς Μεγάλων Δεδομένων NIST οργανώνεται γύρω από πέντε κύριους ρόλους και πολλαπλούς υπο-ρόλους ευθυγραμμισμένους κατά μήκος δύο αξόνων που αντιπροσωπεύουν τις δύο αλυσίδες αξίας των μεγάλων δεδομένων: η Αξία της πληροφορίας (οριζόντιος άξονας) και τεχνολογία πληροφοριών (IT, κατακόρυφος άξονας). Κατά μήκος του άξονα Αξίας Πληροφορίας, η αξία δημιουργείται με τη συλλογή δεδομένων, την ολοκλήρωση, την ανάλυση και την εφαρμογή μετά την αλυσίδα αξίας. Κατά μήκος του άξονα IT, η αξία δημιουργείται μέσω της παροχής δικτύωσης, υποδομή, πλατφόρμες, εργαλεία εφαρμογής και άλλες υπηρεσίες πληροφορικής για τη φιλοξενία και λειτουργώντας τα μεγάλα δεδομένα για την υποστήριξη των απαιτούμενων εφαρμογών δεδομένων. Στη διασταύρωση και των δύο αξόνων είναι ο ρόλος του Παροχέα Εφαρμογών Μεγάλων Δεδομένων, υποδεικνύοντας ότι τα αναλυτικά στοιχεία και τα δεδομένα υλοποίησης παρέχουν την αξία στους ενδιαφερόμενους και στις δύο αλυσίδες αξιών (Chang Wo 2015).

Οι πέντε κύριοι ρόλοι της Αρχιτεκτονικής Αναφοράς Μεγάλων Δεδομένων NIST, που παρουσιάζεται στην Εικόνα 15, αντιπροσωπεύουν τους ρόλους κάθε περιβάλλοντος μεγάλων δεδομένων, και υπάρχουν σε όλες τις επιχειρήσεις (Chang Wo 2015):

- Υπεύθυνος Συστήματος.
- Πάροχος δεδομένων.
- Πάροχος Εφαρμογής Μεγάλων Δεδομένων
- Μεγάλος πάροχος πλαισίου δεδομένων.
- Καταναλωτής δεδομένων.

Οι δύο διαστάσεις που παρουσιάζονται στην Εικόνα 15 που περιλαμβάνουν τους πέντε κύριους ρόλους είναι (Chang Wo 2015):

- Διαχείριση.
- Ασφάλεια και ιδιωτικό απόρρητο.

Αυτές οι διαστάσεις παρέχουν υπηρεσίες και λειτουργικότητα στους πέντε βασικούς ρόλους στις συγκεκριμένες περιοχές και είναι κρίσιμης σημασίας για οποιαδήποτε λύση Big Data.

Υπεύθυνος Συστήματος

Η σύνθεση του συστήματος είναι η αυτόματη ρύθμιση, ο συντονισμός και η διαχείριση των συστημάτων υπολογιστών, το μεσαίο λογισμικό και τις υπηρεσίες. Η ενορχήστρωση εξασφαλίζει ότι οι διαφορετικές οι εφαρμογές, τα δεδομένα και τα συστατικά στοιχεία υποδομής των μεγάλων περιβαλλόντων δεδομένων λειτουργούν από κοινού. Για να το επιτύχει αυτό, ο Υπεύθυνος Συστήματος χρησιμοποιεί ροές εργασίας, αυτοματοποίηση και διαδικασίες αλλαγών διαχείρισης (Chang Wo 2015).

Πάροχος δεδομένων

Ο ρόλος του Παρόχου Δεδομένων είναι να εισάγει νέες πληροφορίες δεδομένων ή τροφοδοσίες πληροφοριών στο σύστημα Big Data για την ανακάλυψη, την πρόσβαση και τον μετασχηματισμό από το σύστημα Big Data. Τα δεδομένα μπορούν να προέρχονται από διαφορετικές πηγές, όπως

δεδομένα ανθρώπινης προέλευσης (κοινωνικά μέσα), αισθητηριακά δεδομένα (ετικέτες RFID) ή συστήματα τρίτων (τραπεζικές συναλλαγές). Ένα από τα βασικά χαρακτηριστικά του Big Data είναι η ποικιλία του, δηλαδή ότι τα δεδομένα μπορούν να έρθουν σε διαφορετικές μορφές από διαφορετικές πηγές. Τα δεδομένα εισόδου μπορεί να έρθουν σε μορφή αρχείων κειμένου, εικόνων, ήχου, weblogs κ.λπ. Οι πηγές μπορεί να περιλαμβάνουν εσωτερικά συστήματα επιχειρήσεων (ERP, CRM, Finance) ή εξωτερικά συστήματα (αγορασμένα δεδομένα, κοινωνικές ροές δεδομένων). Συνεπώς, τα δεδομένα από διαφορετικές πηγές ενδέχεται να έχουν διαφορετικούς λόγους ασφάλειας και προστασίας της ιδιωτικής ζωής (Chang Wo 2015).

Όπως απεικονίζεται στο σχήμα 23, μεταφορές δεδομένων μεταξύ του παρόχου δεδομένων και του παροχέα μεγάλων δεδομένων. Αυτή η μεταφορά δεδομένων συμβαίνει συνήθως σε τρεις φάσεις: έναρξη, μεταφορά και τον τερματισμό. Η φάση έναρξης ξεκινά από ένα από τα δύο μέρη και συχνά περιλαμβάνει κάποιο επίπεδο ελέγχου ταυτότητας. Η φάση μεταφοράς δεδομένων ωθεί τα δεδομένα προς το Big Data Application Provider. Η φάση τερματισμού ελέγχει εάν έχει γίνει η μεταφορά δεδομένων με επιτυχία και καταγράφει την ανταλλαγή δεδομένων (Chang Wo 2015).

Πάροχος εφαρμογής Μεγάλων δεδομένων

Ο πάροχος εφαρμογής μεγάλων δεδομένων είναι το στοιχείο της αρχιτεκτονικής που περιέχει την επιχειρησιακή λογική και λειτουργικότητα που είναι απαραίτητες για τη μετατροπή των δεδομένων στα επιθυμητά αποτελέσματα. Ο κοινός στόχος αυτού του στοιχείου είναι η εξαγωγή της αξίας από τα δεδομένα εισόδου και περιλαμβάνει τις ακόλουθες δραστηριότητες:

- Συλλογή.
- Προετοιμασία.
- Αναλυτική
- Οπτικοποίηση
- Πρόσβαση

Η έκταση και οι τύποι εφαρμογών που χρησιμοποιούνται σε αυτό το στοιχείο της αρχιτεκτονικής αναφοράς ποικίλλουν σημαντικά και βασίζονται στη φύση και τις δραστηριότητες της επιχείρησης. Για τις χρηματοπιστωτικές επιχειρήσεις, οι αιτήσεις μπορούν να περιλαμβάνουν

λογισμικό ανίχνευσης απάτης, εφαρμογές βαθμολογίας πίστωσης ή λογισμικό ελέγχου ταυτότητας. Στις εταιρείες παραγωγής, μπορούν να είναι διαχείριση αποθεμάτων, βελτιστοποίηση της αλυσίδας εφοδιασμού ή λογισμικού βελτιστοποίησης διαδρομής (Chang Wo 2015).

Πάροχος πλαισίου Μεγάλων δεδομένων

Ο μεγάλος Πάροχος δεδομένων έχει τους πόρους και τις υπηρεσίες που μπορεί να χρησιμοποιήσει το Big Data Provider και να παρέχει την βασική υποδομή της αρχιτεκτονικής Big Data. Σε αυτό το στοιχείο, τα δεδομένα αποθηκεύονται και επεξεργάζονται με βάση τα σχέδια που έχουν βελτιστοποιηθεί για Περιβάλλοντα Big Data (Chang Wo 2015).

Ο μεγάλος πάροχος πλαισίου δεδομένων μπορεί να υποδιαιρεθεί περαιτέρω στους ακόλουθους υπο-ρόλους:

- Υποδομή: δικτύωση, υπολογισμός και αποθήκευση
- Πλατφόρμες: οργάνωση και διανομή δεδομένων
- Επεξεργασία: υπολογιστική και αναλυτική

Τα περισσότερα περιβάλλοντα μεγάλων δεδομένων χρησιμοποιούν την κατανεμημένη αποθήκευση και επεξεργασία και το πλαίσιο λογισμικού ανοιχτού κώδικα Hadoop για το σχεδιασμό αυτών των υπο-ρόλων του φορέα παροχής μεγάλων δεδομένων (Chang Wo 2015).

Το *στρώμα υποδομής* αφορά τον εαυτό του με τη δικτύωση, τον υπολογισμό και την αποθήκευση πρέπει να διασφαλίζουν ότι οι μεγάλες και ποικίλες μορφές δεδομένων μπορούν να αποθηκευτούν και να μεταφερθούν με οικονομικά αποδοτικό τρόπο, ασφαλή και κλιμακωτό τρόπο. Στον πυρήνα της, η βασική απαίτηση της αποθήκευσης Big Data είναι ότι είναι σε θέση να χειριστεί πολύ μεγάλες ποσότητες δεδομένων και διατηρεί κλιμάκωση με την ανάπτυξη του οργανισμού και ότι μπορεί να παράσχει τις απαραίτητες λειτουργίες εισόδου / εξόδου ανά δευτερόλεπτο (IOPS) για την παροχή δεδομένων σε εφαρμογές. Το IOPS είναι ένα μέτρο για την απόδοση αποθήκευσης που εξετάζει το ρυθμό μεταφοράς δεδομένων (Chang Wo 2015).

Το *επίπεδο πλατφόρμας* είναι η συλλογή λειτουργιών που διευκολύνει την επεξεργασία υψηλής απόδοσης των δεδομένων. Η πλατφόρμα περιλαμβάνει τις δυνατότητες ενσωμάτωσης, διαχείρισης και εφαρμογής εργασιών επεξεργασίας στα δεδομένα. Σε περιβάλλοντα μεγάλων δεδομένων, αυτό σημαίνει ουσιαστικά ότι η πλατφόρμα πρέπει να διευκολύνει και οργανώνει την κατανεμημένη επεξεργασία σε λύσεις κατανεμημένης αποθήκευσης. Μια από τα πολλές πλατφόρμες για λύσεις Big Data είναι η ανοικτή πηγή Hadoop. Ο λόγος για τον οποίο η Hadoop προσφέρει μια τέτοια επιτυχημένη υποδομή πλατφόρμας οφείλεται στην ενοποιημένη αποθήκευση (κατανεμημένη αποθήκευση) και στην κατανεμημένη επεξεργασία (Chang Wo 2015).

Το *επίπεδο επεξεργασίας* του φορέα παροχής μεγάλων δεδομένων παρέχει τη δυνατότητα αναζήτησης στα δεδομένα. Μέσα από αυτό το επίπεδο, εκτελούνται εντολές που εκτελούν λειτουργίες χρόνου εκτέλεσης στα σύνολα δεδομένων. Συχνά, αυτό γίνεται μέσω της εκτέλεσης ενός αλγορίθμου που εκτελεί μια επεξεργασία. Σε αυτό το επίπεδο, πραγματοποιείται η πραγματική ανάλυση. Διευκολύνει το 'χτύπημα των αριθμών' προκειμένου να επιτευχθούν τα επιθυμητά αποτελέσματα και η αξία των μεγάλων δεδομένων (Chang Wo 2015).

Χρήστης δεδομένων

Παρόμοια με τον παροχέα δεδομένων, ο ρόλος του καταναλωτή δεδομένων στο πλαίσιο της αναφοράς μεγάλων δεδομένων μπορεί να είναι ένας πραγματικός τελικός χρήστης ή άλλο σύστημα. Με πολλούς τρόπους, ο ρόλος αυτός είναι η εικόνα του παροχέα δεδομένων. Οι δραστηριότητες που σχετίζονται με τον ρόλο του Καταναλωτή Δεδομένων περιλαμβάνουν τα ακόλουθα (Chang Wo 2015):

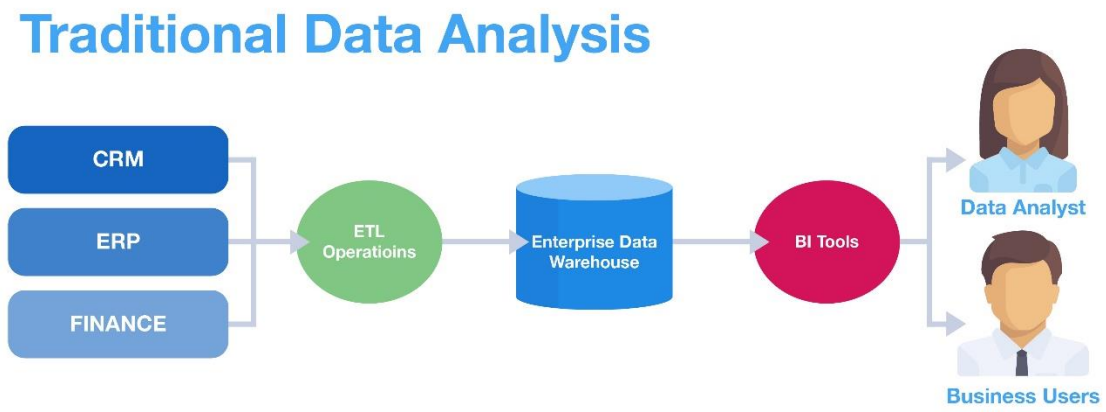
- Αναζήτηση και ανάκτηση
- Μεταφόρτωση
- Τοπική Ανάλυση
- Υποβολή εκθέσεων
- Οπτικοποίηση
- Δεδομένα που πρέπει να χρησιμοποιηθούν για τις δικές τους διαδικασίες.

Ο Καταναλωτής Δεδομένων χρησιμοποιεί τις διεπαφές ή τις υπηρεσίες που παρέχει ο Πάροχος Εφαρμογής Μεγάλων Δεδομένων, για πρόσβαση στις πληροφορίες που τον ενδιαφέρουν. Αυτές οι διεπαφές μπορεί να περιλαμβάνουν δεδομένα αναφοράς, την ανάκτηση δεδομένων και την απόδοση δεδομένων (Chang Wo 2015).

3.2 Κατανεμημένη αποθήκευση δεδομένων και Επεξεργασία

3.2.1 Παραδοσιακή ανάλυση δεδομένων - τοπική αποθήκευση και επεξεργασία

Με την εξέταση αυτής της δομής, μπορεί να γίνει μια ολοκληρωμένη σύγκριση με τον τρόπο που δημιουργούνται μεγάλα περιβάλλοντα δεδομένων. Η παραδοσιακή ανάλυση δεδομένων - όπως εκτελούνται από εκατομμύρια οργανισμούς κάθε μέρα - έχει αρκετά απλό και στατικό σχεδιασμό. Οι περισσότερες επιχειρήσεις δημιουργούν δομημένα δεδομένα με σταθερά μοντέλα δεδομένων μέσω μιας ποικιλίας επιχειρησιακών εφαρμογών, όπως το CRM, το ERP και διάφορα οικονομικά πληροφοριακά συστήματα. Διάφορα εργαλεία ενσωμάτωσης δεδομένων στη συνέχεια χρησιμοποιούν εξαγωγή, μετασχηματισμό και φόρτωση (Extract Transform Load) για να φορτώσει τα δεδομένα από αυτές τις επιχειρησιακές εφαρμογές σε μια κεντρική αποθήκη δεδομένων (Big Data Framework 2018).



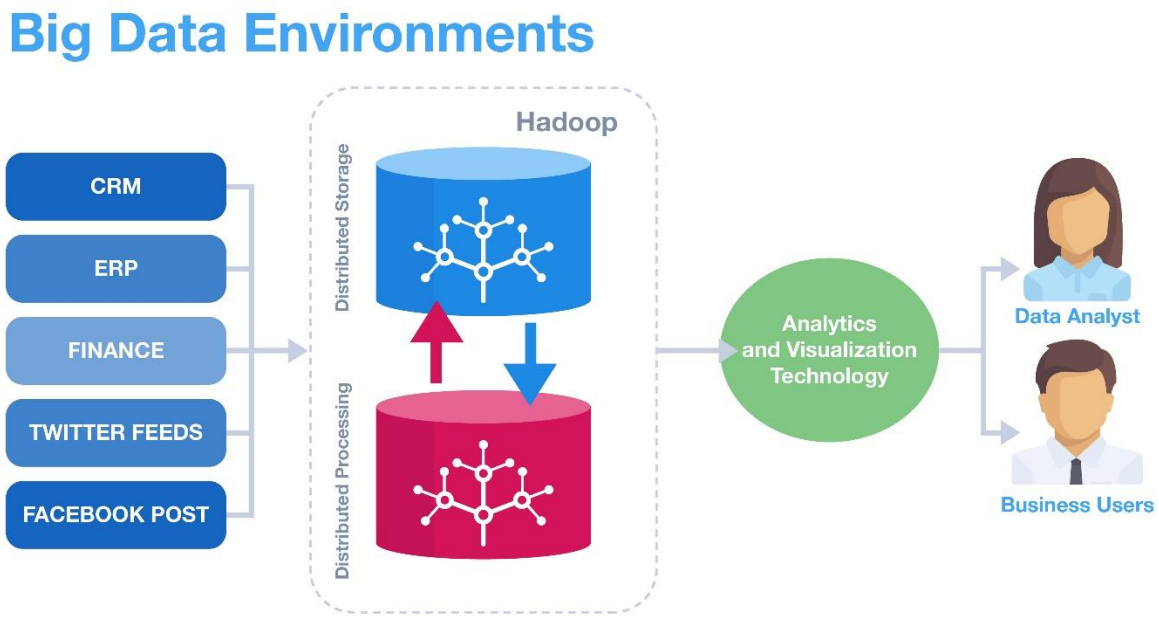
Εικόνα 16: Παραδοσιακή Ανάλυση δεδομένων[1]

Στην αποθήκη δεδομένων, τα διαφορετικά δεδομένα (που προέρχονται από τις διάφορες εφαρμογές) είναι τακτοποιημένα αποθηκεύονται σε βάσεις δεδομένων με δομημένες σειρές και στήλες (παρόμοια με ένα μεγάλο φύλλο Excel). Λόγω αυτής της τακτοποιημένης διάρθρωσης, ένα

εργαλείο ανάλυσης Business Intelligence (παραδείγματα είναι τα SAP Business Objects ή το IBM Cognos) μπορούν στη συνέχεια να εκτελούν ερωτήματα και να παρέχουν αναφορές που παρέχουν τις ζητούμενες πληροφορίες. Οι όγκοι δεδομένων στις αποθήκες δεδομένων σπανίως υπερβαίνουν τα πολλαπλάσια των terabyte, δεδομένου ότι μεγάλοι όγκοι δεδομένων υποβαθμίζουν την απόδοση (Big Data Framework 2018).

3.2.2 Μεγάλα περιβάλλοντα δεδομένων - αποθήκευση και επεξεργασία κατανεμημένων δεδομένων

Τα χαρακτηριστικά όγκου και ποικιλίας διαφοροποιούν τα μεγάλα δεδομένα με την παραδοσιακή ανάλυση δεδομένων. Τα περισσότερα δεδομένα προέρχονται από διάφορες πηγές (όχι μόνο από δεδομένα επιχείρησης) και δεν συμμορφώνεται με δομές σχεσιακών βάσεων δεδομένων. Αυτές οι πηγές δεδομένων υπερβαίνει τα πολλά terabytes (Big Data Framework 2018).



Εικόνα 17:Περιβάλλον Big Data[1]

Προκειμένου να αντιμετωπιστεί το μέγεθος (όγκος) και η διαφορά (ποικιλία) αυτών των δεδομένων, μια διαφορετική αρχιτεκτονική είναι απαραίτητη για να διασφαλιστεί η διατήρηση των επιπέδων επιδόσεων και η επεξεργασία των Big Data φέρνει την πραγματική αξία στην

επιχείρηση. Για την επίτευξη αυτών των στόχων, οι αρχιτεκτονικές δεδομένων συνήθως τηρούν τις ακόλουθες τέσσερις βασικές αρχές σχεδιασμού (Big Data Framework 2018):

1) Η χρήση κόμβων για να επιτρέψει την κλιμακωσιμότητα

2) Η χρήση κατανεμημένων αποθηκευτικών χώρων για την αποθήκευση δομημένων, μη δομημένων ή ημιδομημένων δεδομένων;

3) Η χρήση τεχνικών κατανεμημένης επεξεργασίας για την παράλληλη επεξεργασία.

4) Η χρήση προηγμένης τεχνολογίας ανάλυσης και οπτικοποίησης.

Αυτές οι τέσσερις αρχές σχεδιασμού μιας αρχιτεκτονικής μεγάλων δεδομένων εξασφαλίζουν ότι μπορούν να χρησιμοποιηθούν μεγάλα ποσά δεδομένων καθώς επεξεργάζονται αποτελεσματικά. Για την ανάλυση μεγάλων δεδομένων, δομημένα δεδομένα (CRM, ERP και οικονομικά δεδομένα) και τα μη δομημένα δεδομένα (feeds Twitter και θέσεις Facebook) "Τμήματα", τα οποία στη συνέχεια φορτώνονται σε ένα κατανεμημένο σύστημα αποθήκευσης που αποτελείται από πολλαπλούς κόμβους τρέχοντας σε υλικό βασικών προϊόντων. Για να επεξεργαστούν τα κατανεμημένα δεδομένα, κάθε "μέρος" αναλύεται στη συνέχεια μέσα στην συστοιχία. Αντί να φέρει όλα τα δεδομένα σε μια κεντρική τοποθεσία, η επεξεργασία λαμβάνει χώρα σε κάθε κόμβο ταυτόχρονα και επομένως εργάζονται παράλληλα μεταξύ τους. Η τοπική επεξεργασία δεδομένων στο πλαίσιο των κόμβων ονομάζεται κατανεμημένη επεξεργασία. Τέλος, η τεχνολογία της ανάλυσης και της απεικόνισης μπορεί να χρησιμοποιείται για την εμφάνιση του τελικού αποτελέσματος (Big Data Framework 2018).

Οι τέσσερις αρχές σχεδιασμού ενσωματώνονται στο πλαίσιο λογισμικού ανοικτών πηγών Hadoop, οι οποίες θα συζητηθούν περαιτέρω στην επόμενη ενότητα.

3.3 Αποθήκευση Μεγάλης Κλίμακας Δεδομένων

Ο μαζικός όγκος και η αύξηση των δεδομένων επιβάλλει προηγμένες απαιτήσεις για την αποθήκευση και την διαχείριση. Η αποθήκευση μεγάλων δεδομένων αναφέρεται στην αποθήκευση και διαχείριση μεγάλων σειρών δεδομένων, επιτυγχάνοντας παράλληλα αξιοπιστία

και πρόσβαση διαθέσιμων δεδομένων. Τα σύνολα δεδομένων μεγάλης κλίμακας έχουν ξεχωριστή επίδραση στο σχεδιασμό των συστημάτων αποθήκευσης, όπως καθώς και στους μηχανισμούς αποθήκευσης (Big Data Framework 2018).

Όπως αναλύθηκε στην ενότητα 4.2, στα παραδοσιακά περιβάλλοντα ανάλυσης δεδομένων, η αποθήκευση δεδομένων χρησιμοποιείται για την αποθήκευση και ανάκτηση δεδομένων από CRM, ERP ή συστήματα χρηματοδότησης πραγματοποιείται σε δομημένα Σχεσιακά Συστήματα διαχείρισης βάσεων δεδομένων (RDBMS). Ωστόσο, λόγω της αδόμητης φύσης των Μεγάλων δεδομένων, απαιτούνται διαφορετικά συστήματα αποθήκευσης. Η αποθήκευση στο περιβάλλον Big Data είναι ένα πολύπλοκο θέμα, λόγω των δύο αντίθετων δυνάμεων που ισχύουν. Από τη μία πλευρά, η υποδομή αποθήκευσης πρέπει να παρέχει αποθήκευση πληροφοριών με αξιόπιστο αποθηκευτικό χώρο. Από την άλλη πλευρά, πρέπει να παρέχει μια ισχυρή διεπαφή πρόσβασης για ερώτηματα και ανάλυση των συνόλων Big Data (Big Data Framework 2018).

3.3.1 Συστήματα αποθήκευσης για μαζικά δεδομένα

Υπάρχουν διάφορα συστήματα αποθήκευσης για να καλύψουν τις απαιτήσεις των μαζικών δεδομένων. Οι υπάρχουσες τεχνολογίες μπορούν να ταξινομηθούν είτε ως Απευθείας Αποθήκη (Direct Attached Storage) είτε ως αποθήκευση δικτύου, που μπορεί να υποδιαιρεθεί περαιτέρω στο Network Attached Storage (NAS) και στην περιοχή δικτύων αποθήκευσης (Storage Area Network) (Big Data Framework 2018).

Άμεση προσαρτώμενη αποθήκευση (DAS). Η απευθείας προσαρτημένη αποθήκευση είναι απευθείας ψηφιακή αποθήκευση συνδεδεμένη στον υπολογιστή που έχει πρόσβαση σε αυτόν, σε αντίθεση με την αποθήκευση που έχει πρόσβαση σε έναν υπολογιστή δικτύου. Παραδείγματα DAS περιλαμβάνουν σκληρούς δίσκους, μονάδες SSD, μονάδες οπτικών δίσκων και σε εξωτερικούς δίσκους αποθήκευσης. (Big Data Framework 2018)

Αποθήκευση σε προσαρτημένο δίκτυο (Network Attached Storage). Η αποθήκευση που είναι συνδεδεμένη στο δίκτυο είναι ένας υπολογιστής σε επίπεδο διακομιστή αποθήκευσης δεδομένων συνδεδεμένο σε δίκτυο υπολογιστών που παρέχει πρόσβαση σε δεδομένα σε μια ετερογενή ομάδα πελατών. Το NAS είναι εξειδικευμένο για την εξυπηρέτηση αρχείων είτε από το υλικό, λογισμικό

ή διαμόρφωση. Συχνά κατασκευάζεται ως συσκευή πληροφορικής- έναν εξειδικευμένο ηλεκτρονικό υπολογιστή (Big Data Framework 2018).

Δίκτυο χώρου αποθήκευσης (SAN). Ένα δίκτυο περιοχής αποθήκευσης είναι ένα δίκτυο το οποίο παρέχει πρόσβαση σε ενοποιημένη αποθήκευση δεδομένων σε επίπεδο μπλοκ. Τα SAN χρησιμοποιούνται κυρίως για την ενίσχυση συσκευών αποθήκευσης, όπως συστοιχίες δίσκων, βιβλιοθήκες ταινιών και οπτικά jukeboxes, στα οποία είναι προσβάσιμα έτσι ώστε οι συσκευές να εμφανίζονται στο λειτουργικό σύστημα ως τοπικά συνδεδεμένες συσκευές (Big Data Framework 2018).

Το Direct Attached Storage είναι κατάλληλο μόνο σε μικρή κλίμακα (η αποθήκευση πρέπει να είναι φυσική συνδεδεμένη στον υπολογιστή). Οι περισσότερες επιχειρήσεις χρησιμοποιούν επομένως την αποθήκευση δικτύου (σε κέντρα δεδομένων) για να ικανοποιήσουν τις απαιτήσεις αποθήκευσης τους (Big Data Framework 2018).

3.3.2 Στοιχεία κατανεμημένης αποθήκευσης

Το σύστημα κατανεμημένης αποθήκευσης χρησιμοποιεί δίκτυα υπολογιστών για την αποθήκευση πληροφοριών σε περισσότερους από έναν κόμβο, συχνά με αντιγραφή. Συνήθως χρησιμοποιείται ειδικά για να αναφέρεται είτε σε μια κατανεμημένη βάση δεδομένων όπου οι χρήστες αποθηκεύουν πληροφορίες σε έναν αριθμό κόμβων ή έναν υπολογιστή. (Big Data Framework 2018)

Οι συνιστώσες της κατανεμημένης αποθήκευσης για το Big Data μπορούν να ταξινομηθούν σε τρία βασικά επίπεδα: (Big Data Framework 2018)

1) *Συστήματα αρχείων.* Ένα σύστημα αρχείων χρησιμοποιείται για τον έλεγχο του τρόπου αποθήκευσης και ανάκτησης δεδομένων. Χωρίς το σύστημα αρχείων, οι πληροφορίες που τοποθετούνται σε ένα μέσο αποθήκευσης θα είναι ένα μεγάλο σύνολο δεδομένων χωρίς να υπάρχει κανένας τρόπος να καθορίσει πού ένα κομμάτι πληροφοριών σταματάει και το επόμενο ξεκινά. Χωρίζοντας τα δεδομένα σε τεμάχια και δίνοντας σε κάθε κομμάτι ένα όνομα, οι πληροφορίες είναι εύκολα απομονωμένες και αναγνωρισμένες. Λαμβάνοντας το όνομά της από τον τρόπο που

βασίζονται σε συστήματα πληροφοριών, κάθε ομάδα δεδομένων ονομάζεται "αρχείο". Η δομή και οι λογικοί κανόνες που χρησιμοποιούνται διαχειρίζονται τις ομάδες πληροφοριών και τα ονόματά τους ονομάζονται "συστήματα αρχείων". Σημαντικό αρχείο συστημάτων στα μεγάλα δεδομένα είναι το σύστημα αρχείων Google (GFS) και το Hadoop Distributed Σύστημα αρχείων (HDFS), αμφότερα τα οποία είναι επεκτάσιμα συστήματα κατανεμημένων αρχείων.

2) *Βάσεις δεδομένων*. Οι παραδοσιακές σχεσιακές βάσεις δεδομένων (RDBS) δεν μπορούν να αντιμετωπίσουν τις προκλήσεις στις κλίμακες που απαιτούνται για τα μεγάλα δεδομένα. Για το λόγο αυτό, οι βάσεις δεδομένων NoSQL καθίστανται όλο και πιο δημοφιλείς ως βασική τεχνολογία αποθήκευσης μεγάλων δεδομένων. Η βάση δεδομένων NoSQL παρέχει έναν μηχανισμό αποθήκευσης και ανάκτησης δεδομένων που διαμορφώνεται σε άλλα μέσα εκτός από τις πινακοκεντρικές σχέσεις που χρησιμοποιούνται σε σχεσιακές βάσεις δεδομένων. Συστήματα NoSQL αποκαλούνται επίσης μερικές φορές "όχι μόνο SQL" για να τονίσει ότι μπορεί να υποστηρίζουν SQL-τύπου γλώσσες ερωτήσεων. Σημαντικά παραδείγματα των βάσεων δεδομένων NoSQL περιλαμβάνουν το DynamoDB (Amazon), Voldemort (LinkedIN), BigTable (Google), Cassandra (Facebook), Azure DB (Microsoft), HBase (ανοικτού κώδικα) και MongoDB (ανοιχτού κώδικα).

3) *Μοντέλα προγραμματισμού*. Τα μεγάλα δεδομένα αποθηκεύονται γενικά σε εκατοντάδες ή χιλιάδες εμπορικούς διακομιστές. Προκειμένου να αποκτηθεί πρόσβαση στα δεδομένα που είναι αποθηκευμένα σε αυτούς τους διακομιστές, παράλληλα έχουν αναπτυχθεί μοντέλα προγραμματισμού που αυξάνουν την απόδοση του NoSQL βάσεων δεδομένων. Τα πιο σημαντικά παραδείγματα στα Big Data είναι το MapReduce, το Dryad (που χρησιμοποιείται από τη Microsoft) και το Pregel (χρησιμοποιούνται από την Google).

3.4 Αρχιτεκτονική Ανάλυσης Big Data

Για να επεξεργαστούν μεγάλους όγκους δεδομένων, μπορούν να σχεδιαστούν διαφορετικές αρχιτεκτονικές για την ανάλυση μεγάλων δεδομένων. Η πιο σημαντική διάκριση (από άποψη

αρχιτεκτονικής) είναι η διαφορά μεταξύ της ανάλυσης σε πραγματικό χρόνο και της ανάλυσης εκτός σύνδεσης (Chang Wo 2015):

Ανάλυση σε πραγματικό χρόνο: Η ανάλυση σε πραγματικό χρόνο χρησιμοποιείται κατά κύριο λόγο στο ηλεκτρονικό εμπόριο και στις οικονομικές εφαρμογές, όπου οι αλλαγές δεδομένων πρέπει να υποβάλλονται σε άμεση επεξεργασία. Αφού τα δεδομένα, απαιτούνται να είναι ενημερωμένα. Η ανάλυση συναλλαγών με πιστωτικές κάρτες, για παράδειγμα, θα απαιτούσε αυτόν τον τύπο αρχιτεκτονικής. Οι κύριες αρχιτεκτονικές της ανάλυσης σε πραγματικό χρόνο περιλαμβάνουν τη χρήση συμπλεγμάτων παράλληλης επεξεργασίας, παραδοσιακές σχεσιακές βάσεις δεδομένων και υπολογιστικές πλατφόρμες που βασίζονται στη μνήμη. Παραδείγματα δυνατής επεξεργασίας σε πραγματικό χρόνο περιλαμβάνουν το Greenplum (EMC) και το SAP HANA (Chang Wo 2015).

Ανάλυση χωρίς σύνδεση Η ανάλυση εκτός σύνδεσης χρησιμοποιείται για εφαρμογές που είναι λιγότερο ευαίσθητες στο χρόνο και για τις οποίες η αξία των δεδομένων σε πραγματικό χρόνο είναι λιγότερο επείγουσα. Η επεξεργασία εκτός σύνδεσης (επίσης γνωστή ως επεξεργασία παρτίδων) εισάγει δεδομένα σχετικά με τις καθορισμένες ώρες και στη συνέχεια τα επεξεργάζεται σε συγκεκριμένα διαστήματα. Οι περισσότερες επιχειρήσεις χρησιμοποιούν την αρχιτεκτονική ανάλυσης εκτός σύνδεσης με βάση το Hadoop προκειμένου να μειωθεί το κόστος και να βελτιωθεί η αποτελεσματικότητα της επεξεργασίας δεδομένων. Παραδείγματα εργαλείων εκτός σύνδεσης ς περιλαμβάνουν το Scribe (Facebook), το Kafka (LinkedIN), το TimeTunnel (Tabao) και το Chukwa (ανοικτός κώδικας Hadoop) (Chang Wo 2015).

Αν και η ανάλυση σε πραγματικό χρόνο και η ανάλυση εκτός σύνδεσης παρέχουν ικανοποιητικά αποτελέσματα, οι περισσότερες επιχειρήσεις χρησιμοποιούν ανάλυση εκτός σύνδεσης, εάν η επικαιρότητα των δεδομένων δεν αποτελεί βασική απαίτηση (Chang Wo 2015).

3.5 Πλαίσιο ανοικτού κώδικα Hadoop

Το Hadoop είναι ένα πλαίσιο προγραμματισμού βασισμένο σε Java που υποστηρίζει την επεξεργασία και αποθήκευση εξαιρετικά μεγάλων συνόλων δεδομένων σε κατανεμημένο περιβάλλον πληροφορικής. Είναι μέρος του προγράμματος Apache που χρηματοδοτείται από το

Apache Software Foundation. Ο Hadoop αρχικά δημιουργήθηκε από τον Doug Cutting στο Yahoo! και εμπνευσμένο από τη λειτουργία Map Reduce που αναπτύχθηκε από την Google στις αρχές της δεκαετίας του 2000 για την ευρετηρίαση της διαδικτυακής κυκλοφορίας. Μέχρι τώρα, ο Hadoop έχει μεγαλώσει για να γίνει το τυπικό πλαίσιο λογισμικού για την επεξεργασία μεγάλων δεδομένων και χρησιμοποιείται από τα περισσότερους προμηθευτές πακέτων μεγάλης κλίμακας δεδομένων (Big Data Framework 2018).

Πιο συγκεκριμένα ένα κατανεμημένο υπολογιστικό σύστημα αποτελείται από πολλαπλά λογισμικά και υπολογιστικούς πόρους σε πολλούς υπολογιστές που λειτουργούν ως ένα ενιαίο σύστημα. Οι υπολογιστές που αποτελούν ένα κατανεμημένο σύστημα μπορούν να είναι φυσικά συνδεδεμένοι σε ένα τοπικό δίκτυο ή να είναι γεωγραφικά απομακρυσμένοι και συνδεδεμένοι σε ένα δίκτυο ευρείας περιοχής. Ο στόχος του κατανεμημένου υπολογιστικού συστήματος είναι να καταστήσουν ένα τέτοιο δίκτυο να λειτουργεί ως ένας μόνο υπολογιστής (IBM n.d.).

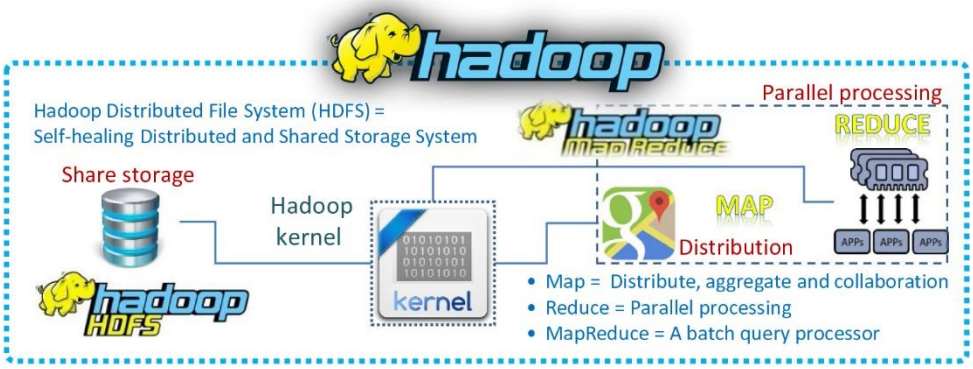
Τα κατανεμημένα συστήματα προσφέρουν πολλά πλεονεκτήματα σε σχέση με τα κεντρικά συστήματα (centralized), συμπεριλαμβανομένων των εξής:

- Scalability (επεκτασιμότητα) Το σύστημα μπορεί εύκολα να επεκταθεί με την προσθήκη περισσότερων μηχανών ανάλογα με τις ανάγκες.
- Redundancy (πλεονασμός)

Κάθε υπολογιστής του δικτύου παρέχει τις ίδιες λειτουργίες. Οπότε, αν υπάρξει κάποιο πρόβλημα με κάποιο μεμονωμένο μηχάνημα, η δουλειά γίνεται από κάποιον άλλον. Επιπλέον, τα υπολογιστικά συστήματα του δικτύου δεν είναι απαγορευτικά δαπανηρά (commodity hardware) (Ξωνίκη 2018).

Η βασική ιδέα για τη δημιουργία της Hadoop καθοδηγείται τόσο από τα συνεχώς αυξανόμενα δεδομένα όσο και από το κόστος των υπολογιστικού υλικού και εξαρτημάτων. Ο στόχος της Hadoop είναι να αξιοποιήσει το υλικό βασικών προϊόντων για μεγάλο φόρτο εργασίας επεξεργασίας, η οποία παλιότερα πραγματοποιούνταν μόνο με ακριβούς υπολογιστές mainframe. Από την προοπτική υποδομής, η Hadoop επιτρέπει την υπολογιστική ικανότητα να εξαλείφεται

παρά να αυξάνεται. Ο πολλαπλασιασμός scale up έχει μια έννοια βελτίωσης της ποιότητας, ενώ η scale out συνεπάγεται την προσθήκη ή την επανάληψη της ίδιας μονάδας σε οριζόντιο επίπεδο .



Εικόνα 18: Ο πυρήνας της Hadoop (Buyya, Calheiros και Dastjerdi 2016)

Το πλεονέκτημα της υιοθέτησης της πλατφόρμας Hadoop είναι μια δωρεάν και ελεύθερη πηγή αποθήκευσης και υπολογιστικής πλατφόρμας. Δημιουργήθηκε για να επιτρέψει την αποθήκευση και την επεξεργασία μεγάλων ποσών δεδομένων χρησιμοποιώντας συστάδες απλού υλικού. Αυτή η δήλωση περιγράφει επίσης τη βασική αρχή της αρχιτεκτονικής Hadoop που αποτελείται από τρία βασικά στοιχεία. HDFS για αποθήκευση αρχείων, Map για τη λειτουργία διανομής και Reduce για λειτουργία παράλληλης επεξεργασίας. Ωστόσο, το βασικό μειονέκτημα της Hadoop είναι ότι επεξεργάζεται όλους τους φόρτους εργασίας σε κατάσταση παρτίδας επειδή η Hadoop είναι ένα γενικό πλαίσιο επεξεργασίας σχεδιασμένο για να εκτελέσει ερωτήματα και άλλες λειτουργίες ανάγνωσης παρτίδας σε μαζικά σύνολα δεδομένων που μπορούν να κλιμακωθούν από δεκάδες terabytes έως petabytes σε μέγεθος. Αυτό σημαίνει ότι η πρώιμη έκδοση της Hadoop δεν μπορεί να χειριστεί ροή και δια δραστικό φόρτο εργασίας. Ο συνοψίζει τα βασικά χαρακτηριστικά της Hadoop (Buyya, Calheiros και Dastjerdi 2016).

Πίνακας 7: Χαρακτηριστικά της Hadoop (Buyya, Calheiros και Dastjerdi 2016)

Γνωρίσματα	Χαρακτηριστικά της Hadoop
Ιδρυτές	Doug Cutting και Michael J. Cafarella
Πρόγονος	Nutch

Μεταγενέστερη έκδοση	YARN ή Hadoop 2.0
Γλώσσα προγραμματισμού	Java
Φιλοσοφία υπολογισμού	Διαίρεση για μεγάλα σύνολα δεδομένων
Αρχές υπολογιστικής διαδικασίας	Φέρνεις τα δεδομένα στον υπολογιστή και όχι
Σύστημα	Ένα κατανεμημένο πλαίσιο προγραμματισμού
Κύρια χαρακτηριστικά	Προσβάσιμα, ανθεκτικά, αυξανόμενα, απλά
Storage-Hadoop distributed file system	στοιχείο αποθήκευσης με αυτό-θεραπεία
Αρχικό υπολογιστικό πρόγραμμα-	Κατανεμημένη, συγκεντρωτική και
Γλώσσα προγραμματισμού της MapReduce	C++
Τύπος διαδικασίας	Παρτίδα
Τύπος hardware	Ετερογενή προϊόντα hardware
Άδεια λογισμικού	Ανοικτό λογισμικό
Αρχικές εφαρμογές	IR και αναζήτηση δείκτη και web crawler
Τύπος λύσης	Λύση τύπου λογισμικού, όχι λύση τύπου
Λύση αύξησης	Scale out και όχι scale up
Τυπικό μέγεθος σετ δεδομένων	Από κάποια GB σε λίγα TB
Δυνατό μέγεθος σετ δεδομένων	Από δεκάδες TB σε λίγα PB
Απλό μοντέλο συνεκτικότητας	Γράφεις ένα και διαβάζεις πολλά
Προεπιλεγμένος παράγοντας αναπαραγωγής	3
Τυπικό μέγεθος του μπλοκ δεδομένων για	64 MB
Μοντέλο άδειας	Μοντέλο Relaxing POSIXa

Κύρια λειτουργικά τμήματα	Mahout, Hive, Pig, HBase, Sqoop, Flume, Chukwa, Pentaho …
Τυπικά διανύσματα	MapR, Cloudera, Hortonworks, IBM, Teradata, Intel, AWS, Pivotal Software, and

Το Hadoop σχεδιάστηκε αρχικά για επεξεργασία δεδομένων κατά συστάδες με στόχο να μπορεί να λειτουργεί σε περιβάλλοντα με χιλιάδες μηχανές. Υποστηρίζοντας ένα τόσο μεγάλο περιβάλλον πληροφορικής θέτει αρκετούς περιορισμούς στο σύστημα. για παράδειγμα, με τόσες μηχανές, το σύστημα θα έπρεπε οι υπολογιστικοί κόμβοι να αποτυγχάνουν. Το Hadoop είναι μια βελτιωμένη εφαρμογή MapReduce με την υποστήριξη για ανοχή σφάλματος, κατανεμημένη αποθήκευση και δεδομένα παράλληλα μέσω δύο πρόσθετων βασικών χαρακτηριστικών του σχεδιασμού: (1) ένα κατανεμημένο σύστημα αρχείων που ονομάζεται Hadoop Distributed File System (HDFS) και (2) μια στρατηγική διανομής δεδομένων που επιτρέπει τον μετασχηματισμό των υπολογισμών στα δεδομένα κατά την εκτέλεση (Buyya, Calheiros και Dastjerdi 2016).

3.5.1 Το σύστημα κατανομής αρχείων Hadoop (HDFS)

Προκειμένου να αναλυθούν οι τεράστιες ποσότητες δομημένων και αδόμητων δεδομένων, τα δεδομένα πρέπει να χωρίζονται σε "μέρη", τα οποία στη συνέχεια φορτώνονται σε ένα κατανεμημένο σύστημα αποθήκευσης αποτελούμενο από πολλαπλούς κόμβους που λειτουργούν με υλικό βασικών προϊόντων (common commodity hardware). Το σύστημα κατανομής αρχείων Hadoop (HDFS) είναι το σύστημα αρχείων που επιτρέπει την αποθήκευση αυτών των στοιχείων δεδομένων σε διαφορετικά μηχανήματα σε ένα σύμπλεγμα. Κατά συνέπεια, το HDFS επιτρέπει την κατανεμημένη αποθήκευση (Big Data Framework 2018).

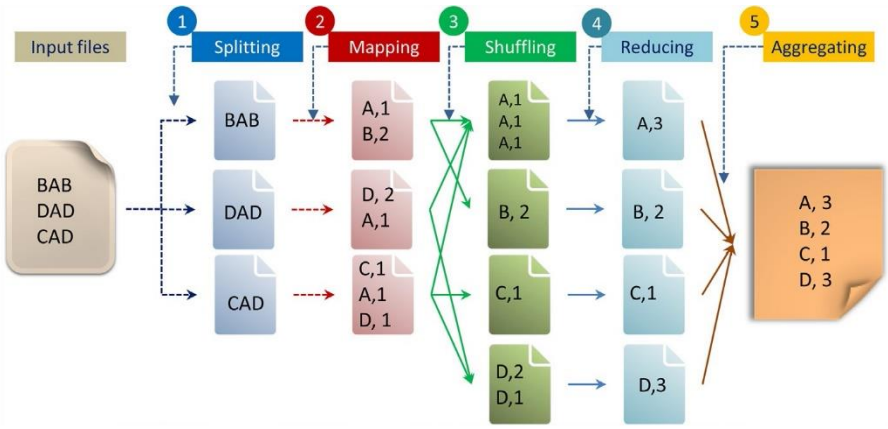
Μία από τις ιδιότητες του πυρήνα του HDFS είναι ότι κάθε ένα από τα μέρη δεδομένων αντιγράφεται πολλαπλές φορές και κατανέμονται σε πολλαπλούς κόμβους εντός του συμπλέγματος. Εάν ένας κόμβος αποτύχει, ένας άλλος κόμβος έχει ένα αντίγραφο αυτού του συγκεκριμένου πακέτου δεδομένων που μπορεί να χρησιμοποιηθεί για επεξεργασία. Εξαιτίας αυτού, τα δεδομένα μπορεί ακόμα να επεξεργαστεί και να αναλυθεί ακόμη και όταν ένας από τους κόμβους αποτύχει λόγω μιας βλάβης υλικού. Αυτό καθιστά το HDFS και το Hadoop πολύ ισχυρό σύστημα (Big Data Framework 2018).

3.5.2 NameNode

Εφόσον το HDFS αποθηκεύει πολλαπλά αντίγραφα των τμημάτων δεδομένων σε διάφορους κόμβους του συμπλέγματος, είναι πολύ σημαντικό να παρακολουθείτε τον τόπο αποθήκευσης των τμημάτων δεδομένων και ποιοι κόμβοι είναι διαθέσιμοι ή έχουν αποτύχει. Το NameNode εκτελεί αυτή την εργασία. Λειτουργεί ως διευκολυντής που επικοινωνεί όπου αποθηκεύονται τα στοιχεία δεδομένων και εάν είναι διαθέσιμα. Το NameNode είναι το κεντρικό στοιχείο ενός συστήματος αρχείων HDFS. Διατηρεί το δέντρο καταλόγου όλων των αρχείων στο σύστημα αρχείων και τα κομμάτια όπου διατηρούνται σε ολόκληρο το σύμπλεγμα τα δεδομένων αρχείου. Δεν αποθηκεύει το δεδομένα αυτών των αρχείων (Big Data Framework 2018).

3.5.3 MapReduce

Το MapReduce είναι ένα μοντέλο προγραμματισμού που χρησιμοποιείται για τη διεκπεραίωση μεγάλου φόρτου εργασίας ενός συνόλου δεδομένων. Σε αντίθεση με την επιτακτική ανάγκη προγραμματισμό (περιγράφοντας τον υπολογισμό ως μια δέσμη δηλώσεων για την αλλαγή της κατάστασης του προγράμματος), η MapReduce θεωρεί τον υπολογισμό ως την αξιολόγηση των μαθηματικών λειτουργιών. Στην ουσία, ο λειτουργικός προγραμματισμός μπορεί να αποφύγει την κατάσταση και απλά να καταγράψει τις καταστάσεις in-and-out.



Εικόνα 19: Πέντε βήματα του μοντέλου MapReduce (Buyya, Calheiros και Dastjerdi 2016)

Η βασική στρατηγική του MapReduce είναι να διαιρέσει και να κατακτήσει. Προκειμένου να εκτελεστούν διαφορετικά δεδομένα, εφαρμογές αποτελεσματικά με το MapReduce στο πλαίσιο

του GFS, οι (Dean and Ghemawat 2008) παρουσίασαν μια διαδικασία ή ένα μοντέλο προγραμματισμού πέντε βημάτων, όπως φαίνεται στην .

Οι Lin et αϊ. απλοποίησαν αυτή τη διαδικασία σε τρία βήματα: map, shuffle και reduce.

- Map: Μη αλληλο-επικαλυπτόμενα κομμάτια από δεδομένα εισόδου (εγγραφές <key,value>) ανατίθενται σε διαφορετικές διεργασίες (mappers) οι οποίες βγάζουν ένα σετ από ενδιάμεσα <key,value> αποτελέσματα.
- Shuffle: Ταξινομεί τις εξόδους της πρώτης φάσης, ομαδοποιώντας τα ζεύγη με βάση τα κλειδιά πριν στείλει το καθένα από αυτά στην επόμενη φάση.
- Reduce: Τα δεδομένα της Map φάσης τροφοδοτούνται σε ένα συνήθως μικρότερο αριθμό διεργασιών (reducers) οι οποίες "συνοψίζουν" τα αποτελέσματα εισόδου σε μικρότερο αριθμό <key,value> εγγραφών

Μια εργασία στο MapReduce δηλαδή, παίρνει μια λίστα με ζεύγη κλειδιού – τιμής ως είσοδο και εξάγει μια λίστα τιμών. Οι χρήστες χρειάζεται μόνο να εφαρμόσουν τις διεπαφές των λειτουργιών map και reduce και μπορούν να αφήσουν στο σύστημα που υιοθετεί το MapReduce να διαχειριστεί όλες τις επικοινωνίες δεδομένων και την παράλληλη επεξεργασία (Foster, Ghani and Jarmin 2017)

Παράδειγμα MapReduce

Υπολογισμός βραβείων National Science Foundation

Έστω ότι έχουμε μια λίστα με τους κύριους ερευνητές του NSF, μαζί με τις πληροφορίες ηλεκτρονικού ταχυδρομείου τους και αναγνωριστικών βραβείων όπως παρακάτω. Στόχος της εφαρμογής είναι να υπολογίσει τον αριθμό των βραβείο που αντιστοιχεί σε κάθε ίδρυμα (Foster, Ghani and Jarmin 2017).

AwardId,FirstName,LastName,EmailAddress

0958723,Roland,Mundil,rmundil@bgc.org

0958915,Randall,Irmis,irmis@umnh.utah.edu

1301647,Zaher,Hani,zh8@nyu.edu

1316375,David,Shuster,dshuster@bgc.org

Παρατηρούμε ότι τα ιδρύματα μπορούν να διακριθούν από τον τομέα διεύθυνσης ηλεκτρονικού ταχυδρομείου τους όνομα. Έτσι, υιοθετούμε μια στρατηγική πρώτης ομαδοποίησης όλων των αναγνωρισμάτων βραβείων ανά όνομα τομέα και, στη συνέχεια, καταμέτρηση του αριθμού των διακριτών βραβείων σε κάθε ομάδα. Για να γίνει αυτό, ρυθμίσαμε πρώτα τη λειτουργία χαρτών για να σαρώσουμε γραμμές εισόδου και να αφαιρέσουμε το ίδρυμα πληροφορίες και αναγνωριστικά βραβείων. Στη συνέχεια, στη φάση reduce, μετράμε απλά μοναδικά αναγνωριστικά στοιχεία για τα δεδομένα, αφού τα πάντα είναι ήδη ομαδοποιημένα ανά φορέα.. Στη φάση του map, η είσοδος θα μετατραπεί σε πλειάδα ιδρυμάτων και ταυτοποιήσεις βραβείων (Foster, Ghani and Jarmin 2017):

"0958723,Roland,Mundil,rmundil@bgc.org"	("bgc.org", 0958723)
"0958915,Randall,Irmis,irmis@umnh.utah.edu"	("utah.edu",958915)
"1301647,Zaher,Hani,zh8@nyu.edu"	("nyu.edu", 1301647)
"1316375,David,Shuster,dshuster@bgc.org"	("bgc.org", 1316375)

Στη συνέχεια, οι πλειάδες θα ομαδοποιηθούν από ιδρύματα και θα μετρηθούν από τη διαδικασία reduce.

("bgc.org", [0958723,1316375])	("bgc.org", 2)
("nyu.edu", [1301647])	("nyu.edu", 1)
("nyu.edu", 1)	("nyu.edu", 1)

Όπως φαίνεται από το παράδειγμα το μοντέλο προγραμματισμού MapReduce είναι πολύ απλό και κατανοητό. Η απλή αρχιτεκτονική του μοντέλου έχει επίσης εμπνεύσει πολλούς νέους προγραμματιστές να αναπτύξουν προηγμένες δυνατότητες, όπως υποστήριξη για κατανεμημένους συστήματα, καταμερισμό δεδομένων και επεξεργασία συνεχούς ροής (Foster, Ghani and Jarmin 2017).

Αρκετές φορές τα ονόματα MapReduce και Hadoop χρησιμοποιούνται σαν να είναι όμοια μοντέλα κάτι το οποίο είναι λάθος. Το MapReduce είναι απλά ένα προγραμματιστικό πρότυπο, το οποίο υποδεικνύει ποια κατηγορία δομής δεδομένων και μετασχηματισμού δεδομένων. Πιο συγκεκριμένα δεν καθορίζει πως θα πρέπει να αποθηκεύονται τα δεδομένα ή πως θα πρέπει να γίνει ο παραλληλισμός για να εκτελεστούν οι υπολογισμοί Το Hadoop αντίθετα αποτελεί μια συγκεκριμένη εφαρμογή του προτύπου MapReduce με ακριβείς προδιαγραφές για το πως πρέπει να αντιμετωπιστούν τα δεδομένα και οι υπολογισμοί εντός του συστήματος (Foster, Ghani and Jarmin 2017).

Μόλις τα στοιχεία δεδομένων αποθηκευτούν σε διάφορους κόμβους του συμπλέγματος, μπορούν να υποστούν επεξεργασία. Το πλαίσιο MapReduce εξασφαλίζει ότι οι εργασίες αυτές ολοκληρώνονται επιτρέποντας την παράλληλη κατανεμημένη επεξεργασία των τμημάτων δεδομένων στους πολλαπλούς κόμβους του συμπλέγματος (Big Data Framework 2018).

Η πρώτη λειτουργία του πλαισίου MapReduce είναι η εκτέλεση μιας διαδικασίας "χάρτη". Ένας από τους κόμβους στο σύμπλεγμα ζητούν τη διαδικασία χάρτη– συνήθως με τη μορφή ενός ερωτήματος Java–προκειμένου να επεξεργαστούν ορισμένα δεδομένα. Ο κόμβος που ξεκινάει τη διαδικασία του Χάρτη φέρει την ετικέτα “Job Tracker”. Στη συνέχεια, ο “Job Tracker” αναφέρεται στο NameNode για να προσδιορίσει ποια δεδομένα είναι απαραίτητα για την εκτέλεση του αιτήματος όπου τα στοιχεία δεδομένων βρίσκονται στο σύμπλεγμα. Μόλις καθοριστεί η θέση των απαραίτητων τμημάτων δεδομένων, ο Job Tracker υποβάλλει το ερώτημα σε μεμονωμένους κόμβους, όπου κάνουν την επεξεργασία. Έτσι, η επεξεργασία πραγματοποιείται τοπικά μέσα σε κάθε κόμβο, καθορίζοντας το βασικό χαρακτηριστικό της κατανεμημένης επεξεργασίας. (Big Data Framework 2018)

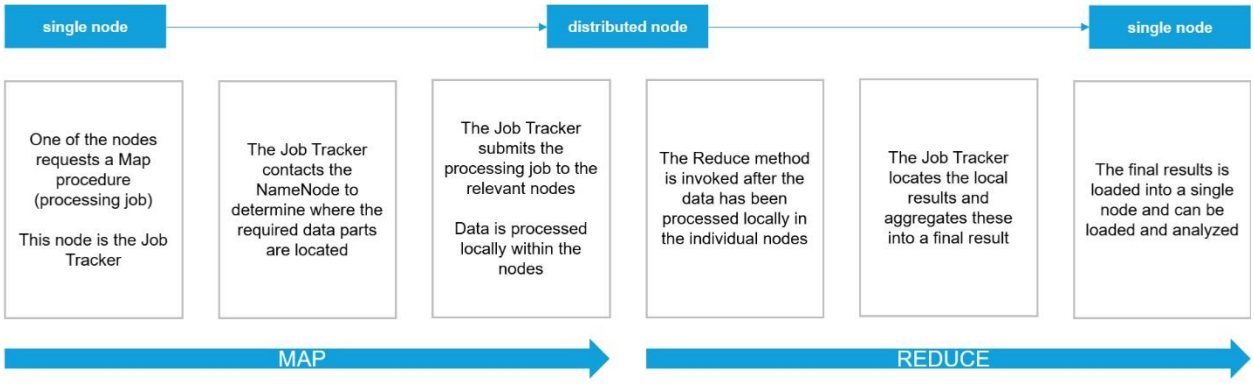
Η δεύτερη λειτουργία του πλαισίου MapReduce είναι η εκτέλεση της μεθόδου “Reduce”. Αυτή η διαδικασία πραγματοποιείται μετά την επεξεργασία. Όταν εκτελείται η εργασία “Reduce”, ο “Job Tracker” εντοπίζει τα τοπικά αποτελέσματα (από τη διαδικασία χάρτη) και συγκεντρώνει τα στοιχεία αυτά μαζί σε ένα τελικό αποτέλεσμα. Αυτό το τελικό αποτέλεσμα είναι η απάντηση στο αρχικό ερώτημα και μπορεί να φορτωθεί σε οποιοδήποτε περιβάλλον αναλύσεων και οπτικοποίησης. Τα βήματα διαδικασιών της MapReduce απεικονίζονται στην Εικόνα 20 (Big Data Framework 2018).

3.5.4 Κόμβος Slave

Οι Κόμβοι Slave είναι οι κόμβοι του συμπλέγματος που ακολουθούν τις οδηγίες από τον “Job Tracker”. Σε αντίθεση με το NameNode, οι Κόμβοι Slave δεν παρακολουθούν τη θέση των δεδομένων (Big Data Framework 2018).

3.5.5 Job Tracker

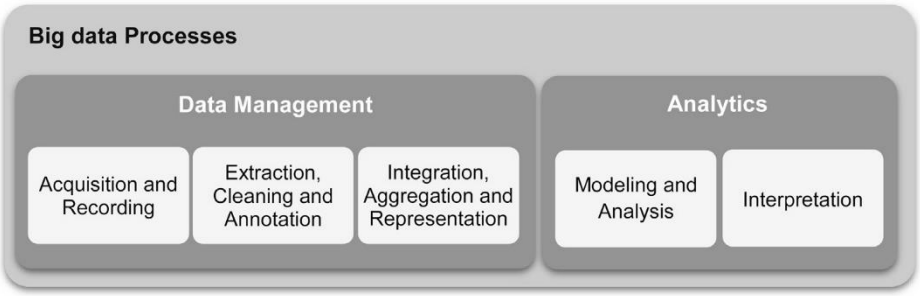
Ο “Job Tracker” - που εισάγεται στο τμήμα MapReduce–είναι ο κόμβος στο σύμπλεγμα που είναι ξεκινά και συντονίζει τις εργασίες επεξεργασίας. Επιπλέον, ο “Job Tracker” επικαλείται “Map” και τη μέθοδο “Reduce” (Big Data Framework 2018).



Εικόνα 20: Διαδικασίες MapReduce (Big Data Framework 2018)

4 Τεχνικές Ανάλυσης

Η δυνητική αξία των μεγάλων δεδομένων θα ανακτηθεί μόνο όταν χρησιμοποιείται για τη λήψη αποφάσεων. Για να καταστεί δυνατή η λήψη αποφάσεων βάσει τεκμηριωμένων στοιχείων, οι οργανισμοί χρειάζονται αποτελεσματικές διαδικασίες για να μετατρέψουν μεγάλο όγκο και ποικιλία δεδομένων σε ουσιαστική γνώση. Η συνολική διαδικασία εξαγωγής πληροφοριών από μεγάλα δεδομένα μπορεί να αναλυθεί σε πέντε στάδια (Labrinidis and Jagadish 2015), που παρουσιάζονται στην Εικόνα 21. Αυτά τα πέντε στάδια αποτελούν τις δύο βασικές υπο-διαδικασίες: Η διαχείριση δεδομένων περιλαμβάνει διαδικασίες και τεχνολογίες υποστήριξης για την απόκτηση και αποθήκευση δεδομένων και την προετοιμασία και ανάκτηση για ανάλυση. Η αναλυτική, από την άλλη πλευρά, αναφέρεται σε τεχνικές που χρησιμοποιούνται για την ανάλυση και την απόκτηση πληροφοριών από μεγάλα δεδομένα. Έτσι, οι μεγάλες αναλύσεις δεδομένων μπορούν να θεωρηθούν ως μια υποδιαδικασία στη συνολική διαδικασία της «εξαγωγής γνώσης» από τα μεγάλα δεδομένα.



Εικόνα 21: Διαδικασίες για την εξαγωγή πληροφοριών από μεγάλα δεδομένα. (Labrinidis and Jagadish 2015)

Στις επόμενες παραγράφους περιγράφετε οι αναλυτικές τεχνικές μεγάλων δεδομένων για δομημένα και μη δεδομένα. Οι ακόλουθες τεχνικές αντιπροσωπεύουν ένα σχετικό υποσύνολο των διαθέσιμων εργαλείων για την ανάλυση μεγάλων δεδομένων.

Η αναλυτική κειμένου (εξόρυξη κειμένου) αναφέρεται σε τεχνικές που εξάγουν πληροφορίες από κείμενα δεδομένων. Η ροή ψηφιακών δεδομένων που παρουσιάζεται στα μέσα κοινωνικής δικτύωσης, στα μηνύματα ηλεκτρονικού ταχυδρομείου, στα blogs, στα ηλεκτρονικά φόρουμ, στις απαντήσεις των ερευνών, στα εταιρικά έγγραφα, στα νέα και στα κέντρα καταγραφής κέντρων είναι παραδείγματα δεδομένων τύπου κειμένου που διατηρούνται από οργανισμούς. Η αναλυτική

κειμένου περιλαμβάνει στατιστική ανάλυση, υπολογιστική γλωσσολογία και μηχανική μάθηση. Η αναλυτική κειμένου επιτρέπει στις επιχειρήσεις να μετατρέπουν μεγάλους όγκους κειμένων που παράγονται από ανθρώπους σε αξιόλογους αριθμούς, υποστηρίζοντας έτσι τη λήψη αποφάσεων βάσει τεκμηριωμένων στοιχείων.

Για παράδειγμα, οι αναλύσεις κειμένων μπορούν να χρησιμοποιηθούν για την πρόβλεψη της χρηματιστηριακής αγοράς βάσει πληροφοριών που προέρχονται από οικονομικές ειδήσεις (Chung 2014). Πιο συγκεκριμενα ο Chung πρότεινε την ανάπτυξη του BizPro, ενός έξυπνου συστήματος που αυτόματα εξάγει και κατηγοριοποιεί τους συντελεστές Επιχειρηματικής Ευφυίας της εταιρείας από ειδησεογραφικά άρθρα. Η BizPro λαμβάνει τα κείμενα των εταιρικών κειμένων όπως τα άρθρα ειδήσεων και οι εκθέσεις των επιχειρήσεων και παράγει αντιπροσωπευτικούς παράγοντες Επιχειρηματικής Ευφυίας (BI) που εξάγονται από τα έγγραφα. Ενσωματώθηκε η εμπειρία από έναν ειδικό BI στο σχεδιασμό και την ανάπτυξη του BizPro. Η διαδικασία περιλαμβάνει την εξαγωγή και την προσθήκη δεικτών στις λέξεις που βρέθηκαν από τα άρθρα ειδήσεων, τη μοντελοποίηση των παραγόντων BI, τον υπολογισμό του όρου "βάρη σημαντικών" και την κατηγοριοποίηση των παραγόντων BI σε μία από τις πέντε κατηγορίες BI (Chung 2014).

- Διοίκηση λειτουργιών
- Οικονομικοί / περιβαλλοντικοί παράγοντες
- Στρατηγική διαχείριση
- Τεχνολογικές αλλαγές
- Νομικά θέματα

4.1 Σημασιολογική Ανάλυση

4.1.1 Ανάλυση κειμένου

Παρουσιάζεται μια σύντομη ανασκόπηση των μεθόδων ανάλυσης κειμένου παρακάτω. Οι τεχνικές *εξαγωγής πληροφοριών* (Information Extraction) εξάγουν δομημένα δεδομένα από αδόμητο κείμενο. Για παράδειγμα, οι αλγόριθμοι IE μπορούν να εξαντλήσουν τις πληροφορίες όπως το όνομα του φαρμάκου, τη δοσολογία και τη συχνότητα από τις ιατρικές συνταγές. Δύο υποεργασίες στην εξόρυξη πληροφοριών είναι η Αναγνώριση Οντοτήτων (Entity Recognition) και η Εξόρυξη Συσχετίσεων (Relation Extraction). Η διαδικασία ER βρίσκει ονόματα σε κείμενο

και τα ταξινομεί σε προκαθορισμένες κατηγορίες, όπως άτομο, ημερομηνία, τοποθεσία και οργάνωση. Ενώ η RE βρίσκει και εξάγει σημασιολογικές σχέσεις μεταξύ των οντοτήτων μέσα στο κείμενο (π.χ. ατόμων, οργανισμών, φαρμάκων, γονιδίων κλπ.). Για παράδειγμα, η φράση " Steve Jobs co-founded Apple Inc. in 1976 ", με ένα σύστημα RE μπορεί να εξάγει σχέσεις όπως ο FounderOf [Steve Jobs, Apple Inc.]. ή το FoundedIn [AppleInc., 1976] (Gandomi and Haider 2014).

Οι *τεχνικές σύνοψης ενός κειμένου* δημιουργούν αυτόματα μια μικρή περίληψη ενός μόνο ή πολλαπλών εγγράφων. Τα αποτελέσματα της σύνοψης συνοδεύουν τις βασικές πληροφορίες στο αρχικό κείμενο. Οι αιτήσεις περιλαμβάνουν επιστημονικά και ειδησεογραφικά άρθρα, διαφημίσεις, μηνύματα ηλεκτρονικού ταχυδρομείου και ιστολόγια. Σε γενικές γραμμές, η σύνοψη ακολουθεί δύο προσεγγίσεις: την εξορυκτική προσέγγιση και την αφαιρετική προσέγγιση. Η πρώτη, δημιουργεί μια περίληψη από τις αρχικές μονάδες κειμένου (συνήθως προτάσεις). Με βάση τη δεύτερη προσέγγιση, η διατύπωση μιας περίληψης περιλαμβάνει τον προσδιορισμό των κυρίαρχων κομματιών ενός κειμένου και τη σύζευξη τους. Η σημασία των κομματιών αυτών αξιολογείται με την ανάλυση της θέσης και της συχνότητας τους στο κείμενο. Οι τεχνικές σύνοψης εξόρυξης δεν απαιτούν «αντίληψη» του κειμένου. Αντίθετα, οι τεχνικές η αφαιρετικής σύνοψης περιλαμβάνουν την εξαγωγή σημασιολογικών πληροφοριών από το κείμενο. Οι περιλήψεις περιέχουν μονάδες κειμένου που δεν εμφανίζονται απαραίτητα στο αρχικό κείμενο. Για να αναλυθεί το αρχικό κείμενο και να δημιουργηθεί η σύνοψη, η αφηρημένη σύνοψη ενσωματώνει τις προηγμένες τεχνικές επεξεργασίας γλωσσών (Natural Language Processing). Ως αποτέλεσμα, τα αφηρημένα συστήματα τείνουν να παράγουν πιο συνεκτικές περιλήψεις από τα εξωστρεφή συστήματα (Hahn & Mani, 2000). Ωστόσο, τα συστήματα εξόρυξης είναι ευκολότερα υιοθετημένα, ειδικά για μεγάλα δεδομένα (Gandomi and Haider 2014).

Οι *τεχνικές Απάντησης Ερωτήσεων* (Question Answering) παρέχουν απαντήσεις σε ερωτήματα που τίθενται στη φυσική γλώσσα. Η Siri της Apple και η Watson της IBM αποτελούν δείγματα τέτοιων εμπορικών συστημάτων QA. Αυτά τα συστήματα έχουν εφαρμοστεί στον τομέα της υγείας, της οικονομίας, το μάρκετινγκ και την εκπαίδευση. Όπως και με την αφηρημένη σύνοψη, τα συστήματα QA βασίζονται σε τεχνικές NLP. Οι τεχνικές QA ταξινομούνται περαιτέρω σε τρεις κατηγορίες: την προσέγγιση ανάκτησης πληροφοριών (information retrieval), που βασίζεται στη

γνώση και την υβριδική προσέγγιση. Τα συστήματα QA βασισμένα σε IR συχνά έχουν τρία υποσυστήματα. Πρώτο υποσύστημα, η *επεξεργασία των ερωτήσεων*, που χρησιμοποιείται για τον προσδιορισμό των λεπτομερειών, όπως ο τύπος ερωτήματος, η εστίαση ερωτήσεων και ο τύπος απάντησης, που χρησιμοποιείται για τη δημιουργία ενός αιτήματος. Δεύτερο είναι η *επεξεργασία κειμένου*, η οποία χρησιμοποιείται για την ανάκτηση σημαντικών αποσπασμάτων, που έχουν ήδη αναφερθεί από ένα σύνολο υφιστάμενων εγγράφων που χρησιμοποιούν το αίτημα που διατυπώνεται στην ερώτηση για επεξεργασία. Το τρίτο υποσύστημα είναι η *επεξεργασία των απαντήσεων*, η οποία χρησιμοποιείται για να εξάγει τις υποψήφιες απαντήσεις από την έξοδο του προηγούμενου στοιχείου, να τις ταξινομεί και να επιστρέφει την υψηλότερη κατάταξη ως αποτέλεσμα του συστήματος QA. Τα συστήματα QA που βασίζονται στη γνώση δημιουργούν μια σημασιολογική περιγραφή της ερώτησης, η οποία στη συνέχεια χρησιμοποιείται για την αναζήτηση δομημένων πόρων. Τα συστήματα QA βασισμένα στη γνώση είναι ιδιαίτερα χρήσιμα για τομείς, όπως ο τουρισμός, η ιατρική και οι μεταφορές, όπου δεν υπάρχουν μεγάλοι όγκοι προκαταρκτικών εγγράφων. Τέτοιοι τομείς στερούνται πλεονασμού δεδομένων, ο οποίος απαιτείται για συστήματα QA που βασίζονται σε IR. Το Siri της Apple είναι ένα παράδειγμα συστήματος QA που εκμεταλλεύεται την προσέγγιση που βασίζεται στη γνώση. Σε υβριδικά συστήματα QA, όπως ο Watson της IBM, ενώ τα αιτήματα αναλύονται σημασιολογικά, οι υποψήφιες απαντήσεις δημιουργούνται χρησιμοποιώντας τις μεθόδους IR (Gandomi and Haider 2014)

Οι *τεχνικές ανάλυσης συναισθημάτων* (εξόρυξης γνώμης) αναλύουν το κείμενό, το οποίο περιέχει τις απόψεις των ανθρώπων απέναντι σε οντότητες όπως προϊόντα, οργανώσεις, άτομα και γεγονότα. Οι επιχειρήσεις καταγράφουν ολοένα και περισσότερα στοιχεία σχετικά με τις πεποιθήσεις των πελατών τους, γεγονός που έχει οδηγήσει στη διάδοση των συναισθηματικών αναλύσεων. Το μάρκετινγκ, τα οικονομικά, η πολιτική και οι κοινωνικές επιστήμες είναι οι κύριοι τομείς εφαρμογής της ανάλυσης των αισθήσεων. Οι τεχνικές ανάλυσης των συναισθημάτων χωρίζονται περαιτέρω σε τρεις υποομάδες, πιο συγκεκριμένα στο επίπεδο εγγράφων, στο επίπεδο προτάσεων και σε επίπεδο οπτικής άποψης. Στο επίπεδο εγγράφου εκφράζεται όποιο αρνητικό ή θετικό συναίσθημα. Η υπόθεση είναι ότι το έγγραφο περιέχει συναισθήματα σχετικά με μια μοναδική οντότητα. Ενώ ορισμένες τεχνικές κατηγοριοποιούν ένα έγγραφο σε δύο τάξεις, αρνητικές και θετικές, άλλες ενσωματώνουν περισσότερες τάξεις συναισθημάτων (όπως το

σύστημα πέντε αστέρων του Αμαζονίου). Οι τεχνικές σε επίπεδο πρότασης προσπαθούν να καθορίσουν την πολικότητα μιας ξεχωριστής εννοίας σχετικά με μια γνωστή οντότητα που εκφράζεται σε μία μόνο πρόταση. Οι τεχνικές επιπέδου πρότασης πρέπει πρώτα να διακρίνουν τις υποκειμενικές αντιλήψεις από τις αντικειμενικές. Ως εκ τούτου, οι τεχνικές σε επίπεδο προτάσεων τείνουν να είναι πιο περίπλοκες σε σύγκριση με τις τεχνικές σε επίπεδο εγγράφου. Οι τεχνικές με βάση την οπτική αναγνωρίζουν όλα τα συναισθήματα μέσα σε μια τεκμηρίωση και προσδιορίζουν τις πτυχές της οντότητας στις οποίες αναφέρεται το κάθε συναίσθημα. Για παράδειγμα, οι αναθεωρήσεις προϊόντων πελατών περιέχουν συνήθως πληροφορίες σχετικά με διαφορετικές πτυχές (ή χαρακτηριστικά) ενός προϊόντος. Χρησιμοποιώντας τεχνικές βασιζόμενες στην οπτική, ο πωλητής μπορεί να αποκτήσει πολύτιμες πληροφορίες σχετικά με τα διαφορετικά χαρακτηριστικά του προϊόντος που θα χάνονταν αν το συναίσθημα ταξινομούνταν μόνο με όρους πολικότητας (Gandomi and Haider 2014)

4.1.2 Ανάλυση ήχου

Η ανάλυση ήχου αναλύει και εξάγει πληροφορίες από μη δομημένα δεδομένα ήχου. Όταν εφαρμόζεται στην ανθρώπινη ομιλούμενη γλώσσα, οι αναλύσεις ήχου αναφέρονται επίσης ως *αναλυτική λόγου*. Δεδομένου ότι αυτές οι τεχνικές έχουν εφαρμοστεί ως επί το πλείστον σε προφορικό ήχο, οι όροι ανάλυση ήχου και ανάλυση ομιλίας χρησιμοποιούνται συχνά ισοδύναμα. Πρόσφατα, τα κέντρα τηλεφωνικής εξυπηρέτησης πελατών και η υγειονομική περίθαλψη είναι οι τομείς πρωταρχικής εφαρμογής των ηχητικών αναλύσεων (Gandomi and Haider 2014).

Τα τηλεφωνικά κέντρα χρησιμοποιούν ανάλυση ήχου για αποτελεσματική ανάλυση χιλιάδων ή και εκατομμυρίων ωρών καταγεγραμμένων κλήσεων. Οι τεχνικές αυτές συμβάλλουν στη βελτίωση της πελατειακής εμπειρίας, στην αξιολόγηση της απόδοσης των αντιπροσώπων, στην αύξηση των ποσοστών των πωλήσεων, στην παρακολούθηση της συμμόρφωσης με διαφορετικές πολιτικές (π.χ. πολιτικές απορρήτου και ασφάλειας), στην ανάδειξη της συμπεριφοράς των πελατών και στον εντοπισμό προβλημάτων προϊόντων ή υπηρεσιών. Τα συστήματα ανάλυσης ήχου μπορούν να σχεδιαστούν για να αναλύσουν μια ζωντανή κλήση, να διατυπώσουν προτάσεις βασισμένες σε πωλήσεις το πελάτη που έκανε στο παρελθόν και να παρέχουν ανατροφοδότηση στους αντιπροσώπους σε πραγματικό χρόνο. Επιπλέον, τα αυτοματοποιημένα κέντρα κλήσεων

χρησιμοποιούν τις πλατφόρμες Interactive Voice Response (IVR) για τον εντοπισμό και τη χειραγώγηση των απογοητευμένων πελατών (Gandomi and Haider 2014).

Στον τομέα της υγείας, οι αναλύσεις ήχου υποστηρίζουν τη διάγνωση και τη θεραπεία ορισμένων ιατρικών παθήσεων που επηρεάζουν τα πρότυπα επικοινωνίας του ασθενούς. Οι αναλύσεις ήχου μπορούν να βοηθήσουν στην ανάλυση των κραυγών ενός βρέφους, οι οποίες περιέχουν πληροφορίες για την υγεία και την συναισθηματική κατάσταση του βρέφους (Patil 2010). Πιο συγκεκριμένα ο ερευνητής Patil το 2010 χρησιμοποίησε φασματογραφική ανάλυση για τη διάγνωση και αντιμετώπιση νεογνικών προβλημάτων. Μια ανάλυση της κανονικής και μη φυσιολογικής βρεφικής κραυγής παρουσιάζεται χρησιμοποιώντας δέκα διακριτές καταστάσεις κραυγής που παρατηρήθηκαν σε φασματογραφήματα. Παρατηρήθηκε οι κλινικές ανωμαλίες στα νεογνά θα μπορούσαν να συσχετιστούν με τις διαφορές στη φασματική ενέργεια της κατανομής και της αρμονικής δομής των φασματογραφιών. Οι τεράστιες ποσότητες δεδομένων που έχουν καταγραφεί μέσω συστημάτων ομιλούμενης κλινικής τεκμηρίωσης είναι άλλος ένας οδηγός για την υιοθέτηση των ηχητικών αναλύσεων στην υγειονομική περίθαλψη.

Η αναλυτική ομιλίας ακολουθεί δύο κοινές τεχνολογικές εξελίξεις: η προσέγγιση που βασίζεται σε μεταγραφή ευρέως γνωστή ως συνεχής αναγνώριση ομιλίας μεγάλου λεξιλογίου, (Large-Vocabulary Continuous Speech Recognition, LVCSR) και η φωνητική προσέγγιση. Αυτά εξηγούνται παρακάτω.

Τα συστήματα *LVCSR* ακολουθούν μια διαδικασία δύο φάσεων: χρήση δεικτών και αναζήτηση. Στην πρώτη φάση, προσπαθούν να μεταγράψουν την ομιλία περιεχομένου του ήχου. Αυτό γίνεται χρησιμοποιώντας αλγορίθμους αυτόματης ομιλίας (Automatic Speech Recognition, ASR) που αντιστοιχούν τους ήχους σε λέξεις. Οι λέξεις αναγνωρίζονται με βάση ένα προκαθορισμένο λεξικό. Εάν το σύστημα αδυνατεί να βρει την ακριβή λέξη στο λεξικό, επιστρέφει μια παρόμοια σε αυτή. Η έξοδος του συστήματος είναι ένας δείκτης αρχείο που μπορεί να αναζητηθεί και περιέχει πληροφορίες σχετικά με την ακολουθία των λέξεων που χρησιμοποιούνται στην ομιλία. Στη δεύτερη φάση, γίνεται χρήση κλασικών μεθόδων το για την εύρεση του όρου αναζήτησης στο αρχείο ευρετηρίου (Gandomi and Haider 2014)

Τα *φωνητικά συστήματα* λειτουργούν με ήχους ή φωνήματα. Τα φωνήματα είναι οι πιο μικρές μονάδες της δεύτερης άρθρωσης με διακριτική αξία. Δεν είναι τα ίδια φορείς σημασίας , αλλά χρησιμεύουν στο να διαφοροποιούν σημασιολογικά τις μονάδες της πρώτης άρθρωσης , τα μορφήματα . Αν π.χ. στη λέξη θέμα [ˈθema] αντικαταστήσουμε το αρχικό [θ] με το [ð] θα προκύψει μια λέξη με διαφορετική σημασία, δέμα [ˈðema]. Τα φωνητικά συστήματα αποτελούνται επίσης από δύο φάσεις: φωνητική χρήση δεικτών και αναζήτηση. Στην πρώτη φάση, το σύστημα μεταφράζει την ομιλία εισόδου σε μια σειρά από φωνήματα. Αυτό έρχεται σε αντίθεση με τα συστήματα LVCSR όπου η ομιλία μετατρέπεται σε μια ακολουθία των λέξεων. Στη δεύτερη φάση, το σύστημα αναζητά την έξοδο της πρώτης φάσης για την φωνητική αναπαράσταση των όρων αναζήτησης (Gandomi and Haider 2014)

4.1.3 Ανάλυση βίντεο

Οι αναλύσεις βίντεο, γνωστές και ως ανάλυση περιεχομένου βίντεο (Video Content Analysis, VCA) περιλαμβάνει μια ποικιλία τεχνικών για την παρακολούθηση, την ανάλυση και την εξαγωγή σημαντικών πληροφοριών από ροές βίντεο. Παρόλο που η ανάλυση βίντεο δεν έχει χρησιμοποιηθεί σε σύγκριση με άλλους τύπους εξόρυξης δεδομένων, έχουν ήδη αναπτυχθεί διάφορες τεχνικές για επεξεργασία σε πραγματικό χρόνο καθώς και προβιντεοσκοπήσεις. Η αυξανόμενη επικράτηση κάμερας κλειστού κυκλώματος (Closed-Circuit Television,CCTV) και η αυξανόμενη δημοτικότητα κοινοποίησης βίντεο είναι οι δύο κορυφαίοι συντελεστές της ανάπτυξης της ηλεκτρονικής ανάλυσης βίντεο. Μια βασική πρόκληση, ωστόσο, είναι το μέγεθος των δεδομένων βίντεο. Ένα δευτερόλεπτο ενός βίντεο υψηλής ευκρίνειας, από την άποψη του μεγέθους, είναι ισοδύναμο με το 2000 σελίδες κειμένου. Τώρα θεωρούν ότι 100 ώρες βίντεο μεταφορτώνονται στο YouTube κάθε λεπτό.

Οι τεχνολογίες μεγάλων δεδομένων μετατρέπουν αυτήν την πρόκληση σε ευκαιρία. Αποφεύγετε την ανάγκη για χειροκίνητη επεξεργασία με μεγάλη ένταση κόστους και κινδύνου, οι τεχνολογίες μεγάλων δεδομένων μπορούν να αξιοποιηθούν για να ξεκαθαρίσουν και να αντλήσουν πληροφορίες από χιλιάδες ώρες βίντεο. Σαν αποτέλεσμα, τα μεγάλα δεδομένα είναι ο τρίτος παράγοντας που συνέβαλε στην ανάπτυξη ανάλυσης βίντεο (Gandomi and Haider 2014).

Η αυτοματοποιημένη ανάλυση βίντεο έχει σημειώσει ταχεία πρόοδο στον τομέα της ασφάλειας. Το κίνητρο είναι απλό. Η ένταση εργασίας είναι δαπανηρή ένας στρατός των φρουρών ασφαλείας αφαιρείται από την κατώτατη γραμμή. Οι κάμερες συμβάλλουν στη μείωση του κόστους, αλλά ποιος παρακολουθεί τις ροές βίντεο; Η ασφάλεια που βασίζεται στην εργασία είναι επίσης αναποτελεσματική - αφοσιωμένο και με κίνητρο προσωπικό δεν μπορεί να παραμείνει εστιασμένη στα καθήκοντα επιτήρησης και παρακολούθησης για περισσότερα από 20 λεπτά. Η αυτοματοποιημένη ανάλυση βίντεο προσφέρει μια πάντα σύγχρονη εναλλακτική λύση για τις παγκόσμιες εργασίες παρακολούθησης. Αρχικά επικεντρώθηκε σε κινούμενα αντικείμενα, η αυτοματοποιημένη ανάλυση αγνόησε σε μεγάλο βαθμό όλες τις υπόλοιπες πληροφορίες σχετικές με τα συμφραζόμενα. Η σημερινή σύγχρονη αυτοματοποιημένη λειτουργία των συστημάτων ανάλυσης βίντεο ενσωματώνει και μαθαίνει από όλους τους διαθέσιμους αισθητήρες και από όλες τις προηγούμενες πληροφορίες σχετικά με τη σκηνή, τους στόχους και τις αναμενόμενες συμπεριφορές για την περαιτέρω βελτίωση της αποτελεσματικότητας. Επιπλέον με το υψηλό κόστος τους, τα συστήματα επιτήρησης με βάση την εργασία τείνουν να είναι λιγότερα αποτελεσματικά από τα αυτόματα συστήματα (Shan , et al. 2012). Οι αναλύσεις βίντεο μπορούν αποτελεσματικά και δραστικά να επιτελέσουν λειτουργίες επιτήρησης, όπως ανίχνευση παραβιάσεων απαγορευμένων ζωνών, εντοπισμός αντικειμένων που έχουν αφαιρεθεί ή παραμείνει ανεπιτήρητα, ανίχνευση καταστροφής σε συγκεκριμένη περιοχή, αναγνώριση ύποπτων δραστηριοτήτων και ανίχνευση παραβίασης κάμερας, για να αναφέρουμε μερικές. , το σύστημα παρακολούθησης μπορεί να ειδοποιεί το προσωπικό ασφαλείας σε πραγματικό χρόνο ή να ενεργοποιεί αυτόματη ενέργεια (π.χ., ηχητική ειδοποίηση, κλειδαριές ή να ανάβει τα φώτα).

Τα δεδομένα που παράγονται από κάμερες CCTV σε καταστήματα λιανικής πώλησης μπορούν να ληφθούν υπόψιν για ζητήματα επιχειρηματικής ευφυΐας. Με κάποιους τρόπους οι εφαρμογές BI είναι λιγότερο δύσκολες από ό, τι εφαρμογές ασφάλειας. Για παράδειγμα, ελεγχόμενες συνθήκες φωτισμού, καλύτερη γνώση των στόχων ενδιαφέροντος και κυρίαρχη χρήση των στατικών φωτογραφικών μηχανών απλοποιούν τη μοντελοποίηση φόντου, της ανίχνευσης στόχων και της ταξινόμησης. Από την άλλη, οι εφαρμογές BI τείνουν να απαιτούν καλύτερες μεθόδους καταμερισμού και παρακολούθησης για την αντιμετώπιση των υψηλότερων πυκνοτήτων στόχων σε τυπικές εσωτερικές εφαρμογές (Shan , et al. 2012).

Όπως σε άλλους τομείς, καλές λύσεις αυτοματοποιημένης ανάλυσης βίντεο για BI είναι (Gandomi and Haider 2014):

- Συνδυασμός αισθητήρων βίντεο και μη αισθητήρων βίντεο για την παροχή εύρωστων λύσεων. Για παράδειγμα, οι εφαρμογές του σημείου πώλησης (Point Of Sale) συσχετίζουν τις πληροφορίες από συσκευές ανάγνωσης γραμμωτού κώδικα, αναγνώριση ραδιοσυχνότητας (RFID) με οπτικές πληροφορίες για την ανίχνευση κλοπής και ορισμένες μορφές " Sweethearting "(ο ταμίας αποφεύγει τη σάρωση των προϊόντων αντικαθιστά την τιμή με μια φθηνότερη και απάτη επιστροφής χρημάτων ή κάρτας δώρου).
- Οι έξυπνοι αλγόριθμοι μπορούν να συλλέξουν δημογραφικές πληροφορίες σχετικά με τους πελάτες, όπως η ηλικία, το φύλο και η εθνικότητα. Ομοίως, οι λιανοπωλητές μπορούν να μετρήσουν τον αριθμό των πελατών, να μετρήσουν την ώρα που μένουν στο κατάστημα, να εντοπίσουν τα πρότυπα κίνησης, να μετρήσουν το χρόνο παραμονής τους σε διαφορετικές ζώνες και να παρακολουθήσουν ουρές σε πραγματικό χρόνο. Πολύτιμες ιδέες μπορούν να αποκτηθούν με τη συσχέτιση αυτών των πληροφοριών με τα πελατογραφικά στοιχεία για τη λήψη αποφάσεων για την τοποθέτηση προϊόντων, την τιμολόγηση, τη βελτιστοποίηση των συνδυασμών, τον σχεδιασμό προώθησης, τις πολλαπλές πωλήσεις, τον σχεδιασμό και τη στελέχωση.

Μια άλλη πιθανή εφαρμογή της ανάλυσης βίντεο στο λιανικό εμπόριο έγκειται στη μελέτη της αγοραστικής συμπεριφοράς των ομάδων. Μεταξύ των μελών της οικογένειας που συνεργάζονται, μόνο ένας αλληλοεπιδρά με το κατάστημα στην ταμειακή υπηρεσία, προκαλώντας τα παραδοσιακά συστήματα να χάσουν δεδομένα σχετικά με τα πρότυπα αγορών άλλων μελών. Οι αναλύσεις βίντεο μπορούν να βοηθήσουν τους λιανοπωλητές να απαντήσουν σε αυτή τη χαμένη ευκαιρία παρέχοντας πληροφορίες σχετικά με το μέγεθος της ομάδας, τα δημογραφικά στοιχεία του ομίλου και την αγοραστική συμπεριφορά των μεμονωμένων μελών (Shan , et al. 2012)

Η αυτόματη δημιουργία ευρετηρίου και ανάκτηση βίντεο αποτελεί άλλο τομέα των εφαρμογών αναλυτικής προβολής βίντεο. Η εκτεταμένη εμφάνιση των online και offline βίντεο έχει επισημάνει την ανάγκη για ευρετηρίαση περιεχομένου πολυμέσων για εύκολη αναζήτηση και ανάκτηση. Η ευρετηρίαση ενός βιντεο πραγματοποιείται με βάση διαφορετικά επίπεδα

διαθέσιμων πληροφοριών σε ένα βίντεο που περιλαμβάνει τα μεταδεδομένα, το soundtrack, τις μεταγραφές και το οπτικό περιεχόμενο του βίντεο. Στη μέθοδο που βασίζεται στα μεταδεδομένα χρησιμοποιούνται συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (Relational Database Management Systems) για αναζήτηση και ανάκτηση βίντεο. Οι αναλύσεις ήχου και οι τεχνικές ανάλυσης κειμένου μπορούν να εφαρμοστούν για την ευρετηρίαση ενός βίντεο που βασίζεται στις συσχετισμένες μουσικές ταινίες και μεταγραφές, αντίστοιχα (Shan , et al. 2012).

Όσον αφορά την αρχιτεκτονική του συστήματος, υπάρχουν δύο προσεγγιστικές αναλύσεις βίντεο:

Αρχιτεκτονική βασισμένη σε διακομιστές. Σε αυτήν τη διαμόρφωση, το βίντεο που καταγράφεται μέσω κάθε κάμερας οδηγείται πίσω σε ένα κεντρικό και αποκλειστικό διακομιστή που εκτελεί τα αναλυτικά βίντεο. Λόγω των ορίων του εύρους ζώνης, το βίντεο που παράγεται από την πηγή συνήθως συμπιέζεται μειώνοντας τους ρυθμούς καρέ και / ή την ανάλυση εικόνας. Η προκύπτουσα απώλεια πληροφοριών μπορεί να επηρεάσει την ακρίβεια της ανάλυσης. Ωστόσο, η προσέγγιση που βασίζεται σε διακομιστές παρέχει οικονομίες κλίμακας και διευκολύνει τη συντήρηση (Shan , et al. 2012).

Συστήματα αρχιτεκτονικής άκρων (Edge-based) . Η σύγχρονη αρχιτεκτονική δικτύου ορίζει συστήματα άκρων ως ανεξάρτητα συστήματα που τοποθετούνται στην εξωτερική περιφέρεια ή στην άκρη του δικτύου. Αυτά τα συστήματα δεν λειτουργούν στα φυσικά πλαίσια των βασικών επιχειρηματικών συστημάτων. Επίσης αποκλείουν όλα τα πακέτα, εκτός από γνωστά και εγκεκριμένα πακέτα, από την είσοδο στα υποδίκτυα όπου λειτουργούν τα βασικά λειτουργικά συστήματα. Η λογική του συστήματος Edge αποκλείει την κυκλοφορία και αναλαμβάνει ότι όλες οι εισερχόμενες κυκλοφορίες είναι επικίνδυνες, εκτός αν αποδεικνύεται διαφορετικά. Το πακέτο δεδομένων πρέπει να αποδείξει ότι είναι ασφαλές πριν μπορέσει να περάσει από ένα σύστημα άκρων. Συνηθισμένα συστήματα άκρων είναι αυτά που προστατεύουν από την ανεπιθύμητη αλληλογραφία, τις υπερφορτίσεις συστημάτων, ύπουλοι ιοί που συνδέονται με ένα ηλεκτρονικό ταχυδρομείο ή άμεσες επιθέσεις στην βασική υποδομή της εταιρείας. Δηλαδή, η ανάλυση βίντεο πραγματοποιείται τοπικά και στα πρωτογενή δεδομένα που συλλέγονται από την κάμερα. Ως αποτέλεσμα, όλο το περιεχόμενο της ροής βίντεο είναι διαθέσιμο για την ανάλυση, επιτρέποντας μια πιο αποτελεσματική ανάλυση περιεχομένου. Ωστόσο, τα συστήματα με βάση τα άκρα είναι

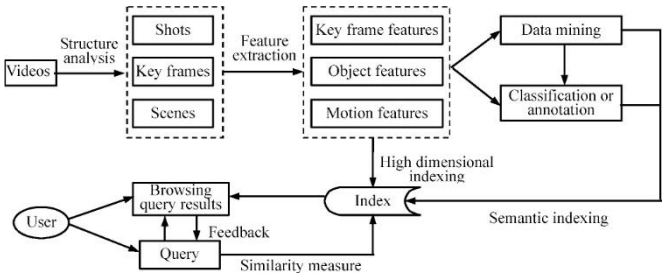
πιο δαπανηρά για να διατηρήσουν και να έχουν χαμηλότερη ισχύ επεξεργασίας σε σύγκριση με τα συστήματα που βασίζονται σε διακομιστές (Shan , et al. 2012).

Μια ολοκληρωμένη ανασκόπηση των προσεγγίσεων και των τεχνικών για την ευρετηρίαση βίντεο παρουσιάζεται σε (Hu, et al. 2011) (Εικόνα 22).

Ανάλυση Δομής

- Ένα *στιγμιότυπο* (shot) είναι μια διαδοχική ακολουθία πλαισίων που έχουν καταγραφεί από το μια δράση κάμερας που λαμβάνει χώρα μεταξύ των λειτουργιών εκκίνησης και διακοπής, που σημαίνουν τα όρια των στιγμιότυπων. Υπάρχουν ισχυροί συσχετισμοί περιεχομένου μεταξύ πλαισίων σε ένα πλάνου. Ως εκ τούτου, πλάνα θεωρούνται οι θεμελιώδεις μονάδες για την οργάνωση του περιεχομένου των ακολουθιών βίντεο και τα πρωταρχικά στοιχεία για υψηλότερο επίπεδο σημασιολογικού σχολιασμού και εργασίων ανάκτησης.
- Υπάρχουν μεγάλες πλεονασμοί μεταξύ των πλαισίων στο ίδιο στιγμιότυπο, επομένως, ορισμένα πλαίσια που αντικατοπτρίζουν καλύτερα τα περιεχόμενα των λήψεων επιλέγονται ως βασικά πλαίσια και συνοπτικά αντιπροσωπεύουν το πλάνο. Τα εξαγόμενα *πλαίσια-κλειδιά* (key framework) πρέπει να περιέχουν όσο το δυνατόν σημαντικότερο περιεχόμενο του στιγμιότυπου και να αποφεύγονται όσο το δυνατόν περισσότερα περιττά σημεία. Τα χαρακτηριστικά που χρησιμοποιούνται για την εξαγωγή βασικών πλαισίων περιλαμβάνουν τα χρώματα (ιδιαίτερα το ιστόγραμμα χρώματος), τις άκρες, σχήματα, οπτική ροή, περιγραφείς κινήσεων MPEG-7 όπως το χρονικό την ένταση κίνησης και τη χωρική κατανομή της κίνησης με διακριτό συντελεστή συνημιτόνου MPEG και διανύσματα κίνησης, δραστηριότητα κάμερας και χαρακτηριστικά που προέρχονται από την εικόνα, και τέλος παραλλαγές που προκαλούνται από την κίνηση της κάμερας.
- Η *κατάτμηση του σκηνικού* (scene) είναι επίσης γνωστή ως κατάτμηση της μονάδας ιστορίας. Γενικά, μια σκηνή είναι μια ομάδα συνεχόμενων λήψεων που είναι συνεπείς με ένα συγκεκριμένο θέμα. Οι σκηνές έχουν υψηλότερο επίπεδο σημασιολογίας από ό, τι τα πλάνα. Οι σκηνές αναγνωρίζονται ή διαχωρίζονται από ομαδοποιημένες διαδοχικές λήψεις

με παρόμοιο περιεχόμενο σε μια σημασιολογική μονάδα. Η ομαδοποίηση μπορεί να βασίζεται σε πληροφορίες από κείμενα, εικόνες ή το κομμάτι ήχου του βίντεο



Εικόνα 22: Γενικό πλαίσιο για την ευρετηρίαση και την ανάκτηση βίντεο βάσει οπτικού περιεχομένου (Hu, et al. 2011)

Χαρακτηριστικά εξαγωγής

Τα *πλαίσια κλειδιά* ενός βίντεο αντικατοπτρίζουν τα χαρακτηριστικά του βίντεο σε κάποιο βαθμό. Οι παραδοσιακές τεχνικές ανάκτησης εικόνας μπορούν να εφαρμοστοούν σε πλαίσια κλειδιά για την επίτευξη της ανάκτησης βίντεο. Το στατικό κλειδί των χαρακτηριστικών καρέ που είναι χρήσιμα για την ευρετηρίαση και την ανάκτηση βίντεο είναι κυρίως ταξινομημένο ως έγχρωμο, βασισμένο σε υφή και με βάση το σχήμα (Hu, et al. 2011).

Τα *χαρακτηριστικά αντικειμένου* περιλαμβάνουν το κυρίαρχο χρώμα, την υφή, το μέγεθος, κ.λπ., των περιοχών εικόνας που αντιστοιχούν στα αντικείμενα. Αυτές οι λειτουργίες μπορούν να χρησιμοποιηθούν για την ανάκτηση βίντεο που ενδέχεται να περιέχουν παρόμοια αντικείμενα. Τα πρόσωπα είναι χρήσιμα αντικείμενα σε πολλά συστήματα ανάκτησης βίντεο (Hu, et al. 2011).

Η *κίνηση* είναι το βασικό χαρακτηριστικό που διαχωρίζει τις ακίνητες εικόνες από ένα δυναμικό βίντεο. Οι πληροφορίες κίνησης αντιπροσωπεύουν το οπτικό περιεχόμενο με χρονική διακύμανση. Τα χαρακτηριστικά κίνησης πλησιάζουν στις σημασιολογικές έννοιες από τα στατικά χαρακτηριστικά των βασικών πλαισίων και από τα αντικειμενικά χαρακτηριστικά. Η κίνηση βίντεο περιλαμβάνει κίνηση φόντου που προκαλείται από την κίνηση της κάμερας και από την κίνηση του προσκηνίου με τη μετακίνηση αντικειμένων (Hu, et al. 2011).

Εξόρυξη δεδομένων ταξινόμηση και σχολιασμός

Η εξόρυξη δεδομένων βίντεο, η ταξινόμηση και ο σχολιασμός βασίζονται σε μεγάλο βαθμό στην ανάλυση δομής βίντεο και στα εξαγόμενα χαρακτηριστικά του βίντεο. Δεν υπάρχουν όρια μεταξύ της εξόρυξης δεδομένων βίντεο, της ταξινόμησης βίντεο, και του σχολιασμού βίντεο (Hu, et al. 2011).

Α. Ο σκοπός της *εξόρυξης δεδομένων* βίντεο είναι να βρεθούν έτοιμες δομές, μοτίβα συμπεριφοράς, κίνησης, κοινά χαρακτηριστικά σκηνών, μοτίβα γεγονότων έτσι ώστε να δημιουργηθούν ευφυείς εφαρμογές βίντεο, όπως η ανάκτηση βίντεο (Hu, et al. 2011).

Β. Το έργο της *ταξινόμησης βίντεο* είναι να βρεθούν κανόνες από βίντεο που χρησιμοποιούν εξαγόμενα χαρακτηριστικά και στη συνέχεια αντιστοιχίζονται τα βίντεο σε προκαθορισμένες κατηγορίες. Η ταξινόμηση είναι ένας σημαντικός τρόπος για να αυξηθεί η αποτελεσματικότητα της ανάκτησης βίντεο. Το σημασιολογικό χάσμα μεταξύ των εξορύξεων πληροφορίας, όπως το σχήμα, το χρώμα και η υφή, καθώς και η ερμηνεία του παρατηρητή αυτών των πληροφοριών, κάνει την ταξινόμηση με βάση το περιεχόμενο βίντεο με βάση το περιεχόμενο πολύ δύσκολη (Hu, et al. 2011).

Γ. Ο σχολιασμός βίντεο είναι η κατανομή του βίντεο σε λήψεις ή τμήματα βίντεο σε διαφορετικές προκαθορισμένες σημασιολογικές έννοιες, όπως πρόσωπο, αυτοκίνητο, ουρανός, άνθρωποι με τα πόδια. Ο σχολιασμός βίντεο είναι παρόμοια διαδικασία με την ταξινόμηση βίντεο, εκτός από δύο διαφορές . 1) Η ταξινόμηση βίντεο έχει διαφορετική έννοια σε σύγκριση με την σχολιασμό βίντεο, αν και μερικές από αυτές τις οι έννοιες θα μπορούσαν να εφαρμοστούν και στα δύο. και 2) η ταξινόμηση βίντεο ισχύει για πλήρη βίντεο, ενώ ισχύει σχολιασμός βίντεο σε τμήματα βίντεο. Σχολιασμός βίντεο και ταξινόμηση βίντεο μοιράζονται παρόμοιες μεθοδολογίες: Πρώτον, χαρακτηριστικά χαμηλού επιπέδου εξάγονται, και έπειτα ορισμένοι ταξινομητές εκπαιδεύονται και χρησιμοποιούνται για τη χαρτογράφηση των χαρακτηριστικών στις ετικέτες ιδέας / κατηγορίας (Hu, et al. 2011).

Αίτηση και επιστροφή

Μόλις ληφθούν οι δείκτες βίντεο, μπορεί να γίνει η ανάκτηση βίντεο βάσει περιεχομένου. Κατά τη λήψη ενός αιτήματος, μια όμοια μέθοδος μέτρησης χρησιμοποιείται, βάσει των δεικτών, για

την αναζήτηση του υποψήφιου βίντεο σύμφωνα με το ερώτημα. Η επιστροφή των αποτελεσμάτων βελτιστοποιούνται με την ανατροφοδότηση της συνάφειας (Hu, et al. 2011).

Οι *τύποι ερωτημάτων* που βασίζονται σε μη εννοιολογικά θέματα περιλαμβάνουν ερώτημα με βάση το παράδειγμα, ερώτημα με σκίτσο και ερώτημα από αντικείμενα. Εννοιολογικά ερωτήματα βίντεο περιλαμβάνουν αιτήματα από λέξεις-κλειδιά και ερωτήματα από φυσική γλώσσα (Hu, et al. 2011).

Τα μέτρα ομοιότητας βίντεο διαδραματίζουν σημαντικό ρόλο στο περιεχόμενο ανάκτησης βίντεο. Οι μέθοδοι μέτρησης ομοιοτήτων βίντεο μπορούν να ταξινομηθούν σε αντιστοίχιση χαρακτηριστικών, αντιστοίχιση κειμένων, βάση αντιστοίχιση οντολογίας και αντιστοίχιση βάσει συνδυασμού. Η επιλογή της μεθόδου εξαρτάται από τον τύπο ερωτήματος (Hu, et al. 2011).

Στην *ανατροφοδότηση σχετικά με τη συσχέτιση* έχουν αποκτηθεί βίντεο που κατατάσσονται είτε από τον χρήστη είτε αυτόματα. Αυτή η κατάταξη χρησιμοποιείται για να βελτιώσει περαιτέρω τις αναζητήσεις. Οι μέθοδοι βελτίωσης περιλαμβάνουν βελτιστοποίηση σημείων αναζήτησης, προσαρμογή βάρους χαρακτηριστικών, και την ενσωμάτωση πληροφοριών. Η ανατροφοδότηση σχετικότητας γεφυρώνει τη διαφορά μεταξύ των σημασιολογικών εννοιών της συνάφειας αναζήτησης και της χαμηλής στάθμης αναπαράστασης περιεχομένου βίντεο. Η ανατροφοδότηση σχετικότητας επίσης αντανακλά τις προτιμήσεις του χρήστη, λαμβάνοντας υπόψη τα σχόλια των χρηστών σχετικά με τα αποτελέσματα που έχουν ήδη αναζητηθεί προηγουμένως. Όπως η ανάδραση σχετικότητας για την ανάκτηση εικόνας, ανατροφοδότηση συνάφειας για την ανάκτηση βίντεο μπορεί να είναι χωριστεί σε τρεις κατηγορίες: άμεση, έμμεση και ψευδοανατροφοδότηση (Hu, et al. 2011).

Συγγραφή βίντεο και αναζήτηση

Η σύνοψη βίντεο αφαιρεί περιττά δεδομένα και κάνει μια αφηρημένη αναπαράσταση ή περίληψη των περιεχομένων, η οποία εκτίθεται στους χρήστες με ένα ευανάγνωστο τρόπο για να διευκολύνεται η περιήγηση. Μια συνοπτική παρουσίαση βίντεο συμπληρώνει την ανάκτηση βίντεο, καθιστώντας την περιήγηση ανάκτησης βίντεο πιο γρήγορα, ειδικά όταν το συνολικό

μέγεθος της ανάκτησης είναι μεγάλο: Ο χρήστης μπορεί να περιηγηθεί στην περίληψη για να εντοπίσετε τα βίντεο που θέλει (Hu, et al. 2011).

Υπάρχουν δύο βασικές στρατηγικές για την περίληψη βίντεο (Hu, et al. 2011).

1) Στατικές περιλήψεις βίντεο: κάθε μία από τις οποίες αποτελείται από μια συλλογή των πλαισίων κλειδιών που εξάγονται από το βίντεο προέλευσης.

2) Δυναμικά «σκαναρίσματα» βίντεο: κάθε μία από αυτές αποτελείται από μια συλλογή των τμημάτων βίντεο (και των αντίστοιχων τμημάτων ήχου), που εξάγονται από το αρχικό βίντεο και στη συνέχεια θα σχηματίσουν ένα βίντεο κλιπ το οποίο είναι πολύ μικρότερο από το αρχικό βίντεο.

4.1.4 Ανάλυση μέσων κοινωνικών δικτύωσης

Οι αναλύσεις κοινωνικών μέσων αναφέρονται στην ανάλυση δομημένων και μη δομημένων δεδομένων από κανάλια κοινωνικών μέσων. Τα κοινωνικά μέσα με ένα πιο γενικευμένο όρο περιλαμβάνουν μια ποικιλία από ηλεκτρονικές πλατφόρμες που επιτρέπουν στους χρήστες να δημιουργούν και να ανταλλάσσουν περιεχόμενο. Τα κοινωνικά μέσα μπορούν να κατηγοριοποιηθούν στους ακόλουθους τύπους: κοινωνικά δίκτυα (π.χ. Facebook και LinkedIn), blogs (π.χ. Blogger και WordPress), μικρά μέσα τύπου blogs (π.χ. Twitter και Tumblr), κοινωνικές ειδήσεις (π.χ. Digg και Reddit), κοινοποίηση μέσων (π.χ., Instagram και YouTube), πληροφορία (π.χ. Wikipedia και Wikihow), ιστοσελίδες ερωτήσεων και απαντήσεων (π.χ. Yahoo! Answers και Ask.com) και ιστοσελίδες κριτικής και αξιολόγησης (π.χ. Yelp , TripAdvisor) (Barbier and Liu 2011). Επίσης, πολλές εφαρμογές για κινητά, όπως το Find My Friend, παρέχουν μια πλατφόρμα για κοινωνικές αλληλεπιδράσεις και ως εκ τούτου χρησιμεύουν ως κανάλια κοινωνικής μέσων (Gandomi and Haider 2014).

Παρόλο που η έρευνα για τα κοινωνικά δίκτυα χρονολογείται από τις αρχές της δεκαετίας του 1920, τα κοινωνικά μέσα ενημέρωσης είναι ένα πεδίο ανάπτυξης που εμφανίστηκε μετά την εμφάνιση του Web 2.0 στις αρχές της δεκαετίας του 2000. Το χαρακτηριστικό κλειδί των σύγχρονων κοινωνικών αναλυτικών μέσων είναι η φύση τους ως προς τα δεδομένα. Η έρευνα για την ανάλυση κοινωνικών μέσων μαζικής ενημέρωσης εκτείνεται σε διάφορους κλάδους, συμπεριλαμβανομένης της ψυχολογίας, της κοινωνιολογίας, της ανθρωπολογίας, της

πληροφορικής, των μαθηματικών, της φυσικής και της οικονομίας. Η αγορά είναι η πρωταρχική εφαρμογή των αναλυτικών μέσων κοινωνικής δικτύωσης για τα επόμενα χρόνια. Αυτό μπορεί να αποδοθεί στην ευρεία και αυξανόμενη υιοθέτηση των κοινωνικών μέσων από τους καταναλωτές παγκοσμίως. Η υιοθέτηση εργαλείων κοινωνικών μέσων ενημέρωσης έχει δημιουργήσει έναν πλούτο δεδομένων κειμένου, ο οποίος περιέχει κρυφές γνώσεις για τις επιχειρήσεις ως ένα εργαλείο ανταγωνιστικού πλεονεκτήματος. Ειδικότερα, τα ενδιαφερόμενα μέλη μπορούν να αναζητήσουν στο τεράστιο ποσό των δεδομένων των κοινωνικών μέσων ενημέρωσης για την ανίχνευση και την ανακάλυψη νέων γνώσεων (π.χ. δημοτικότητα του εμπορικού σήματος) και ενδιαφέροντα πρότυπα, Ακόμα να κατανοήσουν τις ενέργειες των ανταγωνιστών τους και πώς η βιομηχανία αλλάζει και χρησιμοποιεί τα ευρήματα για να βελτιώνει και να επιτυγχάνει ανταγωνιστικό πλεονέκτημα έναντι των ανταγωνιστών τους. Οι υπεύθυνοι λήψης αποφάσεων μπορούν επίσης να χρησιμοποιήσουν τα ευρήματα τους για την ανάπτυξη νέων προϊόντων ή υπηρεσιών και να λαμβάνουν τεκμηριωμένες στρατηγικές και επιχειρησιακές αποφάσεις (He, Zha and Li 2013).

Πιστεύεται ότι η ανταγωνιστική νοημοσύνη μπορεί να βοηθήσει οργανισμούς να συνειδητοποιήσουν τα δυνατά σημεία και τις αδυναμίες τους, έχοντας τελικό στόχο τη βελτίωση της ικανοποίησης των πελατών. Ανταγωνιστική νοημοσύνη ορίζεται ως «η τέχνη του καθορισμού, της συγκέντρωσης και της ανάλυσης πληροφοριών σχετικά με τα προϊόντα του ανταγωνιστή, τις προωθήσεις προϊόντων, τις πωλήσεις κ.λπ. από εξωτερικές πηγές» (Dey, et al. 2011). Μια επιτυχημένη οργάνωση πρέπει να έχει τη δυνατότητα να επεξεργάζεται όλες τις διαθέσιμες πληροφορίες (π.χ. απόψεις πελατών, τιμές προϊόντος από τους ανταγωνιστές, αναθεωρήσεις υπηρεσιών και προϊόντων), να προσδιορίζει τι έχει συμβεί και να προβλέψει τι θα συμβεί στο άμεσο μέλλον.

4.2 Τεχνικές Οπτικοποίησης

Πρόκειται για τεχνικές που χρησιμοποιούνται για τη δημιουργία πινάκων, εικόνων, διαγραμμάτων και άλλων διαισθητικών τρόπων απεικόνισης που βοηθούν στη κατανόηση των δεδομένων μεγάλης κλίμακας Η οπτικοποίηση των Big Data δεν είναι τόσο εύκολη όσο αυτή των παραδοσιακών σχετικά μικρών συνόλων δεδομένων εξαιτίας των 5 Vs που αναφέρθηκαν πιο πάνω. Για αυτό τον λόγο απαιτείται μια επέκταση των παραδοσιακών προσεγγίσεων απεικόνισης.

Έχουν γίνει κάποια βήματα, αλλά ακόμη υπάρχει αρκετός χώρος βελτίωσης. Στην οπτικοποίηση δεδομένων μεγάλης κλίμακας οι περισσότεροι ερευνητές χρησιμοποιούν μια γεωμετρική μοντελοποίηση και εξαγωγή χαρακτηριστικών για να μειώσουν σημαντικά το μέγεθος των δεδομένων πριν από την πραγματική απόδοση. Στη προσπάθεια απεικόνισης των Big Data, σημαντικό ρόλο παίζει η επιλογή του σωστού δείγματος δεδομένων που θα εκπροσωπήσει το σύνολο των δεδομένων (Chen and Zhang 2014).

4.2.1 Χάρτες θερμότητας

Οι χάρτες θερμότητας είναι μια αρκετά αποτελεσματική τεχνική οπτικής ανάλυσης για την έκφραση μοτίβων, συνθέσεις δεδομένων μέσω μερικών ή ολικών σχέσεων και γεωγραφικής κατανομής δεδομένων. Διευκολύνουν επίσης τον εντοπισμό των περιοχών ενδιαφέροντος και την ανακάλυψη των ακραίων τιμών μέσα σε ένα σύνολο δεδομένων. Ο χάρτης θερμότητας είναι μια οπτική, κωδικοποιημένη με βάση το χρώμα αναπαράσταση των τιμών των δεδομένων. Κάθε τιμή παίρνει ένα χρώμα σύμφωνα με τον τύπο ή τη περιοχή που εμπίπτει. Για παράδειγμα ένας χάρτης μπορεί για τις τιμές από 0 ως 3 να αντιστοιχεί το κόκκινο χρώμα, από 4 ως 6 το πορτοκαλί και από 7 ως 10 το πράσινο. Ένας χάρτης θερμότητας μπορεί να είναι στη μορφή γραφήματος ή χάρτη. Ένα διάγραμμα αντιπροσωπεύει ένα πίνακα τιμών στον οποίο κάθε κελί είναι χρωματικά κωδικοποιημένο ανάλογα με τη τιμή. Μπορεί επίσης να αντιπροσωπεύει ιεραρχικές τιμές χρησιμοποιώντας ένθετα ορθογώνια χρωματικού κώδικα (Erl, Khattak and Buhler 2015)

4.2.2 Διαγράμματα χρονοσειρών

Επιτρέπουν την ανάλυση των δεδομένων που καταγράφονται σε περιοδικά χρονικά διαστήματα. Αυτό το είδος της ανάλυσης χρησιμοποιεί τις χρονοσειρές που είναι μια χρονικά διατεταγμένη συλλογή τιμών που καταγράφονται ανά τακτά χρονικά διαστήματα. Ένα παράδειγμα είναι μια χρονοσειρά που περιέχει στοιχεία για τις πωλήσεις που καταγράφονται στο τέλος κάθε μήνα. Η ανάλυση χρονοσειρών βοηθά στην ανακάλυψη χρονικά εξαρτώμενων μοτίβων μέσα στα δεδομένα. Μετά τον εντοπισμό του, το μοτίβο μπορεί να προεκταθεί και στις μελλοντικές προβλέψεις. Οι αναλύσεις χρονοσειρών χρησιμοποιούνται συνήθως για πρόβλεψη εντοπίζοντας μακροπρόθεσμες τάσεις, εποχικά περιοδικά μοτίβα και ακανόνιστες βραχυπρόθεσμες διακυμάνσεις στο σύνολο των δεδομένων. Σε αντίθεση με άλλες αναλύσεις και τεχνικές, εδώ η

μεταβλητή σύγκρισης είναι πάντα ο χρόνος και τα δεδομένα που συλλέγονται είναι εξαρτώμενα από αυτόν (Erl, Khattak and Buhler 2015).

4.2.3 Χωρικά δεδομένα

Τα Γεωγραφικά Πληροφοριακά Συστήματα (GIS) εξελίχθηκαν τα τελευταία χρόνια σαν ένα πολύ χρήσιμο εργαλείο για τη διαχείριση χωρικών προβλημάτων. Ένας βασικός λόγος, που ισχύει αυτό είναι η επεξεργασία πολλαπλών κριτηρίων, που αφορούν στην στήριξη απόφασης για την επιλογή κατάλληλης γης. Ένα τέτοιο Πληροφοριακό Σύστημα δεν είναι χρήσιμο μόνο για την απεικόνιση των δεδομένων, αλλά και για την αξιολόγηση εναλλακτικών επιλογών με βάση τα χωρικά κριτήρια. Όσον αφορά στη αξιολόγηση περιοχών για την ανάπτυξη αιολικών πάρκων, το GIS μπορεί να συνεισφέρει ως ένα εργαλείο στήριξης αποφάσεων, το οποίο στοχεύει στον εντοπισμό οικονομικά βιώσιμων και περιβαλλοντικά εφικτών τοποθεσιών χρησιμοποιώντας ένα μεγάλο αριθμό χωρικών δεδομένων, που σχετίζονται με τεχνικά, οικονομικά, κοινωνικά και περιβαλλοντικά κριτήρια. Για τη διαχείριση τέτοιου είδους πολλαπλών και αντικρουόμενων κριτηρίων γίνεται χρήση κάποιας πολυκριτήριας μεθόδου στήριξης απόφασης, η οποία ενσωματώνεται στο περιβάλλον του GIS με σκοπό να λειτουργήσει σαν ένα ισχυρό εργαλείο, που αξιολογεί την καταλληλότητα των περιοχών για την κατασκευή αιολικών πάρκων (Latinopoulos and Kechagia 2015).

4.3 Στατιστική

Τα μεγάλα δεδομένα είναι ο τομέας της γνώσης που διερευνά τις τεχνικές, τις δεξιότητες και την τεχνολογία που εξάγει πολύτιμες ιδέες από τις τεράστιες ποσότητες δεδομένων. Σε αυτό το κεφάλαιο, η εστίαση θα γίνει κυρίως στις τεχνικές που είναι απαραίτητες για την εξαγωγή αυτών των στοιχείων από τα δεδομένα. Για να βρεθεί «τιμή» σε σύνολα δεδομένων, οι data scientists εφαρμόζουν αλγόριθμους. Αλγόριθμοι είναι σαφείς προδιαγραφές σχετικά με τον τρόπο επίλυσης μιας κατηγορίας προβλημάτων. Οι αλγόριθμοι μπορούν να εκτελούν εργασίες υπολογισμού, επεξεργασίας δεδομένων και αυτοματοποιημένες διαδικασίες αξιολόγησης (Big Data Framework 2018).

Με την εφαρμογή αλγορίθμων στα δεδομένα μεγάλου όγκου, πολύτιμες γνώσεις και πληροφορίες αποκτώνται. Ένα πολύ βασικό παράδειγμα ενός αλγορίθμου - που βρίσκει τη μέγιστη τιμή σε ένα σύνολο δεδομένων - απεικονίζεται στις παρακάτω γραμμές.

```
def find_max (L):
    max=0
    for x in L:
        if x>max:
            max=x
    return max
```

Οι αλγόριθμοι μπορούν και κυμαίνονται από πολύ απλοί με λίγες μόνο γραμμές κώδικα, μέχρι πολύ εξελιγμένους και σύνθετους, με εκατομμύρια γραμμές κώδικα. Η εφαρμογή των αλγορίθμων, και η μετέπειτα χρήση τους για το Big Data, βασίζονται στον επιστημονικό τομέα στατιστικών στοιχείων. Όλοι όσοι ασχολούνται με την επιστήμη των δεδομένων θα πρέπει, συνεπώς, να κατέχουν γνώσεις σχετικά με τις στατιστικές λειτουργίες και τον τρόπο με τον οποίο θα μπορούσαν να εφαρμοστούν σε αλγόριθμους. Σε αυτό το υποκεφάλαιο θα συζητηθούν συνεπώς οι βασικές στατιστικές λειτουργίες και κοινοί αλγόριθμοι που χρησιμοποιούνται σε πακέτα αναλύσεων Big Data (Big Data Framework 2018).

4.3.1 Ταξινόμηση

Η ταξινόμηση χρησιμοποιείται στην εξόρυξη δεδομένων και ο λόγος είναι ότι χαρακτηρίζει τα κύρια δεδομένα σε διακριτές ομάδες με βάση ορισμένα χαρακτηριστικά που ορίζονται στην αρχική κατάσταση. Έτσι ο κύριος ρόλος της ταξινόμησης είναι η πρόβλεψη μελλοντικών γεγονότων και συμπεριφορών. Χαρακτηρίζεται ως μια επιτηρούμενη τεχνική μάθησης, σε

αντίθεση με την ανάλυση κατά συστάδες, καθώς χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης που περιέχει τη μεταβλητή βάσει της οποίας τα δεδομένα ταξινομούνται σε ομάδες και ορίζεται και το χαρακτηριστικό της ομάδας. Πάντως η ταξινόμηση μπορεί να λειτουργήσει αποτελεσματικά και σε πολύ μεγάλα σύνολα δεδομένων. (Πολυμένης 2017)

Γενικά η διαδικασία της ταξινόμησης αποτελείται από δύο βήματα. Το πρώτο είναι η τροφοδότηση του συστήματος με δεδομένα εκπαίδευσης που είναι ήδη κατηγοριοποιημένα ή τους έχουν αποδοθεί ετικέτες (labeled data. Στο δεύτερο, το σύστημα τροφοδοτείται με άγνωστα, αλλά παρόμοια στοιχεία για ταξινόμηση και ανάλογα με τα αποτελέσματα που δημιουργήθηκαν από τα δεδομένα εκπαίδευσης, ο αλγόριθμος κατατάσσει τα δεδομένα που δεν έχουν επισημανθεί με ετικέτες (unlabeled data) (Πολυμένης 2017).

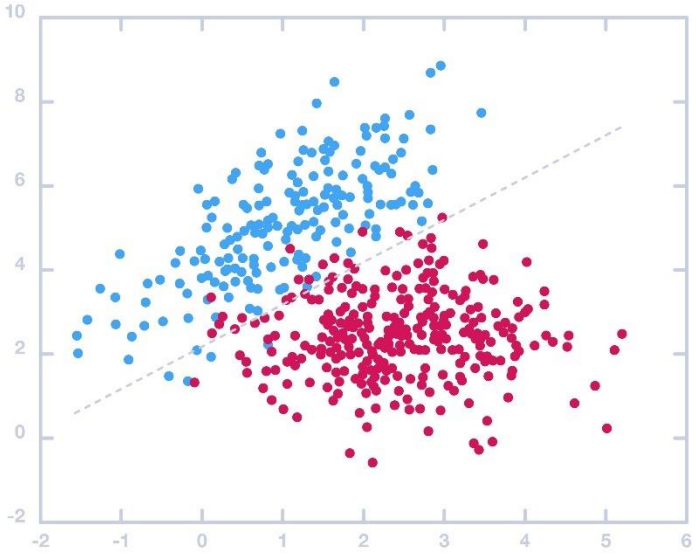
Επειδή ο υπολογιστής "τροφοδοτείται" με δειγματοληπτικά δεδομένα, η ταξινόμηση είναι μια μορφή εποπτείας μηχανικής μάθησης.

Ένας αλγόριθμος ταξινόμησης−που δηλώνεται απλώς−εκτελεί τα ακόλουθα βήματα (Big Data Framework 2018):

1) Ένας υπολογιστής τροφοδοτείται με δειγματοληπτικά δεδομένα που περιέχουν πληροφορίες για την κλάση κάθε δεδομένων. Για παράδειγμα, μαθαίνει να ταξινομεί τα καρότα ως "λαχανικά" και τα πορτοκάλια ως "φρούτα".
2) Μετά την «εκπαίδευση» του μηχανήματος, προσφέρονται νέα δεδομένα ή παρατηρήσεις στον υπολογιστή.
3) Ο υπολογιστής αρχίζει να ταξινομεί από μόνη της. Στο παράδειγμα, βρώσιμα που έχουν παρόμοια χαρακτηριστικά τα καρότα θα φέρουν την ένδειξη "λαχανικά", ενώ στα πορτοκάλια θα επισημαίνονται ως "φρούτα".

Ένα ακόμα παράδειγμα θα ήταν η ανάθεση μηνύματος spam ή μη spam σε μηνύματα ηλεκτρονικού ταχυδρομείου, διάγνωση σε έναν ασθενή όπως περιγράφεται από τα παρατηρούμενα χαρακτηριστικά του ασθενούς (φύλο, αρτηριακή πίεση, παρουσία ή απουσία ορισμένων συμπτωμάτων κ.λπ.) (Big Data Framework 2018).

Ένα παράδειγμα της διαδικασίας ταξινόμησης απεικονίζεται στην παρακάτω Εικόνα 23.



Εικόνα 23: Παράδειγμα γραμμικού ταξινομητή (Big Data Framework 2018)

4.3.2 Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες ομαδοποιεί ένα σύνολο αντικειμένων με τέτοιο τρόπο ώστε αντικείμενα στην ίδια ομάδα (που ονομάζεται σύμπλεγμα) είναι πιο όμοια (με κάποια έννοια) μεταξύ τους σε άλλες ομάδες (ομάδες) (Big Data Framework 2018). Το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση σχετικά με την ύπαρξη ομάδων χαρακτηρίζει την Ανάλυση κατά συστάδες ως μη επιβλεπόμενη μάθηση. Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους. Οι παρατηρήσεις θεωρούνται ως σημεία σε έναν πολυδιάστατο χώρο και αυτή η απόσταση αποτελεί το μέτρο της ομοιότητας τους. Εάν όλα τα γνωρίσματα είναι αριθμητικά, τότε για τον υπολογισμό της ανομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση ή κάποια παραλλαγή της, όπως η απόσταση Manhattan ή η απόσταση Minkowski (Κύρκος 2015).

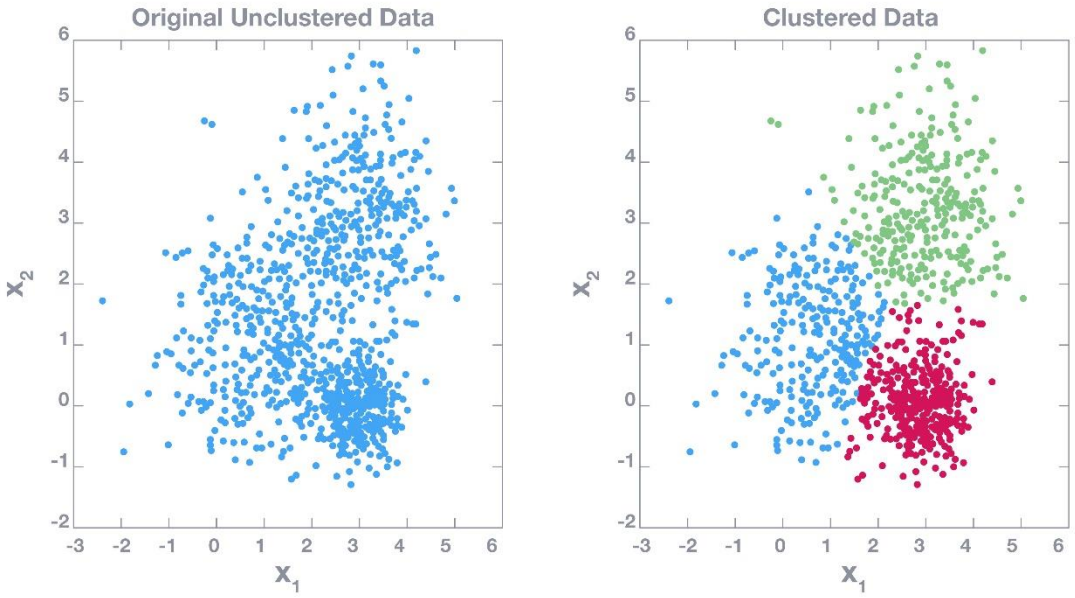
Υπάρχει μια μεγάλη ποικιλία μεθόδων ανάλυσης κατά συστάδες. Οι μέθοδοι αυτές χωρίζονται σε Ιεραρχικές, Διαχωριστικές, μεθόδους βασισμένες στην πυκνότητα, μεθόδους πλέγματος και μεθόδους βασισμένες σε μοντέλα. Οι Ιεραρχικές μέθοδοι δημιουργούν μια ιεραρχία επιπέδων,

κάθε ένα από τα οποία περιλαμβάνει ένα σύνολο συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Η ιεραρχία των επιπέδων και οι αντίστοιχες συστάδες αναπαριστώντα γραφικά με τη χρήση δενδρογραμμάτων. Οι Ιεραρχικές μέθοδοι υποδιαιρούνται σε συσσωρευτικές, οι οποίες δημιουργούν την ιεραρχία μέσα από μια διαδικασία διαδοχικών συγχωνεύσεων, και σε διαιρετικές, οι οποίες δημιουργούν την ιεραρχία μέσω διαδοχικών διασπάσεων. Για τη συγχώνευση ή διάσπαση συστάδων απαιτείται καθορισμός της απόστασης τους. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης των συστάδων. Οι βασικότεροι από αυτούς είναι η μέθοδος της Απλής Σύνδεσης, η μέθοδος της Πλήρους Σύνδεσης, η Σύνδεση Μέσου Όρου, η μέθοδος Ward κλπ. τα αντικείμενα επιμερίζονται σε k συστάδες. Τυπικά, ο αριθμός των συστάδων προκαθορίζεται από τον χρήστη (Κύρκος 2015).

Στη συνέχεια εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράτε με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη, και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος Διαχωριστικής ΑΣ είναι ο k-Means. Στις βασισμένες στην πυκνότητα μεθόδους, ελέγχεται η πυκνότητα των αντικειμένων, και η συστάδα επεκτείνεται, όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι πλέγματος επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα

οποία συγκροτούν ένα πλέγμα, και η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος (Κύρκος 2015).

Εικόνα 24: Παράδειγμα ανάλυσης κατά συστάδες (Big Data Framework 2018)



4.3.3 Εντοπισμός ακραίων τιμών

Μια απόκλιση είναι ένα σημείο παρατήρησης που απέχει από άλλες παρατηρήσεις. Μπορεί να υπάρξει απόκλιση λόγω μεταβλητότητας της μέτρησης ή μπορεί να υποδεικνύει σφάλμα στα δεδομένα. Ειδικά στην ανάλυση των συνόλων των μεγάλων δεδομένων, η ανίχνευση των ακραίων τιμών είναι μια συχνά χρησιμοποιούμενη τεχνική για την ανίχνευση εσφαλμένων ή λανθασμένων σημείων δεδομένων (Big Data Framework 2018).

Τα έκτοπα δεδομένα είναι γενικά σημεία δεδομένων που φαίνεται να είναι απροσδόκητα σε σύγκριση με τα υπόλοιπα τα δεδομένα– δεν εντάσσονται στο πρότυπο των άλλων σημείων δεδομένων. Η κανονική κατανομή μπορεί να χρησιμοποιηθεί για την ανίχνευση των ακραίων τιμών. Μέσα στην τυποποιημένη κατανομή, το 99% των σημείων δεδομένων ταιριάζει σε τρεις τυπικές αποκλίσεις του μέσου όρου. Αν κάποιο ή περισσότερα σημεία δεδομένων είναι συνεπώς

περισσότερα από τρεις τυπικές αποκλίσεις από το μέσο αυτό μπορεί να αποτελεί ένδειξη ότι τα σημεία αυτά είναι εσφαλμένα ή περιέχουν ελαττωματικά δεδομένα (Big Data Framework 2018).

Στους αλγόριθμους μηχανικής μάθησης− που χρησιμοποιούν συχνά τυποποιημένες τιμές− η ανίχνευση απομακρυσμένων τιμών μπορεί να είναι μια πολύ χρήσιμη λειτουργία, επειδή δείχνει ότι υπάρχει μια πολύ μικρή πιθανότητα ότι αυτό το σημείο δεδομένων θα εμφανίζονταν. Η ανίχνευση των ακραίων τιμών είναι μια ευρέως χρησιμοποιούμενη τεχνική, ειδικά στο πλαίσιο του Big Data. Ασφαλιστικές εταιρείες και εταιρίες πιστωτικών καρτών χρησιμοποιούν ανίχνευση ακραίων τιμών που ανιχνεύουν δόλιες αξιώσεις ή συναλλαγές εξετάζοντας δεδομένα που δεν ταιριάζουν στο κανονικό πρότυπο. Ομοίως, αλγόριθμοι ανίχνευσης απομακρυσμένων τιμών χρησιμοποιούνται από τις υπηρεσίες πληροφοριών για την ανίχνευση ανωμαλιών σε ατομικές συμπεριφορές που μπορεί να αποτελέσει απειλή για την εθνική ασφάλεια (Big Data Framework 2018).

4.3.4 Νευρωνικά Δίκτυα

Είναι υπολογιστικά μοντέλα με έμπνευση από τη δομή και λειτουργία των βιολογικών νευρωνικών δικτύων εντός του ανθρώπινου εγκεφάλου που προσπαθούν να βρουν μοτίβα μέσα στα δεδομένα. Είναι κατάλληλα για την εύρεση μη γραμμικών μοτίβων. Μπορούν να χρησιμοποιηθούν και για την αναγνώριση προτύπων αλλά και για βελτιστοποίηση. Μερικές εφαρμογές τους είναι με επιβλεπόμενη μάθηση και άλλες με μη επιβλεπόμενη (Πολυμένης 2017).

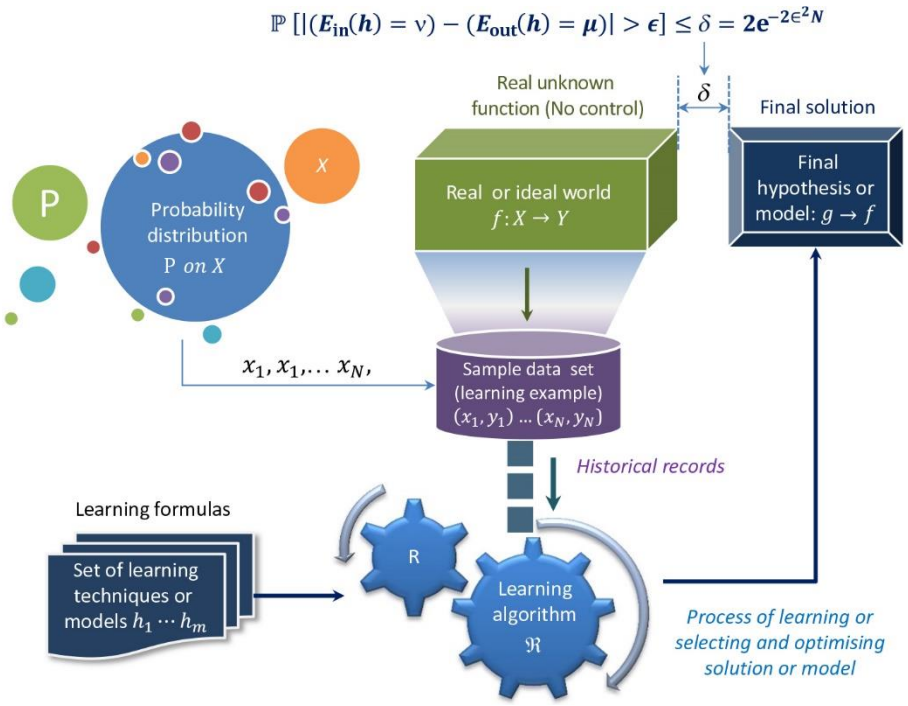
Ένα νευρωνικό δίκτυο αποτελείται από υπολογιστικούς κόμβους που είναι διασυνδεδεμένοι μεταξύ τους. Κάθε ένας από τους κόμβους, δέχεται ένα σύνολο από εισόδους από διαφορετικούς κόμβους, επιτελεί κάποιον υπολογισμό βασισμένο στις εισόδους και παράγει ένα αποτέλεσμα. Αυτό το αποτέλεσμα, είτε τροφοδοτείται σε άλλους κόμβους ως έξοδος, είτε πηγαίνει στο περιβάλλον. Υπάρχουν τρία διαφορετικά είδη κόμβων που είναι οι κόμβοι εισόδου, εξόδου και υπολογιστικοί. Οι πρώτοι δεν κάνουν υπολογισμούς και απλώς υπάρχουν ανάμεσα στο περιβάλλον και τους υπολογιστικούς κόμβους. Οι κόμβοι εξόδου είναι αυτοί που βγάζουν στο περιβάλλον τα τελικά αποτελέσματα του δικτύου. Τέλος οι υπολογιστικοί είναι αυτοί που πολλαπλασιάζουν κάθε στοιχείο εισόδου με μια βαρύτητα και υπολογίζουν το άθροισμα των γινομένων. Αυτό πηγαίνει μετά στη συνάρτηση ενεργοποίησης που έχει ο κάθε κόμβος με τη τιμή

που λαμβάνει η συνάρτηση για τη κάθε είσοδο να είναι και το αποτέλεσμα του κόμβου (Πολυμένης 2017).

4.4 Μηχανική μάθηση (Machine Learning)

Ουσιαστικά η ML είναι μια αυτόματη διαδικασία αναγνώρισης προτύπων από μια μηχανή εκμάθησης. Ο κύριος στόχος της ML είναι η δημιουργία συστημάτων που μπορούν να επιτελέσουν ή να ξεπεράσουν την ικανότητα του ανθρώπινου επιπέδου στον χειρισμό πολλών πολύπλοκων εργασιών ή προβλημάτων. Η ML αποτελεί μέρος της τεχνητής νοημοσύνης (AI). Κατά τη διάρκεια των πρώτων χρόνων της τεχνητής νοημοσύνης, στόχος ήταν η κατασκευή ρομπότ και η προσομοίωση των ανθρώπινων δραστηριοτήτων. Αργότερα, η εφαρμογή της AI έχει γενικευτεί για να λύσει τα γενικά προβλήματα από μια μηχανή. Η δημοφιλής λύση ήταν η τροφή έναν υπολογιστή με αλγόριθμους (ή μια ακολουθία οδηγιών), ώστε να μετατρέπει τα δεδομένα εισόδου σε απαντήσεις στην έξοδο. Αυτό συχνά ονομάζεται σύστημα βασισμένο σε κανόνες ή “Good Old-Fashioned of Artificial Intelligence” (GOFAI), όπως τα συστήματα εμπειρογνωμόνων.

Ωστόσο, δεν είναι εφικτό να βρεθούν κατάλληλοι αλγόριθμοι για πολλά προβλήματα, για παράδειγμα, την αναγνώριση ανθρώπινου χειρόγραφου. Η αρχή της μάθησης από τα δεδομένα είναι παρόμοια τόσο με δοκιμή όσο και με σφάλμα. Αυτό σημαίνει ότι μια δοκιμασία θα μπορούσε να έχει μεγάλο σφάλμα, αλλά αν μπορέσουμε να συγκεντρώσουμε πολλές δοκιμές, το σφάλμα θα μειωθεί σε αποδεκτό επίπεδο ή σύγκλιση. Η απεικονίζει ένα τυπικό παράδειγμα μιας διαδικασίας ML από τα δεδομένα.



Εικόνα 25: Τυπικό παράδειγμα μηχανικής μάθησης (Buyya, Calheiros and Dastjerdi 2016)

Από την έκρηξη εμφάνισης ιστοσελίδων στα τέλη της δεκαετίας του 1990, ο όγκος των δεδομένων γίνεται ολοένα και μεγαλύτερος. Μια λογική ερώτηση είναι πώς θα αντιμετωπιστεί αυτός ο μεγάλος όγκος δεδομένων με σκοπό να βρεθούν χρήσιμα μοτίβα από ένα μεγαλύτερο όγκο δεδομένων. Αυτό οδηγεί στην ανακάλυψη της γνώσης μέσα στις βάσεις δεδομένων, δηλαδή στον όρο εξόρυξη δεδομένων. Με άλλα λόγια, γίνεται σκοπός των ειδικών να «σκάψουν» στη βάση δεδομένων και να ανακαλύψουν το νόημα ή τη γνώση για τη λήψη αποφάσεων. Κάτι τέτοιο επιτυγχάνεται με την ανακάλυψη μοτίβων και τάσεων σε μεγάλα σύνολα δεδομένων. Τα στατιστικά στοιχεία είναι τα ζωτικά εργαλεία για την προσθήκη αξίας στη δειγματοληψία, τη μοντελοποίηση, την ανάλυση, και παρουσίαση. Αυτό οδηγεί στη σύγκλιση των συστημάτων εξόρυξης δεδομένων και του ασαφούς συστήματος κάτω από τη μεγάλη ομπρέλα της Μηχανικής μάθησης. Υπάρχουν διάφοροι ορισμοί για τη Μηχανική Μάθηση κάποιοι ορισμοί συνοψίζονται πιο κάτω (Buyya, Calheiros and Dastjerdi 2016):

- Εκπαιδεύστε το μηχάνημα για να μάθετε αυτόματα και να βελτιώνετε τα αποτελέσματα καθώς παίρνει περισσότερα δεδομένα
- Ανακαλύψτε ή αναγνωρίστε μοτίβα και νοημοσύνη με δεδομένα εισόδου

- Προβλέψτε σε άγνωστες εισόδους
- Το μηχάνημα θα αποκτήσει γνώσεις απευθείας από τα δεδομένα και θα επιλύσει προβλήματα

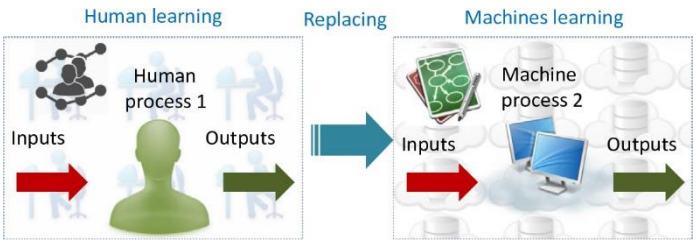
Η ML στηρίζει την εφαρμογή της BDA. Χωρίς ML να επεξεργαστεί όλο και αυξανόμενα τεράστια δεδομένα, η Αναλυτική Μεγάλων Δεδομένων (BDA) θα ήταν αδύνατη. Στόχος της μηχανικής μάθησης είναι να δημιουργήσει συστήματα που φτάνουν στον επίπεδο της ανθρώπινης εκτέλεσης. Συμπερασματικά, η ML είναι το επίκεντρο κάθε BDA. Όλα τα υπόλοιπα στοιχεία μέσα σε ένα πλαίσιο Big Data στοχεύουν στην υποστήριξη της διαδικασίας ML. Όσον αφορά την υπολογιστική υποστήριξη στην BDA, υπάρχουν τέσσερα μεγάλα αρχιτεκτονικά μοντέλα που είναι σε θέση να επεξεργάζονται μεγάλα ποσά δεδομένων σε εύλογο χρόνο (Buyya, Calheiros and Dastjerdi 2016):

• Μαζική παράλληλη επεξεργασία συστήματος βάσεων δεδομένων: Για παράδειγμα, το EMC της Greenplum της και η Netezza της IBM.

• Συστήματα βάσεων δεδομένων μνήμης, όπως το Oracle Exalytics, το HANA της SAP και το Spark

• Μοντέλο επεξεργασίας MapReduce και πλατφόρμες όπως το Hadoop και το Google File System (GFS)

• Συστήματα Bulk Synchronous Parallel (BSP) όπως Apache HAMA και Giraph



Εικόνα 26: Αντικαθιστώντας τον άνθρωπο στην διαδικασία μάθησης (Buyya, Calheiros and Dastjerdi 2016)

Για την εκτέλεση της BDA με τον πλέον οικονομικά αποδοτικό τρόπο, έχει γίνει ένα πέμπτο μοντέλο, cloud computing (CC) και μια προτιμώμενη λύση ειδικά για μικρές και μεσαίες επιχειρήσεις.

Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή των αλγορίθμων που μπορούν να μάθουν από και να κάνουν προβλέψεις στα δεδομένα. Η μηχανική μάθηση στοχεύει να «διδάξει» υπολογιστές για να εκτελέσει ορισμένες λειτουργίες (εκτελώντας αλγόριθμοι μηχανικής μάθησης), έτσι ώστε ο υπολογιστής να είναι σε θέση να λάβει βελτιωμένες αποφάσεις στο μέλλον και να μπορεί να «μαθαίνει» από προηγούμενες καταστάσεις. Η μηχανική μάθηση χρησιμοποιείται ευρέως για την εξόρυξη δεδομένων, η οποία κινείται μέσα σε μεγάλες ποσότητες δεδομένων για να βρεθούν άγνωστα ή κρυμμένα μοτίβα (Big Data Framework 2018).

Η Μηχανική Μάθηση μπορεί να υποδιαιρεθεί σε δύο διαφορετικές κατηγορίες:

1) *Εποπτευόμενη Μηχανική Μάθηση*: Στην εποπτευόμενη μηχανική μάθηση, ένας υπολογιστής μαθαίνει να εκτελεί μια εργασία δεδομένου ότι τροφοδοτείται με μια ετικέτα δεδομένων εκπαίδευσης. Με άλλα λόγια, ο υπολογιστής είναι ο πρώτος που έρχεται αντιμέτωπος με μια σειρά υποθέσεων ", από τα οποία μαθαίνει τι απόφαση πρέπει να κάνει. Όταν εισάγονται νέα δεδομένα από το σύστημα, το σύστημα στη συνέχεια «αποκαλύπτει» ποια απόφαση θα κάνει. Για το λόγο αυτό, η εποπτευόμενη μηχανική μάθηση σχετίζεται κυρίως με τεχνικές ταξινόμησης και παλινδρόμησης. Ένα παράδειγμα εποπτευόμενης μηχανικής μάθησης είναι η διαλογή των μηνυμάτων ηλεκτρονικού ταχυδρομείου σε φάκελο ανεπιθύμητης αλληλογραφίας ή ως κανονικό ταχυδρομείο. Ο υπολογιστής πρώτα πρέπει να «μαθαίνει» ποιο είδος μηνυμάτων θα πρέπει να θεωρείται ανεπιθύμητο, τροφοδοτώντας ένα σύνολο από δεδομένα εκπαίδευσης. Αφού ο υπολογιστής «καταλάβει» αυτό το σετ εκπαίδευσης και παράγει ορισμένους κανόνες από αυτό, μπορεί να ταξινομήσει τα μελλοντικά μηνύματα ηλεκτρονικού ταχυδρομείου από μόνο του.
2) *Μη εποπτευόμενη μηχανική μάθηση*: Εδώ, ένας υπολογιστής τροφοδοτεί δεδομένα και πρέπει να συμπεράνει σχέσεις στα δεδομένα, χωρίς προηγούμενη γνώση σχετικά με το σύνολο δεδομένων. Κάθε σύνολο δεδομένων μπορεί να τροφοδοτηθεί στον υπολογιστή, από το οποίο το μηχάνημα θα προσπαθήσει να βρει συγκεκριμένα μοτίβα και τις σχέσεις εντός των δεδομένων. Συνεπώς, η μη εποπτευόμενη μηχανική μάθηση είναι ιδανική για σκοπούς της εξόρυξης δεδομένων. Οι τεχνικές που σχετίζονται με τη μηχανική μάθηση είναι ανάλυση κατά συστάδες και συσχέτιση. Ένα παράδειγμα μη εποπτείας μηχανικής μάθησης θα ήταν να τροφοδοτεί μεγάλο ποσό ασφαλιστικών απαιτήσεων σε έναν

υπολογιστή. Αλγόριθμοι που βασίζονται στη μάθηση χωρίς επίβλεψη μπορεί να διαπιστώσουν ότι ορισμένες αξιώσεις δεν ταιριάζουν με ένα κανονικό μοτίβο και ως εκ τούτου μπορεί να είναι δόλια. Αυτές οι περιπτώσεις θα πρέπει στη συνέχεια να αξιολογηθούν και να επικυρωθούν από τους ασφαλιστικούς πράκτορες.

5 Διαδικασίες Big Data

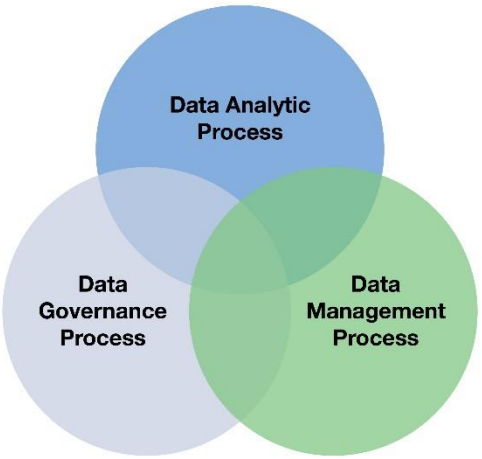
Η εκτέλεση έργων Big Data που προσφέρουν πραγματική αξία στις επιχειρήσεις είναι δύσκολη. Λόγω του όγκου, της ποικιλίας και ταχύτητας των διαθέσιμων πηγών δεδομένων, οι οργανισμοί μπορούν να χαθούν και να δουν μόνο τη συνολική εικόνα και όχι μεμονωμένες καταστάσεις. Σε συνδυασμό με την ατζέντα της διαχείρισης, την πίεση για την επίτευξη αποτελεσμάτων και τους πολύπλοκους αλγορίθμους, πολλοί οργανισμοί αγωνίζονται να επιτύχουν την απαιτούμενη απόδοση της επένδυσης από πρωτοβουλίες ένταξης συστημάτων Big Data (Big Data Framework 2018).

Προκειμένου να αποφευχθούν οι πιθανές παγίδες που φέρνει το Big Data, οι διαδικασίες μπορούν να βοηθήσουν τις επιχειρήσεις να επικεντρώσουν την κατεύθυνση τους. Οι διαδικασίες φέρνουν δομή, μετρήσιμα βήματα και μπορούν αποτελεσματικά να διαχειρίζονται καθημερινά. Επιπλέον, οι διαδικασίες ενσωματώνουν την τεχνογνωσία Big Data στο πλαίσιο της επιχείρησης ακολουθώντας παρόμοιες διαδικασίες και βήματα, ενσωματώνοντάς την ως «πρακτική» του οργανισμού. Η ανάλυση εξαρτάται όλο και λιγότερο από τα άτομα και ως εκ τούτου ενισχύει σημαντικά τις πιθανότητες λήψης αξίας μακροπρόθεσμα (Big Data Framework 2018).

Η εγκατάσταση μεγάλων διαδικασιών δεδομένων στην επιχείρηση μπορεί να είναι αρχικά μια χρονοβόρα εργασία, αλλά σίγουρα παρέχει τα οφέλη μακροπρόθεσμα. Σε αυτή την ενότητα θα συζητήσουμε πως τα μεγάλα δεδομένα και οι διαδικασίες μπορούν να παράσχουν δομή στην ανάλυση των δεδομένων. Οι διαδικασίες μεγάλων δεδομένων μπορούν να υποδιαιρεθούν σε τρεις βασικές υπο-διαδικασίες (Big Data Framework 2018):

- Διαδικασία ανάλυσης δεδομένων (έλεγχος).
- Διαδικασία διακυβέρνησης δεδομένων (συμμόρφωση).
- Διαδικασία διαχείρισης δεδομένων (ποιότητα).

Αν και είναι στενά συνδεδεμένα και ωφέλιμα σε οποιαδήποτε οργάνωση, κάθε μία από τις υπο-διαδικασίες έχει μια διαφορετική εστίαση και λειτουργία: έλεγχος, συμμόρφωση ή ποιότητα Εικόνα 27.



Εικόνα 27: Οι τρεις διαδικασίες των Μεγάλων Δεδομένων Διαδικασία Ανάλυσης Δεδομένων (Big Data Framework 2018)

Η διαδικασία ανάλυσης δεδομένων περιέχει διαδοχικά βήματα που οι επιχειρήσεις εκτελούν προκειμένου να επεξεργαστούν τα μεγάλα δεδομένα. Ιδανικά θα πρέπει, η ίδια διαδικασία να χρησιμοποιείται για κάθε έργο Big Data έτσι ώστε να υπάρχει συνοχή μεταξύ των πρότζεκτ (Big Data Framework 2018).

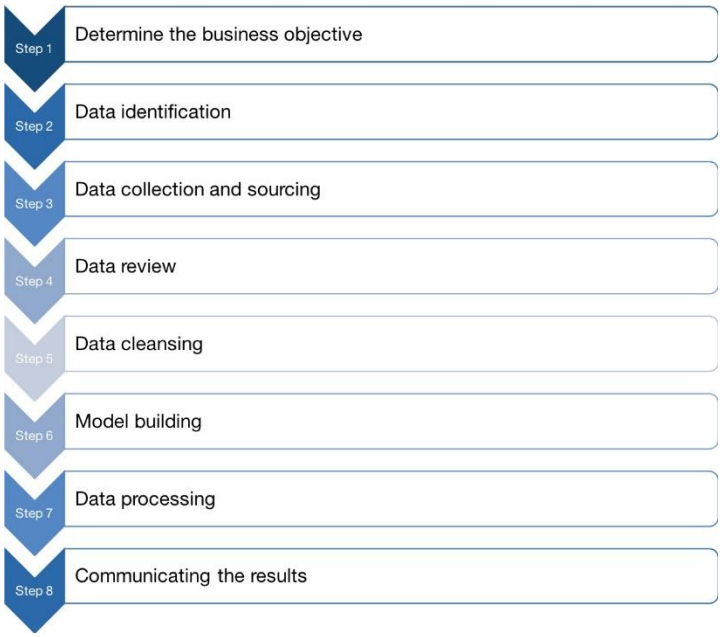
Όπως συμβαίνει με οποιαδήποτε διαδικασία, η διαδικασία ανάλυσης δεδομένων είναι διαδοχική και έχει ξεκάθαρη εκκίνηση (η σκανδάλη) και τέλος (το αποτέλεσμα). Διαχειρίζοντας τα στάδια διαδικασίας της ανάλυσης δεδομένων, οι επιχειρήσεις μπορούν να ελέγξουν καλύτερα τα αποτελέσματα των έργων τους (Big Data Framework 2018).

Βήμα 1ο - Προσδιορίστε τον επιχειρηματικό στόχο

Το πρώτο βήμα της διαδικασίας ανάλυσης δεδομένων πρέπει να καθοριστούν οι επιχειρηματικοί στόχοι του έργου. Λόγω του όγκου των συνολικών δεδομένων στον κόσμο, είναι δυνατό να χάσετε την εστίαση γρήγορα. Το πρώτο βήμα είναι συνεπώς καθοριστικής σημασίας για τον προσδιορισμό του πεδίου εφαρμογής του έργου.

Στα πλαίσια των έργων Big Data, οι επιχειρηματικοί στόχοι (και ως εκ τούτου το υποκείμενο πρόβλημα) μπορούν συχνά να υποδιαιρούνται σε έξι τύπους προβλημάτων. Κάθε ένας από αυτούς

τους τύπους έχει τον δικό του τρόπο αντιμετώπισης του αποτελέσματος του προβλήματος και τον τρόπο με τον οποίο πρέπει τα τελικά αποτελέσματα να ερμηνεύονται (Big Data Framework 2018):



Εικόνα 28: Διαδικασία Ανάλυσης Δεδομένων (Big Data Framework 2018)

1) Περιγραφικός επιχειρησιακός στόχος

2) Διερευνητικός επιχειρηματικός στόχος

3) Επαγωγικός επιχειρηματικός στόχος

4) Προγνωστικός επιχειρηματικός στόχος

5) Αιτιώδης επιχειρησιακός στόχος

6) Μηχανιστικός επιχειρησιακός στόχος

Ένας **περιγραφικός επιχειρησιακός στόχος** έχει ως στόχο να συνοψίσει τα χαρακτηριστικά των διαφορετικών συνόλων δεδομένων, που προέρχονται είτε εντός είτε εκτός της επιχείρησης. Ο στόχος είναι να συγκεντρώσουν και να συνοψίσουν τα δεδομένα προκειμένου να ληφθούν οι

αποφάσεις. Παραδείγματα περιλαμβάνουν τον υπολογισμό του μεριδίου αγοράς μιας εταιρείας σε μια συγκεκριμένη περιοχή (Big Data Framework 2018).

Ένας **ερευνητικός επιχειρηματικός στόχος** στοχεύει να βρει μια σχέση μεταξύ δύο ή περισσότερων διαφορετικά μεταβλητά σύνολα δεδομένων. Ο στόχος αυτού του στόχου είναι να βρεθεί ένα μοτίβο ή μια σχέση στα δεδομένα μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση της απόδοσης. Παραδείγματα θα μπορούσαν να είναι η αναγνώριση των προϊόντων που αγοράζονται μαζί (ανάλυση καλαθιού αγοράς) ή τον προσδιορισμό ενός μοντέλου πωλήσεων που βασίζεται στις καιρικές συνθήκες (Big Data Framework 2018).

Ένας **επιδιωκόμενος επιχειρηματικός στόχος** έχει ως στόχο να βρει χαρακτηριστικά για έναν πληθυσμό μελετώντας μια δειγματοληψία των δεδομένων, όπως αναφέρεται στο τμήμα. Οι επιδιωκόμενοι επιχειρηματικοί στόχοι επικρατούν στοχεύοντας (δυνητικούς) πελάτες σε οργανισμούς μάρκετινγκ και πωλήσεων εντός της επιχείρησης. Τα παραδείγματα περιλαμβάνουν την εύρεση νέων πελατών με βάση τους υπάρχοντες καταλόγους πελατών ή τον προσδιορισμό των περιοχών που θα πρέπει να διαφημιστούν με βάση προηγούμενες συμπεριφορές αγοράς (Big Data Framework 2018).

Ένας **προγνωστικός επιχειρηματικός στόχος** έχει ως στόχο να προβλέψει τις μελλοντικές συμπεριφορές αναλύοντας και βγάζοντας συμπεράσματα για τα συνόλα δεδομένων, όπως η πρόβλεψη των προϊόντων που οι πελάτες είναι πιθανό να αγοράσουν (ανάλυση αγοράς καλαθιού). Παραδείγματα περιλαμβάνουν τον προσδιορισμό ιδανικών περιφερειών ή ιδιότητες για μελλοντικές επενδύσεις (Big Data Framework 2018).

Ένας **αιτιώδης επιχειρηματικός στόχος** στοχεύει να βρει την υποκείμενη σχέση ενός συγκεκριμένου φαινομένου (η αιτία). Αυτός ο τύπος στόχου στοχεύει στην εύρεση της βασικής αιτίας ορισμένων δεδομένων για την καλύτερη κατανόηση των σχέσεων. Ένας αιτιώδης επιχειρηματικός στόχος έχει ως στόχο να μάθει γιατί ορισμένα δεδομένα δημιουργήθηκαν. Παραδείγματα περιλαμβάνουν να μάθετε γιατί οι επιδόσεις των πωλήσεων ήταν υψηλότερες σε κάποιο βαθμό του μήνα ή ποια είναι η πρωταρχική αιτία των αυξημένων ελλείψεων ποιότητας (Big Data Framework 2018).

Ένας **μηχανιστικός επιχειρηματικός στόχος** στοχεύει να βρει πώς οι μεταβλητές επηρεάζουν τα αποτελέσματα των συνόλων δεδομένων. Απαιτεί μια βαθύτερη κατανόηση των υποκείμενων σχέσεων και μοτίβων μέσα στα σύνολα δεδομένων. Παραδείγματα περιλαμβάνουν κατανόηση του τρόπου με τον οποίο επηρεάζεται η απόδοση των πωλήσεων από τις καιρικές συνθήκες ή τον τρόπο με τον οποίο οι συνδυασμοί συστατικών μπορούν να αυξήσουν τα έσοδα (Big Data Framework 2018).

Το πρώτο βήμα της μεγάλης ανάλυσης δεδομένων–ο προσδιορισμός του επιχειρηματικού στόχου– είναι σημαντικό επειδή καθορίζει ποιοι αλγόριθμοι και τεχνικές θα πρέπει να χρησιμοποιηθούν για να λυθεί το πρόβλημα (Big Data Framework 2018).

Βήμα 2ο - Ταυτοποίηση δεδομένων

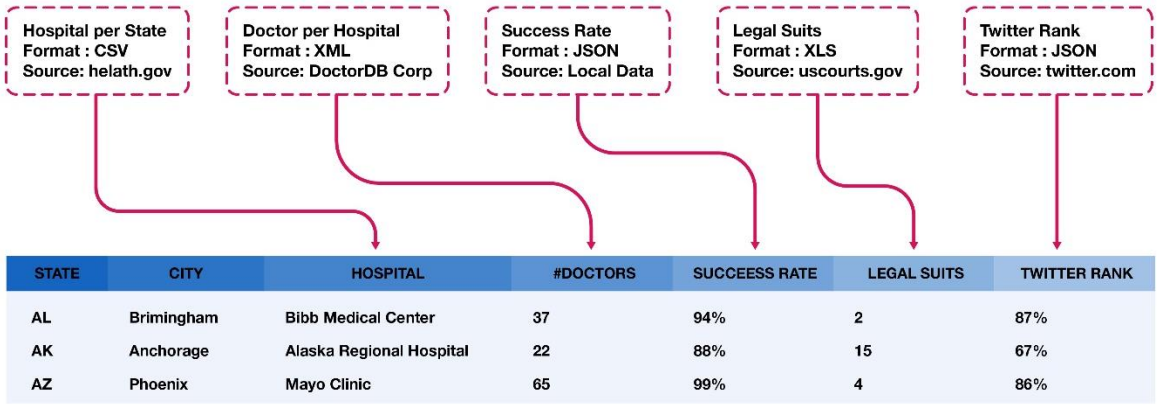
Το δεύτερο βήμα στη διαδικασία ανάλυσης δεδομένων είναι να καθοριστεί ποιες ομάδες δεδομένων πρέπει να επεξεργαστούν. Συχνά, αυτό είναι ένα από τα πιο σημαντικά και δύσκολα βήματα. Πώς να προσδιορίσετε τα σύνολα δεδομένων που είναι απαραίτητα για την ανάλυση του προβλήματος και να δώσει απαντήσεις για τους επιχειρηματικούς σκοπούς (Big Data Framework 2018).

Η περισσότερη ανάλυση δεδομένων ξεκινά με την ταυτοποίηση των ανεπεξέργαστων δεδομένων. Τα ακατέργαστα δεδομένα είναι δεδομένα που δεν υπήρξαν επεξεργασμένα ακόμα και που προέρχεται απευθείας από την πηγή. Οι πηγές δεδομένων μπορούν να περιλαμβάνουν (Big Data Framework 2018):

- Δυαδικά αρχεία από συσκευές μέτρησης (αισθητήρες).
- αρχεία CSV που βρίσκονται σε δημόσιο ιστότοπο.
- Μη μορφοποιημένο φύλλο Excel με πολλαπλές καρτέλες δεδομένων.
- Δεδομένα JSON που προέρχονται από ένα API Twitter.

Προκειμένου να εντοπιστούν τα απαραίτητα δεδομένα για την επίτευξη των επιχειρηματικών στόχων, απαιτείται ένα γράφημα ταυτοποίησης δεδομένων που λειτουργεί τα επεξεργασμένα

δεδομένα προς την ακατέργαστη πηγή δεδομένων. Ένα γράφημα αναγνώρισης δεδομένων απεικονίζεται στην Εικόνα 29 (Big Data Framework 2018).



STATE	CITY	HOSPITAL	#DOCTORS	SUCCEESS RATE	LEGAL SUITS	TWITTER RANK
AL	Brimingham	Bibb Medical Center	37	94%	2	87%
AK	Anchorage	Alaska Regional Hospital	22	88%	15	67%
AZ	Phoenix	Mayo Clinic	65	99%	4	86%

Εικόνα 29: Γράφημα ταυτοποίησης δεδομένων (Big Data Framework 2018)

Ένα γράφημα αναγνώρισης δεδομένων προσδιορίζει πρώτα τα επιθυμητά (επεξεργασμένα) δεδομένα και στη συνέχεια λειτουργεί προς τα πίσω για να προσδιορίσετε πού θα μπορούσαν να ληφθούν τα ακατέργαστα δεδομένα. Στο παραπάνω παράδειγμα, ο στόχος είναι να προσδιοριστεί το καλύτερο νοσοκομείο ανά κράτος, συνδυάζοντας πληροφορίες για τους γιατρούς, αγωγές κακής πρακτικής και Twitter chatter. Το γράφημα αναγνώρισης δεδομένων παραπάνω δείχνει από που προέρχονται τα επεξεργασμένα δεδομένα (Big Data Framework 2018).

Βήμα 3 - Συλλογή δεδομένων και πηγές

Αφού προσδιοριστεί ποια δεδομένα είναι απαραίτητα για την επίτευξη του απαιτούμενου επιχειρηματικού αποτελέσματος, το επόμενο βήμα είναι να διασφαλιστεί ότι τα δεδομένα λαμβάνονται για επεξεργασία. Παρόλο που αυτό φαίνεται ως ένα σχετικά απλό βήμα, στην πράξη είναι συχνά ένα δύσκολο βήμα (Big Data Framework 2018).

Συλλογή δεδομένων

Στις περισσότερες επιχειρήσεις, τα εσωτερικά δεδομένα αποθηκεύονται σε διάφορες φυσικές τοποθεσίες ή κέντρα δεδομένων σε όλη την υφήλιο. Για να χρησιμοποιήσει αυτά τα δεδομένα, ο αναλυτής δεδομένων ή ο επιστήμονας δεδομένων πρέπει να αποκτήσει τα κατάλληλα δικαιώματα

πρόσβασης και να συνεργαστείτε με την ομάδα διαχείρισης δεδομένων. Θα πρέπει να θεσπιστούν μέτρα για την εξασφάλιση της ακεραιότητας των δεδομένων και της εμπιστευτικότητας τους, η προστασία των δεδομένων δεν θα πέσει σε λάθος χέρια. Επιπλέον, οι περισσότερες χώρες έχουν την ιδιωτική ζωή και τις κανονιστικές απαιτήσεις, ώστε τα προσωπικά δεδομένα να μην διασχίζουν τα σύνορα. Αυτό θα μπορούσε να προκαλέσει σημαντικά ζητήματα αν οι ομάδες Big Data, για παράδειγμα, επιδιώξουν να συγκρίνουν τις συμπεριφορές των πελατών από τη Σιγκαπούρη (που βρίσκεται σε ένα κέντρο δεδομένων στη Σιγκαπούρη) με τη συμπεριφορά των πελατών από τις ΗΠΑ (που βρίσκεται σε κέντρο δεδομένων στη Βοστώνη) (Big Data Framework 2018).

Εκτός από τις ανησυχίες για την ασφάλεια και την προστασία της ιδιωτικής ζωής, ένα δεύτερο θέμα στη συλλογή δεδομένων διαδικασία είναι να ασχοληθεί με το μέγεθος, την ποικιλία και ιδιαίτερα τις πτυχές της ταχύτητας των συνόλων δεδομένων. Αν τα δεδομένα ανανεώνονται συχνά ή ακόμα ανανεώνονται σε καθημερινή βάση (για παράδειγμα μια ροή Twitter), η συλλογή περιλαμβάνει αποφάσεις σχετικά με τη συχνότητα των εισαγωγών δεδομένων (εισαγωγές σε πραγματικό χρόνο και κατά παρτίδες) και πώς αντιμετωπίζονται οι προηγούμενες εισαγωγές (Big Data Framework 2018).

Πηγές δεδομένων

Για να αποκτήσετε αξία από τα μεγάλα δεδομένα, τα εσωτερικά σύνολα δεδομένων επιχειρήσεων θα πρέπει να συνδυάζονται με εξωτερικά δεδομένα για παράδειγμα πληροφορίες για τις καιρικές συνθήκες ή μηνύματα Twitter. Μερικά από αυτά τα εξωτερικά σύνολα δεδομένων ενδέχεται να είναι διαθέσιμα δωρεάν, αλλά τα περισσότερα σύνολα δεδομένων θα πρέπει να αποκτηθούν από εξωτερικούς προμηθευτές (Big Data Framework 2018).

Η απόκτηση δεδομένων απαιτεί επίσης τη συμμετοχή των προμηθευτών σύμφωνα με τις διαδικασίες και τους κανονισμούς προμήθειας στις περισσότερες χώρες− με μια ανοιχτή και διαφανή διαδικασία υποβολής προσφορών. Εκτός από το γεγονός ότι αυτές οι διαδικασίες χρειάζονται χρόνο, μεγάλη προσοχή πρέπει να δοθεί ώστε διασφαλιστεί ότι ο πωλητής δεδομένων δεν υπερτιμολογεί την αξία των δεδομένων του ενώ κρύβονται μεγάλα έξοδα επεξεργασίας

δεδομένων στη διαδικασία της προμήθειας δεδομένων. Δηλαδή τα δεδομένα που αγοράζονται πρέπει να βρίσκονται σε μορφής εύκολης επεξεργασίας (Big Data Framework 2018).

Βήμα 4^{ο} - Ανασκόπηση δεδομένων

Αφού έχουν καταστεί διαθέσιμα τα απαιτούμενα σύνολα δεδομένων, ξεκινάει το βήμα αναθεώρησης δεδομένων. Η αναθεώρηση δεδομένων είναι η διαδικασία διερεύνησης των συνόλων δεδομένων σας και συνήθως περιλαμβάνει την εξέταση της δομής και των μεταβλητών των διαφόρων συνόλων δεδομένων. Σε αυτή τη διαδικασία, καθορίζεται εάν έχουν τα σύνολα δεδομένων καταστραφεί, αν υπάρχουν ελλιπείς τιμές ή αν υπάρχουν πολλαπλά (σε σύγκρουση) σύνολα με τις ίδιες μεταβλητές. Μπορεί, για παράδειγμα, να ισχύουν τα δεδομένα πωλήσεων για μια συγκεκριμένη περιοχή από δύο διαφορετικά συστήματα χρηματοδότησης και να έχουν διαφορετικές αξίες (Big Data Framework 2018).

Κύριοι στόχοι της διαδικασίας ανασκόπησης δεδομένων περιλαμβάνουν:

• Για να προσδιοριστούν αν υπάρχουν προβλήματα ή προβλήματα με τα σύνολα δεδομένων.

• Για τον προσδιορισμό των μεταβλητών και τη διανομή των δεδομένων στα σύνολα δεδομένων.

• Για να προσδιοριστεί εάν το σύνολο δεδομένων περιέχουν τιμές που λείπουν ή υπάρχουν κατεστραμμένα δεδομένα.

• Για να προσδιοριστούν αν οι επιχειρηματικοί στόχοι (βήμα 1) μπορούν να υλοποιηθούν με αυτά τα δεδομένα.

Η πραγματικότητα στην πράξη είναι ότι σχεδόν κάθε σύνολο δεδομένων, ακόμη και αν προμηθεύεται από ακριβούς και αξιόπιστους πάροχους δεδομένων, έχει ανακριβείς ή ελλείπουσες τιμές που πρέπει να ληφθούν υπόψη. Η διαδικασία επανεξέτασης δεδομένων είναι, επομένως, ύψιστης σημασίας, διότι ελλιπείς τιμές μπορεί να έχει βαθιές επιπτώσεις στο τελικό αποτέλεσμα.

Σε περίπτωση που υπάρχουν λανθασμένα ή διεφθαρμένα δεδομένα, το σύνολο δεδομένων πρέπει να καθαριστεί στο επόμενο βήμα.

Βήμα 5ο - Καθαρισμός δεδομένων

Ο καθαρισμός δεδομένων είναι η διαδικασία τροποποίησης ή κατάργησης δεδομένων σε μια βάση δεδομένων που είναι εσφαλμένη, ελλιπή, ακατάλληλα μορφοποιημένη ή υπάρχουν διπλότυπα. Ο καθαρισμός των δεδομένων μπορεί να πραγματοποιηθεί διαδραστικά μέσω εργαλείων καθαρισμού δεδομένων ή ως επεξεργασία παρτίδων μέσω δέσμης ενεργειών. Μετά από τον καθαρισμό, ένα σύνολο δεδομένων πρέπει να είναι συνεπές με άλλα παρόμοια σύνολα δεδομένων στο σύστημα και έτοιμο προς χρήση για την επεξεργασία τους. Επιχειρήσεις που ασχολούνται σε τομέα έντασης δεδομένων, όπως τράπεζες, ασφάλιση, λιανικό εμπόριο, τηλεπικοινωνίες ή μεταφορές μπορεί να χρησιμοποιούν εργαλεία καθαρισμού δεδομένων για τη συστηματική εξέταση των δεδομένων για ελαττώματα χρησιμοποιώντας κανόνες, αλγόριθμους και πίνακες αναζήτησης. Συνήθως, ένα εργαλείο καθαρισμού δεδομένων περιλαμβάνει προγράμματα που είναι ικανά να διορθώσουν έναν αριθμό συγκεκριμένων τύπων λαθών, όπως την προσθήκη κωδικών ZIP ή την εύρεση διπλών εγγραφών (Big Data Framework 2018).

Βήμα 6ο - Δημιουργία μοντέλου

Το επόμενο βήμα στη διαδικασία ανάλυσης δεδομένων είναι η δημιουργία ενός στατιστικού μοντέλου που μπορεί να χρησιμοποιείται για την εύρεση του αποτελέσματος στον επιχειρησιακό στόχο. Το μοντέλο είναι η επαναληπτική διαδικασία που προσδιορίζει και βελτιώνει ένα στατιστικό μοντέλο που μπορεί να εφαρμοστεί στο σύνολο καθαρισμένων δεδομένων. Μαθηματικοί τύποι ή μοντέλα που ονομάζονται αλγόριθμοι μπορούν να εφαρμοστούν στα δεδομένα για τον εντοπισμό των σχέσεων μεταξύ των μεταβλητών, όπως η συσχέτιση ή η αιτιώδης συνάφεια. Σε γενικές γραμμές, τα μοντέλα μπορεί να αναπτυχθούν για να αξιολογήσουν μια συγκεκριμένη μεταβλητή που βασίζονται σε άλλες , με κάποιο υπολειπόμενο σφάλμα ανάλογα με την ακρίβεια του μοντέλου (δηλ. Δεδομένα = Μοντέλο + Σφάλμα) (Big Data Framework 2018).

Στον τομέα της πολιτικής, για παράδειγμα, ένας αναλυτής δεδομένων μπορεί να χρησιμοποιήσει ένα δείγμα δημοσκοπήσεων προκειμένου να προβλέψει τα αποτελέσματα μιας εκλογής. Για να γίνει αυτό, ο αναλυτής θα χρειαστεί να δημιουργήσει ένα μοντέλο που μπορεί να εφαρμοστεί στα δεδομένα. Η διαδικασία οικοδόμησης ενός μοντέλου συνεπάγεται την επιβολή συγκεκριμένης δομής των δεδομένων και δημιουργία μιας σύνοψης των δεδομένων. Το στατιστικό μοντέλο είναι

ένα από τα πιο πολύτιμα βήματα στη διαδικασία ανάλυσης δεδομένων, επειδή η ακρίβεια του μοντέλου καθορίζει το τελικό αποτέλεσμα (Big Data Framework 2018).

Βήμα 7ο - Επεξεργασία δεδομένων

Το βήμα της επεξεργασίας δεδομένων αφορά στην εκτέλεση της πραγματικής εργασίας ανάλυσης, η οποία τυπικά περιλαμβάνει τη διεξαγωγή ενός ή περισσότερων στατιστικών αλγορίθμων. Αυτό το βήμα μπορεί να είναι επαναληπτικό, ειδικά εάν η ανάλυση δεδομένων είναι διερευνητική έτσι ώστε η ανάλυση να επαναλαμβάνεται μέχρι το κατάλληλο σχέδιο ή αν η συσχέτιση δεν αποκαλύπτεται. Ανάλογα με τον τύπο της διαδικασίας που απαιτείται, το βήμα επεξεργασίας δεδομένων μπορεί να είναι τόσο απλό όσο ερωτώντας ένα σύνολο δεδομένων σε μέσους όρους, ή διάμεσο. Από την άλλη πλευρά, μπορεί να είναι τόσο περίπλοκο όσο συνδυάζοντας πολλούς σύνθετους αλγορίθμους για την εκτέλεση αλγορίθμων αναγνώρισης προσώπου, αναλύσεις αλληλουχίας πρωτεϊνών DNA ή προβλέψεις χρηματοοικονομικών αγορών. Η διάρκεια του σταδίου επεξεργασίας δεδομένων ποικίλλει ανάλογα με τις απαιτήσεις, όπως αναλύθηκε σε προηγούμενο κεφάλαιο, οι περισσότερες λύσεις Big Data θα χρησιμοποιούν κάποια μορφή διανομής επεξεργασίας (συνηθέστερα το πλαίσιο λογισμικού Hadoop) για τη μείωση των απαραίτητων χρόνων επεξεργασίας (Big Data Framework 2018).

Βήμα 8ο – Μετάδοση των αποτελεσμάτων

Η διαδικασία ανάλυσης Big Data τελειώνει με την μετάδοση των τελικών αποτελεσμάτων. Αν και αυτό είναι το λογικό τελευταίο βήμα κάθε έργου ανάλυσης, η σημασία του δεν μπορεί να υποτιμηθεί. Η επικοινωνία με σαφήνεια είναι απαραίτητη για την ορθή ανάλυση των δεδομένων. Δεδομένου ότι το Big Data και οι υποκείμενοι αλγόριθμοι επεξεργασίας είναι μερικές φορές δύσκολο να το εξηγήσουν οι ηγέτες των επιχειρήσεων, η καλή επικοινωνία είναι απαραίτητη για την επιτυχία και την αξία των Big Data (Big Data Framework 2018).

Να επικοινωνούν σε τακτική βάση (π.χ. ενδιάμεσες αναφορές) και με δομημένο τρόπο (κάθε Παρασκευή) θα παρέχει στις ομάδες και τους υπεύθυνους λήψης αποφάσεων των επιχειρήσεων την απαιτούμενη εμπιστοσύνη ότι ακολουθούνται οι δομημένες διαδικασίες (Big Data Framework 2018).

Ένας από τους καλύτερους τρόπους επικοινωνίας των αποτελεσμάτων οποιουδήποτε έργου Big Data είναι η χρήση των τεχνικών οπτικοποίησης ,που συζητήθηκαν στο τμήμα. Συγκεντρώνοντας τα δεδομένα σε γραφήματα και τα αριθμητικά στοιχεία, γίνεται πιο κατανοητό. Οι τεχνολογίες οπτικοποίησης δεδομένων μπορούν να είναι οι ίδιες πολύ ισχυρές καθώς είναι εύκολο στη χρήση, επιτρέποντας στους αναλυτές δεδομένων να αρθρώνουν γρήγορα και εύκολα τις ιδέες σε όλη την επιχείρηση με άλλους που είναι λιγότερο άνετοι με την ανάλυση δεδομένων (Big Data Framework 2018).

5.1 Διαδικασία διακυβέρνησης δεδομένων

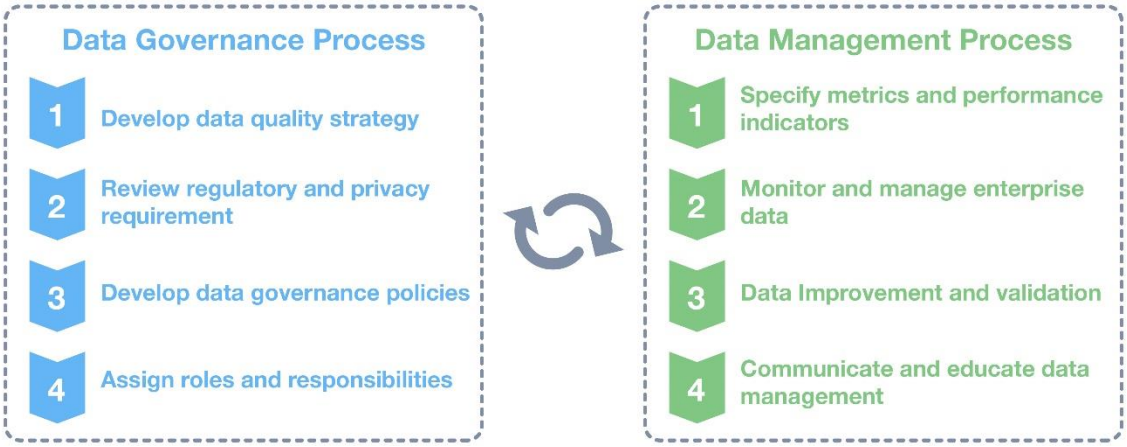
Η διαδικασία διακυβέρνησης των δεδομένων είναι μια καθορισμένη διαδικασία που ακολουθούν οι επιχειρήσεις για να εξασφαλίσουν ότι αυτοί ελέγχουν τα δεδομένα τους καθ 'όλη τη διάρκεια του κύκλου ζωής τους. Δεδομένου ότι το "Big Data" αποτελούν στρατηγικό πλεονέκτημα, οι περισσότεροι οργανισμοί πρέπει να θεσπίσουν μέτρα ελέγχου. Η διαδικασία διακυβέρνησης δεδομένων διασφαλίζει ότι σημαντικά δεδομένα διαχειρίζονται επίσημα σε όλη την επιχείρηση, και τα δεδομένα μπορούν να είναι έμπιστα για τη λήψη αποφάσεων. Συχνά, οι διαδικασίες που χρησιμοποιούνται στη διαχείριση δεδομένων περιλαμβάνουν ανάληψη ευθύνης για κάθε ανεπιθύμητο γεγονός που προκύπτει από την ποιότητα των δεδομένων (Big Data Framework 2018).

Η εστίαση στη διαχείριση των δεδομένων και σε όλες τις διαδικασίες έχει αυξηθεί σε μεγάλο βαθμό τα τελευταία χρόνια, ιδίως λόγω των αυξημένων απαιτήσεων απορρήτου δεδομένων και εμπιστευτικότητας των δεδομένων που έχουν καθοριστεί από τις χώρες. Επομένως, η διαδικασία διακυβέρνησης των δεδομένων δεν χρειάζεται μόνο να ορίσει τις πολιτικές και να αναθέσει ευθύνες σε όλη την επιχείρηση, πρέπει επιπλέον να εξασφαλίσει ότι οι επιχειρήσεις συμμορφώνονται με τους τοπικούς νόμους και κανονισμούς για τα δεδομένα (Big Data Framework 2018).

Υπάρχει στενή σχέση μεταξύ της διαδικασίας διακυβέρνησης και της διαχείρισης δεδομένων, η οποία θα συζητηθεί στην επόμενη ενότητα. Όταν η διαδικασία διακυβέρνησης των δεδομένων αφορά τον καθορισμό των πολιτικών σε στρατηγικό επίπεδο, η διαχείριση δεδομένων εκτελεί και

παρακολουθεί τις πολιτικές αυτές σε επιχειρησιακό επίπεδο. Η συνέργεια μεταξύ των δύο διαδικασιών απεικονίζεται στην Εικόνα 30 (Big Data Framework 2018).

5.1.1 Δραστηριότητες διαδικασίας διακυβέρνησης δεδομένων



Εικόνα 30: Η συνέργεια διακυβέρνησης δεδομένων και διαχείρισης δεδομένων (Big Data Framework 2018)

Η διαδικασία διακυβέρνησης δεδομένων περιλαμβάνει τους ανθρώπους, την οργανωτική δομή και την τεχνολογία που απαιτείται για τη δημιουργία ενός συνεπούς και σωστού χειρισμού των δεδομένων ενός οργανισμού σε όλη την επιχείρηση. Αν και οι στόχοι ενδέχεται να διαφέρουν ανάλογα με τη φύση της επιχείρηση, το απαιτούμενο επίπεδο ελέγχου και τοπικές ρυθμιστικές απαιτήσεις, καθώς και έναν αριθμό καθολικών δραστηριοτήτων διακυβέρνησης δεδομένων είναι οι ίδιες για κάθε οργανισμό (Big Data Framework 2018).

5.1.2 Ανάπτυξη στρατηγικής για την ποιότητα των δεδομένων

Απαιτείται στρατηγική για την ποιότητα των δεδομένων που θα συμφωνεί με την συνολική επιχειρησιακή στρατηγική. Επιπλέον, δηλώνει τη συμμετοχή των ενδιαφερομένων μερών, που κατανοούν το ρόλο των δεδομένων στην επιχείρηση. Η στρατηγική για την ποιότητα των δεδομένων πρέπει να επανεξετάζεται και ενημερώνεται σε ετήσια βάση τουλάχιστον (Big Data Framework 2018).

5.1.3 Έλεγχος των ρυθμιστικών απαιτήσεων και απαιτήσεων απορρήτου

Ορισμένες χώρες (για παράδειγμα οι ΗΠΑ ή η Σιγκαπούρη) έχουν αυστηρούς νόμους σχετικά με την ιδιωτικότητα δεδομένων των πολιτών τους και τις ρυθμιστικές απαιτήσεις για διασυνοριακή μεταφορά δεδομένων. Στις περισσότερες περιπτώσεις, οι χώρες αυτές απαιτούν ελέγξιμες αναφορές και ημερολόγια για να διασφαλίσουν ότι οι οργανισμοί συμμορφώνονται με αυτούς τους νόμους και κανονισμούς. Καθώς τα ρυθμιστικά και τα πρόστιμα μπορεί να είναι πολύ υψηλά (και η φήμη του οργανισμού να έχει καταστραφεί ως άμεσο αποτέλεσμα), υπάρχει μεγάλος λόγος για τις επιχειρήσεις να διασφαλίσουν ότι συμμορφώνονται με τους κανόνες αυτούς. Επομένως, η αναθεώρηση των κανονιστικών απαιτήσεων και των απαιτήσεων προστασίας της ιδιωτικής ζωής αποτελεί αναπόσπαστο μέρος της διαδικασίας διακυβέρνησης δεδομένων που θα πρέπει να διεξάγεται σε μηνιαία βάση για να διασφαλιστεί η απαραίτητη συμμόρφωση (Big Data Framework 2018).

5.1.4 Ανάπτυξη πολιτικών διακυβέρνησης δεδομένων

Η στρατηγική για την ποιότητα των δεδομένων και οι ρυθμιστικές απαιτήσεις πρέπει να μεταφραστούν σε έναν αριθμό των πολιτικών διακυβέρνησης δεδομένων που είναι διαθέσιμες στο κοινό για όλους στην επιχείρηση. Αυτά τα έγγραφα πολιτικής περιέχουν τις αποφάσεις που έχει λάβει η επιχείρηση όσον αφορά τα δεδομένα της ποιοτικής οργάνωσης και να εξηγήσει τις απαιτήσεις για όλους. Αυτές οι πολιτικές είναι χρήσιμες για τη λήψη αποφάσεων, οι εξωτερικές συμβάσεις με τους προμηθευτές και μπορούν να παράσχουν πολύτιμες πληροφορίες για άλλες επιχειρηματικές διαδικασίες (Big Data Framework 2018).

5.1.5 Ανάθεση των ρόλων και των ευθυνών

Η διαδικασία διακυβέρνησης των δεδομένων πρέπει να καθορίζει σαφείς ρόλους και ευθύνες σε όλα τα όρια στην επιχείρηση. Η ανάθεση ρόλων πρέπει να εξασφαλίζει την υπευθυνότητα, την εξουσία και την ευθύνη της εποπτείας, καθώς και της συμμετοχής ανώτερων στελεχών και της διοίκησης να ενθαρρύνουν την επιθυμητή συμπεριφορά στη χρήση των δεδομένων (Big Data Framework 2018).

5.2 Διαδικασία διαχείρισης δεδομένων

Η διαδικασία διαχείρισης δεδομένων είναι μια ξεχωριστή διαδικασία που διασφαλίζει την ποιότητα των δεδομένων σε καθημερινό επίπεδο. Αυτή η διαδικασία εκτελείται βάσει των στρατηγικών οδηγιών. Ο πρωταρχικός στόχος της διαδικασίας διαχείρισης δεδομένων είναι η διασφάλιση της ποιότητας των δεδομένων. Η αξία που μπορούν να ληφθούν με την ανάλυση Big Data εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων εισόδου (Big Data Framework 2018).

Ακόμη και με την πιο εξελιγμένη λύση Big Data, ο γενικός κανόνας «Σκουπίδια μέσα – σκουπίδια έξω» εξακολουθεί να ισχύει. Εάν τα σύνολα δεδομένων είναι διεφθαρμένα ή λανθασμένα, η ανάλυση δεδομένων μπορεί να έχει ως αποτέλεσμα μη έγκυρα αποτελέσματα ή συμπεράσματα (Big Data Framework 2018).

Συνεπώς, οι επιχειρήσεις χρειάζονται τη διαδικασία διαχείρισης δεδομένων για τη συνεχή επαλήθευση, ενημέρωση και καθαρισμό των δεδομένων. Η διαδικασία διαχείρισης δεδομένων που περιγράφεται σε αυτό το κεφάλαιο παρέχει δομημένη και πρακτική προσέγγιση για την εφαρμογή των ακόλουθων ιδεών (Big Data Framework 2018):

- Οι επιχειρήσεις χρειάζονται έναν τρόπο να επισημοποιήσουν τις προσδοκίες τους για τη μέτρηση συμμόρφωσης της ποιότητας των δεδομένων.
- Οι επιχειρήσεις πρέπει να θέσουν μια βάση της ποιότητας των δεδομένων προκειμένου να εντοπίσουν προβλήματα και την ανάλυση των βασικών αιτιών της αποτυχίας των δεδομένων.
- Οι επιχειρήσεις πρέπει να είναι σε θέση να γνωστοποιούν το επίπεδο εμπιστοσύνης τους στην ποιότητα του τα δεδομένα τους.

5.2.1 Δραστηριότητες διαχείρισης δεδομένων

Η διαδικασία διαχείρισης δεδομένων είναι μια πρακτική και λειτουργική διαδικασία που παρακολουθεί καθημερινά την ποιότητα των δεδομένων. Η διαδικασία αποτελείται από τις ακόλουθες δραστηριότητες (Big Data Framework 2018):

5.2.1.1 Καθορίστε μετρήσεις και δείκτες απόδοσης

Για τη διασφάλιση της ποιότητας των δεδομένων καθ 'όλη τη διάρκεια του κύκλου ζωής των δεδομένων, οι εταιρείες προσδιορίζουν μετρήσεις και δείκτες επιδόσεων. Αυτές οι μετρήσεις πρέπει να συνδέονται με τους γενικούς στόχους της εταιρείας και τους στόχους που καθορίζονται στις πολιτικές διακυβέρνησης των δεδομένων (Big Data Framework 2018).

Uniqueness	Uniqueness refers to requirements that data within the enterprise is captured and represented uniquely within the relevant application architectures. Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set.
Accuracy	Accuracy is the extent to which the data that is reflected in the dataset corresponds with the truth. It refers to whether the data values stored for an object are the correct values. To be correct, a data values must be the right value and must be represented in a consistent and unambiguous form.
Consistency	In its most basic form, consistency refers to data values in one data set being consistent with values in another data set. A strict definition of consistency specifies that two data valuesdrawn from separate data sets must not conflict with each other, although consistency does not necessarily imply correctness.
Completeness	An expectation of completeness indicates that certain attributes should be assigned values in a data set. Completeness rules can be assigned to a data to validate that all attributes are present in the data set.
Timeliness	Timeliness refers to the time expectation for accessibility and availability of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use.
Currency	Currency refers to the degree to which information is current with the world that it models. Currency can measure how “up-to-date” information is, and whether it is correct despite possible time-related changes.
Conformance	This dimension refers to whether instances of data are either stored, exchanged, or presented ina format that is consistent with the domain of values, as well as consistent with other similar attribute values.
Integrity	Data integrity is the maintenance and assurance of the accuracy and consistency of data over its entire life-cycle and is a critical aspect to the design, implementation and usage of any system which stores, processes, or retrieves data. Integrity means that the data has not been altered.

Εικόνα 31: Μετρικές δεδομένων και δείκτες απόδοσης (Big Data Framework 2018)

Οι μετρήσεις και οι δείκτες απόδοσης μπορούν να καταγραφούν σε μια ισορροπημένη κάρτα ελέγχου. Η δημιουργία μιας τέτοιας κάρτας ελέγχου παρέχει ένα αποτελεσματικό μέσο για τη συνεχή παρακολούθηση και διαχείριση δεδομένων βάσει βασικών δεικτών απόδοσης. Κοινές μετρήσεις και δείκτες επίδοσης που περιλαμβάνονται στα scorecards για την ποιότητα των δεδομένων απεικονίζονται στην Εικόνα 31 (Big Data Framework 2018).

5.2.1.2 Παρακολουθήστε και διαχειριστείτε δεδομένα επιχείρησης

Με βάση τις μετρήσεις και τους δείκτες απόδοσης που έχουν καθοριστεί στην προηγούμενη δραστηριότητα, τα δεδομένα των επιχειρήσεων πρέπει να παρακολουθούνται. Με αυτοματοποιημένα εργαλεία, τα σύνολα δεδομένων μπορούν να παρακολουθούνται και ευρετηριάζεται για να μετρά την ποιότητα των δεδομένων σε σχέση με τις συγκεκριμένες επιδόσεις δεικτών. Τα αποτελέσματα μπορούν και πάλι να απεικονιστούν σε scorecards ποιότητας δεδομένων. Ένα από τα σημαντικά στοιχεία αυτής της διαδικασίας είναι η δημιουργία των ειδοποιήσεων. Πρέπει να δημιουργούνται ειδοποιήσεις αν διαπιστωθεί ότι τα δεδομένα έχουν καταστραφεί ή αλλάξει (Big Data Framework 2018).

5.2.1.3 Βελτίωση και επικύρωση δεδομένων

Η επόμενη δραστηριότητα της διαδικασίας διαχείρισης δεδομένων είναι η βελτίωση των συνόλων δεδομένων των επιχειρήσεων. Η ισορροπημένη κάρτα βαθμολογίας δεδομένων μπορεί, για παράδειγμα, να υποδείξει ότι υπάρχουν πολλές διπλές εγγραφές σε σύνολα δεδομένων. Η δραστηριότητα βελτίωσης των δεδομένων και επικύρωσης αφορά τον καθαρισμό των συνόλων δεδομένων, προκειμένου να βελτιωθούν οι μετρήσεις και οι δείκτες επίδοσης. Χρησιμοποιώντας κανόνες επαλήθευσης και κανόνες μετασχηματισμού, η ποιότητα των δεδομένων μπορεί να βελτιωθεί όπως απεικονίζεται στην Εικόνα 32 (Big Data Framework 2018).

Εικόνα 32: Εφαρμογή κανόνων επιθεώρησης (Big Data Framework 2018)

5.2.1.4 Μετάδοση και εκπαίδευση σχετικά με τη διαχείριση δεδομένων

Η τελευταία δραστηριότητα της διαδικασίας διαχείρισης δεδομένων είναι η επικοινωνία και η εκπαίδευση της ομάδας να συμμετάσχει ενεργά σε πρωτοβουλίες διαχείρισης δεδομένων. Με τη διασφάλιση ότι οι διαδικασίες διακυβέρνησης ακολουθούνται και τα συστήματα χρησιμοποιούνται σωστά, η ποιότητα των δεδομένων μπορεί να βελτιωθεί σημαντικά. Σε πολλές περιπτώσεις, οι εργαζόμενοι δεν γνωρίζουν τις δομές των δεδομένων και δεν γνωρίζουν την αξία των δεδομένων για τον οργανισμό (Big Data Framework 2018).

Προκειμένου να βελτιωθεί αυτή η γνώση, τα προγράμματα κατάρτισης μπορούν να μειώσουν τα λάθη των χρηστών, να αυξήσουν την παραγωγικότητα και την αύξηση της συμμόρφωσης με τους βασικούς ελέγχους. Η εκπαίδευση απευθύνεται σε βασικές αρχές δεδομένων και στις πρακτικές ποιότητας των δεδομένων που συμπληρώνονται από ειδική εκπαίδευση ρόλων (Big Data Framework 2018).

6 Τα Big Data ως μέσο Στρατηγικής

Οι οργανισμοί χρησιμοποιούν ανταγωνιστική νοημοσύνη (Competitive Intelligence) για να συγκριθούν με άλλους οργανισμούς ("ανταγωνιστική συγκριτική αξιολόγηση"), για τον εντοπισμό κινδύνων και τις ευκαιρίες στις αγορές τους. Η ανταγωνιστική νοημοσύνη ορίζεται ως ο συνδυασμός καθορισμού, συλλογής και ανάλυσης πληροφοριών σχετικά με προϊόντα, πελάτες, ανταγωνιστές και κάθε πτυχής του περιβάλλοντος κατά τη λήψη στρατηγικών αποφάσεων για μια οργάνωση. Σε αντίθεση με τη βιομηχανική κατασκοπεία, η ανταγωνιστική νοημοσύνη θεωρείται ως μια νόμιμη επιχειρηματική πρακτική με επίκεντρο το εξωτερικό επιχειρηματικό περιβάλλον. Περιλαμβάνει τον ορισμό ενός συνόλου διαδικασιών για τη συλλογή πληροφοριών, τη μετατροπή τους σε γνώση και στη συνέχεια, χρησιμοποιώντας αυτό στη λήψη επιχειρηματικών αποφάσεων (Dey, et al. 2011).

Οι περισσότερες επιχειρήσεις σήμερα έχουν ουσιαστικής σημασίας online παρουσία και είναι δυνατόν να συγκεντρώσουν γνώσεις για το τι κάνουν οι ανταγωνιστές και πώς αλλάζει η βιομηχανία. Οι πληροφορίες που συγκεντρώνονται επιτρέπουν στους οργανισμούς να αναγνωρίσουν τα δυνατά σημεία και τις αδυναμίες τους. Οι πληροφορίες που συλλέγονται κατά τη διαδικασία μπορούν να ερμηνεύονται με διαφορετικούς τρόπους από διαφορετικά στελέχη ανάλογα με την ανάλυσή τους (Dey, et al. 2011).

Από την άνοδο της σύγχρονης επιχείρησης, η πληροφορία υπήρξε βασικό στρατηγικό πλεονέκτημα. Οργανισμοί που έχουν ακριβέστερες πληροφορίες για τους πελάτες, πληροφορίες για τη βιομηχανία ή το μάρκετινγκ μπορούν να αξιοποιήσουν αυτές τις πληροφορίες για να ξεπεράσουν τους ανταγωνιστές τους. Με τα χρόνια, πρωτοποριακά συστήματα από εταιρείες όπως η American Airlines (ηλεκτρονικές κρατήσεις), Otis Elevator (προβλέψιμη συντήρηση) και American Supply Hospital (online παραγγελία) ενίσχυσαν δραματικά τα έσοδα και τη φήμη των δημιουργών τους (Big Data Framework 2018).

Η ανάπτυξη των μεγάλων δεδομένων και των λύσεων μεγάλων δεδομένων έφερε την έννοια της πληροφορίας ως στρατηγικού πλεονεκτήματος σε ένα εντελώς νέο επίπεδο. Οι οργανώσεις δεν ανταγωνίζονται μόνο για την ακρίβεια των δεδομένων τους, αλλά και στην ικανότητά τους να επεξεργάζονται μεγάλες ποσότητες (όγκος) με μεγάλη ταχύτητα (ταχύτητα) και από διαφορετικές

πηγές δεδομένων. Έχοντας νέες πληροφορίες πιο γρήγορα από τους ανταγωνιστές προσφέρεται ένα μικρό παράθυρο ευκαιρίας για να ενεργήσει πριν από τον ανταγωνισμό. Σε μια πρόσφατη έρευνα από την IDG, το 78% των επιχειρήσεων συμφωνούν ότι η στρατηγική δεδομένων, συλλογή και ανάλυση των Big Data έχει τη δυνατότητα να αλλάξει θεμελιωδώς τον τρόπο με τον οποίο η επιχείρησή της πραγματοποιεί επιχειρηματικές συναλλαγές τα επόμενα 1 έως 3 χρόνια (Big Data Framework 2018).

Παρόλο που οι περισσότερες επιχειρήσεις συμφωνούν ότι η Αναλυτική Μεγάλων Δεδομένων παρέχει ένα ανταγωνιστικό πλεονέκτημα, πολλοί οργανισμοί παραμένουν ελάχιστα πίσω από την καμπύλη. Οι μελέτες που αφορούν διασταυρωμένες βιομηχανίες δείχνουν ότι κατά μέσο όρο, λιγότερο από το ήμισυ των δομημένων δεδομένων ενός οργανισμού χρησιμοποιούνται ενεργά στη λήψη αποφάσεων και λιγότερο από το 1% των μη δομημένων δεδομένων της αναλύεται ή χρησιμοποιείται καθόλου. Περισσότερο από το 70% των εργαζομένων έχουν πρόσβαση σε δεδομένα που δεν πρέπει, και το 80% του χρόνου των αναλυτών ξοδεύεται απλά προετοιμάζοντας των δεδομένων (Big Data Framework 2018).

Ο λόγος για τον οποίο τόσες πολλές εταιρείες αγωνίζονται να συνειδητοποιήσουν το ανταγωνιστικό τους πλεονέκτημα μέσω του Big Data είναι επειδή δεν έχουν καθορίσει επαρκώς μια στρατηγική μεγάλων δεδομένων. Πολλοί οργανισμοί εξακολουθούν να βασίζονται σε έργα, αντί να ενσωματώνεται στις φλέβες της οργάνωσης (Big Data Framework 2018).

Για να αποφευχθούν αυτές οι παγίδες και να επιτευχθεί ένα μακροπρόθεσμο ανταγωνιστικό πλεονέκτημα, ένα κοινό πλαίσιο Big Data ξεκινά με τον καθορισμό και τη διαμόρφωση μιας στρατηγικής μεγάλων δεδομένων. Κάθε άλλη δραστηριότητα ή διαδικασία που συζητείται περαιτέρω σε ολόκληρο το Πλαίσιο Μεγάλων Δεδομένων πρέπει να σχετίζεται με τη Μεγάλη στρατηγική δεδομένων (Big Data Framework 2018).

6.1 Ανταγωνιστική Νοημοσύνη

Πιο συγκεκριμένα η ανταγωνιστική νοημοσύνη μπορεί να ταξινομηθεί ευρέως σε δύο κατηγορίες ανάλογα με το αν χρησιμοποιείται για μακροπρόθεσμο προγραμματισμό ή βραχυπρόθεσμο σχεδιασμό. Η στρατηγική νοημοσύνη (Strategic Intelligence) επικεντρώνεται σε μακροπρόθεσμα

θέματα που αναλύουν την ανταγωνιστικότητα μιας εταιρείας σε μια καθορισμένη περίοδο στο μέλλον. Η πραγματική χρονολογική σειρά εξαρτάται από το είδος του κλάδου. Ο κύριος στόχος των αναλυτών εδώ είναι να προβλέψουν που θα πρέπει να τοποθετηθεί ο οργανισμός σε x χρόνια και να προσδιορίσουν στρατηγικές για τη μετατροπή τους σε πραγματικότητα. Αυτός ο τύπος ανάλυσης περιλαμβάνει κυρίως τον εντοπισμό των αδυναμιών και των σημάτων προειδοποίησης εντός του οργανισμού. Στρατηγική νοημοσύνη επικεντρώνεται στην παροχή πληροφοριών που μπορούν να επηρεάσουν βραχυπρόθεσμες αποφάσεις. Τις περισσότερες φορές, αυτό σχετίζεται με την ανάλυση της τρέχουσας αγοράς, το μερίδιο και το τοπίο του ανταγωνισμού. Αυτό το είδος νοημοσύνης επηρεάζει άμεσα τη διαδικασία πωλήσεων ενός οργανισμού (Dey, et al. 2011)

Η στρατηγική αυτή νοημοσύνη μπορεί να κατηγοριοποιηθεί ως εξής (Dey, et al. 2011):

1. *Πληροφορίες αγοράς*. Παρέχει πληροφορίες σχετικά με τη δημοτικότητα των ανταγωνιστών όσον αφορά τα προϊόντα τους ή τα εμπορικά σήματα στο σύνολό τους, τα προϊόντα που κυκλοφορούν στην αγορά, το μερίδιο αγοράς των ανταγωνιστών. Περιφερειακές ή οι γεωγραφικές προκαταλήψεις των ανταγωνιστών επίσης εμπίπτουν σε αυτή την κατηγορία. Τα συναισθήματα των καταναλωτών που σχετίζονται με την οργάνωση και οι ανταγωνιστές της ανήκουν επίσης στην κατηγορία αυτή.
2. *Πληροφορίες για τις τιμές*. Παρέχει γνώσεις σχετικά με τις τιμές των ανταγωνιστικών προϊόντων.
3. *Προώθηση*. Παρέχει πληροφορίες σχετικά με την προώθηση τις στρατηγικές και το είδος των δραστηριοτήτων προώθησης που υιοθετούν οι ανταγωνιστές.
4. *Άλλα θέματα* - Οργανωτικές πληροφορίες σχετικά με ανταγωνιστές σαν τη δομή του εργατικού τους δυναμικού, την εσωτερική στροφή εστίαση ή όραση, επιτυχία ή αποτυχία των δοκιμών τους, προσφορά νέων προϊόντων, επενδύσεις τεχνολογίας κ.λπ. συμβάλλουν στην οικοδόμηση ενός προφίλ των ανταγωνιστών που μπορεί να είναι χρήσιμο για τους οργανισμούς.

Ο ανταγωνιστικός προγραμματισμός όταν συγκεντρωθεί και αναλυθεί εγκαίρως μπορεί να βοηθήσει τους οργανισμούς να μειώσουν σημαντικά τους χρόνους αντίδρασης. Η ανταπόκριση μια εταιρίας μπορεί να έχει τη μορφή των προσαρμογών των τιμών, των μεταβαλλόμενων

στρατηγικών μάρκετινγκ, της αναθεώρησης του σχεδίου παραγωγής κλπ. Αρκετά μεγάλες αεροπορικές εταιρείες παρακολουθούν συνεχώς τους ανταγωνιστές και αναπροσαρμόζουν τις τιμές τους σε πολύ σύντομο χρονικό διάστημα (Dey, et al. 2011).

Η ανάλυση αποκαλύπτει ότι υπάρχουν διάφοροι τύποι πόρων Διαδικτύου που παρέχουν διαφορετικά είδη ανταγωνιστικής νοημοσύνης. Πηγές Διαδικτύου όπως η online υπηρεσία του Hoover μπορεί να παρέχει πληροφορίες σχετικά με εταιρικά προφίλ και χρηματοοικονομικές πληροφορίες που ενσωματώνονται μέσα από εκατομμύρια δημόσιες και ιδιωτικές εταιρείες για ποικίλο σύνολο βιομηχανιών. Τα κοινωνικά μέσα ενημέρωσης μπορούν να παρέχουν πληροφορίες σχετικά με το εμπορικό σήμα, τη δημοτικότητα, τα αισθήματα των καταναλωτών και τις κινήσεις των ανταγωνιστών. Ειδήσεις, φόρουμ συζητήσεων και ιστολόγια μπορούν να παρέχουν πληροφορίες σχετικά με άλλα ζητήματα όπως οι επενδύσεις τεχνολογίας, η παρουσίαση προϊόντων ή ανακοινώσεις οραμάτων ανταγωνιστών (Dey, et al. 2011).

Είδος Ανταγωνιστικής Νοημοσύνης	Πηγές Διαδικτύου
Εκδηλώσεις	Ειδήσεις, Ιστοσελίδες εταιριών
Στρατηγικές ανταγωνιστών, Επενδύσεις σε τεχνολογία	Ειδήσεις, Forum συζητήσεων, Ιστολόγια, Ιστοσελίδες πατενταρισμένων προϊόντων
Ψυχολογία καταναλωτών	Ιστοσελίδες αξιολογήσεων, Μέσε κοινωνικής δικτύωσης
Προωθητικές ενέργειες και τιμολόγηση	Twitter, Facebook

Ανταγωνιστική Νοημοσύνη προερχόμενη από το διαδίκτυο

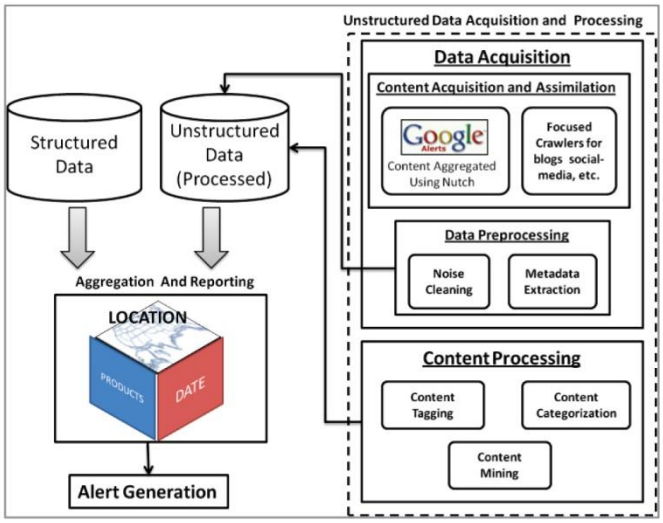
Το διαδίκτυο παρέχει μια χιονοστιβάδα πληροφοριών, που πρέπει να είναι συλλέγονται και υποβάλλονται σε επεξεργασία πριν οι αναλυτές να μπορούν να τις χρησιμοποιήσουν αποτελεσματικά. Σε αυτό το σημείο προτείνεται ο σχεδιασμός ενός ειδικού, εικονικού συστήματος

επιχειρηματικής ευφυΐας που παρέχει ευφυείς διεργασίες στους υπεύθυνους για τη λήψη αποφάσεων έγκαιρα και σε μορφή που τους δίνει τη δυνατότητα να λαμβάνουν αποφάσεις. Ο σχεδιασμός είναι γενικός και σύμφωνα με τους (Dey, et al. 2011) μπορεί να επεκταθεί σε οποιαδήποτε βιομηχανία με τις κατάλληλες προσαρμογές.

Το τμήμα συλλογής και αφομοίωσης περιεχομένου είναι υπεύθυνο για τη συλλογή και την εξαγωγή περιεχομένου από μια σειρά πηγών.

- Χρησιμοποιείται ένα σύνολο εξειδικευμένων εφαρμογών ανίχνευσης ιστότοπων για την εξαγωγή περιεχομένου από μια σειρά προκαθορισμένων πηγών συμπεριλαμβανομένων δημοφιλών ιστότοπων κοινωνικής δικτύωσης. Η λίστα τέτοιων ιστοσελίδων παρέχονται ως εισαγωγή στο σύστημα. Κοινωνικές ιστοσελίδες πολυμέσων όπως το Twitter προσφέρουν APIs (Application Programming Interface)[18], όπως ο κήπος-μάνικα ή πυρκαγιά για τη συλλογή σχετικών tweets (Dey, et al. 2011).
- Προκειμένου να αποφευχθεί η έλλειψη σημαντικών πληροφοριών ακόμα και αν είναι τοποθετημένες σε νέο ή ασήμαντο ιστότοπο, το σύστημα χρησιμοποιεί επίσης το Google Alerts1 για να βελτιώσει την ανακάλεση δεδομένων. Οι Προειδοποιήσεις Google λαμβάνουν ως είσοδο ένα σύνολο λέξεων-κλειδιών ή φράσεων και σε αντάλλαγμα παρέχει στο σύστημα μια λίστα με συνδέσμους, που περιέχουν σχετικό περιεχόμενο. Λέξεις-κλειδιά και φράσεις μπορεί να περιλαμβάνουν ονόματα ανταγωνιστών, προϊόντα ή ονόματα υπηρεσιών κ.λπ. Το σύστημα αναπτύσσει στη συνέχεια τον ανιχνευτή περιεχομένου ανοιχτού κώδικα που ονομάζεται Nutch2 για να εξαγάγει περιεχόμενο από αυτές τις ιστοσελίδες. Επαναλαμβάνεται ως περιοδική διαδικασία, αυτή η μέθοδος εξασφαλίζει ότι ακόμη και ασήμαντα γεγονότα που αναφέρθηκαν στο διαδίκτυο συμπεριλαμβάνονται σίγουρα μόλις αρχίσουν τα «χτυπήματα» των λέξεων-κλειδιών, λόγω της μεθόδου υπολογισμού της συνάφειας της Google (Dey, et al. 2011).

[18] Το API είναι ένα ενδιάμεσο λογισμικό που επιτρέπει την επικοινωνία μεταξύ δύο εφαρμογών. Είναι ο φορέας που παραδίδει κάποιο αίτημά στον πάροχο και στη συνέχεια επιστέφει την απάντηση πίσω στον χρήστη. https://bit.ly/2Ugtj1g (16/4/2019)



Εικόνα 33: Σύστημα συλλογής ανταγωνιστικής νοημοσύνης (Dey, και συν. 2011)

Το τμήμα προ-επεξεργασίας δεδομένων είναι υπεύθυνο για τον καθαρισμό και εξαγωγή σχετικού περιεχομένου από διαφορετικές πηγές. Η αφαίρεση θορύβου περιλαμβάνει δύο βήματα:

- *Λήψη καθαρού περιεχομένου από ιστότοπους*. Κάθε ιστοσελίδα έχει τα δικά της χαρακτηριστικά και περιέχει ανεπιθύμητα υλικό όπως διαφημίσεις ή συνδέσμους σε άλλες άσχετες ιστοσελίδες κ.λπ. Το σύστημα πρέπει να μάθει κανόνες ώστε να εξάγει μόνο σχετικό περιεχόμενο από τις ιστοσελίδες. Αυτοί οι κανόνες έχουν προκύψει από την ανάλυση ιστοτόπων και προσδιορίζοντας τα στατικά και τα δυναμικά δομικά στοιχεία. Τα στατικά δομικά στοιχεία συνήθως περιέχουν συνδέσμους. μενού κλπ. τα οποία δεν έχουν σημασία για το σκοπό συλλογής πληροφοριών. Τα δυναμικά στοιχεία πρέπει να αναλυθούν περαιτέρω για να γίνει διάκριση μεταξύ διαφημίσεων και περιεχομένου. Οι διαφημίσεις περιέχουν συνήθως εικόνες, βίντεο, συνδέσμους κ.λπ. και απορρίπτονται. Το περιεχόμενο κειμένου που δυναμικά αλλάζει μέσα σε μια σελίδα είναι και αυτό που επιλέγεται για περαιτέρω επεξεργασία. Ετικέτες HTML όπως τίτλος, συγγραφέας, by-line κ.λπ. χρησιμοποιούνται για τη δημιουργία σχετικών μετα-δεδομένων (Big Data Framework 2018).
- *Καθαρισμός κειμένου που δημιουργείται από τον καταναλωτή.* Ενώ τα κείμενα ειδήσεων είναι αρκετά καθαρά, περιεχόμενο που συναντάται σε ιστότοπους κοινωνικής δικτύωσης, blogs και φόρουμ συζήτησης είναι γεμάτα με «θόρυβο». Το σύστημα απασχολεί διάφορες

τεχνικές, για να καθαρίσει το σύστημα. Αυτές περιλαμβάνουν την εξάρτηση περιβάλλοντος, την ορθογραφική διόρθωση, την οριοθέτηση προτάσεων, την αφαίρεση περιττής κεφαλαιοποίησης και των ειδικών χαρακτήρων κ.λπ (Big Data Framework 2018).

Το καθαρισμένο περιεχόμενο αποθηκεύεται έπειτα σε κεντρική αποθήκη δεδομένων με όλες τις διαθέσιμες πληροφορίες μετά-δεδομένων όπως τον συγγραφέα, την ημερομηνία δημοσίευσης, τον τόπο αναφοράς, τον τύπο της πηγής, κλπ. Όλα τα πεδία μετά-δεδομένων ενδέχεται να μην ισχύουν για όλους τους τύπους περιεχομένου (Dey, et al. 2011).

Η μονάδα επεξεργασίας περιεχομένου έχει το πιο σημαντικό ρόλο, να αναγνωρίσει δηλαδή και να επισημάνει το σχετικό περιεχόμενο από τη μεγάλη συλλογή. Αποτελείται από αρκετές υπό-μονάδες, οι οποίες εφαρμόζουν μεθόδους μη γραμμικού προγραμματισμού και μεθόδων εξόρυξης κειμένου. Η αφομοίωση περιεχομένου εξαρτάται σε μεγάλο βαθμό από τις επιχειρηματικές απαιτήσεις. Σε αυτή την υπό-μονάδα χρησιμοποιείται γνώση του κλάδου και διάφορα ειδών λεξικών για γλωσσική επεξεργασία. Η φύση των απαιτούμενων πληροφοριών μπορεί να είναι μοναδική για έναν οργανισμό, αν και υπάρχουν αρκετές ομοιότητες σε κάθε επιχειρηματικό τομέα. Ένα από τα βασικά χαρακτηριστικά του προτεινόμενου συστήματος είναι η ευκολία με την οποία μπορεί να προσαρμοστεί στις απαιτήσεις ενός οργανισμού (Dey, et al. 2011).

Το κατηγοριοποιημένο περιεχόμενο είναι εναρμονισμένο και ενοποιημένο για τη δημιουργία αναφορών σε προκαθορισμένα πρότυπα. Η διαδικασία ενοποίησης είναι βασισμένες στη γνώση κάθε οργανισμού. Οι ενοποιημένες αναφορές είναι ποσοτικοποιήσεις εξαγόμενων πληροφοριών και αντιμετωπίζονται ως «επιτομές γνώσης» (intelligence- digests), που παρέχουν στους αναλυτές ένα σημείο πρόσβασης στα υποκείμενα δεδομένα. Πρότυπα αναφοράς καθορίζονται με βάση το συγκεκριμένο πλαίσιο της εταιρείας για τη μετάδοση της πληροφορίας. Η συγχώνευση γίνεται σε τρείς διαστάσεις. Αυτά είναι (i) χρόνος (ii) τοποθεσία και (iii) προϊόν και υπηρεσίες. Όλες οι πληροφορίες που αφορούν ένα μόνο προϊόν ή υπηρεσία συγκεντρώνονται μαζί, ταξινομούνται ανά ώρα και διαχωρίζονται ανά τόπο ενώ παρουσιάζονται στον τελικό χρήστη. Κατά την παρουσίαση πληροφοριών ο βαθμός λεπτομέρειας εξαρτάται από τον τελικό χρήστη. Οι εκθέσεις έχουν στόχο να δώσουν τη δυνατότητα στους υπεύθυνους λήψης αποφάσεων να ενεργήσουν με ενημερωμένο τρόπο (Dey, et al. 2011).

Η ενότητα ειδοποίησης αξιολογεί διάφορα είδη επεξεργασμένων πληροφοριών και δημιουργεί ειδοποιήσεις σχετικά με προκαθορισμένους ενεργοποιητές ή σημαντικά γεγονότα που έχουν αναγνωριστεί ως προοίμια σεναρίων, που απαιτούν άμεσες απαντήσεις. Ένα παράδειγμα ενός γεγονότος ενεργοποίησης είναι η είδηση για το λανσάρισμα προϊόντος από έναν ανταγωνιστή με λεπτομέρειες σχετικά με το προϊόν. Αυτό μπορεί να είναι μια πιθανή προειδοποίηση για μείωση των πωλήσεων ή του μεριδίου αγοράς τα επόμενα τρίμηνα (Dey, et al. 2011).

6.2 Εξαγωγή Πληροφοριών μέσω της Ανταγωνιστικής Νοημοσύνης

Το πρώτο βήμα στην επεξεργασία περιεχομένου αφορά την ταυτοποίηση και την επισήμανση σχετικού περιεχομένου. Το παρόν σύστημα χρησιμοποιεί παρακολούθηση με βάση τα δεδομένα. Οι έννοιες ορίζονται με όρους λέξεων, φράσεων, οντοτήτων και τους συνδυασμούς τους μαζί με τους συντακτικούς και γραμματικούς περιορισμούς που επιβάλλονται στον συνδυασμό. Οργανωτικές πληροφορίες σχετικά με τα προϊόντα, τις υπηρεσίες και τα προϊόντα της οι στρατηγικές μπορούν να αποθηκευτούν πολύ αποτελεσματικά στην οντολογία τομέα περιγράφει μεθοδολογίες για τη δημιουργία οντολογίας τομέα από το αδόμητο κείμενο (Dey, et al. 2011).

Οι πληροφορίες των σχετικών περιεχομένων που συλλέγονται από το κείμενο αναλύονται περαιτέρω για να τους δοθούν ετικέτες ανταγωνιστικής νοημοσύνης. Η ανάθεση ετικέτας βασίζεται στην ανάλυση εξάρτησης ακολουθούμενη από ανάλυση του σχετικού αντικειμένου, του αντικειμένου και της φύσης του ρήματος. Εφόσον το ίδιο ρήμα μπορεί να έχει συνώνυμα, μπορεί να χρησιμοποιηθεί το VerbNet5 για την ταξινόμηση των ρημάτων. Ο μηχανισμός αντιστοίχισης διαφορετικών τύπων ετικετών εξηγείται παρακάτω με παραδείγματα για κάθε κατηγορία (Dey, et al. 2011).

- *Εκδηλώσεις ανθρώπων* - Γεγονότα που σχετίζονται με βασικούς παράγοντες της αγοράς όπως οι Διευθύνοντες Σύμβουλοι ή οι Διευθύνοντες Σύμβουλοι Οικονομικών εταιρειών ονομάζονται «ανθρώπινα γεγονότα». Οι φορείς της αγοράς προσδιορίζονται ως ονομαστικές οντότητες χρησιμοποιώντας τεχνικές ονομασίας οντοτήτων αναγνώρισης (Named Entity Recognition). Υπάρχουν πολλά εργαλεία NER. Ο Dey (2011) χρησιμοποιεί το Stanford NER6. Οι τίτλοι, οι ρόλοι και οι τιμητικές διακρίσεις όπως "κ.", "διευθύνων σύμβουλος", "αντιπρόεδρος" Πρόεδρος ", κλπ. Επιπλέον, το σύστημα εφαρμόζεται με

βάση την εξάρτηση για τον προσδιορισμό της σχετικής πληροφορίας. Ενιαίες σχέσεις που περιλαμβάνουν το τιμητικό και μια ονομαζόμενη οντότητα που συνδέεται από το MOD ή τον μετατροπέα σχέσης. Δυαδικά σχεσιακά αναγνωρίζονται ως γεγονότα ανθρώπων. Τρισδιάστατα σχεσιακά πρότυπα που αποτελούνται από σχεσιακά στοιχεία πληροφοριών που αφορούν πρόσωπα ως υποκείμενα και ονόματα οργάνωσης στο αντικείμενο και τυπικά ρήματα όπως "Εξυπηρετεί", "κατέχει", κλπ. επίσης κατηγοριοποιούνται ως άνθρωποι γεγονότα (Dey, et al. 2011).

- *Στρατηγικές ανταγωνιστών* - Ειδήσεις αγοράς, διπλώματα ευρεσιτεχνίας κ.λπ. παρέχουν πληροφορίες σχετικά με τις επενδυτικές στρατηγικές ανταγωνιστών. Πληροφορίες για λανσάρισμα προϊόντων, συγχωνεύσεις και εξαγορές, νέες επενδύσεις σε τεχνολογίες, το άνοιγμα ενός νέου καταστήματος, τη δέσμευση ενός νέου προμηθευτή κ.λπ. εμπίπτουν στην κατηγορία αυτή. Κάθε μια από αυτές τις δραστηριότητες μπορούν να συσχετιστούν με συγκεκριμένα σύνολα πράξεων ρήματος. Μοτίβα με βάση τη σχέση είναι χρήσιμα για την αναγνώριση των εκδηλώσεις. Για παράδειγμα, τα ουσιαστικά σαν την άνοδο, την πτώση, ανέβηκε μαζί με τις πωλήσεις ως θέμα και τα αριθμητικά στοιχεία στο αντικείμενο είναι ενδεικτικά αυτών των αναφορών. Τα αντικείμενα μπορεί επίσης να περιέχουν τιμές χρημάτων. Ομοίως οι έννοιες γύρω από το επενδυμένο, διορισμένο, πρόσληψη κλπ. συγκεντρώνονται σε αυτήν την κατηγορία (Dey, et al. 2011).
- *Διάθεση και απόψεις των καταναλωτών*. Ιστολόγια, συζητήσεις στα φόρουμ και τα κοινωνικά δίκτυα αφθονούν με απόψεις του καταναλωτή. Όλες οι μεγάλες εταιρείες έχουν Twitter και Facebook για να μπορούν να απευθυνθούν σε ένα μεγαλύτερο κοινό. Η εξόρυξη αυτών των συζητήσεων παρέχει πολύτιμες πληροφορίες. Με βάση αυτές τις πληροφορίες είναι λογικό να δημιουργηθούν αναφορές με τις απόψεις των καταναλωτών για ανταγωνιστικά προϊόντα ή υπηρεσίες (Dey, et al. 2011).
- *Προωθητικές ενέργειες ανταγωνιστών*. Τα προωθητικά γεγονότα μπορούν να ληφθούν πολύ αποτελεσματικά από το διαδίκτυο. Ενώ οι έμποροι λιανικής πώλησης και οι διάφορες μάρκες ανταγωνίζονται μεταξύ τους για την προσοχή των καταναλωτών, καινοτόμα προγράμματα προώθησης όπως Groupon, Foursquare κ.λπ. χρησιμοποιούν την κοινωνική δικτύωση για να εξασφαλίσουν την πολύτιμη συμμετοχή των καταναλωτών. Είναι δυνατό για έναν οργανισμό να κερδίσει πληροφορίες πραγματικού χρόνου σε προωθητικές

ενέργειες ανταγωνιστών και να προσαρμόσει τις βραχυπρόθεσμες και μακροπρόθεσμες στρατηγικές της. Τυπικά ρήματα όπως "προωθεί" ή "ωθεί" παρατηρούνται στο προωθητικό υλικό. Έννοιες που υποδηλώνουν προσφορές περιλαμβάνουν πωλήσεις, έκπτωση, προσφορά, δωρεάν δώρο,% έκπτωση κλπ. Οι προσφορές μπορούν να κατηγοριοποιηθούν περαιτέρω στην προώθηση ή προώθηση προϊόντων (Dey, et al. 2011).

- *Γεγονότα πραγματικού κόσμου.* Γεγονότα που δεν σχετίζονται με την άμεση πληροφόρηση της αγοράς μπορούν επίσης να επηρεάσουν τις επιχειρήσεις με πολλούς διαφορετικούς τρόπους. Ενώ οι καταστροφές δεν μπορούν να προβλεφθούν, η επίδραση των καταστροφών μπορεί να μειωθεί σε πραγματικό χρόνο έχοντας την προσοχή σε εξελισσόμενες καταστάσεις ακόμη και σε απομακρυσμένες περιοχές του κόσμου λόγω του Twitter. Αυτό αποδεικνύεται Έχει ήδη φανεί η άμεση απόκριση χρηστών του Twitter σχετικά με τέτοιες εκδηλώσεις η οποία προηγείται των καλύτερων ειδήσεων πηγών με διαφορά σχεδόν 30 λεπτών. Το Twitter είναι μια ιδιαίτερα σημαντική πηγή πληροφοριών για γεγονότα που δεν έχουν παγκόσμια σημασία, αλλά αφορούν τοπικά θέματα. Αυτά τα είδη γεγονότων, όπως μια πλημμύρα σε μια μικρή τοποθεσία ή μια πυρκαγιά στο εργοστάσιο μπορεί να μην αναφερθεί ακόμη και στα Νέα, αλλά μπορεί να επηρεάσει τις λειτουργίες είτε του ίδιου οργανισμού είτε των ανταγωνιστών του σε σημαντικό βαθμό. Τα tweets μπορούν να αναγνωριστούν και να αναλυθούν ώστε να ληφθούν προληπτικά βήματα και να ελαχιστοποιηθούν οι απώλειες για ένα σενάριο εφοδιαστικής αλυσίδας. Όπως είναι λογικό μοτίβα για αυτή την εφαρμογή δεν μπορούν να προκαθοριστούν. Συχνά τα νέα πρότυπα που παρατηρούνται προστίθενται τακτικά στη συλλογή για τις δράσεις αντίδρασης (Dey, et al. 2011).

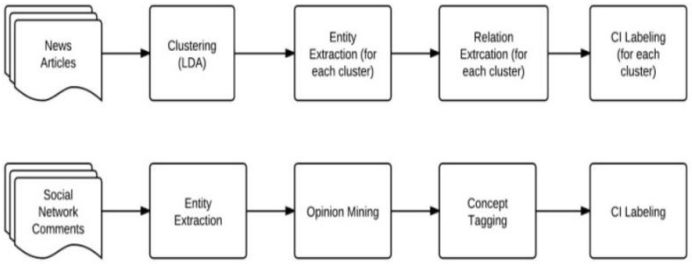
6.3 Εκχώρηση ετικετών ανταγωνιστικής ευφυίας

Στα έγγραφα παραχωρείται μια ετικέτα Competitive Intelligence (CI) με βάση το περιεχόμενό τους, συμπεριλαμβανομένων των οντοτήτων, των σχέσεων, των εννοιών και των απόψεων. Δεδομένου ότι τα άρθρα έχουν μεταβλητό χαρακτήρα το σύστημα ακολουθεί διαφορετικές προσεγγίσεις για την επισήμανσή τους. Για άρθρα ειδήσεων και κείμενων τύπου blog ή του φόρουμ συζητήσεων που είναι εκτεταμένης φύσης και συνήθως γραμματικά πιο συνεπή, χρησιμοποιούνται τεχνικές με βάση το NLP για τον εντοπισμό οντοτήτων και σχέσεων. Οι

ετικέτες Competitive Intelligence βασίζονται στην κατηγορία των σχέσεων που αποκτήθηκαν. Κείμενα κοινωνικού δικτύου από την άλλη είναι συνήθως σύντομα και θορυβώδη στη φύση. Λεκτικές ετικέτες και αυτές που βασίζονται στις έννοιες χρησιμοποιούνται για την ανάθεση ετικετών σε αυτούς τους τύπους έγγραφα. Πιο συγκεκριμένα τα άρθρα ειδήσεων αναλύονται για ανθρώπους-εκδηλώσεις και γεγονότα στρατηγικής ανταγωνιστών, ενώ το περιεχόμενο κοινωνικού δικτύου αφορά προωθητικές ενέργειες και τη γνώμη των καταναλωτών (Dey, et al. 2011).

Ενώ το περιεχόμενο κοινωνικού δικτύου μπορεί να είναι φυσικά ομαδοποιημένο με βάση το πλαίσιο στο οποίο εμφανίζεται (για παράδειγμα, απάντηση σε συγκεκριμένο σχόλιο), η ομαδοποίηση βοηθά στη σύνοψη σχετικών ειδήσεων και στη μείωση των διαστάσεων της ανάλυσης. Το παρών σύστημα χρησιμοποιεί θεματικές συστοιχίες σε άρθρα ειδήσεων για ομαδοποίηση μαζί τους. Τα θέματα εξάγονται χρησιμοποιώντας κατανομή Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003). Κάθε θέμα αντιπροσωπεύει μια ομάδα ειδήσεων που το καθένα μεγιστοποιεί αυτό το θέμα. Ο οντότητες Ονόματα και οι σχέσεις εξάγονται από όλα τα άρθρα ειδήσεων. Οι πιο σημαντικές οντότητες Ονόματος και οι σχέσεις εξ ορύσσονται έπειτα από κάθε σύμπλεγμα χρησιμοποιώντας βάσει του υπολογισμού της μέγιστης εντροπίας. Τα κύρια ρήματα στις σημαντικές σχέσεις χρησιμοποιούνται περαιτέρω για την εκχώρηση μιας ετικέτας CI στο σύμπλεγμα και συνεπώς όλα τα σχετικά αντικείμενα που σχετίζονται σε αυτό (Dey, et al. 2011).

Τα περιεχόμενα του κοινωνικού δικτύου είναι σύντομα και με θόρυβο. Έτσι η επισήμανση αυτών των άρθρων βασίζεται σε κανόνες. Το σύστημα χρησιμοποιεί προηγούμενες γνώσεις σχετικά με τις επιχειρήσεις, τα προϊόντα τους και άλλες έννοιες όπως εκπτώσεις ή προωθήσεις κλπ. για την επισήμανση άρθρων. Η εξόρυξη γνώμης είναι επίσης μια άλλη δραστηριότητα που εκτελείται σε αυτό το περιβάλλον. Για αυτό το λόγο κάθε σχόλιο στη συνέχεια συνδέεται με θετικό και αρνητικό σκορ της κοινής γνώμης μαζί με τις έννοιες προϊόντων που αντιπροσωπεύουν (Dey, et al. 2011).



Εικόνα 34: Προσέγγιση της σήμανσης CI για διαφορετικούς τύπους εγγράφων (Dey, et al. 2011)

Η παρουσιάζει μια προσέγγιση ενώ ο αλγόριθμος που παρουσιάζεται παρακάτω περιγράφει την διαδικασία που εφαρμόζονται οι ετικέτες (Dey, et al. 2011).

ΑΛΓΟΡΙΘΜΟΣ

Έστω D η συλλογή μη δομημένων εγγράφων τη στιγμή t.

$$D = \prod_{d \in N,B,C} d$$

Όπου το N σημαίνει τη συλλογή άρθρων ειδήσεων, το B υποδηλώνει τη συλλογή άρθρων ιστολογίου blog και άρθρων συζήτησης και C υποδηλώνει τη συλλογή σύντομων σχολίων που συλλέχθηκαν από την κοινωνικά μέσα δικτύωσης όπως Twitter, Facebook κ.λπ.

Επεξεργασία Ειδήσεων

- Συλλέξτε όλα τα άρθρα ειδήσεων που έχουν αφορούν την ίδια χρονική σήμανση σε μια ενιαία ομάδα.
- Εφαρμόστε την ομαδοποίηση με βάση την κατανομή κατά Latent Dirichlet σε κάθε ομάδα.
- Κάθε σύμπλεγμα c που αποκτάται για ένα μόνο d έχει εκχωρηθεί στην ίδια σήμανση ημερομηνίας, όπως τα άρθρα που περιέχονται σε αυτό.
- Για κάθε σύμπλεγμα c που λαμβάνεται,
- Αφαίρεση οντοτήτων.
- Εξαγωγή σχέσεων με ReVerb

- Σήμανση κάθε σχέσης ως 1) γεγονός – άνθρωπος 2) ανταγωνιστής – στρατηγική με βάση τα ρήματα που χρησιμοποιούνται
- Βρείτε τις πιο συχνές οντότητες και ετικέτες σχέσεων στο σύμπλεγμα c.
- Σήμανση του συμπλέγματος c με συχνή σχέση ετικετών και οντότητων που υπερβαίνουν συγκεκριμένα όρια.

Επεξεργασία περιεχομένου Blog

Για κάθε ιστότοπο ιστολογίου, προσδιορίστε το ιστολόγιο ως χώρο ανασκόπησης ή συζήτησης. Για χώρους συζήτησης, επεξεργαστείτε άρθρα όπως άρθρα ειδήσεων. Για ιστοσελίδες κριτικής, επεξεργαστείτε άρθρα όπως σχόλια όπως περιγράφονται στη συνέχεια.

Επεξεργασία περιεχομένου κοινωνικών δικτύων

- Για κάθε σχόλιο που συλλέγεται από έναν ιστότοπο κοινωνικής δικτύωσης σχετίζονται με αυτό τα ακόλουθα μεταδεδομένα που σχετίζονται με τον
- συντάκτη
- τον χρόνο της δημοσίευσης
- τίτλος σχόλιου. Αυτό είναι ίδιο με το σχόλιο εάν είναι ένα νέο σχόλιο. Ωστόσο, αν το παρόν σχόλιο είναι μια απάντηση, αντίγραφο ή μια επέκταση ενός προγενέστερου σχολίου, τότε η επικεφαλίδα σχολίου αποθηκεύεται στο αρχικό σχόλιο.
- Περιεχόμενο σχολίου - το πραγματικό κείμενο του σχολίου
- εξωτερικές συνδέσεις. Σε περίπτωση που ένα σχόλιο περιέχει ένα σύνδεσμο εξωτερικής πηγής, ο σύνδεσμος αυτός URL αποθηκεύεται προσωρινά στο πεδίο αυτό. Το παρόν σύστημα δεν επεξεργάζεται το περιεχομένο περαιτέρω.
- Εκτελέστε την εξαγωγή οντοτήτων από το περιεχόμενο των σχολίων.
- Διεξάγετε εξόρυξη γνώμης από το περιεχόμενο των σχολίων.
- Εκτελέστε σήμανση ετικετών βασισμένες σε έννοιες από περιεχόμενο των σχολίων. Ανέθεσε κατηγορία προϊόν ή υπηρεσία προώθησης και μέσω οντολογίας αναζήτησε για έννοιες.
- Ορίστε την ετικέτα CI

- Αντιστοίχιση ετικέτας - "συμβάν προώθησης" εάν τα σχόλια περιέχουν έννοιες σχετικές με την προώθηση
- Αντιστοίχιση ετικέτας ΓΝΩΜΗ - εάν υπάρχει σχόλιο γνώμη
- Άθροισε όλα τα άρθρα σε βάση δεδομένων μαζί με μετα-δεδομένα όπως ο χρόνος και σχετικές ετικέτες όπως ετικέτες οντοτήτων, Ανταγωνιστική Ευφυΐα, ετικέτα συμβάντος και απόψεις.

Τέλος αλγορίθμου

6.4 Σχηματίζοντας μια στρατηγική Μεγάλων Δεδομένων

Σε αυτή την ενότητα θα γίνει μια πρακτική προσέγγιση για τη διαμόρφωση μιας στρατηγικής Big Data. Μια στρατηγική Big Data, δεν μπορεί να ξεχωρίζει από την οργανωτική στρατηγική, και πρέπει να είναι σταθερά ενσωματωμένη σε αυτήν. Όταν συζητάμε για μια στρατηγική μεγάλων δεδομένων, αυτό ουσιαστικά σημαίνει μια επιχειρηματική στρατηγική που περιλαμβάνει τα μεγάλα δεδομένα (Big Data Framework 2018).

Μια στρατηγική Big Data καθορίζει ένα ολοκληρωμένο όραμα σε ολόκληρη την επιχείρηση και ορίζει ένα θεμέλιο για τον οργανισμό να χρησιμοποιεί σχετιζόμενα δεδομένα ή εξαρτώμενα δεδομένα. Μια καλά καθορισμένη και περιεκτική στρατηγική Big Data καθιστά εφικτά τα πλεονεκτήματα τα οποία μπορούν να προσφερθούν στον οργανισμό. Αυτή η στρατηγική ορίζει τα βήματα που πρέπει να εκτελέσει ένας οργανισμός προκειμένου να γίνει "Data Driven Enterprise). Η στρατηγική Big Data περιλαμβάνει 1) μερικές κατευθυντήριες αρχές για την επίτευξη του οράματος μιας επιχείρησης καθοδηγούμενης από δεδομένα, 2) να καθοδηγεί την οργάνωση 3) να επιλέξει συγκεκριμένους επιχειρηματικούς στόχους και 4) να είναι το σημείο εκκίνησης για τον προγραμματισμό των δεδομένων σε ολόκληρη την επιχείρησης (Big Data Framework 2018).

Εκτός από τα οφέλη απόκτησης ενός ανταγωνιστικού πλεονεκτήματος, οι επιχειρήσεις απαιτούν στρατηγική Big Data επειδή ξεπερνά τα οργανωτικά όρια. Χωρίς στρατηγική μεγάλων δεδομένων, οι επιχειρήσεις θα αναγκαστούν να ασχοληθούν με ποικίλες δραστηριότητες που σχετίζονται με δεδομένα, οι οποίες πιθανότατα θα ξεκινήσουν από διαφορετικές επιχειρηματικές μονάδες. Διάφορα τμήματα είναι πιθανό να ξεκινήσουν τις δικές τους αναλύσεις, Επιχειρηματική

Νοημοσύνη ή προγράμματα διαχείρισης δεδομένων, χωρίς να λαμβάνουν υπόψη το συνολικό μακροπρόθεσμο στρατηγικό στόχο (Big Data Framework 2018).

Η κινητήρια δύναμη πίσω από τη διαμόρφωση μιας επιχείρησης Big Data στρατηγική θα πρέπει να είναι ο συνδυασμός είτε του CEO (όταν το Big Data ορίζει την επιχείρηση) είτε του COO (/ CIO (όταν το Big Data βελτιστοποιεί την επιχείρηση). Αυτό αναγνωρίζει ότι τα δεδομένα δεν είναι μόνο ένα στοιχείο IT, αλλά και ένα οργανικό περιουσιακό στοιχείο σε ολόκληρο τον οργανισμό. Μια καλά καθορισμένη στρατηγική μεγάλων δεδομένων για επιχειρήσεις πρέπει να είναι δυνατή για τους οργανισμούς. Για να επιτευχθεί αυτό, οι οργανώσεις μπορούν να ακολουθήσουν την ακόλουθη προσέγγιση σε 5 βήματα για να διατυπώσουν τη Στρατηγική τους, όπως φαίνεται και στην (Big Data Framework 2018).

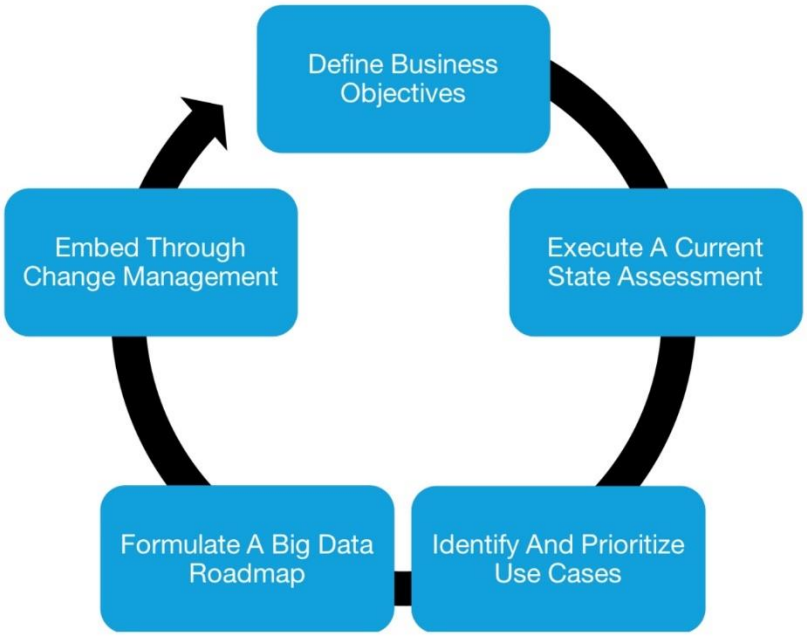
1) Καθορισμός επιχειρηματικών στόχων

2) Εκτέλεση αξιολόγησης της τρέχουσας κατάστασης

3) Προσδιορισμός και προτεραιότητα στις περιπτώσεις χρήσης (Use Cases)

4) Διαμόρφωση ενός χάρτη πορείας για Big Data (Roadmap)

5) Αφομοίωση μέσω των αλλαγών διαχείρισης (Change Management)



Εικόνα 35: Στρατηγική για την ένταξη Αναλυτική Μεγάλων Δεδομένων σε μια Επιχείρηση (Big Data Framework 2018)

Βήμα 1ο : Καθορισμός επιχειρηματικών στόχων

Προκειμένου να χρησιμοποιηθούν τα Big Data σε οποιαδήποτε επιχείρηση, είναι πρώτα απαραίτητο να γίνει μια σύντομη περιγραφή των στόχων μιας επιχείρησης. Τι κάνει μια οργάνωση επιτυχημένη; Τα έσοδα και τα κέρδη είναι συχνά αποτέλεσμα της πραγματοποίησης ή της υπέρβασης των βασικών δεικτών επίδοσης (KPIs). Ξεκινώντας με την κατανόηση του τρόπου με τον οποίο μια επιχείρηση είναι επιτυχής, πριν εξερευνήσετε πώς οι τεχνολογίες και η λύση Big Data θα μπορούσαν να βελτιώσουν τη μελλοντική απόδοση (Big Data Framework 2018).

Η στρατηγική Big Data θα πρέπει να ευθυγραμμιστεί με τους εταιρικούς επιχειρηματικούς και να στοχεύει επιχειρησιακά προβλήματα, καθώς ο πρωταρχικός σκοπός του Big Data είναι να αποκομίζει αξία αξιοποιώντας τα δεδομένα. Ένας τρόπος για να επιτευχθεί αυτό να ευθυγραμμιστεί με τη διαδικασία στρατηγικού προγραμματισμού της επιχείρησης, όπως οι περισσότεροι οργανισμοί που έχουν ήδη εφαρμόσει αυτή τη διαδικασία (Big Data Framework 2018).

Παραδείγματα επιχειρηματικών στόχων που συχνά εμφανίζονται από μια πρόσφατη έρευνα αναφέρθηκαν στην .

Εικόνα 36: Παραδείγματα επιχειρηματικών στόχων (Big Data Framework 2018)

Προκειμένου να προσδιοριστούν οι επιχειρηματικοί στόχοι, η εμπλοκή βασικών επιχειρηματικών φορέων είναι υψίστης σημασίας. Η επιχείρηση πρέπει να έχει αυτούς τους ενδιαφερόμενους να εμπλέκονται από την αρχή και να παρέχουν βασικές πληροφορίες σε συνεχή βάση. Οι βασικοί ενδιαφερόμενοι που πρέπει να εξετάσουν σε αυτό το πρώτο βήμα είναι (Big Data Framework 2018):

• Εκτελεστικοί χορηγοί. Η σημασία της εύρεσης και συνεργασίας με εκτελεστικούς χορηγούς δεν μπορεί να υποτιμηθεί. Η υποστήριξή τους είναι απαραίτητη σε όλες τις φάσεις της διαμόρφωσης της στρατηγικής δεδομένων και την εφαρμογή της.

• Σωστά ταλέντα στην ομάδα. Η συμμετοχή των ανθρώπων με το σωστό ταλέντο και σύνολα δεξιοτήτων είναι απαραίτητα κατά τον καθορισμό των σωστών επιχειρηματικών στόχων. Πρέπει να επιλεχθούν τόσο για το εσωτερικό ταλέντο όσο και την τεχνική συμβουλή τους.

• Πιθανοί κατασκευαστές προβλημάτων. Κάθε έργο ή πρωτοβουλία θα έχει κάποιους «παρακινητές» που είτε σκόπιμα είτε ακούσια αντιτίθενται στην αλλαγή. Γνωρίζοντας ποιοι είναι, και τα κίνητρά τους εκ των προτέρων θα βοηθήσουν αργότερα στη διαδικασία.

Βήμα 2ο : Εκτέλεση αξιολόγησης της τρέχουσας κατάστασης

Σε αυτό το βήμα, η κύρια εστίαση είναι να αξιολογηθούν οι τρέχουσες επιχειρηματικές διαδικασίες, οι πηγές δεδομένων, τα υπάρχοντα δεδομένα, υπάρχων τεχνολογικός εξοπλισμός, δυνατότητες και πολιτικές της επιχείρησης. Ο σκοπός αυτής της άσκησης είναι να βοηθήσει με την ανάλυση διαγνωστικής επιθεώρησης (Gap analysis) της υπάρχουσας κατάστασης και την επιθυμητή μελλοντική κατάσταση.

Για παράδειγμα, εάν το πεδίο εφαρμογής της στρατηγικής δεδομένων είναι να αποκτηθεί μια προβολή 360 μοιρών των πελατών και των υποψήφιων πελατών, η τρέχουσα κατάσταση αξιολόγησης θα συμπεριλάμβανε οποιαδήποτε επιχειρηματική διαδικασία, στοιχεία δεδομένων, συμπεριλαμβανομένης της αρχιτεκτονικής, των δυνατοτήτων της επιχείρησης και πληροφορικής και των πολιτικών που προσεγγίζουν τους πελάτες. Η αξιολόγηση της τρέχουσας κατάστασης διεξάγεται συνήθως με μια σειρά συνεντεύξεων με τους υπαλλήλους που συμμετέχουν στην απόκτηση, διατήρηση και επεξεργασία πελατών (Big Data Framework 2018).

Σε αυτό το στάδιο, είναι επίσης σημαντικό να εντοπιστούν και να αναπτυχθούν ορισμένα μέλη της ομάδας ως πρότυπα δεδομένων. Αυτοί οι άνθρωποι πραγματικά πιστεύουν στην ισχύ των δεδομένων κατά τη λήψη αποφάσεων και μπορεί ήδη να χρησιμοποιούν τα δεδομένα και τις αναλύσεις με ισχυρό τρόπο. Με τη συμμετοχή αυτών των ανθρώπων, ζητώντας τη συμβολή τους, γίνεται ευκολότερη η διατύπωση του χάρτη πορείας σε μεταγενέστερο στάδιο (Big Data Framework 2018).

Βήμα 3ο : Προσδιορισμός και προτεραιότητα στις περιπτώσεις χρήσης (Use Cases)

Στο βήμα 3, οραματίστε πώς προγνωστικές αναλύσεις, περιγραφικές αναλύσεις και τελικά τα γνωστικά τα αναλυτικά στοιχεία μπορούν να βοηθήσουν την επιχείρηση να επιταχύνει, να βελτιστοποιήσει και να μαθαίνει συνεχώς, αναπτύσσοντας περιπτώσεις χρήσης που

ευθυγραμμίζονται με τους επιχειρηματικούς στόχους από το βήμα 1. Καταγράψτε κάθε μία από τις περιπτώσεις χρήσης για να κατανοήσετε πώς μπορούν τα μεγάλα δεδομένα να βοηθήσουν τους στόχους της επιχείρησης, όπως φαίνεται στον (Big Data Framework 2018).

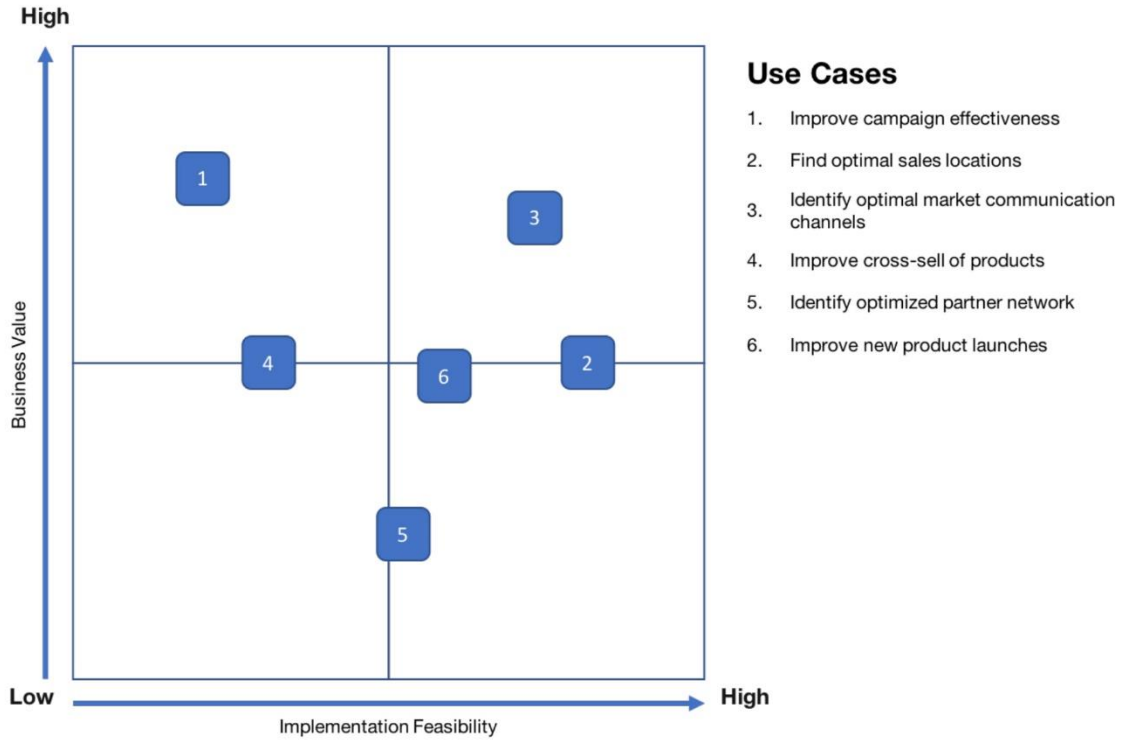
Πίνακας 8: Ανάπτυξη Περιπτώσεων Χρήσης (Big Data Framework 2018)

BUSINESS OBJECTIVE: PRODUCT LAUNCHES IN EMERGING MARKETS	
Identify which products can be launched most successfully in which market by analyzing buying patterns and correlations between product categories in order to find the optimal product-market combinations.	
BUSINESS POTENTIAL	**IMPLEMENTATION RISKS**
• Development of targeted product launches that are aligned with the requirements of each market, so that profit margins are optimized. • Reduction of failed product launches, so that costs are reduced.	• Accuracy and availability of historical purchase data. • Ability to set-up and coordinate localized product launches. • Language barriers. • Policy and governmental barriers in emerging markets.
FINANCIAL GOALS IMPACT	**IMPLEMENTATION CONSIDERATIONS**
• Revenue Growth – 3/5 • Customer Acquisition – 5/5 • Customer Retention – 4/5 • Market Basket Margin – 3/5 • Product Cross Sell – 4/5 • New Product – 4/5 • Financial Goals Score = 3.8 / 5	• Capturing and governing data models and data sets. • Ability to use predictive models to forecast demand. • Analysis capabilities and knowledge. • Implementation Feasibility = 4 / 5

Καλά καθορισμένες περιπτώσεις χρήσης παρέχουν έναν σαφή και αποτελεσματικό τρόπο καθορισμού των τεχνολογιών Big Data και οι λύσεις μπορούν να υλοποιήσουν επιχειρηματικούς στόχους. Μετά την ανάπτυξη των περιπτώσεων χρήσης, το επόμενο βήμα είναι να δοθεί προτεραιότητα σε όλες τις περιπτώσεις χρήσης που βασίζονται στον επιχειρηματικό αντίκτυπό τους, τον προϋπολογισμό και τις απαιτήσεις πόρων. Με τη διεξαγωγή αυτής της άσκησης, οι επιχειρήσεις μπορούν να προσδιορίσουν ποιες πρωτοβουλίες Big Data παρέχουν την μεγαλύτερη επιχειρηματική αξία (Big Data Framework 2018).

Ένας από τους πιο αποτελεσματικούς τρόπους να δοθεί προτεραιότητα στη χρήση υποθέσεων είναι η χρήση ενός πίνακα προτεραιοτήτων. Αυτός ο πίνακας διευκολύνει τη συζήτηση μεταξύ

της επιχείρησης και του τμήματος της πληροφορικής για τον εντοπισμό των "σωστών" περιπτώσεων χρήσης για την έναρξη μιας πρωτοβουλίας μεγάλων δεδομένων– αυτές οι Περιπτώσεις Χρήσης με σημαντική επιχειρηματική αξία και λογική σκοπιμότητα της επιτυχούς εφαρμογής θα μπορούν να εφαρμοστούν. Ο πίνακας προτεραιοτήτων όπως φαίνεται στο παρακάτω είναι ένα εξαιρετικό εργαλείο διαχείρισης για την καθοδήγηση οργανωτικής ευθυγράμμισης και δέσμευσης γύρω από την κορυφαία προτεραιότητα της Χρήσης Υποθέσεων (Big Data Framework 2018).



Εικόνα 37: Πίνακας Προτεραιοτήτων (Big Data Framework 2018)

Βήμα 4ο : Διαμόρφωση ενός χάρτη πορείας για Big Data (Roadmap)

Το επόμενο βήμα είναι πιθανώς η πιο έντονη και αμφιλεγόμενη φάση και χωρίς αμφιβολία θα καταλαμβάνει την πλειοψηφία του χρόνου στη διαμόρφωση της στρατηγικής δεδομένων. Με βάση την τρέχουσα δυνατότητα (βήμα 2) και τις αναγνωρισμένες περιπτώσεις χρήσης με προτεραιότητα (βήμα 3), μπορεί να αναπτυχθεί ο χάρτης πορείας. Ο χάρτης πορείας για τα μεγάλα δεδομένα περιγράφει τα έργα (ή περιπτώσεις χρήσης) να εκτελεστούν πρώτα και ποιες

δυνατότητες (γνώσεις, εργαλεία και δεδομένα) θα αυξηθούν κατά τα επόμενα 3-5 χρόνια (Big Data Framework 2018).

Με την επιθυμητή μελλοντική κατάσταση, ο χάρτης πορείας θα πρέπει να επικεντρωθεί στον εντοπισμό των κενών στα δεδομένα την αρχιτεκτονική, την τεχνολογία και τα εργαλεία, τις διαδικασίες και φυσικά τους ανθρώπους (δεξιότητες, κατάρτιση κ.λπ.). Η αξιολόγηση της τρέχουσας κατάστασης και οι υποθέσεις χρήσης θα παρουσιάσουν πολλές στρατηγικές επιλογές για πρωτοβουλίες και το επόμενο καθήκον είναι να δωθεί προτεραιότητα στις επιλογές αυτές με βάση την πολυπλοκότητα, τον προϋπολογισμό και τα δυνατά οφέλη (Big Data Framework 2018).

Οι χορηγοί και τα ενδιαφερόμενα μέρη θα διαδραματίσουν σημαντικό ρόλο στην ιεράρχηση αυτών των πρωτοβουλιών. Το τελικό αποτέλεσμα αυτής της φάσης είναι ένας οδικός χάρτης για την ανάπτυξη των προτεραιοτήτων πρωτοβουλίας Big Data (Big Data Framework 2018).

Βήμα 5ο : Αφομοίωση μέσω συστήματος Change Management

Παρόλο που τεχνικά δεν αποτελεί μέρος της διατύπωσης της στρατηγικής Big Data, η Διαχείριση Αλλαγών (που περιλαμβάνει την προσπάθεια και τα μυαλά των ανθρώπων) θα έχει βαθιές επιπτώσεις στην επιτυχία ή στην αποτυχία της στρατηγικής Big Data (Big Data Framework 2018).

Το σύστημα Change Management πρέπει να περιλαμβάνει οργανωτικές αλλαγές, πολιτιστικές αλλαγές, τεχνολογικές αλλαγές και αλλαγές στις επιχειρηματικές διαδικασίες. Η Διακυβέρνηση δεδομένων, η οποία ασχολείται με το σύνολο της διαχείρισης της διαθεσιμότητας, της χρηστικότητας, της ακεραιότητας και της ασφάλειας των δεδομένων, καθίσταται ζωτικής σημασίας για την αλλαγή διαχείρισης. Τα κατάλληλα κίνητρα και οι τρέχουσες μετρήσεις πρέπει να είναι καθοριστικής σημασίας οποιουδήποτε προγράμματος διαχείρισης αλλαγών. Περαιτέρω οδηγίες σχετικά με τη Διαχείριση αλλαγών πτυχή των Big Data αναλύεται περαιτέρω στο κεφάλαιο- Λειτουργίες μεγάλων δεδομένων (Big Data Framework 2018).

6.5 Λίστα ελέγχου στρατηγικής Big Data

Στην προηγούμενη ενότητα, καταγράφηκε πώς οι οργανισμοί μπορούν να διαμορφώσουν μια στρατηγική μεγάλων δεδομένων. Στο τέλος, η στρατηγική Big Data θα καταγραφεί σε ένα

έγγραφο έτσι ώστε να μπορεί να εγκριθεί και να μοιραστεί με τον υπόλοιπο οργανισμό (Big Data Framework 2018).

Τα έγγραφα των μεγάλων στρατηγικών δεδομένων πρέπει να περιλαμβάνουν τα ακόλουθα τμήματα (Big Data Framework 2018):

Ιστορικό / Πλαίσιο	Αυτή η ενότητα πρέπει να διαμορφώσει το πλαίσιο που απαιτούσε αρχικά τη Στρατηγική Δεδομένων. Παραδείγματα θα μπορούσαν να είναι: Εταιρική στρατηγική κατεύθυνση, πρωτοβουλία Ψηφιακής Μετατροπής ή συγχωνεύσεις & εξαγορές.
Παράδειγμα υλοποίησης	Ο μοναδικός σκοπός της στρατηγικής δεδομένων είναι να ξεκλειδώσει την αξία της επιχείρησης και αυτό το σημείο θα πρέπει να δομήσει την αξία που θα ξεκλειδωθεί τόσο ποσοτικά όσο και ποιοτικά. Η επιχειρηματική περίπτωση είναι ίσως η πιο δύσκολη, αλλά και απαραίτητη.
Στόχοι	Σε αυτήν την ενότητα προσδιορίζονται συγκεκριμένοι στόχοι σχετικοί με τη στρατηγική δεδομένων και ιδανικά με τρόπο SMART (Specific, Measurable, Agreed upon, Realistic, Time-based).
Εκτέλεση χάρτη πορείας	Αυτή η ενότητα συνδέει τη στρατηγική με την τακτική με έναν χάρτη πορείας για το πώς ηστρατηγική θα εφαρμοστεί σε μια χρονική περίοδο.
Παράγοντες κινδύνου και επιτυχιών	Η στρατηγική θα πρέπει να απευθύνεται άμεσα σε διάφορους παράγοντες κινδύνου και σε παράγοντες επιτυχίας. Ξανά και ξανά, η αλλαγή διαχείρισης αποτελεί είτε κινδύνος ή παράγοντας επιτυχίας, εάν δεν μελετηθεί λεπτομερώς , για αυτό το λόγο βεβαιωθείτε ότι το αντιμετωπίζετε άμεσα σε αυτήν την ενότητα.
Εκτιμήσεις προϋπολογισμού	Ποιο είναι το καλό μιας στρατηγικής αν δεν έχει εκτιμήσεις του προϋπολογισμού. Η εκτίμησης θα πρέπει να είναι ρεαλιστική και όσο το δυνατόν πιο ολοκληρωμένη.

Key Performance Indicators (KPIs) και μετρήσεις	Προκειμένου να διασφαλιστεί ότι η στρατηγική είτε βρίσκεται σε καλό δρόμο είτε πρέπει να προσαρμοστεί, προσδιορίστε τους δείκτες KPI που πρέπει να παρακολουθούνται σε βραχυπρόθεσμη και μακροπρόθεσμη βάση.

7 Λειτουργίες Μεγάλης Κλίμακας Δεδομένων

Αυτό το κεφάλαιο ασχολείται με τις οργανωτικές πτυχές της δημιουργίας μιας πρακτικής Big Data στις επιχειρήσεις. Η ενσωμάτωση μιας πρακτικής Big Data είναι κάτι περισσότερο από την προμήθεια ορισμένων εργαλείων, την εύρεση συνόλων δεδομένων και την πρόσληψη ατόμων με τις κατάλληλες δεξιότητες. Προκειμένου να επιτευχθεί μακροχρόνια αξία από Big Data επενδύσεις, η οργανωτική πτυχή είναι τουλάχιστον εξίσου σημαντική (Big Data Framework 2018).

Οι παλιοί τρόποι εργασίας είναι βαθιά ριζωμένοι, ειδικά εάν υπάρχει μια υποκείμενη δυσπιστία στα Μεγάλα Δεδομένα και τα αναλυτικά στοιχεία. Επομένως, η δημιουργία ενός οργανισμού Big Data είναι εξίσου σημαντική διαχείρισης, καθώς προμηθεύει τις σωστές δεξιότητες, διαδικασίες και τεχνολογία. Τα οφέλη των Μεγάλων Δεδομένων μπορούν να συγκεντρωθούν μόνο αν ευθυγραμμιστεί η θέληση και τα μυαλά των ανθρώπων της οργάνωσης με τη στρατηγική. Όχι μόνο οι άνθρωποι θα πρέπει να αρχίσουν να εργάζονται με διαφορετικό τρόπο, αλλά θα χρειαστεί επίσης να λάβει διαφορετικές αποφάσεις. Στοιχεία που προκύπτουν με ανάλυση μεγάλων δεδομένων θα πρέπει να ενσωματωθούν στην καθημερινή διαδικασία λήψης αποφάσεων προκειμένου να καταστεί η επιχείρηση να οδηγείται από την αξία των δεδομένων (Big Data Framework 2018).

Η ενσωμάτωση μεγάλων δεδομένων στις Επιχειρήσεις αφορά τόσο τη διαχείριση αλλαγών όσο και τα μεγάλα δεδομένα. Οι άνθρωποι προθυμοποιούνται σε μια αλλαγή όταν το καταλαβαίνουν και αισθάνονται μέρος της. Ο σχεδιασμός μιας στρατηγικής BD, άρα και ο ψηφιακός μετασχηματισμός πρέπει επομένως να οδηγείται από το χρήστη και να έχει συμμετοχή από όλα τα επίπεδα της επιχείρησης από την αρχή. Η οργανωτική κουλτούρα, οι οργανωτικές δομές και οι ρόλοι εργασίας έχουν μεγάλο αντίκτυπο στην επιτυχία των πρωτοβουλιών Big Data. Σε αυτό το κεφάλαιο, θα επανεξετάσουμε ορισμένες "καλύτερες πρακτικές" πώς να δημιουργήσετε μια οργάνωση με βάση τα δεδομένα (Big Data Framework 2018).

7.1 Κέντρο Αριστείας Μεγάλων Δεδομένων

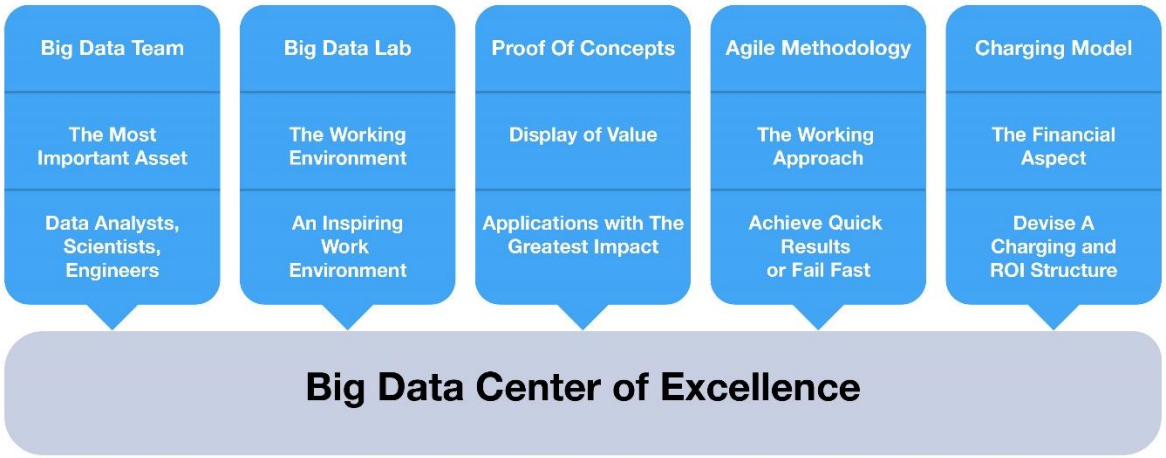
Στις περισσότερες επιχειρήσεις, τα έργα Big Data ξεκινούν όταν ένα στέλεχος γίνεται πεπεισμένο ότι η εταιρεία λείπει από ευκαιρίες στα δεδομένα. Για παράδειγμα, ο CIO μπορεί να έχει ακούσει για τα οφέλη που έχει αποκομίσει ένας ανταγωνιστής σε ένα Big Data project και είναι πρόθυμο να αποκτήσει παρόμοια αποτελέσματα. Σε πολλές οργανώσεις, υπάρχει ένα (μεγάλο) χάσμα μεταξύ της πρώτης έναρξης ενός έργου Big Data και την κλιμάκωση των πλεονεκτημάτων ενός μεγάλου έργου Big Data την επιχείρηση. Προκειμένου να αποκτήσετε μακροπρόθεσμη αξία από τα μεγάλα δεδομένα είναι καθοριστικής σημασίας η δημιουργία ενός κέντρο αριστείας μεγάλων δεδομένων (Big Data Center of Excellence) (Big Data Framework 2018).

Ένα κέντρο αριστείας μεγάλων δεδομένων (BDCoE) είναι μια επιχειρηματική λειτουργία που παίρνει έναν οργανισμό με μηδενική γνώση σε μια πλήρως λειτουργική πρακτική των τεχνολογιών Big Data και διαδικασιών για την επίτευξη ισχυρών επιχειρηματικών αποτελεσμάτων. Με τη χρήση ενός BDCoE ένας οργανισμός αναγνωρίζει νέες τεχνολογίες, μαθαίνει για νέες δεξιότητες και αναπτύσσει τις κατάλληλες διαδικασίες που στη συνέχεια αναπτύσσονται σε όλες τις μονάδες της οργάνωσης (Big Data Framework 2018).

Ένα BDCoE είναι απαραίτητο για την επιτάχυνση της υιοθέτησης μεγάλων δεδομένων από την επιχείρηση με ένα γρήγορο και δομημένο τρόπο. Μειώνει δραστικά τους χρόνους εφαρμογής και ως εκ τούτου το χρονικό πλαίσιο για την ανάπτυξη νέων προϊόντων και υπηρεσιών που βασίζονται σε δεδομένα. Το πιο σημαντικό, διασφαλίζει ότι οι βέλτιστες πρακτικές και οι μεθοδολογίες μοιράζονται μέσω διαφορετικών ομάδων στον οργανισμό. Ένα BDCoE θα πρέπει να είναι μια ζωντανή και εξελισσόμενη οργανωτική λειτουργία που επεκτείνεται και μεγαλώνει όπως οι ανάγκες της επιχείρησης εξελίσσονται (Big Data Framework 2018).

Ένα κεντρικό BDCoE μπορεί να αποτελέσει το θεμέλιο για την καθιέρωση μιας επιχείρησης που βασίζεται σε δεδομένα εκτιμά τα δεδομένα ως το στρατηγικό του πλεονέκτημα. Το BDCoE μπορεί να συνεργαστεί με την επιχείρηση για να προσδιορίσει ποια έργα θα πρέπει να έχουν προτεραιότητα και ποια δεδομένα έχουν στρατηγική σημασία. Ως εκ τούτου, λειτουργεί ως το στρατηγικό αντίγραφο της επιχείρησης να μεταφράσει τρέχουσες και πραγματικές επιχειρηματικές απαιτήσεις σε ζωντανά και ενεργητικά έργα Big Data. Οι άνθρωποι και ο

οργανισμός διαδραματίζει επίσης ζωτικό ρόλο σε αυτή την επιτυχία. Ένα αποτελεσματικό BDCoE αποτελείται από πέντε κύριους πυλώνες που μαζί αποτελούν τη δομή για την απόκτηση αξίας από την κεντρική λειτουργία (Εικόνα 38) (Big Data Framework 2018).



Εικόνα 38: Δομή κέντρου αριστείας μεγάλων δεδομένων (Big Data Framework 2018)

Ομάδες

Το πιο σημαντικό στοιχείο, η ποιότητα των αναλυτών Big Data, Big Data scientists και Μηχανικούς μεγάλων δεδομένων, είναι πρωταρχικής σημασίας για τη δημιουργία επιτυχίας με το Big Data. Στο τέλος, το Big Data είναι τομέας της γνώσης και ότι η γνώση θα προέρχεται από τους ανθρώπους. Οι επαγγελματίες θα πρέπει να είναι πιστοποιημένοι και έμπειροι επαγγελματίες (Big Data Framework 2018).

Εργαστήρια

Τα εργαστήρια μεγάλων δεδομένων αναφέρονται στο περιβάλλον εργασίας του BDCoE. Η προφανής σχέση μεταξύ της επιστήμης των δεδομένων ένα εργαστήριο είναι στο επίκεντρο, αφού το περιβάλλον πρέπει να είναι ένας δημιουργικός χώρος για πειραματισμό και να εκτελέσουν αναλύσεις δεδομένων δοκιμών για να επιτευχθούν τα επιθυμητά αποτελέσματα. Ένα καλά σχεδιασμένο εργαστήριο Big Data περιέχει ανοικτούς χώρους εργασίας που επιτρέπουν την επικοινωνία και τη συνεργασία καθώς και μεμονωμένες δυνατότητες εργασίας. Μια δεύτερη σημαντική απαίτηση για τα μεγάλα εργαστήρια δεδομένων είναι η συμβατότητα του υλικού.

Μεγάλη επεξεργασία δεδομένων. Γενικά, τα μεγάλα εργαστήρια δεδομένων απαιτούν υλικό με αρκετά μεγαλύτερη μνήμη RAM από το συνηθισμένο για την επεξεργασία μεγάλων δεδομένων (Big Data Framework 2018).

Αποδείξεις ιδεών (Proof of Concepts)

Οι αποδείξεις ιδεών (POC) είναι λύσεις βιτρίνας που μπορούν να παρέχονται στις εσωτερικές επιχειρήσεις καθώς και εξωτερικούς πελάτες. Τα POC πρέπει να επιδεικνύουν σαφή απόδοση των επενδύσεων και να παρουσιάζουν με σαφήνεια τις δυνατότητες του BDCoE για την επίτευξη των αποτελεσμάτων (Big Data Framework 2018).

Ευέλικτη μεθοδολογία

Η ευελιξία και η ικανότητα αποτυχίας γρήγορης επιτυχίας είναι απαραίτητα για την επίτευξη του δυναμικού των Μεγάλων Δεδομένων. Μια μεθοδολογία ευέλικτης εργασίας παρέχει τα εργαλεία για την επίτευξη αποτελεσμάτων γρήγορα, συνήθως μέσα σε δύο ή τρεις εβδομάδες. Η δυνατότητα γρήγορης αποτυχίας είναι ένα μεγάλο κλειδί ευκαιρίας– επιχειρηματικοί και τεχνικοί χάρτες πορείας για την παροχή αξίας πρέπει να αλλάζουν πολύ συχνά (Big Data Framework 2018).

Μοντέλα χρέωσης

Στον πυρήνα του BDCoE είναι τα μοντέλα χρέωσης που δικαιολογούν τις μερικές φορές μεγάλες επενδύσεις στους ανθρώπους, τις διαδικασίες και την τεχνολογία του Κέντρου. Προκειμένου να εμφανιστεί μια τιμή πρέπει να σχεδιαστεί μια προσέγγιση για να χρεώνει άλλες υπηρεσίες ή εξωτερικούς πελάτες για υπηρεσίες. Τα μοντέλα ευθύνης μπορούν να σχεδιαστούν με βάση τον αριθμό ή των χρηστών, δεδομένα που έχουν υποστεί επεξεργασία, συχνότητα αναφορών ή συνδρομών. Ένα καλό και αδιαμφισβήτητο μοντέλο χρέωσης θα βοηθήσει σε μεγάλο βαθμό να παρουσιαστεί μια τιμή του BDCoE στην επιχείρηση (Big Data Framework 2018).

7.2 Ρόλοι και αρμοδιότητες των ομάδων

Η πιο σημαντική πτυχή του Big Data είναι οι άνθρωποι που εμπλέκονται. Ενώ υπάρχουν πολλές επιχειρήσεις που σχεδιάζουν να μετατρέψουν τα δεδομένα τους σε αξία, μερικές φορές ξοδεύουν

πάρα πολύ χρόνο για αυτά και όχι αρκετό χρόνο στα άτομα της εξίσωσης. Σε σύντομο χρονικό διάστημα η επιστήμη των δεδομένων είναι πλέον μέρος των επαγγελματικών επιχειρήσεων, έχουν δημιουργηθεί ορισμένοι νέοι ρόλοι που είναι απαραίτητοι στην επιτυχία του Big Data. Κάθε ένας από αυτούς τους ρόλους συμβάλλει στην ομάδα Big Data του BDCoE που εξηγείται στην προηγούμενη ενότητα.

7.2.1 Business Analyst

Ο ρόλος του επιχειρηματικού αναλυτή επικεντρώνεται στη μετατροπή των σχετικών γνώσεων σε πραγματικές επιπτώσεις στις επιχειρήσεις και περιλαμβάνει στοιχεία οργανωτικής αποτελεσματικότητας. Οι θέσεις απασχόλησης αυτής της οικογένειας εργασίας αναφέρουν ευθύνες στον τομέα της αναλυτικής επιχειρηματικής συμβουλευτικής ("λήψη αποφάσεων μέσω της ανάλυσης, συμβουλή στρατηγικών πωλήσεων και μάρκετινγκ, παροχή αναλυτικής υποστήριξης στις επιχειρηματικές πρωτοβουλίες, η αναφορά στρατηγικών πληροφοριών για τους εταίρους") και η διαχείριση έργων επιχειρησιακών αναγκών, να επικοινωνούν αποτελεσματικά την πρόοδο και τα αποτελέσματα, να συνειδητοποιούν τις προτεινόμενες ενέργειες "). Η έρευνά μας δείχνει ότι οι κύριες δεξιότητες για έναν Business Analyst είναι στον τομέα της διαχείρισης έργων και των επιχειρηματικών πρακτικών: Απαιτούνται δεξιότητες διαχείρισης, όπως η αποτελεσματική επικοινωνία, και οικονομία. Οι επιχειρηματικοί αναλυτές είναι η γέφυρα μεταξύ των υπευθύνων λήψης αποφάσεων και των τεχνικών ρόλων: κατά συνέπεια, έχουν επίσης μια πρακτική κατανόηση των αναλυτικών μεθόδων και της σχετικής τεχνολογίας που αφορούν κυρίως τους χρήστες (De Mauro, et al. 2018)

7.2.2 Data scientist

Το επίκεντρο για έναν επιστήμονα δεδομένων είναι τα ίδια τα δεδομένα και οι αναλυτικές μέθοδοι για τη μετατροπή των δεδομένων σε γνώσεις. Οι ρόλοι εργασίας σε αυτήν την οικογένεια περιλαμβάνουν την ευθύνη να «εντοπιστούν τα πρότυπα, να εφαρμοστεί το πλαίσιο και η νοημοσύνη, να εξαχθούν σχετικές πληροφορίες κρυμμένες στις μεγάλες ποσότητες δεδομένων, να σχεδιαστούν και να εφαρμοστούν μοντέλα δεδομένων και στατιστικές μέθοδοι, να ενσωματωθεί η έρευνα και οι βέλτιστες πρακτικές στην αποφυγή προβλημάτων βελτίωση". Οι θέσεις συνήθως αναφέρουν συγκεκριμένες αναλυτικές τεχνικές ("ταξινόμηση, συνεργατικό

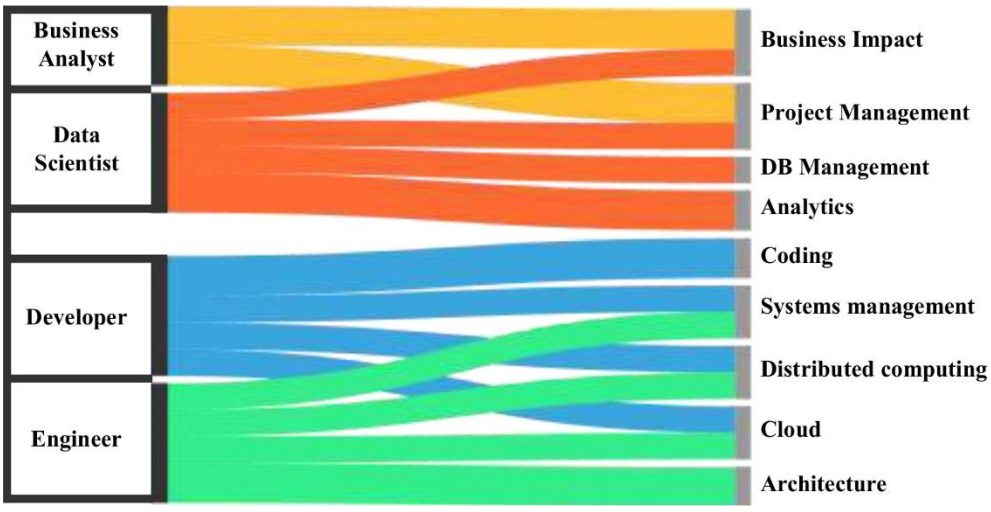
φιλτράρισμα, κανόνες σύνδεσης, νευρωνικά δίκτυα, ευρετικές προσεγγίσεις"), γλώσσες προγραμματισμού ή προγραμματισμού ("Python, SQL, Java, Ruby" Matlab ") που θεωρούνται κρίσιμες για τη θέση. Σύμφωνα με την ανάλυσή μας, το κύριο σύνολο δεξιοτήτων των επιστημόνων δεδομένων είναι σίγουρα αναλυτικά, όπως γνωρίζουν και αξιοποιούν τις Μεγάλες Μεθόδους Δεδομένων όπως και οποιοσδήποτε άλλος στην εταιρεία τους. Οι επιστήμονες δεδομένων πρέπει επίσης να κατανοήσουν το επιχειρηματικό πλαίσιο στο οποίο λειτουργούν και να χρησιμοποιούν τεχνικές διαχείρισης έργων για να αλληλοεπιδρούν αποτελεσματικά με τον υπόλοιπο οργανισμό. Θα πρέπει επίσης να είναι σίγουροι για την πρόσβαση σε εταιρικές αποθήκες δεδομένων και να μπορούν να γράφουν scripts για την αναζήτηση βάσεων δεδομένων (De Mauro, et al. 2018)

7.2.3 Big Data developer

Ο κύριος στόχος των Big Data Developers είναι να σχεδιάσουν, να αναπτύξουν και να τροποποιήσουν το λογισμικό εφαρμογών που βασίζονται σε δεδομένα. Οι περιγραφές εργασίας σε αυτή την οικογένεια αναφέρουν ότι οι υποψήφιοι θα αναπτύξουν πίνακες και λύσεις δεδομένων, θα σχεδιάσουν, θα κατασκευάσουν και θα παραδώσουν νέες εκθέσεις, θα παράγουν πρωτότυπα αποδεικτικά στοιχεία για εφαρμογές πολλαπλών νημάτων, πολλαπλών διακομιστών, θα ενσωματώσουν εφαρμογές τρίτων μερών μέσω διεπαφών προγραμματισμού εφαρμογών ". Οι θέσεις αναφέρονται επίσης στις ευθύνες για τον κύκλο ζωής της εφαρμογής του αναλυτικού προϊόντος, οι οποίες περιλαμβάνουν "σχεδιασμό, ανάπτυξη και υλοποίηση καινοτομιών αυτοματισμού, ανάπτυξη αυτοματοποιημένων σεναρίων δοκιμών, συνεχή υποστήριξη προηγμένων εφαρμογών"). Οι βασικές δεξιότητές τους είναι αναμφισβήτητα δημιουργία κώδικα, αλλά χρειάζονται επίσης μια ισχυρή τεχνογνωσία στη διαχείριση συστημάτων, στο cloud computing και στις κατανεμημένες τεχνολογίες. Οι προγραμματιστές Big Data απαιτούν επίσης μια βασική κατανόηση της διαχείρισης βάσεων δεδομένων, της εταιρικής αρχιτεκτονικής δεδομένων και πρέπει να γνωρίζουν πώς χρησιμοποιούνται τα αναλυτικά στοιχεία στο πλαίσιο της επιχείρησής τους (De Mauro, et al. 2018)

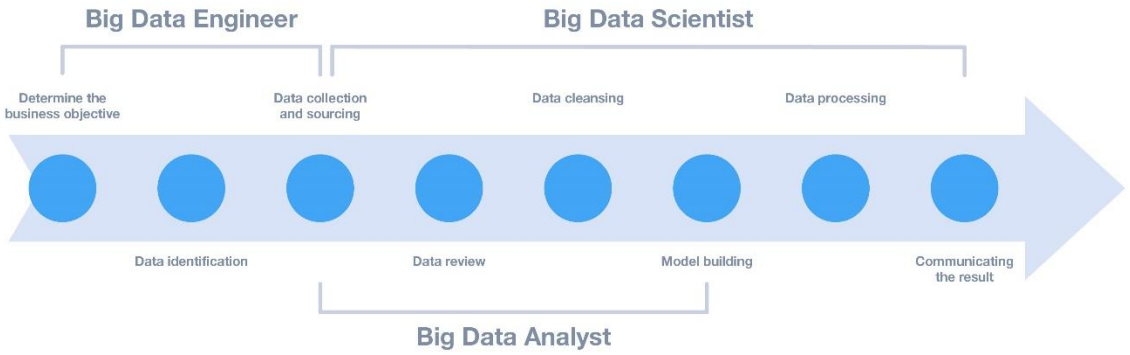
7.2.4 Big Data Engineer

Οι μηχανικοί δεδομένων επικεντρώνονται στην κατασκευή και συντήρηση της υποδομής πλήρους τεχνολογίας, η οποία επιτρέπει την αποθήκευση και επεξεργασία μεγάλων δεδομένων. Οι ρόλοι σε αυτήν την κατηγορία είναι υπεύθυνοι για: "τη διαχείριση της πλατφόρμας διακομιστή επιχειρηματικών αναλύσεων, την υποστήριξη όλων των διαδικασιών για τη φόρτωση και διαχείριση του χώρου αποθήκευσης δεδομένων και την ενσωμάτωση νέων πηγών δεδομένων, τη διασφάλιση της χωρητικότητας, των αντιγράφων ασφαλείας, Συνήθως χρησιμοποιούν συγκεκριμένες αναφορές για τις τεχνολογίες όπως Hadoop, Cassandra, MongoDB, MySQL, Hana, Ceph, GlusterFS, Azure, Amazon Web Services. Το βασικό σύνολο δεξιοτήτων για τους Big Data Engineers σχετίζεται με την αρχιτεκτονική δεδομένων και περιλαμβάνει τις ικανότητες που απαιτούνται για την οικοδόμηση και τη διαχείριση του οικοσυστήματος μεγάλων δεδομένων με βιώσιμο τρόπο. Αυτό περιλαμβάνει την ικανότητα να φροντίζει για την πολυπλοκότητα που συνδέεται με τη διαχείριση των συστημάτων (από την ασφάλεια των πληροφοριών μέχρι την παρακολούθηση της απόδοσης), τον υπολογισμό του νέφους και την κατανεμημένη επεξεργασία. Οι θέσεις εργασίας αναφέρονται επίσης στην ικανότητα αλληλεπίδρασης με βάσεις δεδομένων, στην υιοθέτηση διαδικασιών διαχείρισης έργων και στη γενική κατανόηση του τρόπου με τον οποίο τα δεδομένα μπορούν να υποστηρίξουν τη στρατηγική της εταιρείας (De Mauro, et al. 2018).



Εικόνα 39: Αλλουβιακό διάγραμμα των οικογενειών θέσεων εργασίας Big Data εναντίον ομάδων δεξιοτήτων μεγάλων δεδομένων (De Mauro, και συν. 2018).

Ο Big Data Analyst επικεντρώνεται στην κίνηση και ερμηνεία των δεδομένων, συνήθως με εστίαση στο παρελθόν και το παρόν. Εναλλακτικά, ο Data Scientist μπορεί να είναι κυρίως υπεύθυνος για τη συνοπτική παρουσίαση δεδομένων κατά τρόπο που να παρέχει πρόβλεψη ή μια διορατικότητα στο μέλλον σχετικά με τα πρότυπα που εντοπίστηκαν από προηγούμενα και τρέχοντα δεδομένα. Ο Μηχανικός μεγάλων δεδομένων, τέλος ασχολείται με τη διασφάλιση της υποκείμενης υποδομής Big Data, πριν ξεκινήσει η επεξεργασία. (Big Data Framework 2018).



Εικόνα 40: Big data και ρόλοι εργασίας (Big Data Framework 2018)

7.3 Παράγοντες επιτυχίας

Τα έργα Big Data έχουν ξεκινήσει εδώ και περισσότερο από μια δεκαετία. Ωστόσο, αυτό δεν σημαίνει ότι όλες οι πρωτοβουλίες Big Data έχουν μεγάλη επιτυχία. Διάφορες μελέτες δείχνουν ότι αν και οι επενδύσεις στο Big Data αυξάνονται τα τελευταία χρόνια, πολλές επιχειρήσεις εξακολουθούν να προσπαθούν να αποδείξουν την απόδοση της επένδυσής τους λόγω κακών εφαρμογών. Για να μάθουν από προηγούμενα λάθη, έχουν εντοπιστεί ορισμένοι παράγοντες επιτυχίας που μπορούν να δώσουν μια καλή αρχή στις επιχειρήσεις (Big Data Framework 2018).

1) Δημιουργήστε ένα όραμα για το πώς να δημιουργήσετε αξία: Το πρώτο ορόσημο είναι να αποκτήσετε μια σαφή εικόνα του τι προσπαθεί να επιτύχει ο οργανισμός σας με την εφαρμογή Big Data. Το γεγονός ότι ο οργανισμός συλλέγει terabyte δεδομένων σε καθημερινή βάση δεν έχει νόημα αν δεν υπάρχει σαφής εικόνα με ένα σχέδιο δράσης ως προς το τι θέλει να επιτύχει ο οργανισμός με αυτά τα δεδομένα (Big Data Framework 2018).

2) Για να πετύχετε με τα Big Data, ξεκινήστε τα μικρά: Η δημιουργία δυνατοτήτων από τα μεγάλα δεδομένα απαιτεί χρόνο. Μια καλή αρχική στην επένδυση δεν πρόκειται να παράγει άμεσα αποτελέσματα. Επομένως, συνιστάται μια μικρή αρχή με ελεγχόμενη ανάπτυξη. Κατ 'αρχάς, ορίστε μερικά σχετικά απλά έργα μεγάλων δεδομένων που δεν θα πάρουν πολύ χρόνο ή δεδομένα για να τρέξουν. Για παράδειγμα, ένας διαδικτυακός λιανοπωλητής μπορεί να ξεκινήσει εντοπίζοντας τα προϊόντα που ο κάθε πελάτης είδε έτσι ώστε η εταιρεία να μπορεί να στείλει μια προσφορά παρακολούθησης εάν δεν αγοράσει. Μερικά διαισθητικά παραδείγματα όπως αυτό επιτρέπει στον οργανισμό να δει τι μπορούν να κάνουν τα δεδομένα. Το πιο σημαντικό, αυτή η προσέγγιση αποδίδει αποτελέσματα που είναι εύκολο να δοκιμαστούν για να δουν τι είδους επιστροφές παρέχουν τα μεγάλα δεδομένα (Big Data Framework 2018).
3) Καθιέρωση διαδικασιών μεγάλων δεδομένων από την αρχή: Καθιστά σαφές από την αρχή ποιος είναι υπεύθυνος για το τι. Σχεδιασμός αποτελεσματικής διαχείρισης δεδομένων και διαδικασιών, προσδιορίζοντας ποιος είναι υπεύθυνος για τον ορισμό, τη δημιουργία, την επαλήθευση, επιμέλεια και επικύρωση της επιχείρησης, της πληροφορικής ή του BDCoE (Big Data Framework 2018).
4) Δημιουργία ενός μεγάλου κέντρου δεδομένων αριστείας: Ένα κεντρικό BDCoE παρέχει μια ομοιόμορφη τεχνογνωσία σχετικά με τις πρακτικές και τις τεχνολογίες μεγάλων δεδομένων Η BDCoE μπορεί να συνεργαστεί με την επιχείρηση για να προσδιορίσει ποια έργα θα πρέπει να έχουν προτεραιότητα και ποια δεδομένα έχουν στρατηγική σημασία. Ως εκ τούτου, λειτουργεί ως το στρατηγικό αντίγραφο της επιχείρησης για να μεταφράσει τις τρέχουσες και τις πραγματικές απαιτήσεις των επιχειρήσεων σε ενεργητικά πρότζεκτ (Big Data Framework 2018).
5) Αξιολογήστε την ετοιμότητά σας για Μεγάλα Δεδομένα: Για να προσδιορίσετε τα πιθανά κενά και ενδεχομένως να προκύψουν κίνδυνοι, να διενεργηθεί εκτίμηση της ετοιμότητας μεγάλων δεδομένων. Αυτή είναι η εκτίμηση της ετοιμότητας του περιβάλλοντος πληροφορικής σας και τα σετ δεξιοτήτων που έχουν τα μέλη της για την υλοποίησή Big Data στον οργανισμό και να ενδυναμώσει τα μέλη της υπάρχουσας ομάδας σας ως επιστήμονες δεδομένων σε ολόκληρο τον οργανισμό σας για να χρησιμοποιήσουν τη δύναμη του Big Data (Big Data Framework 2018).

6) Δημιουργήστε ένα τρέχον πρόγραμμα εκπαίδευσης Big Data: Οι γνώσεις και οι δεξιότητες είναι οι πλέον σημαντικό κλειδί για την επιτυχία, αλλά ένα από τα πιο δύσκολα στοιχεία που πρέπει να αποκτηθούν. Εξιδεικευμένο προσωπικό δεδομένων δεν είναι εύκολο να βρεθεί και ακόμα και όταν δημιουργηθεί μια ομάδα, αυτοί χρειάζονται συνεχείς ενημερώσεις σχετικά με τις γνώσεις τους προκειμένου να αναπτυχθούν περαιτέρω. Ρύθμιση ενός σε εξέλιξη εκπαιδευτικού προγράμματος Big Data θα αυξήσει την ικανότητα του οργανισμού και ενσωματώνει μια κουλτούρα συνεχούς μάθησης (Big Data Framework 2018).

8 Τεχνητή Νοημοσύνη

Ο Άλαν Τούρινγκ ήταν επιστήμονας υπολογιστών που ανέπτυξε μία από τις πρώτες θεωρίες για την Τεχνητή Νοημοσύνη. Ένας υπολογιστής κατέχει νοημοσύνη αν κάποιος άνθρωπος ανακριτής καθορίσει ποιος παίκτης – Α ή Β - είναι υπολογιστής και ποιος είναι άνθρωπος, αλλά ο ανακριτής δεν είναι σε θέση να καθορίσει τη διαφορά. Ο ανακριτής περιορίζεται στη χρήση γραπτών ερωτήσεων. Αυτός ο επιχειρησιακός ορισμός της νοημοσύνης είναι τώρα γνωστή ως δοκιμή Turing (Big Data Framework 2018).

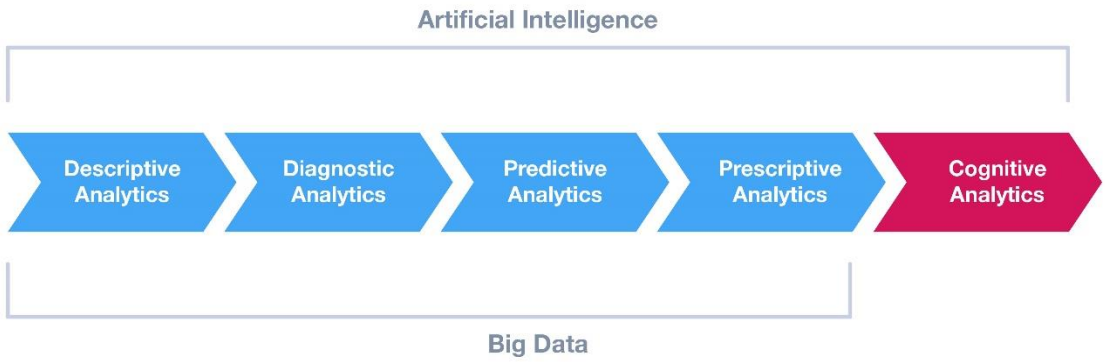
Πώς είναι δυνατόν ένας υπολογιστής να περάσει τη δοκιμή Turing; Οι υποκείμενες τεχνικές που πρέπει να έχει μια μηχανή για να περάσει, ο υπολογιστής θα απαιτήσει τουλάχιστον τις ακόλουθες δυνατότητες (Big Data Framework 2018):

- Φυσική επεξεργασία γλωσσών−ο υπολογιστής πρέπει να είναι σε θέση να μεταφράσει αγγλικά μέσα προκειμένου να επικοινωνούν αποτελεσματικά.
- Αντιπροσώπευση γνώσης−ο υπολογιστής πρέπει να αποθηκεύει τα δεδομένα εισόδου και να ανακτά τα ίδια δεδομένα σε μεταγενέστερο χρόνο.
- Αυτοματοποιημένος συλλογισμός: ο υπολογιστής πρέπει να μπορεί να χρησιμοποιεί τις αποθηκευμένες πληροφορίες να απαντά σε ερωτήσεις και να εξάγει συμπεράσματα. Προκειμένου να επιτευχθεί αυτό, ο υπολογιστής θα πρέπει να εφαρμόσει έναν αλγόριθμο.
- Μηχανική Μάθηση: ο υπολογιστής πρέπει να προσαρμόσει την απόκριση του σε προηγούμενα δεδομένα εισόδου προκειμένου να διατυπωθούν νέες απαντήσεις.

Κάθε ένας από αυτούς τους τέσσερις κλάδους είναι αναπόσπαστο μέρος του τομέα της Τεχνητής Νοημοσύνης, και είναι επομένως, εύκολο να προσδιοριστεί η σχέση μεταξύ μεγάλων δεδομένων και τεχνητής νοημοσύνης. Οι ίδιες στατιστικές τεχνικές και αλγόριθμοι (που αναφέρονται στο κεφάλαιο 4) που εφαρμόζονται για τα μεγάλα δεδομένα χρησιμοποιούνται στη μελέτη της Τεχνητής Νοημοσύνης (Big Data Framework 2018).

Η κύρια διαφορά μεταξύ μεγάλων δεδομένων και τεχνητής νοημοσύνης είναι ότι, όπου η ανάλυση μεγάλων δεδομένων και η αναλυτική σταματούν κυρίως με προγνωστικά και κανονιστικά αναλυτικά στοιχεία, η Τεχνητή Νοημοσύνη συνεχίζει ένα βήμα παραπέρα. Η Τεχνητή Νοημοσύνη

έχει ως στόχο να συμπεριλάβει τις γνωστικές τεχνικές επιστήμης για να αναδιαμορφώνει τον ανθρώπινο εγκέφαλο. Ωστόσο, υπάρχει μεγάλη αλληλοεπικάλυψη μεταξύ Big Data και AI και οι δύο τομείς συνεχίζουν να βελτιώνουν ένας τον άλλον (Big Data Framework 2018).



Εικόνα 41: Μεγάλα Δεδομένα και Τεχνητή Νοημοσύνη (Big Data Framework 2018)

8.1 Γνωστική Αναλυτική

Προκειμένου να διατηρηθεί μια πρακτική άποψη για την Τεχνητή Νοημοσύνη σε ένα επιχειρηματικό πλαίσιο, αυτός ο οδηγός θα να συνεργαστεί με τον επιχειρησιακό ορισμό της ΤΝ και θα επικεντρωθεί στις γνωστικές αναλύσεις ως επέκταση των τεχνικών αναλύσεων Big Data που συζητήθηκαν προηγουμένως σε ολόκληρη την διατριβή. Οι γνωστικές αναλύσεις (cognitive analytics) είναι ο σχεδιασμός και η ανάπτυξη αλγορίθμων που μπορούν να αντανακλούν την ανθρώπινη λήψη αποφάσεων, με βάση το περιβάλλον και τα εξατομικευμένα χαρακτηριστικά (Big Data Framework 2018).

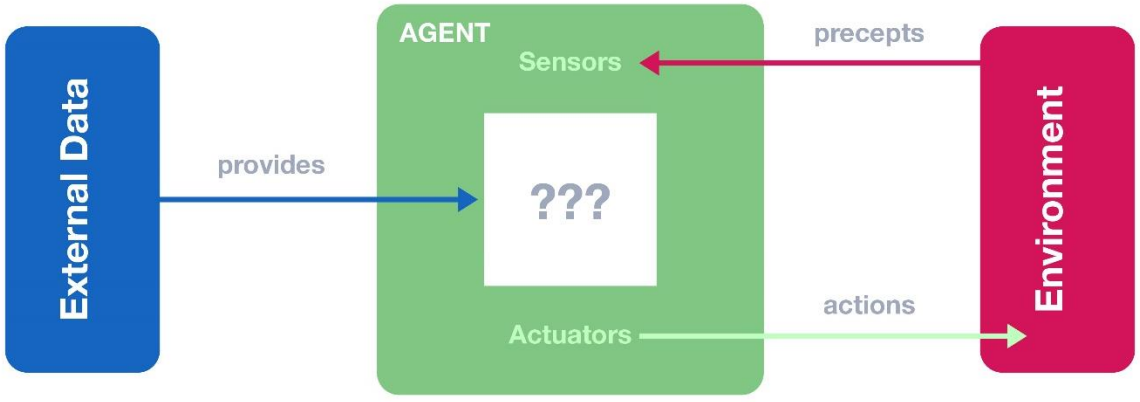
Οι γνωστικές αναλύσεις διαφοροποιούνται από άλλες μορφές αναλύσεων λόγω δύο κύριων λόγων (Big Data Framework 2018):

1) Οι γνωστικές αναλύσεις παίρνουν αποφάσεις με βάση το αντιληπτό περιβάλλον. Το περιβάλλον μπορεί να είναι διαφορετικό σε οποιαδήποτε δεδομένη στιγμή της ημέρας και πρέπει να εξεταστεί με βάση την ειδική κατάσταση. Για την ανίχνευση του αντιληπτού περιβάλλοντος, τα δεδομένα εισόδου πρέπει να καταγράφονται μέσω αισθητήρων.
2) Οι γνωστικές αναλύσεις λαμβάνουν αποφάσεις βασισμένες σε εξατομικευμένα χαρακτηριστικά. Οι αλγόριθμοι μαθαίνουν από τον συγκεκριμένο χρήστη, προκειμένου να

προσαρμόσει τη λήψη αποφάσεων σε αυτό το συγκεκριμένο άτομο. Στο παράδειγμα των θερμοστατών μάθησης, η θερμοκρασία σε δύο διαφορετικά σπίτια θα έχουν διαφορετικά μοντέλα θέρμανσης βάσει των χαρακτηριστικών των χρηστών.

Προκειμένου να επιτευχθούν αυτά τα δύο βασικά χαρακτηριστικά της Τεχνητής Νοημοσύνης, οι γνωστικές αναλυτικές χρειάζονται την ανάπτυξη λογικών πρακτόρων. Ένας πράκτορας είναι ακριβώς κάτι που ενεργεί (από τη λατινική "agere", που σημαίνει "να κάνεις"). Ένας ορθολογικός πράκτορας είναι αυτός που ενεργεί έτσι ώστε να επιτύχει το καλύτερο αποτέλεσμα ή, όπου υπάρχει αβεβαιότητα, το καλύτερο αναμενόμενο αποτέλεσμα. Ένας ορθολογικός πράκτορας προσπαθεί να μιμείται τις ορθολογικές αποφάσεις που λαμβάνουν οι άνθρωποι. Οι γνωστικές αναλύσεις επομένως, επικεντρώνεται στο σχεδιασμό και την ανάπτυξη λογικών παραγόντων (Big Data Framework 2018).

Όπως φαίνεται στην Εικόνα 43, ένας πράκτορας αντιλαμβάνεται δεδομένα από ένα συγκεκριμένο περιβάλλον (κυκλοφορία σε αυτόματα αυτοκίνητα ή ομιλία στην περίπτωση του Siri) μέσω ενός ή περισσοτέρων αισθητήρων. Ο πράκτορας στη συνέχεια επεξεργάζεται αυτά τα δεδομένα (με κάποιο είδος αλγορίθμου) και στη συνέχεια ακολουθεί μια συγκεκριμένη πορεία δράσης. Η απόφαση είναι αυτόνομη, και παρόμοια με την απόφαση ενός ανθρώπου που θα αντιμετώπιζαν παρόμοιες συνθήκες (Big Data Framework 2018).



Εικόνα 42: Λογικός πράκτορας (Big Data Framework 2018)

Η εξωτερική πηγή δεδομένων (Μεγάλα Δεδομένα) παρέχει δεδομένα εισόδου ή αναφοράς στον λογικό πράκτορα με τη σειρά να κάνει τους υπολογισμούς του. Δεδομένου ότι αυτές οι εξωτερικές πηγές δεδομένων μπορούν επίσης να ενημερώνονται σε πραγματικό χρόνο (για παράδειγμα, οι προβλέψεις καιρού), οι αποφάσεις ενδέχεται να διαφέρουν ανά άτομο. Αν εξεταστεί ο λογικός νοήμον πράκτορα για ένα αυτοκίνητο με αυτο-οδήγηση για παράδειγμα. Το περιβάλλον παρέχει χιλιάδες σήματα εισόδου στον λογικό πράκτορα. Αυτά τα σήματα εισόδου θα μπορούσαν να είναι το χρώμα των φαναριών, η απόσταση από το αντικείμενο μπροστά, ταχύτητα του αυτοκινήτου πίσω, κλπ. Ο πράκτορας συνδυάζει αυτά τα δεδομένα εισόδου με εξωτερικά δεδομένα από άλλες πηγές. Αυτά τα εξωτερικά δεδομένα θα μπορούσαν να είναι τα οχήματα, το ιστορικό οδήγησης (εξατομικευμένα δεδομένα) ή δεδομένα από εξωτερικές βάσεις δεδομένων, όπως ο καιρός πρόβλεψη (εάν αναμένεται βροχή από το χιόνι, η ταχύτητα του αυτοκινήτου μειώνεται). Με βάση τα δεδομένα εισόδου των αισθητήρων και τα εξωτερικά δεδομένα, ο λογικός πράκτορας λαμβάνει μια απόφαση. Σε αυτό το παράδειγμα, αυτό σημαίνει το αυτοκίνητο να αυξάνει ή μειώνει την ταχύτητα, τα φρένα, να αλλάζει ταχύτητες ή αλλάζει διαφορετική κατεύθυνση. Η απόφαση του ορθολογικού πράκτορα θα έχει άμεσο αντίκτυπο στο περιβάλλον γιατί πρέπει να εξασφαλίσει ότι όλοι μπορούν να οδηγούν με ασφάλεια (Big Data Framework 2018).

8.2 Δυνατότητες της Τεχνητής Νοημοσύνης

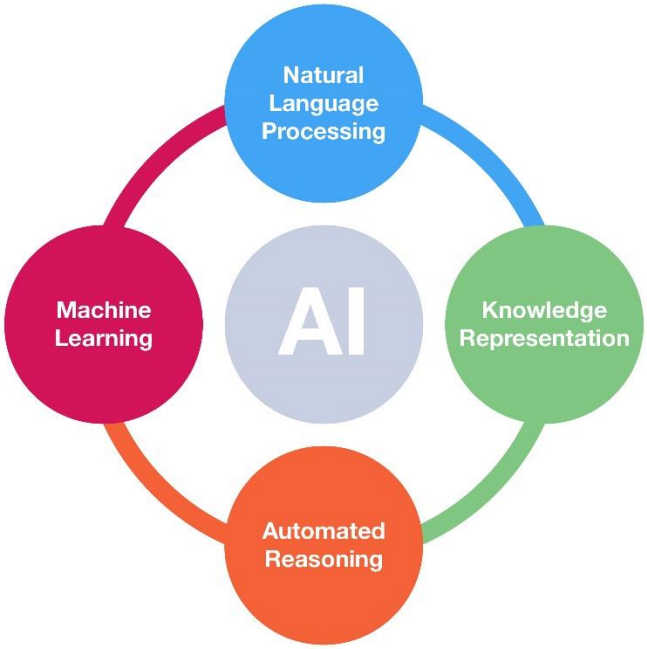
Η Τεχνητή Νοημοσύνη συνδυάζει τις δυνατότητες ενός μεγάλου αριθμού δυνατοτήτων μέσω του σχεδιασμού και την εφαρμογή ορθολογικών παραγόντων. Η Εικόνα παρουσιάζει τις τέσσερις βασικές δυνατότητες της Τεχνητής Νοημοσύνης, οι οποίες και αναλύονται παρακάτω.

8.2.1 Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing) είναι ο τομέας που καθορίζει τις αλληλεπιδράσεις μεταξύ των υπολογιστών και (φυσικής) ανθρώπινης γλώσσας, έτσι ώστε οι άνθρωποι να μπορούν να αλληλοεπιδρούν με τον υπολογιστή. Οι προκλήσεις στην επεξεργασία της φυσικής γλώσσας περιλαμβάνουν την αναγνώριση ομιλίας, την κατανόηση των προτάσεων και φραγμών γλώσσας ή διαλέκτου (Big Data Framework 2018).

Οι περισσότερες λύσεις τεχνητής νοημοσύνης θα χρειαστεί να χρησιμοποιήσουν κάποια μορφή NLP για να διευκολύνουν τη μεταφορά δεδομένων από το περιβάλλον στον ορθολογικό

παράγοντα. Στο παράδειγμα της αναγνώρισης ομιλίας στις λειτουργίες του τηλεφωνικού κέντρου, το NLP πρέπει να ανιχνεύσει τη γλώσσα του ατόμου, να ανιχνεύσει τις ακολουθίες της λέξης και να ανιχνεύσει ενδεχομένως τα συναισθήματα με τον τρόπο που διαδίδεται το μήνυμα. Λόγω του μεγέθους των διαφορετικών συνδυασμών που είναι δυνατοί στο NLP, η ανάπτυξη των εφαρμογών NLP βασίζονται σε μεγάλα περιβάλλοντα μεγάλων δεδομένων. Το αγγλικό λεξικό αποτελείται από περίπου 170.000 λέξεις και ο αριθμός των κινεζικών λέξεων είναι περίπου 370.000. ο αριθμός των συνδυασμών που μπορούν να γίνουν με αυτά είναι σχεδόν άπειρος. Η ποσότητα των δεδομένων απαραίτητη για την αποθήκευση και την επεξεργασία αυτών των συνδυασμών αποτελούνται από πολλαπλά zettabytes (Big Data Framework 2018).



Εικόνα 43: Οι τέσσερις βασικές δυνατότητες της Τεχνητής Νοημοσύνης (Big Data Framework 2018)

Οι βασικές προκλήσεις στο NLP έχουν να κάνουν με τη σύνταξη και τη σημασιολογία. Η σύνταξη είναι ο τρόπος με τον οποίο είναι οι προτάσεις και οι συνδυασμοί λέξεων δίνουν νόημα σε μια πρόταση. Από την άλλη πλευρά, η σημασιολογία εξετάζει το γεγονός ότι (συνδυασμοί) λέξεων μπορούν να έχουν διαφορετική σημασία όταν χρησιμοποιούνται σε διαφορετικά πλαίσια (Big Data Framework 2018).

8.2.2 Αναπαράσταση της γνώσης

Η αναπαράσταση της γνώσης είναι το πεδίο της Τεχνητής Νοημοσύνης αφιερωμένο στην παρουσίαση πληροφοριών για τον κόσμο σε μια μορφή που ένα σύστημα υπολογιστή μπορεί να χρησιμοποιήσει για να λύσει πολύπλοκα καθήκοντα. Η αναπαράσταση της γνώσης ενσωματώνει ευρήματα από την ψυχολογία για το πώς οι άνθρωποι λύνουν προβλήματα και να εκπροσωπήσουν τη γνώση για να σχεδιάσουν λογικές δηλώσεις που κάνουν πολύπλοκα συστήματα ευκολότερα σχεδιασμένα και κατασκευασμένα. Ως εκ τούτου, βασίζεται σε μεγάλο βαθμό στην εφαρμογή της λογικής για να διαμορφώσουν το σκεπτικό (Big Data Framework 2018).

Δεδομένου ότι τα περισσότερα σύνολα δεδομένων είναι ετερογενή όσον αφορά τον τύπο, τη δομή και την προσβασιμότητά τους, αυτά θέτουν προκλήσεις για τα συστήματα ηλεκτρονικών υπολογιστών να ερμηνεύουν με συστηματικό τρόπο. Η αναπαράσταση γνώσης συμβάλλει στον εντοπισμό των αποθηκευμένων δεδομένων και του τρόπου με τον οποίο μπορούν να ανακτηθούν αργότερα όταν απαιτείται επεξεργασία. Συγκεκριμένα, στοχεύει στην οικοδόμηση συστημάτων που γνωρίζουν τον κόσμο τους και είναι σε θέση να ενεργούν με ενημερωμένο τρόπο, όπως κάνουν οι άνθρωποι. Ένα σημαντικό μέρος αυτών των συστημάτων είναι ότι η γνώση αντιπροσωπεύεται συμβολικά και ότι οι διαδικασίες λογικής είναι σε θέση να εξαχθούν τις συνέπειες μιας τέτοιας γνώσης σαν νέες συμβολικές αναπαραστάσεις (Big Data Framework 2018).

8.2.3 Αυτοματοποιημένος συλλογισμός

Η αυτοματοποιημένη συλλογιστική πορεία στην Τεχνητή Νοημοσύνη είναι η ικανότητα γνώσης που αφορά τον εαυτό της με την κατανόηση των δυνατοτήτων συλλογιστικής σε συστήματα υπολογιστών. Ο στόχος της αυτοματοποιημένου συλλογισμού είναι να σχεδιάσουμε συστήματα ηλεκτρονικών υπολογιστών που μπορούν να αιτιολογούν εντελώς αυτόματα (χωρίς ανθρώπινη συμμετοχή). Αυτοματοποιημένη συλλογιστική είναι απαραίτητη στο σχεδιασμό οποιουδήποτε τεχνητού Συστήματος Πληροφοριών για να μιμηθεί τη διαδικασία που συμβαίνει στον ανθρώπινο εγκέφαλο. Δεδομένης των συνθηκών που μπορούν να παρατηρηθούν ή να αισθανθούν, το σύστημα πληροφορικής πρέπει να φτάσει στο καλύτερο πιθανό συμπέρασμα ακολουθώντας μια (αυτοματοποιημένη) διαδικασία σκέψης (Big Data Framework 2018).

8.2.4 Μηχανική μάθηση

Η μηχανική μάθηση, όπως παρουσιάζεται νωρίτερα, είναι μία από τις θεμελιώδεις δυνατότητες που απαιτούνται τόσο για την ανάλυση Big Data όσο και για την Τεχνητή Νοημοσύνη. Ο στόχος της μηχανικής μάθησης είναι να σχεδιαστεί ένα σύστημα που βελτιώνεται και βελτιώνεται με την πάροδο του χρόνου. Ακριβώς όπως απομνημονεύουν οι άνθρωποι πληροφορίες ή σχέσεις όταν τους παρουσιάζονται, έτσι μπορούν να μάθουν τα συστήματα υπολογιστών από προηγούμενες αλληλεπιδράσεις. Κοινή αλγόριθμοι μηχανικής μάθησης (ταξινόμηση, παλινδρόμηση και συσσωμάτωση) συζητήθηκαν στο κεφάλαιο 4 (Big Data Framework 2018).

8.3 Deep Learning

Η βαθιά εκμάθηση είναι ένας τομέας ενδιαφέροντος που έχει αναπτυχθεί πολύ γρήγορα την τελευταία δεκαετία εξαιτίας του αυξημένου ενδιαφέρον για την Τεχνητή Νοημοσύνη. Εν ολίγοις, η Deep Learning είναι μια προηγμένη μορφή μηχανικής μάθησης που χρησιμοποιεί τις αναπαραστάσεις δεδομένων μάθησης (σε αντίθεση με τους αλγορίθμους) που χρησιμοποιείται ειδικά για την Τεχνητή Νοημοσύνη (Big Data Framework 2018).

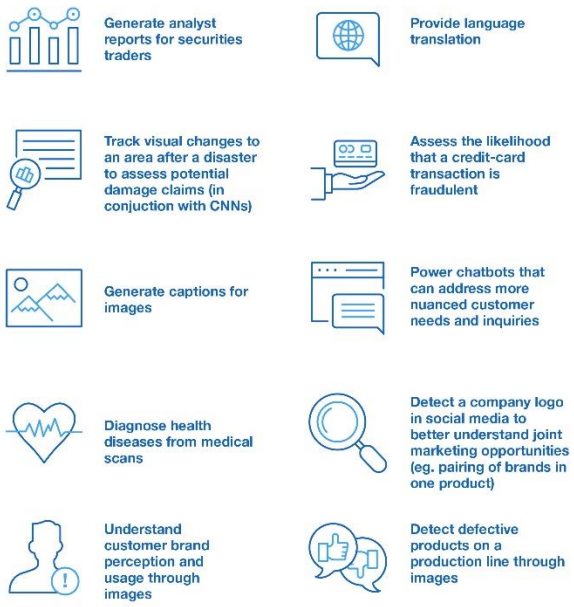
Η βαθιά εκμάθηση είναι ένας τύπος μηχανικής μάθησης που μπορεί να επεξεργαστεί ένα ευρύτερο φάσμα πηγών δεδομένων, απαιτεί λιγότερη επεξεργασία δεδομένων από τους ανθρώπους και μπορεί συχνά να παράγει πιο ακριβή αποτελέσματα από τις παραδοσιακές προσεγγίσεις εκμάθησης μηχανών (αν και απαιτεί μεγαλύτερη ποσότητα δεδομένων). Στη βαθιά μάθηση, αλληλοσυνδεόμενα στρώματα λογισμικού με βάση τον υπολογιστή, γνωστή ως "νευρώνες", σε ένα νευρωνικό δίκτυο. Το δίκτυο μπορεί να καταναλώνει τεράστια ποσά δεδομένων εισόδου και να τα επεξεργάζεται μέσω πολλαπλών στρώσεων που μαθαίνουν όλο και πιο πολύπλοκα χαρακτηριστικά των δεδομένων σε κάθε στρώμα. Το δίκτυο μπορεί στη συνέχεια να αποφασίσει για τα δεδομένα, να μάθει αν η απόφασή του είναι σωστή, και να χρησιμοποιήσει ό, τι έχει μάθει για να κάνει διαπιστώσεις σχετικά με νέα δεδομένα. Για παράδειγμα, μια φορά μαθαίνει τι μοιάζει με ένα αντικείμενο, μπορεί να αναγνωρίσει το αντικείμενο σε μια νέα εικόνα (Big Data Framework 2018).

Οι συμβατικές τεχνικές εκμάθησης μηχανών που χρησιμοποιούνται στην ανάλυση Big Data είναι περιορισμένες στην ικανότητά τους να επεξεργάζονται δεδομένα στην ακατέργαστη μορφή τους.

Στο παράδειγμα των συστημάτων αναγνώρισης προσώπου, τα ανεπεξέργαστα δεδομένα (δηλ. φωτογραφίες ατόμων) πρέπει να μετασχηματιστούν σε φορείς χαρακτηριστικών που μπορούν να συγκριθούν με άλλους φορείς στο σύστημα. Εάν υπάρχει αντιστοιχία μεταξύ δύο παραγόντων χαρακτηριστικών, το άτομο έχει αναγνωριστεί. "Για να μεταφράσει αυτό το« αρχείο δεδομένων» σε ένα χρήσιμο διάνυσμα, απαιτείται προσεκτική μηχανική και εκτεταμένη τεχνογνωσία στον τομέα (Big Data Framework 2018).

Η Deep Learning λύνει αυτό το πρόβλημα χρησιμοποιώντας την εκμάθηση δεδομένων μάθησης. Αυτό επιτρέπει σε μια μηχανή να τροφοδοτείται με δεδομένα και να ανακαλύπτει αυτόματα τις αναπαραστάσεις που απαιτούνται για ανίχνευση ή ταξινόμηση. Για να επιτευχθεί αυτό, η Deep Learning κατατάσσει τα ακατέργαστα δεδομένα σε έναν αριθμό (χρησιμοποιώντας τον αλγόριθμο backpropagation) και στη συνέχεια συγκρίνει αυτά τα επίπεδα. Χρησιμοποιώντας αυτή την τεχνική, γίνεται πιο αποτελεσματική η κατάτμηση μεγάλων συνόλων δεδομένων σε δομημένες πληροφορίες που μπορούν να αναλυθούν (Big Data Framework 2018).

Η βαθιά εκμάθηση χρησιμοποιείται κατά κύριο λόγο για την επεξεργασία εικόνων, βίντεο, ομιλίας και ήχου. Παραδείγματα των καταστάσεων στις οποίες χρησιμοποιείται η βαθιά εκμάθηση απεικονίζονται στην Εικόνα 44



Εικόνα 44: Παραδείγματα Deep Learning (Big Data Framework 2018)

9 Μελέτη Περίπτωσης: Πρόβλεψη σφαλμάτων βασισμένη στην ανάλυση μεγάλης κλίμακας δεδομένων για τον προγραμματισμό της παραγωγής

9.1 Σκοπός μελέτης

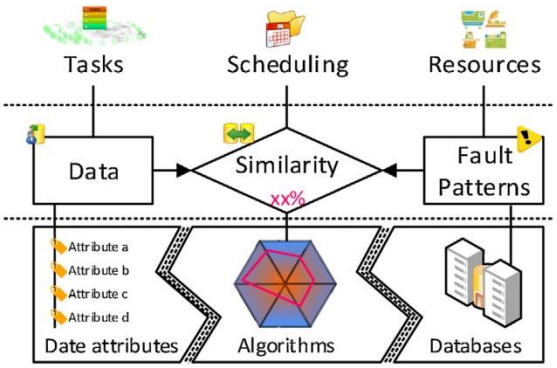
Η σύγχρονη μεταποιητική βιομηχανία χαρακτηρίζεται από υψηλή ποιότητα και ταχεία παράδοση για τα τρέχοντα συστήματα παραγωγής με στόχο την καλή απόδοση, υψηλή σταθερότητα και υψηλή επαναληψιμότητα. Μεταξύ πολλών άλλων, η προβλεψιμότητα, προσφέρει καλές δυνατότητες για μελλοντική μεταποίηση, όπως η έξυπνη κατασκευή και ευφυή παραγωγή. Δύο πιθανά σενάρια υπάρχουν σχετικά με την προβλεψιμότητα: (1) πριν από τη λήψη αποφάσεων του προγραμματισμού παραγωγής, μια αποτελεσματική πρόβλεψη πιθανών βλαβών που μπορεί να αποφευχθεί, η καθυστέρηση εργασίας ή άλλη περιττή απώλεια. και (2) κατά τη διάρκεια της μηχανουργικής κατεργασίας, όπου τα πρότυπα των σημάτων παρακολούθησης σε πραγματικό χρόνο μπορεί να δείχνουν πιθανά ελαττώματα, οδηγώντας στην αναδιάταξη των υπόλοιπων εργασιών. Επομένως, η πρόβλεψη σφάλματος στη γραμμή παραγωγής παίζει σημαντικό ρόλο στη βελτιστοποίηση την κατανομή των πόρων, τη μείωση του κόστους κατασκευής και τη βελτίωση της παραγωγικής αποδοτικότητας (Ji and Wang 2017)

Σήμερα, πολύ δυναμικές και ολοκληρωμένες μέθοδοι είναι η αφορούν στον προγραμματισμό, γεγονός που καθιστά δυνατή την ταχεία αναδιάρθρωση όταν υπάρχουν σφάλματα κατά τη μηχανική κατεργασία. Ωστόσο, τα σημερινά συστήματα προγραμματισμού δεν διαθέτουν πρόβλεψη όσον αφορά τα σφάλματα ή τις πιθανές βλάβες προγραμματισμένων ή συνεχιζόμενων εργασιών στην παραγωγική διαδικασία. Είναι επίσης μια μεγάλη πρόκληση για τον τρόπο πρόβλεψης πιθανών βλαβών και ποια είναι τα πρότυπα σφάλματος πριν από τον προγραμματισμό. Τα τελευταία χρόνια, αναδύθηκε η αναλυτική μεγάλων δεδομένων, η οποία προσφέρει καλές δυνατότητες για την πρόβλεψη σφάλματος (Ji and Wang 2017).

9.2 Αναλυτική μεγάλων δεδομένων στην πρόβλεψη σφάλματος

Σε μια πραγματική γραμμή παραγωγής ενός εργοστασίου, οι κύριες εργασίες περιλαμβάνουν τον προγραμματισμό εργασιών, τον χειρισμό υλικών, τη μηχανική κατεργασία και την επιθεώρηση.

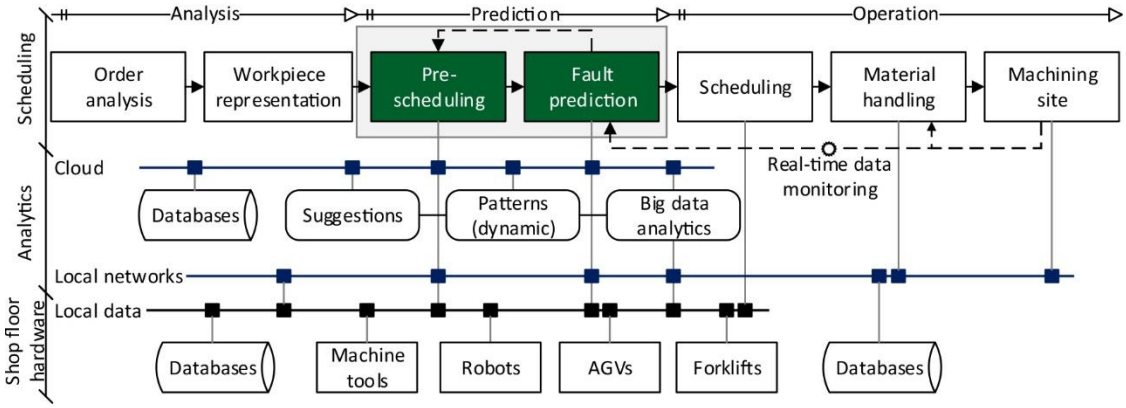
Βασικά, ο προγραμματισμός εργασιών εξετάζει τις διαθέσιμες συσκευές, όπου ο κύριος στόχος είναι το κόστος και ο χρόνος. Εντούτοις, οι προβλέψεις συνθηκών των μηχανών και η κατάσταση κατεργασίας δεν εξετάζονται λόγω του περιορισμού των παραδοσιακών τεχνικών. Στην παρούσα μελέτη περίπτωσης εξετάζονται τα πιθανά σφάλματα και σφάλματα που μπορούν να προβλεφθούν πριν από τον προγραμματισμό και κατά τη διάρκεια της μηχανουργικής κατεργασίας. Η Εικόνα 45 απεικονίζει την έννοια της πρόβλεψης σφαλμάτων ανάλυσης μεγάλων δεδομένων, όπου εξετάζονται δύο σενάρια: (1) οι προγραμματισμένες εργασίες, που αντιπροσωπεύονται από ένα σύνολο χαρακτηριστικών δεδομένων, συγκρίνονται με τα μοντέλα εξορύξεων εξόρυξης δεδομένων από τις βάσεις δεδομένων του εργοστασίου και (2) τα δεδομένα πραγματικού χρόνου συλλέγονται και επεξεργάζονται ως ομάδα χαρακτηριστικών δεδομένων για τις συνεχιζόμενες εργασίες και συγκρίνονται με τα δεδομένα που έχουν εξ ορυχθεί μέσα από πρότυπα βλαβών από τις βάσεις δεδομένων. Βάσει των συγκρίσεων των δύο σεναρίων, η ομοιότητα μπορεί να επιτευχθεί. Συνεπώς, η ομοιότητα παρέχει μια αναφορά για τον προγραμματισμό ή την αναδιάταξη των εργασιών. Εντός του πλαισίου, τα ιστορικά δεδομένα που σχετίζονται με την παραγωγή συλλέγονται πριν από την πρόβλεψη και χρησιμοποιούνται στην ανάλυση δεδομένων για τη δημιουργία προτύπου βλάβης. Οι εργασίες αντιπροσωπεύονται από τα σχετικά χαρακτηριστικά γνωρίσματα και η ομοιότητα ή η διαφορά υπολογίζεται συγκρίνοντας τα μοντέλα που προσφέρουν μοτίβα ελλαττωματικών με την προγραμματισμένη εργασία. Ακόμα η πιθανότητα ρίσκου των προγραμματισμένων εργασιών μπορεί να υπολογιστεί. Αυτή η πιθανότητα προσφέρει μια αναφορά στον τελικό υπολογισμό της λήψης απόφασης.



Εικόνα 45: Άποψη πρόβλεψης σφαλμάτων για τον προγραμματισμό (Ji and Wang 2017)

9.3 Αρχιτεκτονική Συστήματος

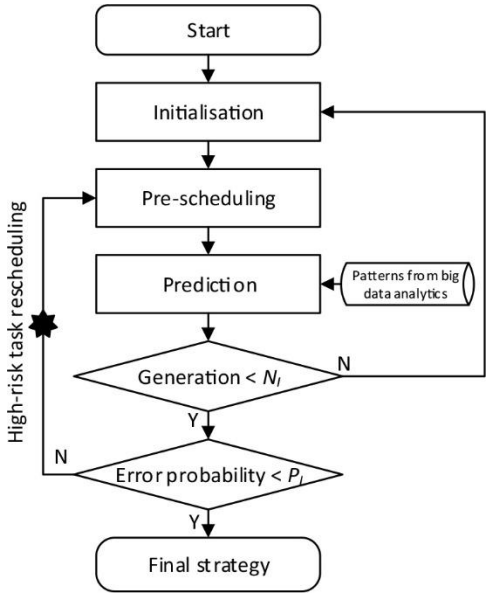
Η έρευνα αυτή αντιμετωπίζει δύο προκλήσεις: (1) να αποφευχθεί τυχόν εσφαλμένη αντιστοίχιση των εργασιών επεξεργασίας στις μηχανές πριν από τον προγραμματισμό και (2) να αποφευχθούν τυχόν σφάλματα κατά τη διάρκεια της μηχανουργικής κατεργασίας μέσω πραγματικής χρονικής παρακολούθησης και αναδιάταξης. Για το λόγο αυτό, προτείνεται μια μεγάλη ανάλυση βασισμένη στην ανάλυση δεδομένων που βασίζεται στην πρόβλεψη σφαλμάτων. Η Εικόνα 46απεικονίζει την αρχιτεκτονική συστήματος τριών βημάτων σε οριζόντια διεύθυνση και τρία επίπεδα σε κατακόρυφη κατεύθυνση. Το βήμα της ανάλυσης συνίσταται στην ανάλυση παραγγελιών και στην αντιπροσώπευση του τεμαχίου. Το βήμα πρόβλεψης περιέχει προ-προγραμματισμό και πρόβλεψη σφαλμάτων. Αυτά τα δύο βήματα έχουν εισαχθεί πριν από τον προγραμματισμό για την πιθανή πρόβλεψη σφαλμάτων κατεργασίας λόγω των συνθηκών της μηχανής και την κατάσταση κατεργασίας μέσω παρακολούθησης σε πραγματικό χρόνο. Στο τελευταίο βήμα λειτουργίας, πραγματοποιείται προγραμματισμός, χειρισμός υλικού και κατεργασία. Κατά τη διάρκεια της επεξεργασίας, συλλέγονται δεδομένα σε πραγματικό χρόνο και διαβιβάζονται στη μονάδα προειδοποίησης σφάλματος για προβλέψεις σφαλμάτων. Επιπλέον, αν υπάρχουν ακόμα μη προβλέψιμα σφάλματα κατά τη διάρκεια της κατεργασίας, ένα μήνυμα λάθους στέλνεται στον προ-προγραμματισμό της μονάδας, και το ημιτελές έργο διαμορφώνεται εκ νέου για την πρόληψη της βλάβης (Ji and Wang 2017).



Εικόνα 46: Αρχιτεκτονική αναλυτικής μεγάλων δεδομένων για τον προγραμματισμό της γραμμής παραγωγής (Ji and Wang 2017)

9.4 Πρόβλεψη βλάβης

Η Εικόνα 47 δείχνει τη ροή εργασίας της πρόβλεψης σφάλματος, συμπεριλαμβανομένης της αρχικοποίησης, τον προ-προγραμματισμού, την πρόβλεψη, κρίσης κριτηρίων τερματισμού και αξιολόγησης πιθανότητας σφάλματος. Οι αρχικές παράμετροι ρυθμίζονται κατά την αρχικοποίηση, π.χ. τον μέγιστο αριθμό δημιουργιών υπολογισμού και την πιθανότητα. Στη συνέχεια υπολογίζονται στη διαδικασία προ-προγραμματισμού τα σχέδια εργασιών (χρησιμοποιώντας τις ίδιες διαδικασίες με μια κανονική διαδικασία προγραμματισμού) , με βάση την οποία υπολογίζεται η πιθανότητα σφάλματος για την προγραμματισμένη εργασία στην διαδικασία πρόβλεψης. Εάν ορισμένες πιθανότητες σφάλματος είναι μεγαλύτερες από την αρχική P_I, αυτές οι εργασίες υψηλού κινδύνου θα επαναπρογραμματιστούν. Ωστόσο, τα σχέδια των ενεργειών μπορεί να μην έχουν ακόμα καθοριστεί όταν ο αριθμός των γενεών φτάσει στην αρχικά καθορισμένη τιμή κατωφλίου N_I. Σε μια τέτοια περίπτωση, οι τιμές των αρχικών παραμέτρων πρέπει να ρυθμιστούν ξανά (Ji and Wang 2017).



Εικόνα 47: Ροή εργασιών της πρόβλεψης σφάλματος (Ji and Wang 2017)

Η πρόβλεψη σφάλματος αποτελεί ίσως τον πυρήνας των διαδικασιών της παραγωγής, καθώς πιθανά πρότυπα σφάλματος εξορύσσονται από τις αναλύσεις μεγάλης κλίμακας δεδομένων. Μαζί, παρέχουν μια αναφορά για τη λήψη αποφάσεων προγραμματισμού. Τα πιθανά πρότυπα από την αναλυτική μεγάλων δεδομένων έχουν ως εξής:

- Πιθανές βλάβες του μηχανήματος και του τεμαχίου εργασίας: αναφέρεται στην πιθανότητα βλάβης μιας τάξης εργαλειομηχανών ή ενός εργαλείου μηχανής πριν από τον προγραμματισμό της εργασίας και πιθανά σφάλματα κατά τη διάρκεια επεξεργασίας δεδομένων σε πραγματικό χρόνο.
- Κατάσταση συντήρησης της μηχανής και του τεμαχίου: που σχετίζονται με το χρόνο χρήσης του. Ωστόσο, αν προγραμματιστεί ένα ακατάλληλο έργο για το μηχάνημα, τα μηχανήματα θα πρέπει να συντηρούνται εκ των προτέρων.
- Η ποιότητα του κατεργαζόμενου τεμαχίου και του εργαλείου εργαλειομηχανών: η ποιότητα του μηχανήματος συνδυάζεται με την κατάσταση του μηχανήματος, π.χ. ακαμψία, επαναληψιμότητα και σταθερότητα. Επομένως, εάν η κατάσταση του μηχανικού εργαλείου δεν είναι κατάλληλη για την κατεργασία ενός τεμαχίου εργασίας, η εργασία δεν πρέπει να εκχωρηθεί στο μηχάνημα.

9.5 Χειρισμός μεγάλης κλίμακας δεδομένων

Οι αλγόριθμοι που χρησιμοποιούνται σε μεγάλες αναλύσεις δεδομένων είναι οι εξής (Ji and Wang 2017):

- Ανάλυση κατά συστάδες
- Ανάλυση κατά παράγοντες
- Ανάλυση συσχέτισης και εξάρτησης
- Ανάλυση παλινδρόμησης
- Δοκιμή A/B
- Εξόρυξη δεδομένων

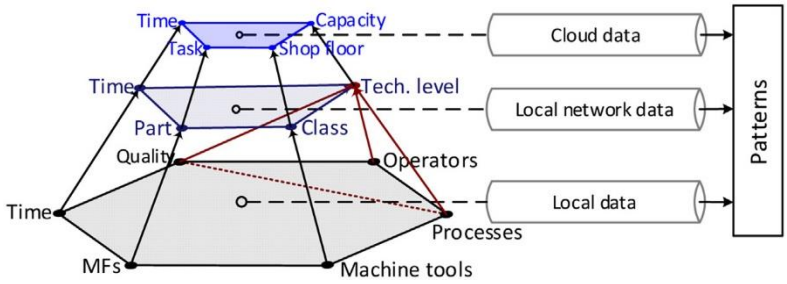
Εδώ εισάγονται οι βάσεις δεδομένων. Λαμβάνοντας υπόψη το μέγεθος των δεδομένων, τα δεδομένα θα πρέπει να διαχωρίζονται και να αποθηκεύονται σε διαφορετικές βάσεις δεδομένων: στα 1) τοπικά δεδομένα, 2) τοπικά δεδομένα δικτύου και 3) δεδομένα νέφους, όπως φαίνεται στην Εικόνα 48 (Ji and Wang 2017).

Οι τοπικές πληροφορίες αποθηκεύονται σε κατανεμημένους υπολογιστές, συμπεριλαμβανομένων τόσο δεδομένων ιστορικού όσο και πραγματικού χρόνου από

συστήματα παρακολούθησης, και την κάλυψη των πληροφοριών σχετικά με τα χαρακτηριστικά επεξεργασίας, τα εργαλειομηχανές, τις διαδικασίες, τους χειριστές, τη μέτρηση της ποιότητας και τον χρόνο. Η ανάλυση τοπικών δεδομένων εξετάζει τα πρότυπα κάθε εργαλείου εργαλειομηχανών σε σχέση με τα ιδιαίτερα χαρακτηριστικά κάθε μηχανήματος (Ji and Wang 2017).

Τα δεδομένα του τοπικού δικτύου αναφέρονται στα ιστορικά δεδομένα που καλύπτουν την κλάση των εργαλειομηχανών, τα τμήματα, το χρόνο και την τεχνική ικανότητα (ένα είδος αξιολόγησης που βασίζεται στην ποιότητα των διαδικασιών, του χειριστή και του μηχανικού στο τοπικό επίπεδο δεδομένων). Η ανάλυση του τοπικού δικτύου δεδομένων επικεντρώνεται στα πρότυπα των τάξεων εργαλειομηχανών σε σχέση με τις επιδόσεις τους (Ji and Wang 2017).

Τα δεδομένα cloud αναφέρονται επίσης στα ιστορικά δεδομένα που αφορούν τον χρόνο, την παραγωγική ικανότητα και τις απαιτήσεις. Η ανάλυση δεδομένων cloud εφαρμόζεται για να αποκτηθεί το πρότυπο μιας παραγωγικής διαδικασίας, η οποία χρησιμοποιείται για την αξιολόγηση της παραγωγικής ικανότητας του καταστήματος (Ji and Wang 2017).



Εικόνα 48: Επίπεδα δεδομένων (Ji και Wang, 2017)

9.6 Χαρακτηριστικά δεδομένων

Τα χαρακτηριστικά των μεγάλων δεδομένων περιλαμβάνουν τις πληροφορίες σχετικά με τα εργαλεία μηχανών, τα τεμάχια εργασίας, τις διαδικασίες μηχανουργικής κατεργασίας, το χρόνο μηχανουργικής κατεργασίας, τα αποτελέσματα μηχανουργικών κατεργασιών και τους χειριστές (βλέπε Εικόνα 19). Επίσης, οι λεπτομέρειες αυτών των πληροφοριών καλύπτουν όλους τους παράγοντες που επηρεάζουν τις κατασκευαστικές εργασίες. Ένα δείγμα φύλλου δεδομένων

παρουσιάζεται στον Πίνακας 9. Οι λεπτομέρειες των χαρακτηριστικών ημερομηνίας περιλαμβάνουν (Ji and Wang 2017):

Πίνακας 9: Δείγμα πίνακα δεδομένων ((Ji and Wang 2017)

No.	Workpiece							
	Quantity	MFs	Volume per MF (cm^3)	Hardness (HRC)	Yield strength (MPa)	Density (g/cm^3)	Roughness Ra (μm)	Accuracy (mm)
000001	20	Face	100	35	690	7.9	1.0	±0.02

Time		Machine tool					
When	Duration (min)	No.	Type	Structure	X\Y\Z error (mm)	A\B\C error (mm)	XYZ power (kw)
14:30	30	M15	4-axis	XYZBC	0.0005\0.0008\0.0007	0\0.0015\0.001	2.8\3.0\3.5

Machine tool				Machining process				
Spindle power (kw)	Maintenance date	Fault date	Fault type	Cutting speed (m/min)	Feed (mm/min)	Cutting depth (mm)	Cutting width (mm)	Fluid type/pressure
26	20160305	20160228	A	50	2000	0.2	20	F01/30 bar

	Machining process							
Tool type	Tool mate.\coat	Tool radius (mm)	Tool length(mm)	Entrance angle (°)	Rake angle (°)	Flank angle (°)	X\Y\Z cutting force (N)	X\Y\Z vibration (m/s^2)
IT	K2\TiCN	30	50	60	10	10	400\400\300	40\35\20

Machining result					Human factor	
Error (mm)	Ra (μm)	White layer (μm)	Dark layer (μm)	Hardened layer (μm)	Operator	Level
0.01	0.8	No	No	No	O001	T02

- Χαρακτηριστικά δεδομένων του τεμαχίου εργασίας: οι πληροφορίες για το τεμάχιο του περιλαμβάνουν την ποσότητα των εξαρτημάτων, τις γεωμετρίες εξαρτημάτων, τα μερικά υλικά και τις απαιτήσεις μηχανουργικής επεξεργασίας, κλπ. Εδώ, οι γεωμετρίες και η ποσότητα εξαρτημάτων μπορούν να αναπαρασταθούν με χαρακτηριστικά μηχανουργικής κατεργασίας και την ποσότητα τους. Μερικά υλικά αναφέρονται στα προφίλ υλικού σε σχέση με τη μηχανική κατεργασία υλικού, π.χ. τη σκληρότητα, την ευθραυστότητα και την ελαστικότητα κ.λπ., ώστε κάθε υλικό να μπορεί να εκπροσωπείται από μια σειρά παραμέτρων (Ji and Wang 2017).
- Χαρακτηριστικά δεδομένων του χρόνου κατεργασίας: ο χρόνος κατεργασίας αναφέρεται στην διάρκεια της περιόδου μηχανουργικής κατεργασίας που σχετίζεται με τη χρήση μιας εργαλειομηχανή, καθώς και ο χρόνος κατεργασίας, όπου ο χρόνος κατεργασίας περιλαμβάνει δύο περιπτώσεις: 1) ανοιχτή παραγωγική διαδικασία: οι περιβαλλοντικοί παράγοντες αλλάζουν με την πάροδο του χρόνου, π.χ. η θερμοκρασία το πρωί είναι χαμηλότερη από το απόγευμα. 2) κλειστή παραγωγική διαδικασία: οι ελεγχόμενοι περιβαλλοντικοί παράγοντες περιλαμβάνουν τη θερμοκρασία, την υγρασία και την περιεκτικότητα σε σκόνη κτλ (Ji and Wang 2017).

- Χαρακτηριστικά δεδομένων των εργαλειομηχανών: οι πληροφορίες των εργαλειομηχανών περιλαμβάνουν συνήθως τον αριθμό των εργαλειομηχανών, τους τύπους των εργαλειομηχανών, τις δομές εργαλειομηχανών, την ισχύ του άξονα, τον γραμμικό άξονα και τον περιστροφικό άξονα, τα σφάλματα κάθε άξονα και την κατανάλωση ενέργειας κάθε στοιχείου (Ji and Wang 2017).
- Χαρακτηριστικά δεδομένων των διαδικασιών κατεργασίας: οι διεργασίες μηχανικής κατεργασίας ανταποκρίνονται στις συνθήκες κοπής, τα εργαλεία κοπής, τα εξαρτήματα κοπής και τα φυσικά δεδομένα της διαδικασίας κοπής. Οι συνθήκες κοπής συμπεριλαμβάνουν τις παραμέτρους, το κοπτικό υγρό και την πίεση του. Τα εργαλεία κοπής περιλαμβάνουν τύπους εργαλείων, υλικά εργαλείων, και γεωμετρικές παράμετροι εργαλείων. Τα αξεσουάρ κοπής περιλαμβάνουν τους τύπους των εξαρτημάτων και τις παραμέτρους τους. Τα φυσικά δεδομένα αποτελούνται από δύναμη κοπής, κραδασμούς κοπής και θερμοκρασία κοπής κτλ (Ji and Wang 2017).
- Χαρακτηριστικά δεδομένων των αποτελεσμάτων μηχανουργικής κατεργασίας: τα αποτελέσματα της μηχανουργικής κατεργασίας σημαίνουν τις ιδιότητες από την άποψη των γεωμετρικών σφαλμάτων και της επιφανειακής ολοκλήρωσης. Τα γεωμετρικά σφάλματα περιλαμβάνουν γεωμετρικά και διατομικά σφάλματα έναντι των ονομαστικών ανοχών. Η ακεραιότητα της επιφάνειας υποδηλώνει την τραχύτητα της επιφανείας, την μορφολογία της επιφάνειας και τις ιδιότητες του υποστρώματος (π.χ., λευκό στρώμα, σκουρόχρωμο στρώμα, στρώμα παραμόρφωσης κόκκων και υπολειμματικό φορτίο) (Ji and Wang 2017).
- Χαρακτηριστικά δεδομένων ανθρώπινων παραγόντων: οι ανθρώπινοι παράγοντες αφορούν τους χειριστές, ιδιαίτερα τους εργάτες που εκτελούν τις διεργασίες χειροκίνητα. Οι δεξιότητες των χειριστών είναι οι βασικοί παράγοντες που σχετίζονται στενά με τα αποτελέσματα της μηχανουργικής κατεργασίας (Ji and Wang 2017).

9.7 Καθαρισμός δεδομένων

Τα δεδομένα υψηλής ποιότητας αυξάνουν την ακρίβεια της πρόβλεψης. Η ποιότητα των δεδομένων εξαρτάται από ένα σύνολο κριτηρίων ποιότητας (Ji and Wang 2017):

1. Εγκυρότητα: Οι μεταβλητές δεδομένων είναι το βασικό κομμάτι και υπάρχουν πολλοί τύποι περιορισμών όπως τύπος δεδομένων, εύρος, υποχρεωτικό, μοναδικότητα, set-membership, ξένο κλειδί, , και επικύρωση διασταυρούμενου πεδίου.
2. Ακρίβεια: αναφέρεται στο βαθμό συμμόρφωσης μιας μετρούμενης τιμής σε μια τυπική ή μια πραγματική τιμή
3. Πληρότητα: αναφέρει όλα τα απαιτούμενα μέτρα που πρέπει να είναι γνωστά
4. Συνεκτικότητα: ο βαθμός στον οποίο ένα σύνολο μέτρων είναι ισοδύναμο σε όλα τα συστήματα και
5. Ομοιομορφία: ο βαθμός στον οποίο ένα σύνολο δεδομένων μετρήθηκε με τη χρήση των ίδιων μονάδων μέτρησης σε όλα τα συστήματα.

Ο καθαρισμός των δεδομένων είναι η διαδικασία για τη διόρθωση ή την αφαίρεση ανακριβών αρχείων από βάσεις δεδομένων. Οι βασικές διαδικασίες καθαρισμού δεδομένων συνίστανται στον έλεγχο των δεδομένων, στον προσδιορισμό της ροής εργασίας, στην εκτέλεση ροής εργασίας, στην μετα-επεξεργασία και στον έλεγχο. Εδώ, στη διαδικασία ελέγχου των δεδομένων, εφαρμόζονται οι στατιστικές μέθοδοι για την ανίχνευση ανωμαλιών και αντιφάσεων στις βάσεις δεδομένων. Με σκοπό να αποκτήσει πληροφορίες σχετικά με τις υπάρχουσες ανωμαλίες στην μετάδοση δεδομένων, η ανίχνευση και η εξάλειψη πραγματοποιείται με μια σειρά λειτουργιών στα δεδομένα και στις προδιαγραφές ροής εργασιών. Αφού εκτελεστεί η ροή εργασίας του καθαρισμού δεδομένων, τα αποτελέσματα ελέγχονται για να επαληθεύσουν την ορθότητα κατά τη διάρκεια της μετα-επεξεργασίας και του ελέγχου. Οι μέθοδοι καθαρισμού δεδομένων είναι η ανάλυση , ο μετασχηματισμός δεδομένων, η ενίσχυση ακεραιότητας περιορισμών, η διπλή εξάλειψη και οι στατιστικές μέθοδοι (Ji and Wang 2017).

9.8 Ενοποίηση δεδομένων

Οι βάσεις δεδομένων κατασκευάζονται για κάθε επιμέρους εργαλειομηχανή, όπου τα δεδομένα σε πραγματικό χρόνο από τους αισθητήρες ενσωματώνονται στα εργαλεία του μηχανήματος συλλέγονται και αποθηκεύονται στις βάσεις δεδομένων, μαζί με τα ιστορικά δεδομένα. Στη συνέχεια, οι τοπικές βάσεις δεδομένων αποκτώνται για κάθε κλάση μηχανής ενσωματώνοντας τις βάσεις δεδομένων εργαλειομηχανών. Με βάση τις βάσεις δεδομένων των κλάσεων αυτών, μπορεί να δημιουργηθεί η βάση δεδομένων του εργοστασίου. Σε αυτή τη διαδικασία, η ενσωμάτωση

δεδομένων είναι η λειτουργία κλειδιών από βάσεις δεδομένων σε αποθήκες δεδομένων και περιλαμβάνει τον συνδυασμό δεδομένων που διαμένουν σε διαφορετικές πηγές και παρέχει στους χρήστες μια ενιαία προβολή αυτών των δεδομένων. Η ενσωμάτωση δεδομένων γίνεται με αυξανόμενη συχνότητα καθώς ο όγκος και η ανάγκη ανταλλαγής υφιστάμενων δεδομένων αυξάνεται ραγδαία (Ji and Wang 2017).

Κατά τη διαδικασία της προσαρμογής των δεδομένων, πρέπει να ακολουθείται ένας ειδικός κανόνας που σχετίζεται με την παραγωγική διαδικασία, δηλαδή να διατηρούνται τα μοναδικά μοτίβα της μηχανής για τον προγραμματισμό της μηχανής και τα κοινά μοντέλα των κατηγοριών μηχανών χρησιμοποιείται για προγραμματισμό υψηλού επιπέδου (Ji and Wang 2017).

9.9 Διαχειριστές διεργασιών

Μεταξύ των μεθόδων ανάλυσης μεγάλων δεδομένων, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως και μπορεί να ικανοποιήσει πλήρως τις απαιτήσεις της πρόβλεψης σφάλματος για τον προγραμματισμό της παραγωγής. Γενικά, υπάρχουν δύο τύποι προσεγγίσεων που χρησιμοποιούνται στην εξόρυξη δεδομένων, ταξινόμηση και ομαδοποίηση. Κατά την ταξινόμηση, η εργασία πρέπει να αντιστοιχηθεί σε στιγμιότυπα με προκαθορισμένες κλάσεις, ενώ στην ομαδοποίηση, η εργασία πρέπει να ομαδοποιείται με συναφή σημεία δεδομένων χωρίς να επισημανθούν. Η ταξινόμηση θεωρείται ως παράδειγμα εποπτευόμενης μάθησης, ενώ η ομαδοποίηση θεωρείται ως παράδειγμα μη εποπτευόμενης μάθησης. Ως εκ τούτου, η ταξινόμηση είναι κατάλληλη πρόβλεψη για σφάλματα στην παραγωγή. Υπάρχουν πολλοί αλγόριθμοι που χρησιμοποιήθηκαν ως ταξινομητές, π.χ. δέντρο απόφασης , ταξινομητής δικτύου Naïve Bayesian, Bayesian δίκτυο, τεχνητό νευρωνικό δίκτυο, μηχανισμός υποστήριξης vectors [SVM], «τεμπέλης μαθητής», γενετικός αλγόριθμος, τραχύ σύνολο και ασαφές σύνολο (Ji and Wang 2017).

Υπάρχουν και άλλες σημαντικές διαδικασίες για την εφαρμογή πρόβλεψης σφαλμάτων βασισμένης σε δεδομένα μεγάλης κλίμακας, π.χ. την ακρίβεια και τα μέτρα σφάλματος για τον ταξινομητή και τον προγνώστη, την αξιολόγηση της ακρίβειας, τις μεθόδους του συνόλου (για τη βελτίωση της ακρίβειας) και την επιλογή του μοντέλου. Με την προσεκτική παρακολούθηση των διαδικασιών, μπορεί να διασφαλιστεί η ακρίβεια της πρόβλεψης σφάλματος (Ji and Wang 2017).

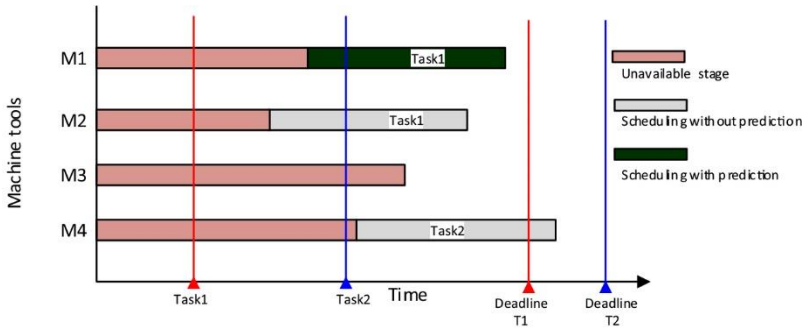
9.10 Εφαρμογή μοντέλου

Μια απλουστευμένη περίπτωση proof-of-concept απεικονίζεται για να δείξει την πορεία της προτεινόμενης μεθόδου. Η υπόθεση αυτή επικεντρώνεται αποκλειστικά στην εκτίμηση βλάβης στην αρχή του προγραμματισμού. Οι διαδικασίες της εκτίμησης σφαλμάτων σχετικά με την παρακολούθηση και συντήρηση σε πραγματικό χρόνο είναι παρόμοιες με την πρόβλεψη στην αρχή του προγραμματισμού, τουλάχιστον στο μεγάλο επίπεδο ανάλυσης δεδομένων. Σε ένα μια γραμμή παραγωγής υπάρχουν τέσσερις μηχανές οι M1-M4. Έχουν τοποθετηθεί στη γραμμή τα καθήκοντα που έχουν δημιουργηθεί. Δύο MFs (μια για κάθε εργασία) επιλέγονται μαζί με σκληρότητα υλικού, απαιτούμενη ανοχή και τραχύτητα και προθεσμία παράδοσης (βλ. Πίνακας 10). Ο προγραμματισμός αφορά κυρίως τη διαθεσιμότητα των μηχανημάτων. Σε αυτή την περίπτωση, οι χρήσεις των δύο εργαλειομηχανών απεικονίζονται στο διάγραμμα Gantt (Εικόνα 49). Οι εργασίες 1 και 2 θα αντιστοιχούν στα M2 και M4, αντιστοίχως (Ji and Wang 2017).

Πίνακας 10: Λεπτομέρειες μηχανών (Ji και Wang, 2017)

Details of two machining tasks.

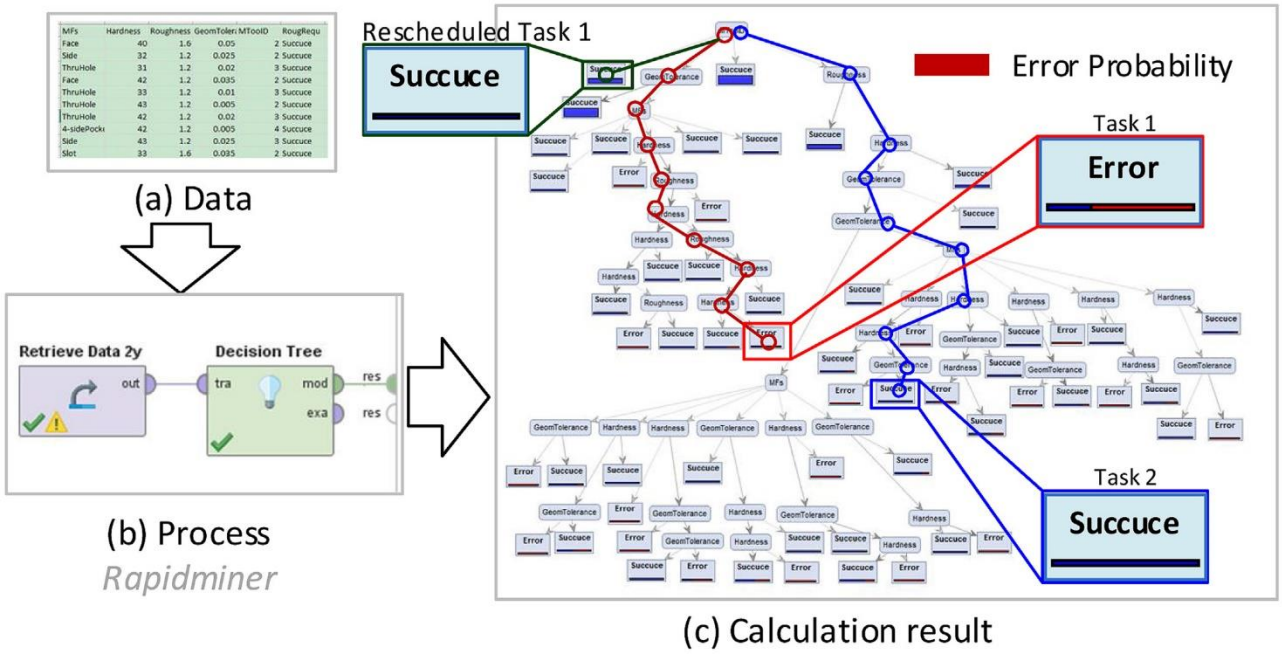
Task	MF	Hardness (HRC)	Tolerance (mm)	Ra (μm)	Deadline
1	Side	31	0.005	1.2	T1
2	Slot	45	0.05	0.8	T2



Εικόνα 49: Διάγραμμα Gantt για δύο μηχανές (Ji and Wang 2017)

Λόγω του γεγονότος ότι δεν υπάρχουν διαθέσιμα σημαντικά δεδομένα κατά τη διεξαγωγή της μελέτης περίπτωσης, παράγεται τυχαία ένα φύλλο υποθετικής φύσεως για τη δοκιμή του συστήματος. Τα δεδομένα εισάγονται στο RapidMiner όπου εφαρμόζονται δέντρα αποφάσεων. Η διαδικασία λειτουργίας και τα αποτελέσματα περιγράφονται στο . Σύμφωνα με τα αποτελέσματα,

εάν η εργασία 1 είναι διατεταγμένη στο M2, η πιθανότητα σφάλματος είναι πάνω από 50% (υποδεικνύεται από την κόκκινη γραμμή στην Εικόνα 50, ενώ είναι εφικτό εάν η εργασία 2 είναι διατεταγμένη στο M4. Επομένως, ο κίνδυνος του τρέχοντος σχεδίου είναι υπερβολικά υψηλός, υποδεικνύοντας την έγκαιρη παράδοση. Το εργαλείο που εκτελεί την εργασία 1 αναπροσαρμόζεται έτσι στο M1 όπως φαίνεται στην Εικόνα 49. Σύμφωνα με την πρόβλεψη, η πιθανότητα σφάλματος μειώνεται στο 0% για τη νέα διάταξη (που φαίνεται στην πράσινη γραμμή στην) (Ji and Wang 2017).



Εικόνα 50: Αποτελέσματα δένδρου απόφασης (Ji and Wang 2017)

10 Συμπεράσματα

Βρισκόμαστε σε μια εποχή όπου τα δεδομένα έχουν διαφορετικές μορφές και διαφορετικά μεγέθη για αυτό το λόγο είναι απαραίτητη η συνεχής εξεύρεση νέων τεχνικών ανάλυσης για να γίνεται πιο εύκολη και αποδοτική η επεξεργασία τους.

Η γνώση είναι δύναμη για κάθε οργανισμό και προσδίδει ανταγωνιστικό πλεονέκτημα. Οι σύγχρονες επιχειρήσεις καλούνται να επιλέξουν την πιο αποδοτικά οικονομική λύση με στόχο την αξιοποίηση της πληροφορίας που θα οδηγεί στην έγκαιρη και όσο το δυνατό πιο σωστή λήψη απόφασης προσφέροντας τη βέλτιστη αξία στην ίδια την επιχείρηση.

Το πρώτο συμπέρασμα είναι ότι η επιχείρηση πρέπει να έχει ξεκάθαρη ανάγκη για την ένταξη Αναλυτικής Μεγάλων Δεδομένων καθώς και τι αναμένει από την υλοποίηση της. Επίσης δεν πρέπει να υποτιμηθούν οι αλλαγές που θα χρειαστούν να γίνουν ώστε να πραγματοποιηθεί αυτό το τεράστιο πρότζεκτ. Αλλαγές όπως η δομή της πληροφορικής και η δημιουργία αποθηκών δεδομένων, διαχείρισης δεδομένων, Business Intelligence, αναλύσεις πρόβλεψης.

Το δεύτερο και σημαντικότερο συμπέρασμα για μένα είναι η αξία που θα πρέπει να δίνεται στην επιλογή κατάλληλης ομάδας που θα αναλάβει να ενσωματώσει αυτό το νέο έργο σε υπάρχοντα συστήματα και καθιερωμένες διαδικασίες. Μια κατάλληλη ομάδα που αποτελείται από τους σωστούς παράγοντες ειδικοτήτων και background θα καταφέρουν να κάνουν τον οργανισμό leader και όχι follower, ακόμα και στην Ελλάδα της κρίσης.

Πλέον όντας σε θέση να διαχειριστούν οι οργανισμοί αποδοτικά μεγάλες ροές δεδομένων θα είναι σε θέση να ανοίγονται αγορές content – based. Η δυνατότητα αυτή θα μπορούσε να επόμενο στάδιο να χρησιμοποιηθεί στον τομέα την Τεχνητής Νοημοσύνης και της Επιχειρηματικής Ευφυίας για την αναγνώριση προτύπων σε τρισδιάστατο χώρο.

Τέλος θα ήθελα να κλείσω με τη φράση «Processed data is information. Processed information is knowledge Processed knowledge is Wisdom» (Ankala V. Subbarao).

11 Βιβλιογραφία

Stokes , Mark. *China's Nuclear Warhead Storage and Handling System.* 10 March 2010. https://project2049.net/2010/03/12/chinas-nuclear-warhead-storage-and-handling-system/ (πρόσβαση November 15, 2018).

Aaltonen, Aleksi, και Niccolo Tempini. «'Everything Counts in Large Amounts: A Case Study of the Mechanisms of Data-based Production.» *Journal of Information Technology* 29, αρ. 1 (2014): 97-100.

Barbier, Geoffrey, και Huan Liu. «Data Mining in Social Media.» Στο *Social Network Data Analytics*, του/της Charu Aggarwal, 508. NY: Springer, 2011.

Bholat, David. «Big Data and central banks.» *Big Data & Society*, 2015: 1-6.

Big Data Framework. «ENTERPRISE BIG DATA PROFESSIONAL .» *Official Reference Guide Version 1.4*, 2018: 121.

Blei, David, Andrew Ng, και Michael Jordan. «Latent Dirichlet Allocation.» *Journal of Machine Learning Research*, 2003: 993-1022.

Buyya, Rajkumar, Rodrigo N Calheiros, και Amir Vahid Dastjerdi. *Big Data Principles and Paradigms.* Cambridge: Morgan Kaufmann, 2016.

Chang Wo. *NIST Big Data Interoperability Framework.* National Institute of Standards and Technology, Department of Commerce, 2015, 62.

Chen, Daniel, David Preston, και Morgan Swink. «How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management.» *Journal of Management Information Systems* 34, αρ. 4 (2015): 4-39.

Chen, Philip, και Chun-Yang Zhang. «Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.» *Information Sciences* 275 (2014): 314-347.

Chung, Wingyan. «BizPro: Extracting and categorizing business intelligence factors from textual news articles.» *International Journal of Information Management*, 2014: 272-284.

Constantiou , Ioanna, and Jannis Kallinikos. "New games, new rules: Big data and the changing context of strategy." *Journal of Information Technology*, 2015: 44-57.

Constantiou, Ioanna, και Jannis Kallinikos . «New games, new rules: Big Data and the changing context of strategy.» *Journal of Information* 30 (2015): 44-57.

Copi, Irving M, Carl Cohen, και Kenneth McMahon. *Introduction to Logic*. USA: Pearson Education, 2014.

De Mauro, Andrea, Marco Greco, Grimaldi Michele, και Paavo Ritala. «Human resources for Big Data professions: A systematic classification of job roles.» *Information Processing & Management*, 2018: 807-817.

Dean, Jeffrey, και Sanjay Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters.* : 6th Symposium on Operating Systems Design and Implementation, USENIX Association, 2008.

Dey, Lipika, Mirajul Haque, Arpit Khurdiya, και Gautam Shroff. «Acquiring Competitive Intelligence from Social Media.» *In Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data*, 2011: 9.

Easton, Ian M, και Russel Hsiao. *The Chinese People's Liberation Army's.* China, Beijing: The Chinese People's Liberation Army's UAV Projec, 2013.

Elliott, Timo. *timoelliott.* 5 07 2013. https://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html (πρόσβαση 07 09, 2019).

Erl, Thomas, Wajid Khattak , και Paul Buhler . *Erl Thomas, Khattak Wajid , Buhler Paul, 2015, Big Data Fundamentals: Concepts,*. Crawfordsville: The Prentice Hall, 2015.

Foote, Keith. *A Review of Different Database Types: Relational versus Non-Relational.* 21 December 2016. https://www.dataversity.net/review-pros-cons-different-databases-relational-versus-non-relational/.

Foster, Ian, Rayid Ghani, και Ron S Jarmin. *Big data and social science: A practical guide to methods and tools FosterIan, GhaniRayid, JarminRon S, et al. (eds), Big data and social science: A practical guide to methods and tools.* Μοντάζ: Chapman & Hall/. Τόμ. 45. Boca Raton: CRC Press, 2017.

Gandomi, Amir, και Murtaza Haider. «Beyond the hype: Big data concepts, methods, and analytics.» *International Journal of Information Management*, 2014: 137-144.

Gao, Jing, Andy Koronios, και Sven Selle. «Towards A Process View on Critical Success Factors in Big Data Analytics Projects.» *Big Data Process CSF*, 2015.

Govindan, Kannan, T.C.E Cheng, Nishikant Mishra, και Nagesh Shukla. «Big data analytics and application for logistics and supply chain management.» *Transportation Research Part E: Logistics and Transportation*, 2018: 343-349.

Günther, Wendy Arianne, Mohammad Rezazade Mehrizi, Marleen Huysman, και Frans Feldberg. «Debating big data: A literature review on realizing value from big data.» *Journal of Strategic Information Systems* 26, αρ. 3 (2017): 191-209.

He, Wu, Shenghua Zha, και Ling Li. «Social media competitive analysis and text mining: A case study in the pizza industry.» *International Journal of Information Management*, 2013: 464-472.

Hu, Weiming, Nianhua Xie, Li Li, Xianglin Zeng, και Stephen Maybank. «A Survey on Visual Content-Based Video Indexing and Retrieval.» *IEEE Transactions on Systems* 41 (2011): 797-819.

IBM. *What is distributed computing.* χ.χ. https://www.ibm.com/support/knowledgecenter/en/SSAL2T_8.2.0/com.ibm.cics.tx.doc/concepts/c_wht_is_distd_comptg.html (πρόσβαση 2019).

Issenberg, Sasha. *MIT Technology Review.* 19 December 2012. https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/ (πρόσβαση July 2019).

Ji, Wei, και Lihui Wang. «Big data analytics based fault prediction for shop floor scheduling.» *Journal of Manufacturing Systems* 43 (2017): 187-194.

Jin, Xiaolong, Benjamin W Wah, Xueqi Cheng, και Yuanzhuo Wang . «Significance and Challenges of Big Data Research.» *Big Data Research*, 2015: 59-64.

Kalil, Tom. «The White House.» 29 March 2012. https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal.

Labrinidis, Alexandros, και H Jagadish. «Challenges and Opportunities with Big Data.» *Proceedings of the VLDB Endowment*, 2015: 2032-2033.

Laney, Doug. «Application Delivery Strategies.» 6 February 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (πρόσβαση July 01, 2019).

Latinopoulos, Dionysis, και K Kechagia . «A GIS-based multi-criteria evaluation for wind farm site selection. A regional scale application in Greece.» *Renewable Energy: An International Journal* 78 (2015): 550-560.

Madsen, Anders Koed. «Between technical features and analytic capabilities: Charting a relational affordance space for digital social analytics.» *Big Data & Society*, 2015.

Madsen, Anders Koed. «Between technical features and analytic capabilities: Charting a relational affordance space for digital social analytics.» *Big Data & Society*, 2015: 1-15.

Marr, Bernard. *A Brief History of Big Data Everyone Should Read.* 24 February 2015. https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr/?originalSubdomain=au (πρόσβαση June 17, 2019).

Nazrul, Syed Sadat. *CAP Theorem and Distributed Database Management Systems.* 24 April 2018. https://towardsdatascience.com/cap-theorem-and-distributed-database-management-systems-5c2be977950e.

Newell, Sue, και Marco Marabelli. «Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'.» *The Journal of Strategic Information Systems* 24 (2015): 3-14.

Patil , Hemant. «"Cry Baby": Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant's Cry.» *Advances in Speech Recognition*, 2010: 323-348.

Press, Gil. *Forbes.* 03 May 2013. https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1202a3df65a1 (πρόσβαση June 17, 2019).

Shan , Caifeng, Fatih Porikli, Tao Xiang, και Shaogang Gong. *Video Analytics for Business Intelligent.* Berlin: Springer, 2012.

Sharma, Rajeev, Sunil Mithas, και Atreyi Kankanhalli. «Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations.» *European Journal of Information Systems*, 2014: 433-441.

Shollo, Arisa, και Robert D. Galliers. «Towards an understanding of the role of business intelligence systems in organisational knowing.» *Information Systems Journal* 26, αρ. 4 (2015): 339-367.

Tamm, Toomas, Peter Seddon, και Graeme Shanks . «PATHWAYS TO VALUE FROM BUSINESS ANALYTICS.» *Knowledge Management and Business Intelligence*, 2013.

Tweed, Thomas A. *Crossing and Dwelling: A Theory of Religion.* USA: Harvard University Press, 2008.

van der Vlist, Fernando. «Accounting for the Social: Investigating Commensuration and Big Data Practices at Facebook.» *Big Data & Society*, 2016.

Wikipedia. χ.χ. https://en.wikipedia.org/wiki/Zhou_Kehua (πρόσβαση November 18, 2018).

Winshuttle . χ.χ. https://www.winshuttle.com/big-data-timeline/ (πρόσβαση June 17, 2019).

Zikopoulos, Paul, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, και David Corrigan. *Harness the Power of Big Data The IBM Big Data Platform.* McGraw-Hill Education, 2012.

Κύρκος , Ευστάθιος. *Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων.* Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.

Ξωνίκη, Μιχαήλ. *Αναλυτική Μεγάλων Δεδομένων με χρήση Hadoop.* Μεταπτυχιακή Διατριβή, Θεσσαλονίκη: ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ, 2018.

Πολυμένης , Ιορδάνης. *ΤΑ Δεδομένα Μεγάλης Κλίμακας: Τεχνικές και Εργαλεία Ανάλυσης τους, και η Προσφορά τους ως Υπηρεσία του Υπολογιστικού Νέφους.* Θεσσαλονίκη: Πανεπιστήμιο Μακεδονίας, 2017.