ΣΤΡΑΤΙΩΤΙΚΗ ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ
Τμήμα Στρατιωτικών Επιστημών

Ε Λ Λ Η Ν Ι Κ Η Δ Η Μ Ο Κ Ρ Α Τ Ι Α
ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΑΚΑΔΗΜΑΪΚΟΥ ΕΤΟΥΣ 2016-17
ΣΧΕΔΙΑΣΗ & ΕΠΕΞΕΡΓΑΣΙΑ
ΣΥΣΤΗΜΑΤΩΝ (SYSTEMS ENGINEERING)
(ΠΔ 96 /2015/ΦΕΚ 163Α'/20.08.2014)

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Τα Μεγάλα Δεδομένα: Ανάλυση, Σπουδαιότητα και Εφαρμογές

Διατριβή που υπεβλήθη για την μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης

Υπό:
Ευάγγελος Μηλάτος
Α.Μ.: 2015018009

Η Μεταπτυχιακή Διατριβή του Εύαγγελου Μηλάτου…….. εγκρίνεται:

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Καθηγητής ΟΝΟΜΑΤΕΠΩΝΥΜΟ (Επιβλέπων) , Δρ. Δάρας Νικόλαος

Καθηγητής ΟΝΟΜΑΤΕΠΩΝΥΜΟ , Δρ. Ματσατσίνης Νικόλαος

Επίκουρος Καθηγητής ΟΝΟΜΑΤΕΠΩΝΥΜΟ , Δρ. Παπαδάκης Νικόλαος

ΣΕΛΙΔΑ ΣΚΟΠΙΜΑ ΚΕΝΗ

Έτος 2019

ΣΕΛΙΔΑ ΣΚΟΠΙΜΑ ΚΕΝΗ

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στον Καθηγητή μου κ. Δρ. Νικόλαο Δάρα για την δυνατότητα που μου έδωσε να υλοποιήσω τον μεγάλο μου στόχο να σπουδάσω σε αυτό το πολύ σημαντικό τμήμα, όπως επίσης να τον ευχαριστήσω για τον πολύτιμο χρόνο που διέθεσε για την περάτωση της παρούσας εργασίας. Οι σημαντικές υποδείξεις και συμβουλές του με κατεύθυναν σ' ένα σωστό τρόπο σκέψης πάνω απ' όλα και μου προσέφεραν σημαντικά εφόδια για την μετέπειτα ζωή μου.

Θα ήθελα να ευχαριστήσω ακόμα, όλους του καθηγητές του Πολυτεχνείου και της Στρατιωτικής Σχολής Ευελπίδων για τις πολύτιμες γνώσεις που μου προσέφεραν όλα αυτά τα χρόνια.

Τέλος, θέλω να εκφράσω ένα τεράστιο ευχαριστώ στην οικογένεια μου, για την στήριξη και την εμπιστοσύνη που μου έδειξε όλα αυτά τα χρόνια των σπουδών μου. Πέραν όμως από την πολύτιμη αυτή στήριξη, μου έδωσαν όλα τα εφόδια ώστε να γίνω ένας σωστός Άνθρωπος και αυτό είναι κάτι που δεν μαθαίνεται, αλλά μεταδίδεται.

Περιεχόμενα

Περίληψη

Η παρούσα μεταπτυχιακή διατριβή αναλύει τον όρο «Big Data» και διερευνά τη μέθοδο της επιλεκτικής προτεραιότητάς για τις μεγάλες ποσότητες των δεδομένων κατά την επεξεργασία των μεγάλων δεδομένων «Big Data». Μετά τον ορισμό των συνόλων των εστιακών δεδομένων, εισάγεται η έννοια του προγράμματος της επιλογής δεδομένων που καθορίζει το ποσό δεδομένων που μπορεί να λάβει υπόψη ο υπολογιστικός επεξεργαστής. Στη συνέχεια, καθορίζονται οι σχέσεις των συλλεγόμενων δεδομένων επιλογής και της ορθολογικής επιλογής για τα ποσά των δεδομένων. Ακολούθως, θεωρείται η περίπτωση πολλών επεξεργαστών δεδομένων και αποδεικνύεται ότι υπάρχουν πυρήνες και ισορροπίες των αντιθέσεων, η μελέτη των οποίων μπορεί να παρέχει χρήσιμες πληροφορίες.

1. Εισαγωγή

1.1 Τεκμηρίωση

Σκοπός της παρούσας εργασίας είναι η τεκμηρίωση μιας ποσοτικής συστηματικής μοντελοποίησης για την επεξεργασία της μεγάλης ροής δεδομένων. Δεδομένου ότι, σύμφωνα με επίσημους υπολογισμούς, η συνολική παγκόσμια ροή δεδομένων βαίνει ολοένα και αυξανόμενη, είναι σαφές ότι ο συνεχώς αυξανόμενος όγκος δεδομένων σύντομα θα προκαλέσει μεγάλες δυσκολίες στην αποτελεσματική επεξεργασία πληροφοριών και θα κάνει εξαιρετικά δύσκολο έργο επεξεργασίας της ροής δεδομένων. Γι' αυτό θα ερευνήσουμε τη μέθοδο της επιλεκτικής προτεραιότητάς για τις μεγάλες ποσότητες των δεδομένων κατά την επεξεργασία των μεγάλων δεδομένων «Big Data». Μετά τον ορισμό των συνόλων των εστιακών δεδομένων, εισάγεται η έννοια του προγράμματος της επιλογής δεδομένων που καθορίζει το ποσό δεδομένων που μπορεί να λάβει υπόψη ο υπολογιστικός επεξεργαστής. Στη συνέχεια, καθορίζονται οι σχέσεις των συλλεγόμενων δεδομένων επιλογής και της ορθολογικής επιλογής για τα ποσά των δεδομένων. Ακολούθως, θεωρείται η περίπτωση πολλών επεξεργαστών δεδομένων και αποδεικνύεται ότι υπάρχουν πυρήνες και ισορροπίες των αντιθέσεων, η μελέτη των οποίων μπορεί να παρέχει χρήσιμες πληροφορίες.

Προκειμένου να ξεπεραστεί επειγόντως αυτό το εμπόδιο, μια καλή ιδέα φαίνεται να είναι η κατάλληλη επιλογή των ποσών δεδομένων. Σε αυτή την κατεύθυνση, το παρόν έγγραφο μελετά ένα λογικό ερώτημα που τίθεται και μπορεί να αποτελέσει κεντρικό αντικείμενο συζήτησης σε επακόλουθες επιπρόσθετες επιστημονικές μελέτες. Η ερώτηση αφορά την προτίμηση των επιλογών και των προτεραιοτήτων στην επεξεργασία μεγάλων δεδομένων. Παρομοίως, εάν κάθε μία από μια ομάδα επεξεργαστών δεδομένων προτιμά να περιορίζεται σε διαφορετικά σύνολα δεδομένων από μια συλλογή μεγάλων δεδομένων, τότε πόσο οι διαφορετικές προτεραιότητες της επεξεργασίας θα μπορούσαν να οδηγήσουν σε καταστάσεις ισορροπίας ή αντιθέσεις; Το κίνητρο για απάντηση σε αυτή την ερώτηση εμπνέεται από την οικονομική θεωρία των πυρήνων και των ισορροπιών και η σχετική ανάλυση προσαρμόστηκε στο πλαίσιο για τον μαθηματισμό μιας μεγάλης οικονομίας που παρουσίασε ο W. Hildenbrand στο βιβλίο του [1].

1.2 Ιστορική αναδρομή

Την Δεκαετία 1950 για την διαχείριση και επεξεργασία μεμονωμένων αρχείων χρησιμοποιούνταν οι γνωστές για την εποχή εκείνη ταινίες και κάρτες. Οι τεχνολογικές όμως εξελίξεις στις συσκευές μαζικής αποθήκευσης και η αύξηση της υπολογιστικής ισχύος, θέτουν τις προϋποθέσεις για την ανάπτυξη των συστημάτων διαχείρισης δεδομένων ώστε να αντικαταστήσουν τα συστήματα διαχείρισης αρχείων.

Την Δεκαετία του 1960 τα πρώτα συστήματα διαχείρισης βάσεων δεδομένων ξεκίνησαν να κάνουν την εμφάνισή τους με σκοπό ένα κοινό οργανωτικό πλαίσιο με στόχο την διαχείριση τους, τα οποία δεδομένα μέχρι τότε αποθηκεύονταν σε μεμονωμένα αρχεία. Το 1964, ο Charles Bachman στέλεχος της General Electric πρότεινε ένα δικτυωτό μοντέλο δεδομένων (network data model) στο οποίο οι εγγραφές των δεδομένων ήταν συνδεδεμένες μεταξύ τους με τέτοιο τρόπο ώστε να σχηματίζουν τεμνόμενα σύνολα δεδομένων. Τα πρώτα λοιπόν συστήματα διαχείρισης βάσεως δεδομένων στηρίχθηκαν σε αυτό το δικτυωτό μοντέλο. Το 1965 η εταιρία IBM και η διεύθυνση διαστήματος της North American Aviation ανέπτυξαν από κοινού το ιεραρχικό μοντέλο δεδομένων. Σε αυτό το μοντέλο, τα δεδομένα παριστάνονταν ως δενδροειδής δομές μέσα σε μια ιεραρχία εγγράφων. Το Σύστημα Διαχείρισης Πληροφοριών (information management system-IMS) [2] της IBM που κυκλοφόρησε το 1969 ήταν βασιζόμενο στο ιεραρχικό μοντέλο δεδομένων. Από τα δικτυωτά και ιεραρχικά συστήματα βέβαια, μόνο τα IMS παραμένουν σε χρήση έως σήμερα. Την δεκαετία του 1970, ο ορισμός του σχεσιακού μοντέλου δεδομένων έγινε για πρώτη φορά εν έτη 1970 από τον Edgar Codd της IBM με τίτλο «System R4 Relational». Στην αρχή δεν ήταν ξεκάθαρο κατά πόσο ένα σύστημα σχεσιακό που θα βασιζόταν στο σχεσιακό μοντέλο θα μπορούσε να πετύχει εμπορικά. Έτσι μέχρι και το 1979 όλες οι εμπορικές υλοποιήσεις βάσεων δεδομένων βασίζονταν είτε στην δικτυωτή είτε στην ιεραρχική προσέγγιση.

Το 1976 το μοντέλο οντοτήτων-σχέσεων (ER-Entity Relationship) προτάθηκε από τον Ταϊβανέζο επιστήμονα Η/Υ P.P. CHEN για να περιγράψει με γραφικά σύμβολα τα δεδομένα ως συσχετίσεις (σχέσεις), οντότητες και γνωρίσματα. Αναπτύχθηκαν οι έννοιες της διαχείρισης συναλλαγών (transaction management) από τον Jim Gray [3]. Οι τάσεις που

άρχιζαν να εμφανίζονται τότε, εκείνη την περίοδο, αφορούσαν τα αντικειμενοστραφή συστήματα, την αρχιτεκτονική πελάτη - διακομιστή και τις κατανεμημένες βάσεις. Οι εγκαταστάσεις των σχεσιακών συστημάτων αυξάνονται με γοργούς ρυθμούς, με πρώτα τα συστήματα της Oracle. Εμφανίζονται επιτυχώς τα σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων και στους γνωστούς τότε, προσωπικούς υπολογιστές. Η Dbase εξελίχθηκε μέχρι τις μέρες μας σε Paradox και η πιο γνωστή στο κοινό, την Microsoft Access. Την δεκαετία 1990 εμφανίζονται τα πρώτα εμπορικά αντικειμενοστραφή συστήματα Βάσεων δεδομένων και η σύνδεση των βάσεων δεδομένων στο διαδίκτυο. Διαδίδεται ευρύτατα η τεχνολογία που επιτρέπει την επικοινωνία των χρηστών με βάσεις δεδομένων μέσω διαδικτύου (όπως HTML, ASP, XML).

Το 1991 το διαδίκτυο, ο παγκόσμιος ιστός όπως τον ξέρουμε, γεννιέται. Το πρωτόκολλο μεταφοράς υπερκειμένων (HTTP) γίνεται το βασικό μέσον διαμοιρασμού πληροφοριών. Το 1995 η SUN βγάζει στην κυκλοφορία την πλατφόρμα Java. Η Java, που ανακαλύφθηκε το 1991, γίνεται η δεύτερη πιο διαδεδομένη γλώσσα μετά την γνωστή C. Κυριαρχεί στις εφαρμογές διαδικτύου και καθιερώνεται στις μεσαίου επιπέδου εφαρμογές. Αυτές οι εφαρμογές είναι η πηγή καταγραφής και αποθήκευσης της κίνησης του διαδικτύου. Το παγκόσμιο σύστημα εντοπισμού (GPS) γίνεται πλήρως λειτουργικό. Το GPS είχε αναπτυχθεί αρχικά από την DAPRA (υπηρεσία προγραμμάτων προηγμένης έρευνας και άμυνας του Αμερικανικού στρατού) για στρατιωτικές εφαρμογές στις αρχές της δεκαετίας του ’70. Σήμερα η τεχνολογία αυτή είναι παρούσα, από εφαρμογές πλοήγησης αυτοκινήτων, πλοίων και αεροπλάνων μέχρι την στόχευση πυραύλων με εξαιρετική μεγάλη ευκρίνεια. Το 1998 ο Carlo Strozzi αναπτύσσει μια ανοιχτού κώδικα βάση δεδομένων και την αποκαλεί NoSQL. Δέκα χρόνια αργότερα, η πρωτοβουλία ανάπτυξης βάσεων δεδομένων NoSQL που θα μπορεί να επεξεργάζεται μεγάλα και αδόμητα σύνολα δεδομένων, κερδίζει έδαφος. Ιδρύεται η Google απο τους Larry Page και Sergey Brin οι οποίοι ξεκίνησαν σε μια εργασία μιας μηχανής αναζήτησης του πανεπιστημίου STANFORD με την ονομασία BackRub.

Το 2001 ξεκινά η λειτουργία του Wikipedia. Μια ηλεκτρονική εγκυκλοπαίδεια πληθώρας Πηγών άρθρων που φέρνει την επανάσταση στον τρόπο με τον οποίο οι άνθρωποι αναζητούν αλλά και καταγράφουν πληροφορίες. Μέχρι το 2017 η Wikipedia έφθασε τα 5.700.000 στην αγγλική γλώσσα μόνο. Το 2002 το Ινστιτούτο Ηλεκτρολόγων και

Ηλεκτρονικών Μηχανικών (IEEE) ορίζει την πρώτη έκδοση των προδιαγραφών Bluetooth. Το Bluetooth είναι μια ασύρματη τεχνολογία που μεταφέρονται δεδομένα μεταξύ των ηλεκτρονικών συσκευών μεταξύ τους σε κοντινές αποστάσεις. Η εξέλιξη αυτών των προδιαγραφών και η υιοθέτησή τους οδήγησε σε μια νέα σειρά φορητών συσκευών και επέτρεψε την επικοινωνία μεταξύ της συσκευής αυτής και ενός άλλου υπολογιστή. Το 2003 σύμφωνα με μελέτες του IDC (International Data Corporation) που εδρεύει στη Μασαχουσέτη των ΗΠΑ, ο όγκος των δεδομένων που δημιουργήθηκε το 2004 ξεπερνά εκείνον που είχε δημιουργηθεί σε ολόκληρη την ιστορία της ανθρωπότητας μέχρι εκείνη τη στιγμή.

Εκτιμάται οτι 1.8 zettabytes δεδομένων δημιουργήθηκαν μόνο το 2010. (1.8 zettabytes ισοδυναμούν με 220 δισεκατομμύρια ταινίες υψηλής ευκρίνειας HD, διάρκειας 2 ωρών περίπου). Ξεκινά η λειτουργία του LinkedIn, του δημοφιλούς μέσου κοινωνικής δια δικτύωσης για επαγγελματίες. Το 2014 η ιστοσελίδα είχε περίπου 290 εκατομμύρια χρήστες. Η υπηρεσία κοινωνικής δικτύωσης Facebook, ιδρύεται απο τον Mark Zuckerberg με άλλους συμφοιτητές του στο Cambridge της Μασσαχουσέτης. Έως το 2017, η ιστοσελίδα είχε πάνω απο 1.8 δισεκατομμύρια χρήστες.

Το 2005 το πιο σημαντικό ερευνητικό έργο που απασχόλησε τα Big Data, ήταν το Apache Hadoo, δημιουργείται απο τους Doug Cutting και Mike Caferella. Ο διάσημος κίτρινος ελέφαντας, έγινε οικεία λέξη σχεδόν όλων των στρατηγικών μεγάλων δεδομένων. Το Εθνικό Επιστημονικό Συμβούλιο των ΗΠΑ, προτείνει στο Εθνικό Ίδρυμα Επιστημών την δημιουργία μιας επαγγελματικής κατηγορίας για «έναν επαρκή αριθμό υψηλής ποιότητας καταρτισμένων κατάλληλα επιστημόνων» που θα διαχειριστούν την αυξανόμενη συλλογή των ψηφιακών πληροφοριών.

Το 2009 ο αριθμός των ηλεκτρονικών συσκευών που είναι συνδεδεμένα στο διαδίκτυο ξεπερνά τον παγκόσμιο πληθυσμό που είναι 6,6 δις.

Το πρωτόκολλο IPv4 βασίστηκε σε έναν 32μπιτο αριθμό που μας δείχνει ότι δύναται να υπάρξουν 232 ή 4,5 δισεκατομμύρια μοναδικές διευθύνσεις διαθέσιμες. Το γεγονός αυτό μαρτυρά την αλματώδη αύξηση των συσκευών που είναι συνδεδεμένα στο διαδίκτυο. Γι' αυτό το λόγο έχει αναπτυχθεί η νέα έκδοση IPv6 που θα πολλαπλασιάσει τις διαθέσιμες μοναδιαίες διευθύνσεις.

3.1 zettabytes δεδομένων θα δημιουργηθούν το 2019 αλλά μόνο το 4% όσων θα μπορούσαν να χρησιμοποιηθούν για μεγάλα δεδομένα θα έχουν αναλυθεί. Προβλέπεται από το IDC ότι μέχρι το 2020 ο ψηφιακός κόσμος θα κατέχει 42 zettabytes, 59 φορές των συνολικό αριθμό των κόκκων άμμου από όλες τις παραλίες του πλανήτη.

Το «Harvard Business Review» αναφέρει το επάγγελμα του αναλυτή δεδομένων ως «την πιο δύσκολη εργασία του 21ου αιώνα».

2. Τι εννοούμε με τον όρο «Μεγάλα Δεδομένα»

2.1 Εισαγωγή

Ο όρος "Big Data" (μεγάλα δεδομένα) χρησιμοποιείται για να περιγράψει δεδομένα τα οποία χαρακτηρίζονται από έναν εξαιρετικά μεγάλο όγκο, ο οποίος καθίσταται ιδιαίτερα δύσκολος στην εξόρυξη, αποθήκευση, διαχείριση και ανάλυση τους από τις παραδοσιακές εφαρμογές διαχείρισης βάσεων δεδομένων που υπάρχουν σήμερα. Ωστόσο, ο ορισμός αυτός εμπεριέχει μια υποκειμενικότητα όσον αφορά τον ελάχιστο όγκο που πρέπει να έχουν τα δεδομένα, ώστε να μπορούν να θεωρηθούν "μεγάλα δεδομένα". Λαμβάνουμε υπόψιν ότι, καθώς η τεχνολογία εξελίσσεται μέσα στο χρόνο, ο όγκος των πακέτων δεδομένων, που χαρακτηρίζονται ως μεγάλα δεδομένα αυξάνεται επίσης. Πρέπει να σημειωθεί κιόλας ότι ο ορισμός μπορεί να διαφέρει ανάλογα τον τομέα, ανάλογα με το είδος των λογισμικών, που είναι διαθέσιμα και ποια είναι τα συνήθη μεγέθη των πακέτων δεδομένων σε κάθε επιστημονικό κλάδο που θέλουμε να αναλύσουμε. Με αυτές τις επισημάνσεις, τα μεγάλα δεδομένα σε πολλούς τομείς σήμερα κυμαίνονται από μερικές δεκάδες terabytes έως τα πολλαπλάσια τους petabytes. Ψηφιακά δεδομένα συναντάμε πλέον παντού, σε κάθε οικονομία, σε κάθε οργανισμό και χρήστη της ψηφιακής τεχνολογίας. Τα μεγάλα δεδομένα έλκουν όλο και περισσότερο το ενδιαφέρον των επαγγελματιών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται και αυτοί να ωφεληθούν από την

αξιοποίησή τους αλλά και οι ίδιοι οι επιχειρηματίες – επιστήμονες (γιατροί, μηχανικοί κτλ). Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή. Όπως πχ ο Νόμος του «Moore» στην πληροφορική [4].

Ο Νόμος του Moore, ο οποίος ειπώθηκε για πρώτη φορά από τον συνιδρυτή της Intel Gordon Moore, αναφέρει ότι η πυκνότητα των τρανζίστορ στα τσιπ (ο αριθμός δηλαδή των τρανζίστορ ανά μονάδα επιφάνειας) διπλασιάζεται κάθε περίπου δύο χρόνια. Με άλλα λόγια, η ποσότητα υπολογιστικής ισχύος που μπορούν να αγοραστούν με το ίδιο χρηματικό ποσό διπλασιάζεται περίπου κάθε δύο χρόνια. Το cloud computing αναφέρεται στη δυνατότητα πρόσβασης σε υψηλή ποσότητα κλιμακούμενης (scalable) υπολογιστικής δύναμης μέσω του Διαδικτύου, συχνά σε τιμές χαμηλότερες από αυτές που θα απαιτούνταν για την εγκατάσταση στον υπολογιστή κάποιου αποθηκευτικού υλικού, διότι οι πόροι διαμοιράζονται σε πολλούς χρήστες. Η δυνατότητα παραγωγής, επικοινωνίας, μερισμού και πρόσβασης δεδομένων έχει εκτοξευθεί από την αύξηση του αριθμού των ατόμων, συσκευών και αισθητήρων, που συνδέονται σήμερα σε ψηφιακά δίκτυα. Κάποιος μπορεί να αγοράσει μια μονάδα δίσκου με ικανότητα να αποθηκεύσει όλη τη μουσική του κόσμου.

Επίσης, τα μέσα εξόρυξης από τα δεδομένα σημειώνουν σημαντική βελτίωση, καθώς τα διαθέσιμα λογισμικά για την εφαρμογή τεχνικών αυξανόμενης πολυπλοκότητας συνδυάζονται με την αυξανόμενη υπολογιστική ισχύ. Το 2017, περισσότερα από 4,5 δισεκατομμύρια άνθρωποι, ή το 70 τοις εκατό του παγκόσμιου πληθυσμού, χρησιμοποιούσαν κινητά τηλέφωνα, και περίπου 45 τοις εκατό από αυτούς τους ανθρώπους είχαν smartphones, των οποίων η διείσδυση αυξάνεται κατά περισσότερο από 23 τοις εκατό το χρόνο. Περισσότερα από 70 εκατ. δικτυωμένοι κόμβοι αισθητήρων βρίσκονται πλέον στους κλάδους μεταφορών, αυτοκινητοβιομηχανίας, επιχειρήσεων κοινής ωφέλειας, καθώς και σε τομείς του λιανικού εμπορίου. Ο αριθμός αυτών των αισθητήρων αυξάνεται σε ποσοστό άνω του 33% .[5] Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή δεδομένων .Αυτός είναι ο λόγος που τα μεγάλης κλίμακας δεδομένων έχουν γίνει πρόσφατη περιοχή των στρατηγικών επενδύσεων για τους IT οργανισμούς .Αν και είναι σαφές ότι οι νέες τεχνολογίες και νέες μορφές προσωπικής επικοινωνίας οδήγησαν στην τάση των μεγάλης κλίμακας δεδομένων , θεωρούν ότι ο παγκόσμιος πληθυσμός του διαδικτύου

αυξήθηκε κατά 6,5% από το 2010-2011 και τώρα αντιπροσωπεύει πάνω από δισεκατομμύρια ανθρώπους. Αυτό μπορεί να φαίνεται μεγάλο ,αλλά υποδηλώνει ότι η συντριπτική πλειοψηφία του παγκόσμιου ιστού έχει ακόμα να συνδεθεί .Ενώ μπορεί να είναι ότι ποτέ δεν θα φτάσουμε 100% του παγκόσμιου πληθυσμού σε απευθείας σύνδεση (λόγω των περιορισμένων διαθέσιμων πόρων , κόστος των αγαθών ,και την περιορισμένη υλική ευελιξία),όλο και περισσότερο είναι εκείνα που είναι σε απευθείας σύνδεση περισσότερο από ποτέ .

Μόλις λίγα χρόνια πριν ήταν λογικό να σκεφτείς ότι πολλά είχαν μια επιφάνεια εργασίας και ίσως ένα laptop στην διάθεση τους. Το 2013, η ανθρωπότητα δημιούργησε πάνω από 1,4 τρισεκατομμύρια GB δεδομένων. Ο όγκος των δεδομένων αναμένεται να αυξηθεί 60 φορές μέχρι το 2021. Η Google λαμβάνει πάνω από 2.400.000 ερωτήματα αναζήτησης κάθε λεπτό. 72 ώρες βίντεο προστίθενται στο YouTube κάθε λεπτό ενώ υπάρχουν 374 νέοι χρήστες του Ιντερνέτ κάθε λεπτό. Οι χρήστες του Twitter στέλνουν πάνω από 130.000 tweets κάθε λεπτό (που είναι πάνω από 170 εκατομμύρια ανά ημέρα). Η IDC προβλέπει ότι η αγορά για την τεχνολογία των μεγάλης κλίμακας δεδομένων και υπηρεσίες θα φτάσει τα $16,9 δισεκατομμύρια μέχρι το 2018 με αύξηση 40% πάνω από τον ορίζοντα της πρόβλεψης.

2.2 Η σπουδαιότητα των μεγάλων δεδομένων

Η χρήση των μεγάλων δεδομένων προσφέρει τεράστιες ανεκμετάλλευτες δυνατότητες δημιουργικής αξίας. Οργανισμοί σε πολλούς κλάδους και πολλές επιχειρηματικές λειτουργίες μπορούν να αξιοποιήσουν μεγάλα δεδομένα με σκοπό τη βελτίωση της κατανομής και του συντονισμού των πόρων τους, τον περιορισμό της σπατάλης, την αύξηση της διαφάνειας και της λογοδοσίας, και την ανάδειξη νέων ιδεών και αντιλήψεων. Τα μεγάλα δεδομένα δημιουργούν αξία με διάφορους τρόπους. Οι σημαντικότεροι είναι οι κάτωθι:

Δημιουργία Διαφάνειας (transparency). Η εύκολη και έγκαιρη πρόσβαση σε μεγάλα δεδομένα από τους ενδιαφερόμενους φορείς παρέχει ευκαιρίες δημιουργίας τεράστιας αξίας.

Συχνά, τέτοιες ευκαιρίες προκύπτουν σε περιπτώσεις όπου παρατηρείται έλλειψη συμφωνίας κινήτρων για δημιουργία διαφάνειας δεδομένων. Για παράδειγμα, στον δημόσιο τομέα, υπάρχουν περιπτώσεις όπου το προσωπικό διαφόρων υπηρεσιών σπαταλά σημαντικό ποσοστό του χρόνου τους για να εντοπίσουν πληροφορίες σε άλλες κυβερνητικές υπηρεσίες, χρησιμοποιώντας μη-ψηφιακά μέσα (π.χ. σε έντυπους καταλόγους ή τηλεφωνώντας), και στη συνέχεια για να πάρουν τις πληροφορίες αυτές έπρεπε να επισκεφθούν την πηγή της πληροφορίας για να λάβουν τα στοιχεία με φυσικά μέσα, (πχ. οπτικοί δίσκοι). Τέτοιου είδους σπατάλη έχει μειωθεί σημαντικά σε οργανισμούς, που αξιοποιούν τα μεγάλα δεδομένα για να ψηφιοποιήσουν την πληροφορία αυτή, χρησιμοποιώντας τα διαθέσιμα δίκτυα, και αναπτύσσοντας εργαλεία ευκολότερης εύρεσης της αναζητούμενης πληροφορίας. Ωστόσο, ακόμη και σε τομείς, που έχουν υιοθετηθεί οι νέες τεχνολογίες και τα μεγάλα δεδομένα, υπάρχουν σημαντικά κίνητρα για υψηλότερη απόδοση, υπάρχουν περιθώρια αύξησης διαφάνειας και ανταλλαγής μεγάλων δεδομένων. Στον τομέα της μεταποίησης, πολλές εταιρείες χρησιμοποιούν τα μεγάλα δεδομένα για τη βελτίωση στην απόδοση της Έρευνας και Τεχνολογίας (π.χ. πολύπλοκες προσομοιώσεις) και στη διαχείριση της αλυσίδας εφοδιασμού τους. [5]

2.3 Εντοπισμός αναγκών, μεταβλητότητας και αύξηση απόδοσης

Όλο και περισσότερες εταιρείες ψηφιοποιούν και αποθηκεύουν μια αυξανόμενη ποσότητα εξαιρετικά λεπτομερών δεδομένων σχετικά με τις συναλλαγές. Όλο και περισσότεροι αισθητήρες ενσωματώνονται σε φυσικές συσκευές - από τον εξοπλισμό της γραμμής παραγωγής έως σε αυτοκίνητα και σε κινητά τηλέφωνα- οι οποίοι μετρούν διαδικασίες, χρήση προϊόντων, και ανθρώπινες συμπεριφορές. Επίσης, ατομικά οι καταναλωτές δημιουργούν και μοιράζονται μια τεράστια ποσότητα δεδομένων μέσω του blogging, των ενημερώσεων κατάστασης, και την ανάρτηση των φωτογραφιών τους. Μεγάλο μέρος των δεδομένων αυτών μπορεί τώρα να συγκεντρώνεται σε πραγματικό ή σχεδόν πραγματικό χρόνο. Η δυνατότητα πρόσβασης σε όλα τα δεδομένα αυτά και, σε ορισμένες περιπτώσεις, η δυνατότητα διαχείρισης των συνθηκών δημιουργίας τους, παρέχουν έναν πολύ

διαφορετικό τρόπο λήψης αποφάσεων, τον οποίο εισάγει πιο πολύ η επιστήμη στη Διοίκηση. Ένας οργανισμός, που είναι προσανατολισμένος στα δεδομένα λαμβάνει αποφάσεις με βάση τα εμπειρικά αποτελέσματα, και τα οφέλη μιας τέτοιας προσέγγισης έχουν αποδειχθεί και από την ακαδημαϊκή έρευνα. Οι ηγέτες σε πολλούς τομείς έχουν ήδη αρχίσει να χρησιμοποιούν ελεγχόμενες έρευνες για τη λήψη καλύτερων αποφάσεων. Για παράδειγμα, στον τομέα της υγείας εκπονούνται μελέτες συγκριτικής αξιολόγησης αποτελεσματικότητας σε ολόκληρο τον πληθυσμό, καθώς εντοπίζονται επαρκή κλινικά δεδομένα για τον εντοπισμό και την κατανόηση των πηγών της μεταβλητότητας σε θεραπείες και αποτελέσματα και έτσι βοηθιούνται οι υπεύθυνοι για τη λήψη αποφάσεων στη χάραξη κατευθυντήριων γραμμών, που εξασφαλίζουν ότι οι αποφάσεις για τη θεραπεία βασίζονται στην ορθότερη επιστήμη. Οι πωλητές, κυρίως εκείνοι, που δραστηριοποιούνται διαδικτυακά, προσαρμόζουν τις τιμές και τις προσφορές τους σε μια προσπάθεια εντοπισμού του βέλτιστου συνδυασμού κυκλοφορίας και πωλήσεων. Ωστόσο, δεν είναι πάντα δυνατόν η κατασκευή μιας ελεγχόμενης έρευνας και ο «χειρισμός» μια ανεξάρτητης μεταβλητής. Μια εναλλακτική είναι η εύρεση «φυσικών πειραμάτων», που εντοπίζουν την υπάρχουσα μεταβλητότητα στις μετρήσεις απόδοσης. Η κατανόηση των αιτιών αυτής της μεταβλητότητας μπορεί στη συνέχεια να συμβουλέψει τους υπευθύνους διαχείρισης να λάβουν αποφάσεις και να βελτιώσουν την απόδοση. Στο δημόσιο τομέα, εντοπίζονται υπηρεσίες με τεράστιες αποκλίσεις στην παραγωγικότητα και την ακρίβεια του έργου, οι οποίες εκτελούν σχεδόν πανομοιότυπα καθήκοντα. Η γνωστοποίηση και μόνο αυτής της πληροφορίας μπορεί να έχει ως αποτέλεσμα σημαντική αύξηση απόδοσης στις υστερούσες υπηρεσίες και χωρίς χρηματικό αντίκρισμα ως κίνητρο. [5]

2.4 Κατάτμηση του πληθυσμού για την προσαρμογή δράσεων

Οι πολιτικές στοχευμένων υπηρεσιών ή του μάρκετινγκ για να ανταποκρίνονται στις ανάγκες των ατόμων είναι ήδη οικείες σε εταιρείες προσανατολισμένες προς την ιδιωτική κατανάλωση. Η ιδέα της κατάτμησης της αγοράς και της ανάλυσης των πελατών τους μέσω συνδυασμών χαρακτηριστικών όπως δημογραφικά στοιχεία, μετρήσεις αγορών πελατών, και

αγοραστικές συμπεριφορές είναι ευρέως καθιερωμένες. Επιχειρήσεις όπως οι ασφαλιστικές εταιρείες, οι οποίες βασίζονται σε αποφάσεις αβεβαιότητας έχουν χρησιμοποιήσει επί μακρόν μεγάλα δεδομένα για την τμηματοποίηση. Ωστόσο, καθώς η τεχνολογία εξελίσσεται, πολλές εταιρείες αποκτούν τη δυνατότητα να τμηματοποιούν και να αναλύουν σε πραγματικό χρόνο τα συλλεγόμενα δεδομένα. Ακόμη και στο δημόσιο τομέα, που η τάση είναι να αντιμετωπίζονται όλες οι δομές με τον ίδιο τρόπο, η χρήση μεγάλων δεδομένων για τμηματοποίηση αρχίζει να εφαρμόζεται. Για παράδειγμα, οι φορολογικές υπηρεσίες όπου οι φορολογούμενοι τμηματοποιούνται από μια σειρά παραγόντων όπως το εισόδημα, το ποσοστό φερεγγυότητάς τους και το πιστωτικό ιστορικό τους για την επιλογή μέσων κατάλληλων για περαιτέρω ελέγχους. [5].

2.5 Αντικατάσταση / υποστήριξη της λήψης αποφάσεων με αυτοματοποιημένους αλγόριθμους

Εξελιγμένα analytics μπορεί να βελτιώσουν σημαντικά τη λήψη αποφάσεων, την ελαχιστοποίηση της αβεβαιότητας, και την ανάδειξη πολύτιμων πληροφορίων. Τα μεγάλα δεδομένα παρέχουν την πρώτη ύλη, που απαιτείται είτε για την ανάπτυξη αλγορίθμων, είτε για τη λειτουργία τους. Για παράδειγμα, φορολογικές υπηρεσίες, που εφαρμόζουν και χρησιμοποιούν αυτοματοποιημένες μηχανές αβεβαιότητας που χρησιμοποιούν μεγάλα δεδομένα για τον εντοπισμό υποψηφίων, που χρήζουν περαιτέρω διερεύνησης. Οι αλγόριθμοι μεγάλων δεδομένων στον τομέα της λιανικής μπορούν να αριστοποιήσουν τις 16 διαδικασίες λήψης αποφάσεων, επιτρέποντας την αυτόματη ρύθμιση καταλόγων και τιμολογώντας σε πραγματικό χρόνο και σε καταστήματα και σε OnLine πωλήσεις. Οι κατασκευαστικές εταιρείες μπορούν να προσαρμόσουν τις γραμμές παραγωγής τους αυτόματα, για βελτιστοποίηση αποδοτικότητας, μείωση σπατάλης, και αποφυγή επικίνδυνων συνθηκών. Σε ορισμένες περιπτώσεις, εταιρείες δεν αυτοματοποιούν απαραίτητα τις αποφάσεις, αλλά τις διευκολύνουν μέσω της ανάλυσης των μεγάλων δεδομένων, που είναι πολύ περισσότερα από τα δεδομένα, που είναι διαχειρίσιμα από ένα άτομο χρησιμοποιώντας ένα υπολογιστικό φύλλο. Ορισμένοι οργανισμοί λαμβάνουν ήδη πιο αποτελεσματικές αποφάσεις αναλύοντας ολόκληρα σύνολα δεδομένων από πελάτες και

εργαζόμενους, ή ακόμα και από αισθητήρες ενσωματωμένους σε προϊόντα. [5] Καινοτόμα νέα επιχειρηματικά μοντέλα, προϊόντα και υπηρεσίες

Τα μεγάλα δεδομένα επιτρέπουν στις επιχειρήσεις όλων των ειδών την ανάπτυξη νέων προϊόντων και υπηρεσιών, την ενίσχυση των υφιστάμενων, και την εισαγωγή εντελώς νέων επιχειρηματικών μοντέλων. Στον τομέα της Υγείας, η ανάλυση των κλινικών δεδομένων και δεδομένων τη συμπεριφοράς των ασθενών έχει οδηγήσει σε προγράμματα προληπτικής φροντίδας, στοχευμένα στις κατάλληλες ομάδες ατόμων. Επίσης, στο λιανικό εμπόριο, οι υπηρεσίες σύγκρισης τιμών σε πραγματικό χρόνο δίνουν στους καταναλωτές πλήρη εικόνα των τιμών σε βαθμό, που ποτέ πριν δεν απολάμβαναν και δημιουργούν σημαντικό πλεόνασμα για αυτούς. Άλλες εταιρείες χρησιμοποιούν δεδομένα που λαμβάνονται από αισθητήρες ενσωματωμένους σε προϊόντα για τη δημιουργία καινοτόμων μετά την πώληση προσφορών υπηρεσιών, όπως η προληπτική συντήρηση και για τη δημιουργία βάσης για την ανάπτυξη της επόμενης γενιάς προϊόντων. [5]

2.6 Περιπτώσεις χρήσης μεγάλης κλίμακας δεδομένων

Υπάρχουν πολλά παραδείγματα περιπτώσεων χρήσης των μεγάλων κλίμακας δεδομένων σε κάθε βιομηχανία που μπορεί να φανταστεί κανείς. Ορισμένες επιχειρήσεις έχουν γίνει πιο δεκτικές στις τεχνολογίες και έχουν ενσωματώσει πιο γρήγορα την ανάλυση δεδομένων στην καθημερινότητα της επιχείρησης σε σχέση με άλλες. Αυτό είναι προφανές ότι οι επιχείρησης που αγκαλιάζουν την τεχνολογία όχι μόνο θα δουν σημαντικά πρωτοποριακά πλεονεκτήματα, αλλά θα είναι σημαντικά πιο ευέλικτες και πιο προσαρμοστικές στις προσφορές τους. Παραδείγματα χρήσης των μεγάλων κλίμακας δεδομένων περιλαμβάνουν:

• Οι χρηματοπιστωτικές υπηρεσίες υιοθετούν υποδομές ανάλυσης μεγάλων δεδομένων για να βελτιώσουν τις αναλύσεις των πελατών τους για το μετοχικό κεφάλαιο, ασφάλιση, υποθήκη, ή πίστωση. [6]

• Αεροπορικές εταιρείες και εταιρείες οδικών μεταφορών χρησιμοποιούν μεγάλης κλίμακας δεδομένων για να παρακολουθήσουν την κατανάλωση καύσιμων και τα πρότυπα

κυκλοφορίας στους στόλους τους σε πραγματικό χρόνο για να βελτιώσουν την αποτελεσματικότητα και την εξοικονόμηση κόστους.

• Οι υγειονομικής περίθαλψης υπηρεσίες διαχειρίζονται και κάνουν κοινή χρήση ηλεκτρονικών μητρώων ασθενών από πολλαπλές πήγες —εικόνες, θεραπείες, και δημογραφικά στοιχεία. Επιπλέον, οι φαρμακευτικές εταιρείες και οι ρυθμιστικοί οργανισμοί δημιουργούν λύσεις μεγάλης κλίμακας δεδομένων για την παρακολούθηση της αποτελεσματικότητας των φαρμάκων και για να παρέχουν πιο αποτελεσματική και πιο σύντομη ανάπτυξη φαρμάκων. [6]

Εταιρίες μέσων ενημέρωσης και ψυχαγωγίας αξιοποιούν τις υποδομές της μεγάλης κλίμακας δεδομένων για να βοηθήσουν με την λήψη αποφάσεων γύρω από τον πελάτη και για να παρέχει πιο εστιασμένο μάρκετινγκ. [6] Υπάρχουν περιπτώσεις χρήσης και συγκεκριμένα παραδείγματα των μεγάλης κλίμακας δεδομένων για κάθε βιομηχανία και εταιρεία . Ως εκ τούτου, έστω και αν αυτήν την περίοδο η επιχείρηση σας δεν χρησιμοποιεί λύσεις μεγάλων δεδομένων, είναι πιθανόν οι ανταγωνιστές σας να χρησιμοποιούν. Το πραγματικό ερώτημα είναι πως μπορείς να βελτιστοποιήσεις καλυτέρα το περιβάλλον σου ώστε να δημιουργήσεις μια πιο γρήγορη αποτελεσματική λύση που σου δίνει ανταγωνιστικό πλεονέκτημα.

2.7 Επισκόπηση των σημαντικότερων ορισμών για τα μεγάλα δεδομένα

Από το 2011 το ενδιαφέρον για το χώρο των μεγάλων δεδομένων έχει αυξηθεί με εκθετικό βαθμό. Σε αντίθεση με την συντριπτική πλειοψηφία των ερευνών σχετικά με την επιστήμη των υπολογιστών, τα Μεγάλα δεδομένα έλαβαν μεγάλη δημοσιότητα και ενδιαφέρον από τα μέσα ενημέρωσης. Τίτλοι όπως "Μεγάλα δεδομένα: το μεγαλύτερο αγαθό ή καταπάτηση της ιδιωτικότητας [7] " και "Τα Μεγάλα δεδομένα ανοίγουν πόρτες αλλά ίσως πάρα πολλές (Big Data Is Opening Doors, but Maybe Too Many, 2013)" λένε

πολλά ως προς την αντίληψη που επικρατεί για τα μεγάλα δεδομένα. Από την αρχή γίνεται σαφές ότι τα Μεγάλα δεδομένα σχετίζονται με σημαντικά τεχνικά αλλά και κοινωνικό-τεχνικά θέματα αλλά ο ακριβής ορισμός τους δεν είναι αρκετά σαφής. Πρόσφατη βιβλιογραφία που κάνει χρήση του όρου συναντάται σε πολλά και διαφορετικά πεδία με αποτέλεσμα την ύπαρξη πολλών, διφορούμενων και συχνά αντιφατικών ορισμών σχετικά με τον όρο Μεγάλα δεδομένα. Τα Μεγάλα δεδομένα σχετίζονται κυρίως με δύο ιδέες: την αποθήκευση δεδομένων και την ανάλυση δεδομένων. Σε αντίθεση με αυτό το ξαφνικό ενδιαφέρον για τα μεγάλα δεδομένα, οι έννοιες αυτές δεν είναι καινούργιες στον επιστημονικό κόσμο. Αυτό στο στοιχείο ωστόσο, αναδεικνύει το ερώτημα του πώς τα Μεγάλα δεδομένα θεωρούνται σημαντικά διαφορετικά από τις τυπικές τεχνικές επεξεργασίας δεδομένων. Δεν χρειάζεται ιδιαίτερη διορατικότητα για να καταλάβουμε ότι για να βρούμε την απάντηση σε αυτό το ερώτημα πρέπει απλώς να εξεταστεί περαιτέρω ο όρος μεγάλο δεδομένα. Ο όρος "Μεγάλα" υποδηλώνει σημαντικότητα, πολυπλοκότητα και πρόκληση. Δυστυχώς όμως ο όρος "Μεγάλα" περιέχει ποσοτικό χαρακτηριστικό και εδώ έγκειται η δυσκολία για την εξαγωγή ενός ορισμού. Ένας εκ των πιο διαδεδομένων ορισμών περιλαμβάνεται σε έκθεση του Meta (σήμερα Gartner) το 2001 (3d data management:Controlling data volume,velocity and variety, 2001). Η έκθεση της Gartner δεν κάνει καμία αναφορά στη φράση «μεγάλα δεδομένα», ωστόσο, η έκθεση αυτή θεωρείτε βασικός ορισμός των Μεγάλα δεδομένα. Η Gartner πρότεινε έναν ορισμό που περιλάμβανε τα "τρία Vs (Volume, Velocity, Variety)": τον όγκο, την ταχύτητα και την ποικιλία. Πρόκειται για έναν ορισμό που εστιάζει στο μέγεθος. Η έκθεση επισημαίνει το αυξανόμενο μέγεθος των δεδομένων, το αυξανόμενο ποσοστό παραγωγής τους και το αυξανόμενο εύρος των μορφών που εφαρμόζονται. Όπως είναι σύνηθες στη βιβλιογραφία των μεγάλων δεδομένων, τα ευρήματα που παρουσιάζονται στον ορισμό της Gartner είναι εντελώς αποσπασματικά και δεν παρέχεται καμία ποσοτικοποίηση των μεγάλων δεδομένων. Ο ορισμός αυτός έχει επαναληφθεί από τη NIST (Nist Big Data program, 2013) και τη Gartner το 2012 (M.A Beyer and D.Laney, 2012) και διευρυνθεί από την IBM (IBM, 2013) για να συμπεριλάβει και ένα τέταρτο V: την πιστότητα (Veracity). Η Oracle υποστηρίζει ότι τα μεγάλα δεδομένα είναι το αποτέλεσμα από την ένταξη πρόσθετων πηγών δεδομένων για να αυξήσουν τις ήδη υπάρχουσες λειτουργίες. Αξίζει να σημειωθεί ότι ο ορισμός της Oracle εστιάζει στην υποδομή. Σε αντίθεση με ορισμούς που εκφράστηκαν από άλλους, η Oracle

δίνει έμφαση σε μια σειρά από τεχνολογίες όπως: NoSQL, Hadoop, HDFS, R και σχεσιακές βάσεις δεδομένων. Έτσι, παρείχαν και έναν ορισμό και μια λύση για τα μεγάλα δεδομένα. Παρόλο που ο ορισμός αυτός είναι σχετικά πιο εύκολο να υιοθετηθεί σε σχέση με άλλους, υστερεί ωστόσο στην ποσοτικοποίηση. Σύμφωνα με τον ορισμό της Oracle δεν είναι σαφές ως προς το πότε ακριβώς ο όρος μεγάλα δεδομένα εντοπίζεται στην πράξη και παρέχει περισσότερο μία έννοια ότι «θα τα καταλάβετε όταν τα δείτε". Η Intel είναι μία από τις λίγες επιχειρήσεις που παρέχουν ποσοτικά στοιχεία στη βιβλιογραφία τους. Η Intel συσχετίζει τα μεγάλα δεδομένα με οργανισμούς που "δημιουργούν κατά μέσο όρο 350 terabytes (TB) δεδομένων εβδομαδιαίως". Αντί να δώσει έναν ορισμό όπως έκαναν οι προαναφερθέντες οργανισμοί, περιγράφει τα μεγάλα δεδομένα ποσοτικοποιώντας τις εμπειρίες των επιχειρηματικών εταίρων της. Επισημαίνει ότι οι οργανισμοί οι οποίοι μελετήθηκαν ασχολούνται εκτενώς με μη-δομημένα δεδομένα και δίνουν έμφαση στη διεξαγωγή αναλύσεων των δεδομένων τους τα οποία παράγονται με ρυθμό 500 terabytes ανά εβδομάδα. Τέλος, ισχυρίζεται ότι ο πιο σύνηθες τύπος δεδομένων που συναντάται είναι οι επιχειρηματικές συναλλαγές που είναι αποθηκευμένες σε σχεσιακές βάσεις δεδομένων (σύμφωνα με τον ορισμό της Oracle), και ακολουθούν τα έγγραφα, τα e- mail, τα blogs και τα social media. Η Microsoft παρέχει ένα ιδιαίτερα περιεκτικό ορισμό: "Μεγάλα δεδομένα" είναι ο όρος που χρησιμοποιείται όλο και περισσότερο για να περιγράψει τη διαδικασία εφαρμογής σημαντικής υπολογιστικής ισχύς - την τελευταία λέξη της μηχανικής μάθησης και της τεχνητής νοημοσύνης - σε μαζικά και εξαιρετικά πολύπλοκα σύνολα πληροφοριών.Ο ορισμός αυτός καθιστά σαφές ότι τα μεγάλα δεδομένα απαιτούν σημαντική υπολογιστική ισχύ. Η σημασία της υπολογιστικής ισχύς αναφέρθηκε και σε προηγούμενους ορισμούς, αλλά δεν ορίστηκε με ακρίβεια. Επιπλέον, ο ορισμός αυτός εισάγει δύο τεχνολογίες: την μηχανική μάθηση και την τεχνητή νοημοσύνη που είχαν αγνοηθεί από προηγούμενους ορισμούς. Αυτό, ως εκ τούτου, εισάγει την ιδέα ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες που είναι ζωτικής σημασίας συστατικά του τελικού ορισμού. 20 Η Google Trends αναφέρει τους ακόλουθους όρους σε σχέση με τα μεγάλα δεδομένα: ανάλυση δεδομένων, Hadoop, NoSQL, Google, IBM, και Oracle. Από αυτούς τους όρους μια σειρά από τάσεις είναι εμφανείς. Πρώτον, ότι τα μεγάλα δεδομένα είναι άρρηκτα συνδεδεμένα με την ανάλυση δεδομένων και την εξαγωγή γνώσης του από τα δεδομένα. Δεύτερον, είναι σαφές ότι υπάρχουν μια σειρά από σχετιζόμενες τεχνολογίες όπως φαίνεται

και από τον το ορισμό της Microsoft, δηλαδή τις NoSQL και Apache Hadoop. Τέλος, είναι προφανές ότι υπάρχει ένας αριθμός οργανισμών, κυρίως βιομηχανικών οργανισμών που σχετίζονται με μεγάλα δεδομένα. Όπως επισημαίνεται από το Google Trends, υπάρχουν μια σειρά από τεχνολογίες που συχνά αναφέρονται ότι εμπλέκονται με τα μεγάλα δεδομένα. Αποθήκες δεδομένων NoSQL όπως Amazon, Dynamo, Cassandra, Couch DB, Mongo DB κ.ά. παίζουν κρίσιμο ρόλο στην αποθήκευση μεγάλου όγκου μη δομημένων και ιδιαίτερα μεταβαλλόμενων δεδομένων. Για τη χρήση των χώρων αποθήκευσης δεδομένων NoSQL υπάρχει μια σειρά εργαλείων ανάλυσης και μέθοδο, συμπεριλαμβανομένων των MapReduce, NLP, στατιστικού προγραμματισμού, της μηχανικής μάθησης και την οπτικοποίηση πληροφοριών. Η εφαρμογή μίας από αυτές τις τεχνολογίες από μόνη της δεν είναι επαρκής για να αξιολογήσει τη χρήση του όρου μεγάλα δεδομένα. Αντίθετα, άλλες τάσεις δείχνουν ότι είναι ο συνδυασμός μιας σειράς τεχνολογιών και η χρήση σημαντικών συνόλων δεδομένων που εξηγούν τον όρο. Οι τάσεις αυτές δείχνουν τα μεγάλα δεδομένα σαν μια τεχνική κίνηση η οποία ενσωματώνει ιδέες, νέες και παλιές και σε αντίθεση με άλλους ορισμούς παρέχει λίγες αναφορές ως προς τις κοινωνικές και επιχειρηματικές επιπτώσεις. Καθώς οι προαναφερθέντες ορισμοί βασίζονται σε ένα συνδυασμό μεγέθους, πολυπλοκότητας και τεχνολογίας, ένα λιγότερο κοινός ορισμός βασίζεται μόνο στην πολυπλοκότητα. Η μέθοδος για ένα Ολοκληρωμένο σε Γνώση

Περιβάλλον- αποδίδεται συχνά στην κοινότητα ανοικτού κώδικα, εισάγοντας μια αντιφατική ιδέα: "Τα Μεγάλα δεδομένα μπορεί να είναι πολύ μικρά και δεν είναι όλα τα σύνολα δεδομένων μεγάλα" (Roberd Hillard, 2012). Αυτό είναι ένα επιχείρημα υπέρ της πολυπλοκότητας και υπέρ της άποψης ότι το μέγεθος δεν είναι το κυρίαρχο στοιχείο.

Δεδομένης της συνεχώς αναπτυσσόμενης φύσης της επιστήμης των υπολογιστών ο ορισμός αυτός δεν είναι τόσο πολύτιμος όσο μπορεί να φαινόταν αρχικά. Ο ισχυρισμός ότι τα μεγάλα δεδομένα είναι δεδομένα που αμφισβητούν και προκαλούν τις υφιστάμενες πρακτικές δεν είναι κάτι νέο. Ο ορισμός αυτός υποδηλώνει ότι τα δεδομένα είναι "μεγάλα" σε σχέση με το ισχύον πρότυπο υπολογισμού. Η εφαρμογή πρόσθετων υπολογισμών ή ακόμη η ανάπτυξη των υπαρχόντων υπόσχεται να συρρικνώσει τα μεγάλα δεδομένα. Ο ορισμός αυτός μπορεί να χρησιμεύσει μόνο ως ένα σύνολο από συνεχώς ανανεωνόμενων στόχων και υποστηρίζει ότι τα μεγάλα δεδομένα υπήρχαν ανέκαθεν, και πάντα θα υπάρχουν. Παρά το εύρος και τις διαφορές που υπάρχουν σε καθένα από τους προαναφερθέντες

ορισμούς υπάρχουν μερικά σημεία ομοιότητας. Αξίζει να σημειωθεί ότι όλοι οι ορισμοί κάνουν τουλάχιστον ένα από τα παρακάτω ισχυρισμούς:

Μέγεθος: ο όγκος των συνόλων δεδομένων είναι ένας κρίσιμος παράγοντας. Πολυπλοκότητα: η δομή, η συμπεριφορά και οι μεταθέσεις των συνόλων δεδομένων είναι ένας κρίσιμος παράγοντας. Τεχνολογίες: τα εργαλεία και οι τεχνικές που χρησιμοποιούνται για την επεξεργασία ενός μεγάλου και πολύπλοκου συνόλου δεδομένων είναι ένας κρίσιμος παράγοντας. Οι ορισμοί που αναφέραμε παραπάνω περιλαμβάνουν τουλάχιστον έναν από αυτούς τους παράγοντες και πολλοί από αυτούς περιλαμβάνουν δύο. Μια προέκταση αυτών των παραγόντων θα ήταν επομένως να υποθέσουμε τα εξής:

Μεγάλα δεδομένα είναι ένας όρος που περιγράφει την αποθήκευση και ανάλυση μεγάλων ή πολύπλοκων συνόλων δεδομένων χρησιμοποιώντας μια σειρά από τεχνικές που περιλαμβάνουν, αλλά δεν περιορίζονται στις ακόλουθες: NoSQL, MapReduce και μηχανές μάθησης.

ΤΟ ΠΡΟΤΥΠΟ ΤΩΝ 3V

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, ένας από τους πιο γνωστούς ορισμούς για τα Μεγάλα Δεδομένα διατυπώθηκε πολύ εύστοχα από την Gartner με βάση το αποκαλούμενο πρότυπο 3V. Σύμφωνα με αυτή την προσέγγιση, τρία είναι τα κύρια χαρακτηριστικά των μεγάλων δεδομένων: ο όγκος (Volume), η ταχύτητα (Velocity) και η ποικιλομορφία (Variety). Ο όγκος αναφέρεται στο μέγεθος των διαθέσιμων δεδομένων, η ποικιλία αναφέρεται στο μεγάλο εύρος διαφορετικών τύπων δεδομένων που πρέπει να διαχειριστούμε και η ταχύτητα αναφέρεται στον ρυθμό που τα δεδομένα παράγονται και επεξεργάζονται. Όγκος (Volume): Τα τελευταία χρόνια παρατηρούμε μια εντυπωσιακή αύξηση του όγκου των δεδομένων που καλούμαστε να αποθηκεύσουμε. Διαμορφώνεται ένα νέο περιβάλλον όπου δεν κυριαρχούν πια τα δεδομένα κειμένου. Αυτή η νέα πραγματικότητα, χαρακτηρίζεται από τεράστιες ποσότητες δεδομένων σε μορφή video, ήχου και εικόνων όχι μόνο επιστημονικής προελεύσεως αλλά έχοντας αξιοσημείωτη πηγή τα μέσα κοινωνικής δικτύωσης. Είναι απαραίτητο πια στοιχείο για έναν οργανισμό επιχείρηση να έχει στη διάθεσή του μεγάλο αποθηκευτικό χώρο. Εκ των πραγμάτων, η τεράστια αύξηση

των δεδομένων, επιτάσσει την επανεξέταση εφαρμογών και αρχιτεκτονικών καθώς οι τυπικές μέθοδοι αποδεικνύονται ανεπαρκείς. Ξεπερνώντας σιγά σιγά τα προβλήματα εύρεσης επαρκούς χώρου αποθήκευσης, νέα ζητήματα αναδύονται όπως η ανάγκη συσχέτισης των μεγάλων δεδομένων και η δυνατότητα αλίευσης πολύτιμης αξίας.

Ταχύτητα (Velocity): Καθώς τα δεδομένα ρέουν με καταιγιστικό ρυθμό αναδύεται η ανάγκη άμεσης αντίδρασής μας σε αυτά. Η ταχεία αντίδραση μας ώστε να αντιμετωπίσουμε την ταχύτητα των δεδομένων αποτελεί ιδιαίτερη πρόκληση για τους περισσότερους οργανισμούς. Η ταχύτητα αναφέρεται στον πολύ γρήγορο ρυθμό εμφάνισης νέων δεδομένων αλλά και ανανέωσης των υπαρχόντων. Επίσης, σχετίζεται με τον αναγκαίο χρόνο επεξεργασίας των εισερχομένων στο σύστημα δεδομένων μέσω προηγμένων εφαρμογών, με την ανάλυση των δεδομένων αυτών, τον εντοπισμό των σχέσεων μεταξύ δεδομένων και την εξαγωγή πληροφοριών από τα δεδομένα μέσω συσχετίσεων και συμπερασμάτων. Αναφορικά με το ρυθμό εμφάνισης νέων δεδομένων, ο φόρτος εργασίας είναι δοσοληψίες (OLTP) και το πρόβλημα είναι πώς το σύστημα θα υποδεχθεί, θα φιλτράρει, θα διαχειριστεί και θα αποθηκεύσει τα δεδομένα που ρέουν με πολύ γρήγορο ρυθμό. Τα συμβατικά συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων αδυνατούν να καλύψουν τις ανάγκες, αφού αναπόφευκτα επεξεργάζονται πολύ μεγάλη επιβάρυνση στα δεδομένα για λόγους κλειδώματος, logging, και latching σε πολυνηματικές εφαρμογές.

Ο ρυθμός ανανέωσης των υπαρχόντων, σχετίζεται με το χρόνο που χρειάζεται ώστε να αντλήσουμε πληροφορία από τα εισερχόμενα δεδομένα (stream analysis / mining). Αξίζει να σημειωθεί ότι δεν είναι αρκετό να μπορούμε να αναλύσουμε τα δεδομένα και να αντλήσουμε πληροφορία σε πραγματικό χρόνο, αλλά είναι εξίσου σημαντικό να εκτελούμε και λειτουργίες που ενεργοποιούνται από αυτά, σε πραγματικό χρόνο. Μια εφαρμογή παρακολούθησης των τιμών του χρηματιστηρίου θα ήταν επιθυμητό όχι απλώς να καταγράφει τις διακυμάνσεις των τιμών αλλά και να παρέχει τη δυνατότητα αγοραπωλησίας μιας μετοχής.

Ποικιλομορφία (Variety): Το τρίτο γνώρισμα των μεγάλων δεδομένων είναι η ποικιλομορφία. Σήμερα καλούμαστε να αποθηκεύσουμε, να συνδυάσουμε και να επεξεργαστούμε δεδομένα από πολλές διαφορετικές πηγές π.χ κινητά δίκτυα επικοινωνιών,

tablets, κάμερες, κοινωνικά δίκτυα, εταιρείες εμπορίας πληροφοριών (data brokers) κτλ. Τα δεδομένα αυτά εισρέουν σε οποιαδήποτε μορφή, δομημένα δεδομένα, αριθμητικά δεδομένα αποθηκευμένα σε παραδοσιακές βάσεις, πληροφορίες που δημιουργούνται από εμπορικές εφαρμογές, αδόμητα έγγραφα κειμένου, email, video, ήχου, δεδομένα χρηματιστηριακών συναλλαγών και εμπορικών συναλλαγών. Έτσι, έχουμε να κάνουμε όχι μόνο με διαφορετικούς τύπους δεδομένων, αλλά και με διαφορετική δομή μεταξύ ίδιων τύπων. Γεννάται έτσι η απαίτηση να ενσωματωθούν δεδομένα με αυστηρή δόμηση (structured), ημιδομημένα (semi-structured) και αδόμητα (unstructured). Ακολούθως, ακόμα και αν οι πηγές μας χρησιμοποιούν αυστηρή δόμηση των δεδομένων, πιθανόν να είναι ετερογενή, δηλαδή η δόμηση της μιας να μην είναι συμβατή με κάποια άλλη, να χρησιμοποιούν διαφορετική σημασιολογία κλπ. Συνεπώς προκύπτει ότι τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων που απαιτούν αυστηρή δομή στα δεδομένα τους, δεν μπορούν να ανταποκριθούν σε αυτές τις νέες προκλήσεις.

Εγκυρότητα (Veracity): Πρόσφατα, πολλοί ερευνητές τονίζουν ολοένα και περισσότερο μια ακόμα πτυχή προσθέτοντας ένα τέταρτο «v»: την έννοια veracity (πιστότητα, εγκυρότητα). Η προσέγγιση αυτή δεν σχετίζεται τόσο με τα ιδιαίτερα χαρακτηριστικά των μεγάλων δεδομένων αλλά κυρίως με το πώς πρέπει να πραγματοποιείται η χρήση τους έτσι ώστε να επιτυγχάνεται ο αναγκαίος βαθμός εμπιστοσύνης για την αξιοπιστία των δεδομένων. Το να διαθέτεις πολλά δεδομένα σε διαφορετικούς όγκους ρέοντα σε υψηλές ταχύτητες δεν ωφελεί σε κάτι αν τα δεδομένα δεν είναι ακριβή. Εσφαλμένα δεδομένα μπορούν να προκαλέσουν πολλά προβλήματα και στους οργανισμούς όσο και στους καταναλωτές.

Συνεπώς, οι οργανισμοί πρέπει να διασφαλίζουν ότι τα δεδομένα είναι σωστά, καθώς και ότι οι αναλύσεις που πραγματοποιήθηκαν στα δεδομένα είναι σωστές. Ειδικά σε αυτοματοποιημένες διαδικασίες λήψης αποφάσεων, όπου κανένας άνθρωπος δεν συμμετέχει πια, πρέπει να υπάρχει σιγουριά ότι τόσο τα στοιχεία και οι αναλύσεις είναι σωστές. Η κύρια υπόσχεση των μεγάλων δεδομένων είναι η λήψη καλύτερων αποφάσεων βασιζόμενοι σε δεδομένα. Η ιδέα φαίνεται ελκυστική αλλά υπάρχει μια προειδοποίηση: είναι τα στοιχεία αρκετά αξιόπιστα για να στηρίξουν τις αποφάσεις μας πάνω σε αυτά; σε ποιο βαθμό

μπορούμε να εμπιστευθούμε τα δεδομένα; Πολλές φορές το μεγαλύτερο ποσοστό του χρόνου ενός έργου αφορά τη διαλογή και καθαρισμό των δεδομένων. Το θέμα είναι ότι τα περισσότερα δεδομένα στην εποχή των μεγάλων δεδομένων είναι αβέβαια.

Αξία (Value): Επίσης, υπάρχει μια ακόμα παράμετρος η οποία μπορεί να ληφθεί υπόψη κατά την εξέταση μεγάλων δεδομένων, ένα πέμπτο V: Η Αξία (Value). Η πρόσβαση σε μεγάλα δεδομένα αν δε μπορούμε να τη μετατρέψουμε σε αξία είναι άχρηστη. Για αυτό πολλοί υποστηρίζουν ότι η αξία είναι το πιο σημαντικό V των μεγάλων δεδομένων. Οι οργανισμοί καλούνται να επιλέξουν την πιο αποτελεσματική από πλευράς κόστους λύση με στόχο την αξιοποίηση της πληροφορίας που θα οδηγεί στην έγκαιρη και όσο το δυνατό πιο σωστή λήψη αποφάσεων, δίνοντας το μεγαλύτερη δυνατόν αξία στην επιχείρηση.

2.7.1 Τεχνολογίες μεγάλων δεδομένων

Η αναγκαιότητα διαχείρισης των δεδομένων σε εφαρμογές μεγάλων δεδομένων επέβαλε τη δημιουργία μίας νέας γενιάς συστημάτων, μοντέλων και προγραμματιστικών εργαλείων όπως: Map Reduce, Hadoop και οικοσύστημα αυτού, NoSQL, κ.α., τεχνολογίες που επιτρέπουν την παράλληλη επεξεργασία δεδομένων σε μεγάλη κλίμακα και με ανεκτικό στα σφάλματα τρόπο [9].

Το MapReduce αποτελεί το σπουδαιότερο εργαλείο που έχει αναπτυχθεί για την ανάλυση μεγάλων δεδομένων. Είναι ένα προγραμματιστικό μοντέλο, μαζί με τη σχετική υλοποίηση για δημιουργία και επεξεργασία πολύ μεγάλων συνόλων δεδομένων. Αναπτύχθηκε στη Google από τους Jeffrey Dean και Sanjay Ghemawat το 2004 [8] Έναυσμα για τη δημιουργία του ήταν το μεγάλο πλήθος υπολογισμών που εκτελούνταν ημερησίως στη Google σε πολύ μεγάλο όγκο εισερχόμενων δεδομένων. Εξαιτίας του τεράστιου αριθμού χρηστών, ο όγκος των εισερχόμενων δεδομένων επίτασσε τη χρήση κατανεμημένων συστημάτων με εκατοντάδες ή και χιλιάδες υπολογιστές ώστε να είναι εφικτό η επεξεργασία να ολοκληρωθεί μέσα σε λογικά χρονικά πλαίσια. Το MapReduce είναι ένα απλό αλλά συγχρόνως πολύ δυνατό πλαίσιο το οποίο κάνει χρήση ενός αλγορίθμου ο οποίος παραλληλοποιεί και κατανέμει το σύνολο του όγκου που πρόκειται να επεξεργαστεί, μοιράζοντας κομμάτια του σε πολλούς υπολογιστές για επεξεργασία. Σε πρώτη φάση

μοιράζει το σύνολο του όγκου των εργασιών σε πολλαπλούς υπολογιστές, οι οποίοι εκτελούν τα κομμάτια που τους αναθέτουν ταυτόχρονα (φάση map) και ακολούθως όλα τα αποτελέσματα συγκεντρώνονται και αναλύονται συνολικά πριν επιστραφούν (φάση reduce). Το MapReduce ουσιαστικά επιτρέπει στον προγραμματιστή να εκτελεί τις απαιτούμενες εργασίες γράφοντας 2 συναρτήσεις: τη συνάρτηση map και τη συνάρτηση reduce.

Η επίλυση των προβλημάτων υλοποιείται σε 2 στάδια. Η συνάρτηση map δέχεται σαν είσοδο ένα ζεύγος κλειδί-τιμή και παράγει σαν έξοδο ένα ζεύγος κλειδί-τιμή. Η έξοδος της συνάρτησης map, ταξινομημένη με βάση το κλειδί, είναι η είσοδος της συνάρτησης reduce. Η συνάρτηση reduce εκτελείται μετά την συνάρτηση map. Η συνάρτηση reduce παίρνει σαν είσοδο την έξοδο της συνάρτησης map στην μορφή κλειδί- ενδιάμεσες τιμές και την επεξεργάζεται. Συνήθως για κάθε κλειδί έχουμε μία τιμή στην έξοδο. Για την επίλυση κάποιου προβλήματος με το Map Reduce, ο προγραμματιστής πρέπει να υλοποιήσει τουλάχιστον την συνάρτηση map. Κάποιες απλές εργασίες μπορούν να υλοποιηθούν μόνο με την χρήση της συνάρτησης map. Το Hadoop είναι ένα λογισμικό ανοιχτού κώδικα που υποστηρίζει κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων (petabytes) και παρέχει μια υλοποίηση του MapReduce.

Το Hadoop βασίστηκε στο Google Map Reduce framework και το Google File System (GFS). Είναι ένα έργο του Apache Software Foundation που αναπτύσσετε και χρησιμοποιείτε από ανθρώπους από όλο τον κόσμο και κυρίως την Yahoo!. Σήμερα είναι η πιο διαδεδομένη υλοποίηση του MapReduce και χρησιμοποιείται για διδακτικούς σκοπούς σε αρκετά πανεπιστήμια του κόσμου, αλλά και σε μεγάλους οργανισμούς ανά το παγκόσμιο για την επεξεργασία μεγάλων δεδομένων εισόδου. Κάποιοι από τους οργανισμούς, που διατηρούν clusters για εκτέλεση Hadoop εργασιών είναι: Yahoo!, Amazon, AOL, Alibaba, Cornell University Web Lab, ETH Zurich Systems Group, Facebook, Google, IBM, New York Times κ.α. Το Hadoop κατανέμει τα δεδομένα και την ανάλυσή τους σε ομάδες υπολογιστών (Clusters) ώστε να επεξεργαστούν ταυτόχρονα τα δεδομένα, εξοικονομώντας έτσι χρόνο και πόρους. Πιο συγκεκριμένα, προωθεί τα δεδομένα και το πρόβλημα στον υπολογιστή master της ομάδας, και αυτός στη συνέχεια θα κατακερματίσει το πρόβλημα σε μικρότερα προβλήματα και κάθε νέο πρόβλημα θα το προωθήσει με τη σειρά του σε κάθε ένα από τους υπόλοιπους υπολογιστές της ομάδας. Κάθε υπολογιστής της ομάδας θα επιλύσει το δικό του μικρότερο πρόβλημα και θα επιστρέψει τη λύση στον master

υπολογιστή ο οποίος θα συνδυάσει τις λύσεις των υποπροβλημάτων για να βρει τη λύση στο αρχικό πρόβλημα.

Το Hadoop στηρίζεται σε αυτή την διαδικασία ενώ επιπλέον προτέρημα είναι ότι επιτυγχάνει να ανακτήσει δεδομένα σε περίπτωση που ένας υπολογιστής της ομάδας πάθει ζημιά και να μεταβιβάσει το υποπρόβλημα σε άλλον υπολογιστή. Τα σημαντικότερα πλεονεκτήματα του Hadoop μπορούν να συνοψιστούν ως εξής: Επεκτασιμότητα: Δυνατότητα αξιόπιστης αποθήκευσης και επεξεργασίας μέχρι και petabytes δεδομένων Οικονομία Πόρων: Κατανομή δεδομένων και επεξεργασίας σε ομάδες υπολογιστών που αποτελούνται από έως και χιλιάδες κοινούς υπολογιστές. Αποδοτικότητα: Με την κατανομή των δεδομένων, η επεξεργασία γίνεται ταυτόχρονα σε όλους τους κόμβους, παρέχοντας ταχεία εκτέλεση των εργασιών. Κριτική στο Hadoop *Παρά τα πλεονεκτήματά του και τη θόρυβο που γίνεται γύρω από αυτό, το Hadoop δεν σημειώνει εξίσου υψηλό βαθμό δημοτικότητας από όλους τους επιστήμονες δεδομένων. Στην πράξη, δεν είναι λίγοι εκείνου που το χρησιμοποίησαν και το εγκατέλειψαν. Σε έρευνα για τα εμπόδια των big data analytics, το Paradigm σημειώνει ότι το 76% των επιστημόνων δεδομένων που χρησιμοποιούσαν Hadoop, απάντησαν ότι έχει «σημαντικούς περιορισμούς». Μία αιτία απογοήτευσης είναι το κόστος του Hadoop.*

Πολλοί οργανισμοί επιλέγουν το Hadoop διότι θεωρούν ότι θα εξοικονομήσουν χρήματα επειδή είναι ανοικτού κώδικα, ωστόσο διαπιστώνουν το αντίθετο. Στην πραγματικότητα αναγκάζονται να πληρώνουν για επιπλέον υπηρεσίες της Hadoop και για πρόσληψη προγραμματιστών και αναλυτών. Οι οργανισμοί που δοκίμασαν το Hadoop και αντιμετωπίσανε προβλήματα ενδεχομένως να αποδειχθούν τα πρώτα θύματα του πρώτου κύματος της μόδας του Hadoop. Η σταδιακή ωρίμανση των μεγάλων δεδομένων και της τεχνολογίας analytics, ταυτόχρονα με καλύτερα εκπαιδευμένους χρήστες, θα καταστήσουν πιο εφικτή τη δυνατότητα εξεύρεσης καλύτερης λύσης για analytics.

Το Hadoop αποτελείται από τα εξής βασικά δομικά στοιχεία:

• Το **Hadoop Common Utilities** που περιέχει βασικές βιβλιοθήκες και λειτουργίες που απαιτούνται από τα υπόλοιπα στοιχεία.

• Το **Hadoop Distributed File System** (**HDFS**) που διαχειρίζεται την αποθήκευση κατανεμημένων δεδομένων.

• Το **Hadoop YARN Framework**, το οποίο αποτελεί μία πλατφόρμα διαχείρισης πόρων. Ουσιαστικά, είναι υπεύθυνο για τη διαχείριση των υπολογιστικών πόρων σε συστάδες και για τον προγραμματισμό των εφαρμογών των χρηστών.

• Το **Hadoop Map-Reduce** που αποτελεί υλοποίηση του μοντέλου Map-Reduce για κατανεμημένη επεξεργασία μεγάλης κλίμακας δεδομένων.

Ένα υπολογιστικό σύστημα που εκτελεί την εφαρμογή Hadoop αποτελείται από υπολογιστικές συστάδες **(clusters**) οι οποίες απαρτίζονται από εμπορικό υλικό (**commodity hardware**). Η δομή του Hadoop βασίζεται στην υπόθεση ότι οι αστοχίες υλικού (hardware failures), δηλαδή οι δυσλειτουργίες στα ηλεκτρονικά στοιχεία των υπολογιστικών συστημάτων - είναι συχνές κατά τη διαχείριση μεγάλου όγκου δεδομένων και οφείλει η ίδια η εφαρμογή να τις διαχειρίζεται αποδοτικά.18

Ο πυρήνας του Hadoop αποτελείται από ένα τμήμα αποθήκευσης, γνωστό ως Hadoop Distributed File System (HDFS) και ένα τμήμα επεξεργασίας, που βασίζεται στο μοντέλο Map-Reduce που αναφέρθηκε προηγουμένως. Το Hadoop χωρίζει τα δεδομένα σε μεγάλα τμήματα (blocks) και τα κατανέμει μεταξύ διαφόρων υπολογιστικών κόμβων που συνιστούν το υπολογιστικό σύστημα. Στη συνέχεια, μεταφέρει τον κώδικα που πρόκειται να εκτελεστεί στους κόμβους ώστε να πραγματοποιηθεί παράλληλη, δηλαδή ταυτόχρονη επεξεργασία των δεδομένων στους κόμβους αυτούς. Ουσιαστικά, διενεργείται αξιοποίηση της ιδιότητας της τοπικότητας των δεδομένων (data locality) και οι κόμβοι διαχειρίζονται τα επιμέρους δεδομένα στα οποία έχουν πρόσβαση.19

Η χρήση της παραλληλίας, αποτελεί κλασσική προσέγγιση βελτίωσης της αποδοτικότητας εφαρμογών λογισμικού. Μάλιστα αποδεδειγμένα, η χρήση της τοπικότητας των δεδομένων κατ' αυτό τον τρόπο από εμπορικά συστήματα, παρέχει καλύτερα αποτελέσματα από αυτά που προσφέρουν εξελιγμένοι υπερυπολογιστές (supercomputers), οι οποίοι βασίζονται σε παράλληλα συστήματα αρχείων (parallel file systems), όπου ο υπολογισμός και τα δεδομένα διαμοιράζονται μέσω υψηλής ταχύτητας δικτύου.20

Τέλος, αξίζει να σημειωθεί πως αν και η βασική δομή του Hadoop συνίσταται από τα στοιχεία που ήδη αναφέρθηκαν, συχνά χρησιμοποιούνται επεκτάσεις από την Apache που εμπλουτίζουν τις δυνατότητες του Hadoop, αναλόγως την περίσταση, οι σημαντικότερες από τις οποίες είναι: Apache HBase, Apache Pig, Apache Hive, Apache Phoenix, Apache Spark, Apache ZooKeeper, Apache Flume, Apache Sqoop, Apache Storm.21

Apache Spark

Αποτελεί επίσης ελεύθερου κώδικα λογισμικό για επεξεργασία Big Data. Δημιουργήθηκε αρχικά στο Πανεπιστήμιο Berkeley, της California και στη συνέχεια παραχωρήθηκε αφιλοκερδώς στην Apache Software Foundation. Δημιουργήθηκε μετά το Hadoop και ουσιαστικά προσφέρει κάποια πλεονεκτήματα, τα οποία θα αναλυθούν στη συνέχεια. Το Spark προσφέρει στον προγραμματιστή μία διεπαφή (Interface) επικεντρωμένη σε μία δομή δεδομένων, γνωστή ως **Ελαστικό Κατανεμημένο Σύνολο Δεδομένων** (**Resilient Distributed Dataset** ή **RDD**) και πρόκειται για μια συλλογή κατανεμημένων αντικειμένων σε ένα σύνολο υπολογιστικών κόμβων η οποία διασφαλίζει αποτελεσματική διαχείριση αστοχιών υλικού, όπως ακριβώς το Hadoop.22

Ο λόγος δημιουργίας του Spark είναι ορισμένοι δομικοί περιορισμοί που επιβάλλονται από το μοντέλο Map-Reduce του Hadoop. Συγκεκριμένα, απαιτείται γραμμική ροή δεδομένων ως είσοδος από το δίσκο σε κατανεμημένα συστήματα, κατάλληλη επεξεργασία σύμφωνα με τις συναρτήσεις Map και Reduce και τέλος γραμμικού χρόνου αποθήκευση των δεδομένων στο δίσκο. Αντίθετα, το Spark παρέχει τη δυνατότητα πραγματοποίησης των υπολογισμών σε διαμοιραζόμενη μνήμη, όπου η ταχύτητα είναι σημαντικά μεγαλύτερη από την αντίστοιχη σε δίσκο. Με τον τρόπο αυτό, καθίσταται δυνατή η εφαρμογή επαναληπτικών αλγορίθμων που πραγματοποιούν πολλαπλές φορές πρόσβαση στα δεδομένα σε κάθε επανάληψη, χωρίς αυτό να συμβαίνει εις βάρος του χρόνου υπολογισμού, αφού ο χρόνος πρόσβασης σε δεδομένα μνήμης είναι ταχύτερος και «πλησιέστερα» στον επεξεργαστή των υπολογιστικών κόμβων.23 Σύμφωνα μάλιστα με πληροφορίες που βρίσκονται στην επίσημη ιστοσελίδα του Spark, είναι δυνατόν να εκτελεστούν εφαρμογές έως και 100 φορές ταχύτερα

στη μνήμη και έως 10 φορές ταχύτερα στο δίσκο, εν συγκρίσει με το Hadoop. Είναι δυνατόν μάλιστα να εκτελεστεί πάνω στον πυρήνα του Hadoop.

Το Apache Spark πέραν του τρόπου διαχείρισης των Big Data, προσφέρει τις εξής βασικές επεκτάσεις:

• **Spark SQL**: Επιτρέπει ερωτήματα (queries) σε δεδομένα με χρήση SQL, σε συνδυασμό με τις γλώσσες προγραμματισμού Java, Scala, Python και R.

• **Spark Streaming**: Καθιστά εφικτή την επεξεργασία δεδομένων σε ροή, δηλαδή δεδομένων που εισέρχονται στο σύστημα ενώ βρίσκονται ήδη σε εξέλιξη υπολογισμοί στα προηγούμενα δεδομένα. Αυτό το χαρακτηριστικό είναι πολύ σημαντικό, καθώς στο Hadoop δεν μπορούν να προστίθενται νέα δεδομένα κατά τη διάρκεια της επεξεργασίας, αλλά πρέπει να είναι διαθέσιμο όλο το σύνολό τους όταν εκκινεί μία Map-Reduce διαδικασία. Υποστηρίζονται οι γλώσσες προγραμματισμού Java, Scala και Python.

• **MLlib**: Πρόκειται για μία βιβλιοθήκη μηχανικής μάθησης (**Machine Learning Library**) η οποία δίνει τη δυνατότητα εκτέλεσης αλγορίθμων αυτού του είδους έως και 100 φορές ταχύτερα από το Hadoop.

• **GraphX**: Παρέχει ένα **API** (**Application Programming Interface**) για τα δεδομένα σε μορφή γραφημάτων, επιτρέποντας μάλιστα υπολογισμούς με χρήση επαναληπτικών αλγορίθμων με αποδοτικό τρόπο.

Talend

Το Talend είναι μια πλατφόρμα ανοικτού κώδικα, βασισμένη στο μοντέλο Hadoop που προσφέρει μια σειρά προϊόντων διαχείρισης Big Data. Το βασικότερο στοιχείο είναι το Master Data Manage-ment (MDM), το οποίο έχει τη δυνατότητα να επεξεργάζεται δεδομένα σε πραγματικό χρόνο, να αξιοποιεί άλλες εφαρμογές, να ενσωματώνει τα δεδομένα τους και να εκτελεί διάφορες διαδικασίες, όπως εκτιμήσεις της ποιότητας των Big Data.

Παρέχεται δωρεάν και προσφέρει αρκετές δυνατότητες, καθιστώντας το καλή επιλογή για πολλές ανάγκες.

Χρήση NoSQL Databases

Οι 10Gen, Cloudera και Amazon ήταν οι πρώτες εταιρίες που διαμόρφωσαν πλατφόρμες εκμετάλλευσης Big Data με δυνατότητα υποστήριξης του Apache Hadoop και τεχνολογιών για μη σχεσιακές βάσεις δεδομένων (**NoSQL Databases**), ενώ στην πορεία εμφανίστηκαν και άλλες, όπως οι Amazon, DataStax, Neo Technologies, Hortonworks, Platfora, 10Gen και CouchBase.

Οι σχεσιακές βάσεις δεδομένων, δομημένες βάσει της γλώσσας SQL, ήταν για πολλά χρόνια ο πιο δημοφιλής τρόπος διαχείρισης δεδομένων για οργανισμούς και επαγγελματίες στον τομέα της τεχνολογίας. Με την έλευση των Big Data, τα οποία χαρακτηρίζονται τόσο από το μεγάλο μέγεθος, όσο και από την ποικιλομορφία στη δομή τους, καθίσταται απαραίτητη η δυνατότητα επεξεργασίας δε-δομένων σε μεγάλη κλίμακα με σκοπό την εξαγωγή ενιαίων συμπερασμάτων. Στο πρόβλημα της διαχείρισης των δεδομένων αυτών, τα συστήματα που βασίζονται στην SQL δεν είναι δυνατόν να προσφέρουν αυτόνομα τη λύση.

Το πρόβλημα αντιμετωπίζεται μέσω της χρήσης NoSQL βάσεων δεδομένων (NoSQL databases), οι οποίες παρέχουν δυνατότητες δυναμικής διαχείρισης δεδομένων (dynamic data management), αυξη-μένης ευελιξίας (flexibility) και κλιμάκωσης (scalability) εν συγκρίσει με τις σχεσιακές βάσεις δεδο-μένων. Τα χαρακτηριστικά τους, τις καθιστούν ιδανικές για διαχείριση μεγάλου μεγέθους, μη-ομοιογενών δεδομένων τα οποία ανανεώνονται συχνά και πολλές φορές μεταβάλλονται οι τύποι (formats) των πεδίων (fields) των δεδομένων, πέραν των ίδιων των δεδομένων. Η διαδικασία αυτή είναι εξαιρετικά χρονοβόρα και σε πολλές περιπτώσεις ανέφικτη σε σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων.

Στο σημείο αυτό είναι απαραίτητο να διευκρινιστεί το γεγονός πως παρά τα οφέλη που παρέχουν οι NoSQL βάσεις δεδομένων, δεν είθισται να χρησιμοποιούνται αυτόνομα για τη διαχείριση Big Data. Όπως αναφέρει ο επιστήμονας μεγάλων δεδομένων (Big Data Scientist) William McKnight, «οι NoSQL βάσεις δεδομένων πρέπει να χρησιμοποιούνται

μόνο όταν δεν είναι δυνατή η χρήση σχεσιακών βάσεων», ώστε να βελτιστοποιείται η αξιοποίηση των δεδομένων.

2.8 Εργαλεία εκκαθάρισης των δεδομένων

Σε πολλές περιπτώσεις είναι χρήσιμο να πραγματοποιείται εκκαθάριση και μετατροπή των μη-δομημένων δεδομένων σε δομημένα, ιδιαίτερα όταν αυτά προέρχονται από πηγές του διαδικτύου και δια-θέτουν μεγάλη μορφολογική ποικιλία.

OpenRefine

Το OpenRefine είναι ένα εργαλείο ελεύθερου κώδικα το οποίο χρησιμοποιείται για την εκκαθάριση μεγάλων συνόλων δεδομένων. Παρέχει διάφορες επιλογές εκκαθάρισης (Reconciliation Services) και μία προγραμματιστική διεπαφή (Application Programming Interface) για τη μετατροπή των δεδομένων εισόδου σε πιο χρήσιμη και ευκολότερα αξιοποιήσιμη μορφή.

Data Cleaner

Το Data Cleaner χρησιμοποιείται για την εκκαθάριση των δεδομένων. Στηρίζεται στην παραδοχή ότι η διαχείριση των δεδομένων είναι χρονοβόρα και επίπονη διαδικασία.

Η λειτουργία του βασίζεται στα εξής ακολουθιακά βήματα:

- Εισαγωγή δεδομένων
- Συγχώνευση συνόλων δεδομένων

- Επανακατασκευή ελλιπών δεδομένων
- Προτυποποίηση
- Κανονικοποίηση
- Απαλοιφή διπλότυπων
- Επιβεβαίωση και Επέκταση
- Εξαγωγή δεδομένων

Εργαλεία Απεικόνισης

Η ποικιλία των Big Data σε συνδυασμό με τον τεράστιο όγκο τους, καθώς και όλες οι ιδιομορφίες που συνήθως τα συνοδεύουν, καταδεικνύουν την ανάγκη αποτελεσματικής απεικόνισής τους. Για το λόγο αυτό είναι ιδιαίτερα σημαντική η συνεισφορά των εργαλείων που παρέχουν αυτή τη δυνατότητα και κατατάσσονται στην κατηγορία Οπτικής Αναλυτικής (**Visual Analytics**). Μέσω αυτών τα Big Data γίνονται κατανοητά από τον ενδιαφερόμενο, δίχως να απαιτείται στις περισσότερες περιπτώσεις, τεχνική γνώση.

Ακολουθούν οι σημαντικότερες εφαρμογές που προσφέρονται:

Tableau

Το Tableau είναι ένα εργαλείο απεικόνισης δεδομένων με κύρια στόχευση την εύκολη δημιουργία διαγραμμάτων, χωρίς να είναι απαραίτητη η γνώση προγραμματισμού. Η μεγαλύτερη καινοτομία έγκειται στη δυνατότητα αξιοποίησης δεδομένων που βρίσκονται στο διαδίκτυο, χωρίς να είναι απαραίτητη η λήψη τους, μέσω μίας διεπαφής που παρέχει η εφαρμογή. Γενικά, θεωρείται η πιο ευέλικτη και πλούσια σε δυνατότητες εφαρμογή για την απεικόνιση στατιστικών στοιχείων και Big Data.

Υπάρχουν πέντε εκδόσεις της εφαρμογής με διαφορετικές λειτουργίες και δυνατότητες υποστήριξης. Για τους αρχάριους προτείνεται η έκδοση Tableau Public, η οποία παρέχεται δωρεάν και ευνοεί την εξοικείωση με την εφαρμογή και τη χρήση της.

CartoDB

Δίχως να προσφέρει τόσες δυνατότητες γραφημάτων, όσες το Tableau, το CartoDB είναι μία εφαρμογή η οποία δημιουργεί χάρτες και χρησιμοποιείται κυρίως για αναπαράσταση πληροφορίας με κριτήριο την τοπικότητα των φαινομένων. Ενδείκνυται για μελέτη Big Data, καθώς διαχειρίζεται πολλών ειδών δεδομένα και τύπους αρχείων, παρέχει τη δυνατότητα ενιαίων αποτελεσμάτων και διαθέτει πρότυπα σύνολα δεδομένων με τα οποία είναι εύκολο να εξοικειωθεί κανείς.

Σύμφωνα με αυτά, το CartoDB, αποτελεί εξαιρετική επιλογή για χρήση σε χάρτες.

Chartio

Το Chartio προσφέρει τη δυνατότητα συνδυασμού των πηγών δεδομένων και της εκτέλεσης ερωτημάτων (queries) στο πρόγραμμα περιήγησης. Το σημαντικότερο πλεονέκτημα της εφαρμογής αυτής είναι η ταχύτητα με την οποία μπορεί να λειτουργήσει, καθώς σε ελάχιστο χρόνο μπορεί να εισάγει δεδομένα από διαφορετικές πηγές και να τα αξιοποιήσει χωρίς να απαιτούνται γνώσεις SQL ή άλλων πολύπλοκων γλώσσες προγραμματισμού. Χρησιμοποιείται συνήθως για την εξαγωγή απλούστερων διαγραμμάτων.

Plot.ly

Το τελευταίο σχεδιαστικό εργαλείο που θα παρουσιαστεί είναι το Plot.ly, το οποίο είναι επίσης εύκολο στη χρήση, δίχως να απαιτεί τεχνικές γνώσεις, και μέσω αυτού γίνεται άμεσα να δημιουργηθούν δυσδιάστατα, αλλά και τρισδιάστατα γραφήματα, μία δυνατότητα που δεν είναι εξίσου εύκολο να υλοποιηθεί στις προηγούμενες εφαρμογές.

2.9 Γλώσσες Προγραμματισμού

Για την επεξεργασία και ανάλυση των Big Data, υπάρχει ανάγκη επιλογής των κατάλληλων γλωσσών προγραμματισμού. Οι κυριότερες γλώσσες που χρησιμοποιούνται στα Big Data Analytics είναι οι R, Python και Scala, ενώ σε κάποιες εφαρμογές χρησιμοποιούνται οι Java, RegEx και XPath.

Python

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού η οποία δημιουργήθηκε το 1990. Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λίγες γραμμές κώδικα. Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές εργασίες και για την ταχύτητα εκμάθησης της.

R

Η R είναι γλώσσα προγραμματισμού και περιβάλλον που παρέχει στον χρήστη τη δυνατότητα να κάνει υπολογιστική στατιστική και γραφήματα. Παρέχει τα απαραίτητα εργαλεία προκειμένου να υλοποιηθεί μια στατιστική ανάλυση, όπως:

- δημιουργία τυχαίων δειγμάτων
- διακριτές και συνεχείς μεταβλητές (Poisson, Gamma, Exponential)
- έλεγχοι υποθέσεων
- στατιστικά τεστ (Kolmogorov-Smirnoff)
- δημιουργία γραφημάτων (ιστόγραμμα, qq plot, pie chart, bar chart)

Scala

Η Scala είναι μια γλώσσα προγραμματισμού πολλαπλών παραδειγμάτων που σχεδιάστηκε για να ενσωματώσει χαρακτηριστικά του αντικειμενοστρεφούς και του

συναρτησιακού προγραμματισμού. Το όνομα Scala προέρχεται από την αγγλική φράση "scalable language", που δηλώνει ότι έχει σχεδιαστεί για να μπορεί να μεγαλώνει παράλληλα με τις ανάγκες των χρηστών της.

Java

Η γνωστή σε όλους μας πλέον Java είναι μια αντικειμενοστρεφής γλώσσα προγραμματισμού που σχεδιάστηκε από την εταιρεία πληροφορικής Sun Microsystems. Ένα από τα βασικά πλεονεκτήματά της έναντι των περισσότερων άλλων γλωσσών είναι η ανεξαρτησία του λειτουργικού συστήματος και της πλατφόρμας.

3. Τεχνικές για την ανάλυση των BIG DATA

Η αύξηση των δεδομένων που παράγονται κάθε χρόνο είναι τεράστια και οι εταιρίες προσπαθούν να συλλέξουν όσα περισσότερα μπορούν για την καλύτερη και αποτελεσματικότερη ανάλυσή τους. Όμως αυτός ο όγκος των δεδομένων έφερε στο προσκήνιο αδυναμία στην ανάλυσή τους καθώς δεν υπήρχαν οι τεχνικές εκείνες που θα μπορούσαν να τα αναλύσουν αποτελεσματικά, ούτε η υπολογιστική ισχύς και ακόμη και ο διαθέσιμος αποθηκευτικός χώρος δεν επαρκούσε. Έτσι από την πρώτη δεκαετία της νέας χιλιετίας, πολλές εταιρίες ασχολήθηκαν με την δημιουργία νέων τεχνικών και τεχνολογιών που θα μπορούσαν να αντιμετωπίσουν αυτήν την πρόκληση που προκύπτει από την τεράστια αύξηση των διαθέσιμων δεδομένων.

Οι τεχνικές και οι τεχνολογίες που αναπτύχθηκαν και αναπτύσσονται έχουν ως σκοπό τους την συγκέντρωση των δεδομένων από διαφορετικές πηγές, την ανάλυση των δεδομένων όσο γίνεται ταχύτερα και αποτελεσματικότερα και τέλος την οπτικοποίηση τους και την εξαγωγή συμπερασμάτων προς υλοποίηση. Αυτές οι τεχνολογίες και τεχνικές αναπτύχθηκαν από διάφορους κλάδους όπως αυτούς της στατιστικής, της επιστήμης των υπολογιστών, των εφαρμοσμένων μαθηματικών και των οικονομικών. Από αυτό προκύπτει ότι μια εταιρία για

να έχει το μέγιστο κέρδος από την ανάλυση των Big Data θα πρέπει να έχει εργαζομένους τόσο με προγραμματιστικές ικανότητες, όσο και με την ικανότητα να αντιλαμβάνονται στατιστικά και οικονομικά μοντέλα και όρους που θα προκύπτουν από την ανάλυση των datasets. Μερικές από αυτές τις τεχνικές και τεχνολογίες αναπτύχθηκαν από ακαδημαϊκούς, και άλλες από επιχειρήσεις που είχαν άμεσο συμφέρον από την εκμετάλλευση των Big Data, όπως εταιρίες με online συναλλαγές σαν την amazon. Κάποιες από αυτές τις τεχνικές και τις τεχνολογίες που αναπτύχθηκαν τα τελευταία χρόνια, αναπτύχθηκαν σε μια εποχή όπου υπήρχε πολύ μικρότερη πρόσβαση σε μεγάλες ποσότητες δεδομένων. Πολλές από αυτές όμως εξελίχθηκαν και έχουν προσαρμοστεί στα δεδομένα της εποχής με μεγάλη επιτυχία και μπορούν να εφαρμοστούν σε πολύ μεγάλα datasets.

Πολλά από τα εργαλεία που είναι πλέον στην διάθεση μας για να αναλύσουμε τα Big Data είναι εξειδικευμένα σε κλάδους και υπηρεσίες. Δηλαδή πολλές φορές οι τεχνικές πρέπει να παραμετροποιούνται ανάλογα με την εταιρία που τις χρησιμοποιεί και τον στόχο που έχει από τηνανάλυση των Big Data. Σε αυτό το κεφάλαιο θα αναλύσουμε μερικές τεχνικές και τεχνολογίες που εφαρμόζονται στις μέρες μας, αν και αυτές εξελίσσονται ή δημιουργούνται νέες που αντικαθιστούν τις υπάρχουσες καθώς γράφω αυτήν την εργασία. Στην συνέχεια θα παραθέσουμε τις σημαντικότερες τεχνικές και τεχνολογίες που υπάρχουν στις μέρες μας για την ανάλυση των Big Data.

Οι λόγοι που αναπτύχθηκαν τεχνικές για την καλύτερη αξιοποίηση των Big Data είναι:

- Για την παραγωγή αξιόπιστων αποτελεσμάτων από την ανάλυση ογκωδέστατων datasets.
- Για την γρηγορότερη ανάλυση τους.
- Για την πρόβλεψη μοτίβων συμπεριφορών και την αποφυγή λανθασμένων στρατηγικών.

Στη συνέχεια θα αναλύσουμε ορισμένες από αυτές τις τεχνικές που χρησιμοποιούνται στις μέρες μας. Πολλές από αυτές τις τεχνικές είναι αποτέλεσμα εξέλιξης προηγουμένων τεχνικών που χρησιμοποιούσαν παλιότερα οι εταιρίες όταν ο όγκος των δεδομένων ήταν μικρότερος, και προσαρμόστηκαν στις απαιτήσεις των Big Data. Καθώς γράφονται αυτές οι γραμμές, ολοένα και περισσότερες και περισσότερο αξιόπιστες τεχνικές αναδύονται καθώς

έχουν δοθεί πολλά κονδύλια για έρευνα στο συγκεκριμένο τομέα και νέες βελτιωμένες τεχνικές παράγονται συνεχώς. Οι περισσότερες από αυτές έχουν πολλά στοιχεία από την επιστήμη των υπολογιστών σε συνδυασμό με την επιστήμη της στατιστικής. Οι ποιο σημαντικές από αυτές είναι:

3.1 DATA MINING

Το Data Mining περιλαμβάνει μια σειρά από στατιστικά στοιχεία καθώς και την δυνατότητα να μάθει στον υπολογιστή να ξεχωρίζει μοτίβα μελετώντας υπάρχων datasets και μετά την ανάλυση μας παρέχει με πληροφορίες που θα ήταν αδύνατον να βρούνε οι εργαζόμενοι μόνοι τους χωρίς την ανάλυση [10]. Στην ιδανική περίπτωση, το Data Mining προβλέπει μοτίβα πελατών και παρέχει πληροφορίες προς αξιοποίηση στις ενδιαφερόμενες εταιρίες. Ουσιαστικά αυτό που μας παρέχει σαν γνώση είναι όχι 'ποια είναι η σχέση μεταξύ διαφημίσεων και πωλήσεων' αλλά 'ποια συγκεκριμένη διαφήμιση, ή συγκεκριμένο προϊόν πρέπει να δείξω σε έναν καταναλωτή που ψωνίζει στο διαδίκτυο εκείνη την στιγμή'. Ένα άλλο ενδιαφέρον στοιχείο είναι ότι πέρα από ατομικές προβλέψεις, κατηγοριοποιεί τους καταναλωτές ανάλογα με τις καταναλωτικές συναλλαγές που είχαν στο παρελθόν, και έχοντας αυτήν την πληροφορία μια εταιρία προσαρμόζει σε κάθε group καταναλωτών διαφορετική στρατηγική στο marketing.

Με το Data Mining εφαρμόζουμε μια τεχνική που χρησιμοποιούσαν ήδη οι οικονομολόγοι, οι στατιστικολόγοι, οι μετεωρολόγοι και αυτή αφορούσε την ιδέα ότι μοτίβα δεδομένων μπορούν να προκύψουν, αν αναλυθούν τα δεδομένα. Το Data Mining μας δίνει την δυνατότητα αυτά τα μοτίβα να προκύπτουν από ανάλυση ογκωδέστατων datasets που μεγαλώνουν καθημερινά όπως οι καταναλωτικές συνήθειες των πελατών μιας αλυσίδας μαγαζιών. Αυτή η δυνατότητα να προκύπτουν μοτίβα πρόβλεψης συνηθειών των ανθρώπων βάζει το Data Mining στην πρώτη γραμμή επιλογής για εταιρίες πωλήσεων ή εταιρίες διαφημίσεων κυρίως αλλά μπορεί να εφαρμοστεί σχεδόν παντού. Έχει εκτιμηθεί ότι η ποσότητα των δεδομένων που αποθηκεύονται σε βάσεις δεδομένων σε όλο τον κόσμο, διπλασιάζεται κάθε 20 μήνες, οπότε γίνεται κατανοητό ότι σε αυτήν την τρομακτική αύξηση των δεδομένων, το Data Mining γίνεται αναπόσπαστο κομμάτι για την ανάλυση των Big Data.

Το θετικό είναι ότι παρόλη την τεράστια αύξηση των δεδομένων, πλέον υπάρχουν ακόμα και κατ' οίκον τα μηχανήματα εκείνα τα οποία διαθέτουν την υπολογιστική ισχύ για ανάλυση και εξαγωγή αποτελεσμάτων, οπότε περισσότερος κόσμος έχει την δυνατότητα να ωφεληθεί από το Data Mining. Καθώς τα δεδομένα παράγονται σε εξωφρενικό ρυθμό και με μεγάλη πολυπλοκότητα, το Data Mining είναι το σημαντικότερο όπλο που διαθέτουμε για την ανακάλυψη και την κατανόηση κρυμμένων μοτίβων στον ωκεανό των νέων δεδομένων, τα οποία οδηγούν σε νέες επιχειρηματικές ιδέες και σε εμπορικά πλεονεκτήματα έναντι των ανταγωνιστών. Το Data Mining ουσιαστικά σχετίζεται με την επίλυση προβλημάτων από την ανάλυση δεδομένων που βρίσκονται ήδη σε βάσεις δεδομένων ανεξαρτήτου μεγέθους.

Ένα παράδειγμα και να γίνει ποιο κατανοητή η χρησιμότητα του, ας υποθέσουμε ότι μια εταιρία θέλει να μάθει κατά πόσο οι πελάτες της ψωνίζουν από την δικιά της αλυσίδα καταστημάτων και με τη συχνότητα ψωνίζουν από άλλες. Αφορά δηλαδή την πίστη που έχουν οι πελάτες της στο δικό της brand name σε μια άκρως ανταγωνιστική αγορά. Το κλειδί για την επίλυση του συγκεκριμένου προβλήματος βρίσκεται σε βάσεις δεδομένων που έχουν στοιχεία με τις προγενέστερες επιλογές του πελάτη σε συνδυασμό με τα προφίλ των πελατών. Μπορούμε να αναλύσουμε μοτίβα συμπεριφοράς πελατών που είχαν αγοράσει στο παρελθόν προϊόντα ή υπηρεσίες από την επιχείρηση μας, και μέσω της ανάλυσης να δούμε κατά πόσο προτιμήσαν να ξαναγοράσουν από την επιχείρηση μας ή αγόρασαν προϊόντα ανταγωνιστικών εταιριών. Μόλις ολοκληρωθεί η ανάλυση μέσω του Data Mining και βρεθούν τα χαρακτηριστικά που ψάχνουμε, μπορούμε να τα εφαρμόσουμε σε τωρινούς πελάτες και να εξακριβώσουμε την πιθανότητα να προτιμήσουν κάποιο άλλο προϊόν. Αυτό το group των πελατών στην συνέχεια στοχεύετε από την επιχείρηση μας με συγκεκριμένη στρατηγική στο marketing, πράγμα που θα ήταν πολύ δαπανηρό άμα γινόταν στο σύνολο των πελατών μας. Η στρατηγική αυτή μπορεί να αφορά την έκπτωση σε ένα προϊόν που δείχνουν να μην απολαμβάνουν ή την προσφορά ενός άλλου προϊόντος μαζί με αυτό για να γίνει ποιο ελκυστικό. Έτσι μεγαλώνει η πιθανότητα οι πελάτες αυτοί να το αγοράσουν το προϊόν ενώ θα ήταν σχεδόν βέβαιο ότι οι περισσότεροι από αυτούς θα είχαν στραφεί σε άλλη επιχείρηση αν δεν είχε προηγηθεί ανάλυση των δεδομένων. Αυτό προσφέρει τεράστια κέρδη στις επιχειρήσεις και αυτό το απλό παράδειγμα καταδεικνύει την χρησιμότητα της ανάλυσης των Big Data στην σημερινή, πελατοκεντρική κοινωνία μας, και για ποιον λόγο έχει προκληθεί τόσος θόρυβος γύρο από αυτά.

Παρακάτω θα παραθέσουμε μερικές από τις Data Mining τεχνικές που χρησιμοποιούνται στις μέρες μας όπως: Association rule learning, Cluster analysis και Classification. (Galit Shmueli, 2017)

3.1.1 Association rule learning

Αυτή η τεχνική μας δίνει την δυνατότητα να ανακαλύψουμε ζευγάρια μεταβλητών που μπορεί να σχετίζονται με ένα αποτέλεσμα. Αυτό συμβαίνει με διαδοχικά τεστ και παραμετροποιήσεις στους αλγορίθμους μέχρι να βγει ένα επιθυμητό μοτίβο. Όταν λέμε ένα επιθυμητό μοτίβο αυτό σημαίνει ο αλγόριθμός μας θα έχει κάποια κατώτερα όρια, τα οποία αν δενικανοποιηθούν δεν επιστρέφουν αποτέλεσμα. Αυτό μας βοηθάει να διώχνουμε τις περιττές μεταβλητές που θα επηρέαζαν την ευστοχία της πρόβλεψης του αλγορίθμου. Αυτή η τεχνική χρησιμοποιείται κατά κόρων στο Data Mining από εταιρίες πωλήσεων προϊόντων όπως η Amazon, από supermarkets και γενικότερα από εταιρίες πωλήσεων που θέλουν να βρουν μοτίβα πίσω από τις καταναλωτικές συνήθειες των πελατών.

Ένα παράδειγμα χρησιμότητας της συγκεκριμένης τεχνικής είναι ότι μια αλυσίδα supermarket χρησιμοποιώντας την συγκεκριμένη τεχνική κατέληξε στο συμπέρασμα ότι άντρες καταναλωτές που έγιναν γονείς πρόσφατα, αγόραζαν πάνες για το μωρό τους μαζί με μπύρες για την προσωπική τους διασκέδαση. Έτσι τα supermarket βρίσκοντας τέτοια μοτίβα στους καταναλωτές τους, είτε δημιουργούν μια προσφορά μεταξύ δυο προϊόντων, είτε φέρνουν τα προϊόντα στο supermarket σε παραπλήσιους διαδρόμους.

3.1.2 Clustering

Αυτή η τεχνική χρησιμοποιείται συνήθως όταν έχουμε έναν τεράστιο αριθμό από δεδομένα τα οποία προσπαθούμε να ομαδοποιήσουμε βάση κοινών χαρακτηριστικών και γνωρισμάτων τα οποία δεν ξέραμε προηγουμένως έτσι ώστε να πετύχουμε συγκεκριμένα αποτελέσματα.

Ένα παράδειγμα που είχα αναφέρει προηγουμένως είναι ότι από όλο το πελατολόγιο μιας εταιρίας πωλήσεων, ο αλγόριθμος χωρίζει τους καταναλωτές σε groups βάση των

προτιμήσεών τους ώστε να μπορεί η εταιρία να προσαρμόσει το marketing της, στις συνήθειες του κάθε group.

Βλέπουμε ότι με έναν απλό αλγόριθμο Data Mining με την τεχνική του Clustering ο υπολογιστής έχει την δυνατότητα να βρίσκει μοτίβα και με αυτά να χωρίζει τους ανθρώπους σε groups. Με την ίδια ακριβώς τεχνική, με πολύ καλύτερους αλγόριθμους, και με datasets χιλιάδων καταναλωτών, οι εταιρίες χωρίζουν τους ανθρώπους βάση αγορών και συνηθειών και εξάγουν τα δικά τους αποτελέσματα προς υλοποίηση.

3.1.3 Classification

Η τεχνική του Classification, σε αντίθεση με το Clustering, χρησιμοποιείται όταν έχουμε συγκεκριμένα αποτελέσματα από ένα dataset, τα groups είναι ήδη χωρισμένα βάση των αποτελεσμάτων, και βάση αυτών των στοιχείων όταν έρχεται ένα νέο data point ο αλγόριθμος το κατατάσσει στο group με τα περισσότερα κοινά χαρακτηριστικά. Αυτή η τεχνική ονομάζεται και supervised learning καθώς σε αντίθεση με το clustering (unsupervised learning) που δεν υπήρχαν διαμορφωμένα groups, εδώ τα groups υπάρχουν και μαθαίνουμε τον αλγόριθμο να σκέφτεται παρέχοντας του ένα κομμάτι των αποτελεσμάτων από το dataset.

Για να γίνει πιο κατανοητό, έστω ότι έχουμε ένα dataset το οποίο στις στήλες του έχει τα στοιχεία που χρειάζονται για να αποφανθεί ένας γιατρός αν μια γυναίκα έχει εμφανίσει κακοήθη ή καλοήθη όγκο στον μαστό. Από αυτά τα στοιχεία τα οποία υπάρχουν στο dataset, μαθαίνουμε στον αλγόριθμο μέσω της τεχνικής του Classification να βρίσκει τα μοτίβα εκείνα μεταξύ των στοιχείων που υποδεικνύουν αν ένας όγκος είναι καλοήθης ή κακοήθης, και όταν θα εισάγουμε τα στοιχεία μιας νέας ασθενούς, ο αλγόριθμος έχοντας δημιουργήσει μοτίβα από τα στοιχεία που τα παρείχαμε, κατατάσσει την ασθενή σε ένα από τα δύο groups. Ουσιαστικά με αυτήν την τεχνική δημιουργούνται αποτελέσματα πρόβλεψης, τα οποία θα έπαιρναν πάρα πολύ χρόνο στους γιατρούς χωρίς την τεχνική αυτή. Τα αποτελέσματα αυτά, συνοδεύονται από ένα ποσοστό πρόβλεψης δηλαδή ότι κατά 98,7% η ασθενής Χ ανήκει στο group με τους καλοήθεις όγκους. Αυτή η πρόβλεψη βοηθάει το προσωπικό να γλιτώσει τον χρόνο ψαξίματος και να ψάξει στοχευμένα για το αν η πρόβλεψη ήταν αληθής η όχι. (το 98.7% θα ήταν ένα υπέροχο νούμερο σε πρόβλεψη μετοχής στο

χρηματιστήριο, αλλά σε ένα τόσο ευαίσθητο θέμα οι γιατροί πρέπει να είναι 100% σίγουροι πριν παραδώσουν τα αποτελέσματα).

3.2 Machine learning and statistics

Όπως προδίδει και το όνομα του το Machine Learning, που οι τεχνικές του συνδυάζονται πολλές φορές με την στατιστική και το Data Mining, είναι ένα είδος τεχνητής νοημοσύνης (AI) το οποίο δίνει την δυνατότητα στους ηλεκτρονικούς υπολογιστές να μάθουν μοτίβα και συνήθειες χωρίς να έχουν προγραμματιστεί γι' αυτόν τον σκοπό. Ο Data Analyst που θα χρησιμοποιήσει την τεχνική του Machine Learning, ουσιαστικά εστιάζει στην δημιουργία προγραμμάτων τα οποία μπορούν να αλλάξουν και να προσαρμόσουν τα αποτελέσματα τους κάθε φορά που δέχονται νέα δεδομένα [11].

Σαν έννοια το Machine Learning δεν είναι καινούργιο, αλλά ουσιαστικά τώρα με την υπάρχουσα τεχνολογία γίνεται χρήσιμο. Η ιδέα υπήρχε από το 1970, ότι δηλαδή οι υπολογιστές μπορούν να μάθουν να κάνουν συγκεκριμένα πράγματα χωρίς να έχουν προγραμματιστεί γι' αυτά (πρώιμη θεωρία τεχνητής νοημοσύνης). Η εφαρμογή του στις μέρες μας, όπως γενικότερα η άνοδος των Big Data, προέρχεται από την δυνατότητα που έχουμε πλέον να χρησιμοποιούμε πολύ δυνατούς υπολογιστές με πολύ μικρό κόστος (για την δουλειά που παράγουν) σε συνδυασμό με τις τεράστιες ποσότητες δεδομένων που έχουμε διαθέσιμες. Η τεχνική του Machine Learning μοιάζει πολύ με εκείνη του Data Mining και πολλές φορές ο διαχωρισμός τους δεν είναι εύκολα διακριτός. Και οι δύο οι τεχνικές ψάχνουν στα δεδομένα που τους έχουμε τροφοδοτήσει και προσπαθούν να βρουν συσχετίσεις. Ωστόσο, σε αντίθεση με το Data Mining που εξάγει δεδομένα για να βγουν αποτελέσματα μη διακριτά σε ανθρώπους λόγω του τεράστιου όγκου των δεδομένων, το Machine Learning χρησιμοποιεί αυτά τα δεδομένα για να εντοπίσει μοτίβα και προσαρμόζει τα αποτελέσματα του προγράμματος ανάλογα.

Ένα απλό παράδειγμα χρήσης Machine Learning, είναι το Facebook, το χρησιμοποιούν οι αναλυτές της εταιρίας αλγορίθμους που εξατομικεύουν για κάθε χρήστη την ροή των ειδήσεων ή των φίλων που θα βλέπει όταν θα ανοίγει την σελίδα. Αν ας πούμε ένας χρήστης της εφαρμογής σταματάει συχνά να διαβάσει μια δημοσίευση ή να πατήσει like

σ' ένα συγκεκριμένο φίλο ή φίλη, τότε ο αλγόριθμος συσχετίζει αυτήν την κίνηση με τον χρήστη και την επόμενη φορά που θα ανοίξει την σελίδα, αν ο φίλος του έχει δημοσιεύσει κάτι καινούργιο, αυτό θα είναι στην κορυφή της σελίδας ώστε να το δει πρώτο. Το λογισμικό πολύ απλά χρησιμοποιεί στατιστική ανάλυση των κινήσεων κάθε ατόμου και όταν έχει συλλέξει αρκετά δεδομένα τότε μπορεί βάση των προηγούμενων κινήσεων, να προβλέπει τι ενδιαφέρει τον κάθε χρήστη συγκεκριμένα και να προσαρμόζει ανάλογα την ροή των δημοσιεύσεων. Αν στην περίπτωση που σταματήσει ο χρήστης μας να διαβάζει ή να πατάει like στον συγκεκριμένο φίλο, τότε αυτά τα νέα δεδομένα προσαρμόζονται σαν νέο μοτίβο στον αλγόριθμό και ξανά αλλάζει την ροή των δημοσιεύσεων που θα βλέπει. (rouse, 2016)

Το Machine Learning χρησιμοποιείται παντού επειδή πλέον πολλές εταιρίες και οργανισμοί έχουν στην διάθεση τους τεράστιες ποσότητες αχρησιμοποίητων δεδομένων που αν αναλυθούν σωστά θα τους αποφέρουν τεράστια κέρδη.

Παρακάτω θα αναλύσω μερικές τεχνικές που βασίζονται στο Machine Learning και στην στατιστική όπως Regression, Natural Language Processing, A/B testing και Spatial analysis.

Το Regression αφορά μια σειρά τεχνικών που είχαν χρησιμοποιηθεί αρχικώς στην στατιστική και αφορούσαν την μεταβολή μια μεταβλητής (label) όταν άλλαζαν οι τιμές συσχετιζόμενων μεταβλητών. Αυτό υιοθετήθηκε από τους αναλυτές που χρησιμοποιούν το Machine Learning, μαθαίνοντας στον υπολογιστή πως η label επιθυμητή μεταβλητή αλλάζει όταν τις παρέχουμε συνεχώς νέα δεδομένα τα οποία μπορεί να είναι και real-time (μεταβολή μιας μετοχής).

Συνήθως χρησιμοποιείται στο κλάδο των οικονομικών από εταιρίες και ανθρώπους που ασχολούνται με το χρηματιστήριο, τις επενδύσεις, τις τιμές των ακινήτων και πως αυτές επηρεάζονται σε καθημερινή βάση. Ένα παράδειγμα λειτουργίας είναι ότι αναλυτές που ασχολούνται με την διακύμανση των μετοχών, έχουν αλγορίθμους στους οποίους έχουν περάσει τα δεδομένα της διακύμανσης μιας μετοχής του τελευταίου διμήνου, μαθαίνοντας στον υπολογιστή να συσχετίζει την άνοδο ή την πτώση μιας μετοχής συγκρίνοντας την με μια σειρά από μεταβλητές, και έτσι σε κάθε νέα συνεδρίαση όπου νέα real-time δεδομένα παρέχονται στον αλγόριθμο, βάση των συσχετίσεων που είχε κάνει από

τα αποτελέσματα που του είχαμε δώσει γιατο τελευταίο δίμηνο, δίνει μια πρόβλεψη για την διακύμανση της μετοχής κατά την διάρκεια της ημέρας. (investopedia, 2016)

3.2.1 Natural Language Processing

Αυτό είναι ένα γενικότερο πεδίο που έχει στοιχεία από την επιστήμη των υπολογιστών, την τεχνητή νοημοσύνη, και της υπολογιστικής γλωσσολογίας και πραγματεύεται τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας και ειδικότερα με την κατανόηση από τους υπολογιστές μέσω προγραμμάτων της δομής, της σύνταξης και ειδοποιών διαφορών κάθε γλώσσας, έτσι ώστε να μπορεί να αντιλαμβάνεται ο υπολογιστής τις εντολές που του δίνουμε όταν λαμβάνει σαν δεδομένα κείμενα από κάθε γλώσσα.

Ένα τέτοιο παράδειγμα χρησιμοποιώντας μια τεχνική που ανήκει στο πεδίο του Natural Language Processing είναι το **Sentiment analysis** όπου αυτό χρησιμοποιείται για τον εντοπισμό και την εξαγωγή πληροφοριών από διάφορα κείμενα. Βασικός σκοπός της ανάλυσης αυτής είναι να προσδιορίζει σ' ένα κείμενο αν ο τόνος των σχολίων είναι θετικός ή αρνητικός. Αυτό χρησιμοποιείται κατά κόρων από εταιρίες πωλήσεων όπως η Amazon, όπου χρησιμοποιώντας αυτήν την τεχνική κατηγοριοποιούν τις κριτικές σε θετικές ή αρνητικές, χωρίς να χρειαστεί να διαβαστούν αυτοτελώς οι κριτικές. Αυτό έχει ως αποτέλεσμα την εξαγωγή συμπερασμάτων πάνω στα υπό πώληση προϊόντα, δηλαδή κατά πόσο ανταποκρίθηκαν θετικά ή αρνητικά στις προσδοκίες των καταναλωτών (οι αλγόριθμοι είναι φτιαγμένοι έτσι ώστε ο υπολογιστής να μπορεί να κατανοεί ακόμα και το αν ένα σχόλιο ήταν μετρίως αρνητικό ή απόλυτα αρνητικό και θετικό αντίστοιχα), αλλά επίσης και σε μεγάλη εξοικονόμηση χρόνου, καθώς αν δεν έκανε ο αλγόριθμος την συγκεκριμένη δουλεία, θα χρειαζόντουσαν δεκάδες εργατοώρες για την προσπέλαση όλων των σχολίων σε όλα τα προϊόντα.

Μια από τις ποιο δύσκολες εφαρμογές είναι το **machine translation** όπου δεν είναι τίποτα άλλο από μετάφραση από την μια γλώσσα στην άλλη. Ακούγεται πολύ εύκολο και είναι σε μερικές περιπτώσεις όταν οι γλώσσες είναι φτιαγμένες από τον ίδιο κορμό (Γαλλικά, Ιταλικά, Ισπανικά), αλλά όταν οι γλώσσες διαφέρουν πάρα πολύ (Αγγλικά σε σχέση με Ελληνικά ή Κινέζικα), οι διαφορές στην γραμματική, στο συντακτικό, στην σήμανση των λέξεων, στους ιδιωματισμούς κλπ.) καθιστά πολύ δύσκολη την διαδικασία 'εκμάθησης' του

υπολογιστή. Παράδειγμα είναι το google translate όπου όπως έχετε δει η μετάφραση από τα αγγλικά στα ελληνικά είναι μέτρια στην καλύτερη περίπτωση.

Τέλος μια ακόμη δύσκολη, και υπό διαρκής εξέλιξη τεχνική του Natural Language Processing, είναι το **speech recognition,** δηλαδή η δυνατότητα που έχουμε να μιλάμε στον υπολογιστή μέσω μικροφώνου και αυτό να το γράφει σε κείμενο. Οι δυσκολίες εδώ πέρα βρίσκονται στο γεγονός ότι στον προφορικό μας λόγο ο κάθε ανθρωπος έχει τις δικές του ιδιαιτερότητες (χροιά φωνής, καθαρότητα λόγου, ταχύτητα λόγου), όπου ο υπολογιστής πρέπει σε κάθε περίπτωση να επεξεργαστεί με τον ίδιο τρόπο, και επίσης είναι δύσκολο να καταλάβει πότε σταματάμε μια πρόταση επειδή συνήθως στην ομιλία μας δεν υπάρχουν σχεδόν καθόλου παύσεις μεταξύ των διαδοχικών λέξεων. Επίσης σε κάθε γλώσσα, οι ήχοι που αντιπροσωπεύουν τα συνεχόμενα γράμματα δεν είναι ίδιοι (π.χ. Γερμανικά σε σχέση με Αγγλικά) οπότε η μετατροπή σε χαρακτήρες γίνεται μια πολύ δύσκολη διαδικασία (στην συγκεκριμένη τεχνική έχουν επενδύσει αρκετά κονδύλια πολύ μεγάλες εταιρίες και θεωρείται το νέο μεγάλο επίτευγμα στον χώρο της τεχνητής νοημοσύνης). (Wikipedia, 2017)

3.2.2 Statistics

Όπως και το Natural Language Processing, και η στατιστική είναι ένα γενικότερα πεδίο με πολλές τεχνικές στο Machine Learning. Οι περισσότερες από τις τεχνικές αυτές προϋπήρχαν πριν την χρήση των υπολογιστών για ανάλυση δεδομένων, και εξελίχθηκαν για να μπορούν να υλοποιηθούν σε ένα υπολογιστικό περιβάλλον. Ουσιαστικά η επιστήμη αυτή αφορά την συλλογή, οργάνωση και ερμηνεία των δεδομένων καθώς επίσης τον σχεδιασμό πειραμάτων για την αποτελεσματικότερη ανάλυση των δεδομένων. Οι περισσότερες από τις τεχνικές που χρησιμοποιούνται προσπαθούν να καθορίσουν αν η σχέση μεταξύ δύο μεταβλητών σ' ένα dataset είναι τυχαία, ή αν υπάρχει κάποια σχέση μεταξύ των μεταβλητών και αν ναι, από ποιους παράγοντες προκλήθηκε και ποιοι παράγοντες την επηρεάζουν [12].

Μια πρώτη τεχνική ονομάζεται **A/B testing** [13]. Με αυτήν, συγκρίνουμε ένα group δεδομένων που έχουμε δημιουργήσει εμείς, με διάφορα test groups, έτσι ώστε να καθορίσουμε τις αλλαγές που χρειάζεται να γίνουν ώστε να βελτιώσουμε το αρχικό μας group. Για παράδειγμα αυτή η τεχνική χρησιμοποιείται συνεχώς από web developers ώστε

να κάνουν την ιστοσελίδα ποιο προσιτή στον κόσμο και να την επιλέγουν και βάση της αισθητικής. Πραγματοποιούνε διάφορα πειράματα όπως, τι περίγραμμα είναι περισσότερο θεμιτό στο κοινό, τι εικόνες προτιμώνται, τι χρώμα να είναι το φόντο κλπ. Τα Big Data μας επιτρέπουν να τρέχουμε χιλιάδες πειράματα ταυτόχρονα και να αναλύουν τα αποτελέσματα ώστε να βρίσκουν το βέλτιστο αποτέλεσμα.

4. Εφαρμογές μεγάλων δεδομένων

Καθώς η τεχνολογία των μεγάλων δεδομένων διαρκώς εξελίσσεται, ολοένα και πιο πολλοί οργανισμοί αντιλαμβάνονται την ανάγκη και ευκαιρία αξιοποίησης τους. Ορισμένα "κέρδη" από τις εφαρμογές της ανάλυσης μεγάλων δεδομένων είναι:

- Ανάλυση της συμπεριφοράς και των προτιμήσεων των χρηστών και ανάλογη προσαρμογή της πολιτικής προβολής και διαφήμισης
- Εντοπισμός μοτίβων που αφορούν στην ανταπόκριση ασθενών σε ιατρικές θεραπείες
- Εντοπισμός αναζητουμένων ή/και υποψηφίων εγκληματιών μέσω της ανίχνευσης υπόπτων τηλεπικοινωνιακών, αγοραστικών και μετακινήσεων συμπεριφορών
- Εφαρμογές για κινητές συσκευές: Τα τελευταία χρόνια οι κινητές συσκευές αποκτούν όλο και μεγαλύτερο μερίδιο της αγοράς για τις διαδικτυακές υπηρεσίες. Οι υπηρεσίες αυτές για να ανταποκριθούν κατάλληλα, θα πρέπει να εξασφαλίσουν υψηλή διαθεσιμότητα αλλά και τη δυνατότητα για άμεση επεξεργασία μεγάλου πλήθους δεδομένων. Τα δύο αυτά χαρακτηριστικά υποδεικνύουν τα big data ως την καλύτερη λύση.

Εφαρμογές με υψηλό βαθμό παραλληλοποίησης: Η σημερινή έκρηξη δεδομένων έχει ως συνέπεια την ανάγκη για γρήγορη επεξεργασία τεράστιου όγκου πληροφορίας. Αυτό με τις συμβατικές τεχνολογίες είναι αδύνατον να επιτευχθεί. Στο clouding ωστόσο, με την ύπαρξη εικονικά απεριόριστου hardware (on demand), μπορούμε να μοιράσουμε τα δεδομένα μας σε εκατοντάδες διαφορετικούς υπολογιστές για να επιτύχουμε γρηγορότερη επεξεργασία τους. Μάλιστα, οι πάροχοι διαθέτουν έτοιμα περιβάλλοντα (MapReduce, Hadoop) για τη διευκόλυνση αυτών των εφαρμογών. Επιχειρησιακές εφαρμογές (business applications): Μεγάλα συστήματα (ERP, CRM κλπ.) που χρησιμοποιούνται για το

στρατηγικό σχεδιασμό των επιχειρήσεων, είναι συνήθως πολύ απαιτητικά σε υπολογιστική ισχύ.

Η λύση του cloud λοιπόν φαντάζει μονόδρομος, από τη στιγμή που το μέγεθος των δεδομένων προς επεξεργασία για τις μεγάλες επιχειρήσεις αυξάνεται εκθετικά. Πολλές είναι οι εταιρείες οι οποίες κινηθήκαν με ταχύτητα και κατάφεραν να επιτύχουν επιχειρηματικούς στόχους μέσα από την ανάλυση μεγάλων δεδομένων. Πολλά παραδείγματα εταιρειών που με χρήση μεγάλων δεδομένων βελτίωσαν υπηρεσίες μεταφορών, γεωργίας, χρηματοοικονομικών και παιδείας και μπορούν να χωριστούν σε διάφορες κατηγορίες ανάλογα με την περίπτωση. Μέσω των analytics μεγάλων δεδομένων είναι εφικτή η παρακολούθηση της εξέλιξης των πωλήσεών.

4.1 Τηλεπικοινωνίες

Τηλεπικοινωνίες : Στα τηλεφωνικά κέντρα συλλέγονται πολύ μεγάλες ποσότητες αδόμητων και δομημένων δεδομένων. Η χαρτογράφηση και ταξινόμηση των κλήσεων παρέχει τη δυνατότητα εντοπισμού σφαλμάτων και αδυναμιών στις σχετικές υποδομές [14]. Εμπόριο Η eBay, μία από τις μεγαλύτερες ηλεκτρονικές πλατφόρμες δημοπρασιών στον κόσμο, καταγράφει συναλλαγές με περισσότερους από 108 εκατ. πελάτες ετησίως ενώ εισπράττει πάνω από 250 εκατ αιτήματα στον ιστοχώρο της ημερησίως. Επίσης, στο εμπόριο ηλεκτρονικών συσκευών, εκτιμάται ότι πωλείται ένα κινητό τηλέφωνο κάθε 5 δευτερόλεπτα [15]. Είναι εμφανώς λοιπόν ο όγκος πληροφορίας που αποθηκεύεται τόσο για τους πελάτες και για τα προϊόντα και η δυνατότητα εξόρυξης γνώσης αναφορικά με τις συνήθειες και τάσεις και η διαμόρφωση σχετικών πολιτικών. Αλλά ακόμα και στο λιανεμπόριο γίνεται αξιοποίηση μεγάλων δεδομένων μελετώντας πληροφορίες από ΜΜΕ για τον εντοπισμό του βέλτιστου σημείου εγκατάσταση ενός καταστήματος. Αντιμετώπιση καταστροφών Μέσα από τη συγκέντρωση πληροφορίας που μπορεί να προέρχεται είτε από ΜΜΕ είτε από κοινό δίνεται η δυνατότητα χαρτογράφησης του τόπου όπου λαμβάνει χώρα μια καταστροφή, αξιολόγησης της σοβαρότητάς της και χάραξη της άριστης διαδρομής για την τάχιστη άφιξη των σχετικών υπηρεσιών. Η γρήγορη πρόγνωση του τυφώνα Irene στην Φλόριντα το 2011 ελαχιστοποίησε τις συνέπειες αφού έδωσε το χρόνο για τη λήψη όλων των αναγκαίων μέτρων και βασίστηκε στην ανάλυση μεγάλων γεωχωρικών δεδομένων

4.2 Ο κλάδος του πετρελαίου

Ο κλάδος του πετρελαίου: Θεωρείται από τους πρώτους όπου άρχισε να ασχολείται με τα μεγάλα δεδομένα. Πετρελαϊκές εταιρείες και κυβερνήσεις κάνουν χρήση και ανάλυση τεράστιων ποσοτήτων δεδομένων που είναι διαθέσιμα σχετικά με σεισμική δραστηριότητα σε όλη την υφήλιο με σκοπό την εξερεύνηση και εξόρυξη πετρελαίου.

4.3 Μεταφορές

Μεταφορές: Η εταιρεία διεθνών ταχυμεταφορών UPS, άρχισε να καταγράφει και να τις κινήσεις πακέτων και συναλλαγών από τις αρχές της δεκαετίας 1980. Σήμερα συγκεντρώνει δεδομένα για 18,5 εκατ. πακέτα ημερησίως, για 9,3 εκατ πελάτες, με μέσω όρο 42,8 εκ αιτημάτων παρακολούθησης πακέτου καθημερινά. Το μεγαλύτερο τμήμα των μεγάλων δεδομένων που διαθέτει προέρχονται από τηλεματικούς αισθητήρες τοποθετημένους στα οχήματά της. Τα δεδομένα αξιοποιούνται τόσο για την καθημερινή εποπτεία και μέτρηση της αποδοτικότητας αλλά και για τη διαμόρφωση της βέλτιστης δομής των δρομολογίων.

4.4 Διαδίκτυο

Διαδίκτυο: Ιστότοποι όπως το facebook και το twitter συγκεντρώνουν πάνω από 27 και 14 terabytes δεδομένων αντίστοιχα. Η Google μέσω των διάφορων εφαρμογών της (mail, google drive, google earth κ.α) συγκεντρώνει δεδομένα όγκου πλέον των 90 terabytes ημερησίως. Η ανάλυση των δεδομένων των χρηστών τους είναι ο οδηγός της διαμόρφωσης της στρατηγικής τους στόχευσης.

4.5 Υγεία

Υγεία: Στον τομέα της υγειονομικής περίθαλψης συνεργάζονται επιχειρηματικοί κλάδοι από τις φαρμακευτικές εταιρίες, τις εταιρίες παραγωγής ιατρικού εξοπλισμού, παρόχους, ασθενείς κλπ. και κάθε μια από αυτές τις συνδεόμενες μεταβλητές παράγει τεράστια ποσότητα δεδομένων. Κάθε μια από αυτές έχει διαφορετικά συμφέροντα και επιχειρηματικά κίνητρα, και στην πλειοψηφία τους τα δεδομένα που παράγονται δεν αναλύονται. Επίσης πολλά από τα κλινικά δεδομένα δεν έχουν ψηφιοποιηθεί ακόμα οπότε είναι αδύνατη η εκμετάλλευσή τους. Επιπλέον, παρατηρείται μια αύξηση 5% τον χρόνο στα έξοδα που αφορούν τον συγκεκριμένο κλάδο, και αυτό επιβαρύνει τον προϋπολογισμό των κυβερνήσεων και διογκώνει το δημόσιο χρέος. Σε αυτό έρχεται να προστεθεί και η παγκόσμια γήρανση του πληθυσμού όπως και επίσης νέες, ποιο δαπανηρές θεραπείες. Γενικότερα ο κλάδος της υγείας έχει μείνει αρκετά πίσω σε σχέση με άλλους κλάδους στην βελτίωση των λειτουργικών εξόδων, όπως επίσης και στην ενσωμάτωση τεχνολογιών για την βελτίωση τους. Αυτά τα προβλήματα δημιουργούν μια τεράστια ευκαιρία από την εκμετάλλευση των δεδομένων αυτών για την δημιουργία κέρδους από την ανάλυση τους, αν αυτά τα δεδομένα ψηφιοποιηθούν, συνδεθούν μεταξύ τους και χρησιμοποιηθούν σε αυτά κατάλληλες τεχνικές.

Τα Big Data που μπορούμε να αναλύσουμε στον τομέα της υγείας χωρίζονται σε τέσσερις κατηγορίες:

- **Φαρμακευτικές έρευνες και δαπάνες** (κλινικές μελέτες φαρμάκων, κόστος παρασκευής και πώλησης, αποτελεσματικότητα κόστους/θεραπείας κλπ.)

- **Κλινικά δεδομένα ασθενών** (ηλεκτρονική απεικόνιση ιστορικού υγείας, ψηφιακό αρχείο από ακτινογραφίες)

- **Υγειονομικές δαπάνες των ιατρείων αλλά και των ασθενών**

- **Δεδομένα που αφορούν την συμπεριφορά των ασθενών και τις προτιμήσεις τους κατά την διαρκεια της νοσηλείας**

Πολλά από αυτά τα πολύ σημαντικά δεδομένα δεν έχουν ψηφιοποιηθεί ακόμη και δεν υπάρχει εκμετάλλευση των δεδομένων. Αυτά τα δεδομένα χρειάζεται να αναλυθούν καθώς θα παρέχουν άμεσα αποτελέσματα. Μερικά παραδείγματα από το κέρδος της ανάλυσης τους είναι:

- Στοχευμένη ανάλυση ιστορικού υγείας για κάθε περίπτωση ασθενούς, ώστε να παρέχεται η καλύτερη εξατομικευμένη θεραπεία.

- Real-time συλλογή δεδομένων από ασθενείς με χρόνιες παθήσεις, όπου παρακολουθείται αν οι ασθενείς έχουν κάνει αυτά που προβλεπόντουσαν, και άμεση αναπροσαρμογή θεραπείας ή φαρμάκων όπου κρίνεται απαραίτητο, ανάλογα με την πρόοδο του ασθενούς.

- Ανάλυση του προφίλ των ασθενών, για τον εντοπισμό ατόμων που θα ωφεληθούν από προληπτική φροντίδα ή αλλαγές στις διατροφικές συνήθειες και στον τρόπο ζωής, αν βρεθεί ότι υπάρχουν αρκετές πιθανότητες (μέσω αλγορίθμων πρόβλεψης) να αναπτύξουν στο μέλλον μια ασθένεια.

- Αποτελεσματικότητα θεραπειών/φαρμάκων βάση απόδοσης/κόστους, και μείωση τιμών όπου τα αποτελέσματα δεν είναι τα θεμιτά. [16].

4.6 Χρηματοπιστωτικές υπηρεσίες

Χρηματοπιστωτικές υπηρεσίες: Από το 2015 οι μεγάλες τράπεζες εφάρμοσαν σε μεγάλη κλίμακα την ανάλυση των αχρησιμοποίητων δεδομένων, και τα αποτελέσματα που είδαν θεωρήθηκαν τόσο σημαντικά, ώστε το 2015 να θεωρείται χρονιά σταθμός στον τραπεζικό τομέα και στις χρηματοπιστωτικές αγορές. Στις μέρες μας καταβάλλεται μεγάλη προσπάθεια για το μέγιστο κέρδος από τα Big Data, καθώς συνεχίζεται η αναδιοργάνωση των διαδικασιών, που τόσα χρόνια παρέμεναν ίδιες. Βλέπουμε ότι οι τράπεζες και όσοι ασχολούνται με επενδύσεις και χρηματιστήριο προσπαθούν να προσαρμοστούν στην νέα εποχή για να παραμείνουν ανταγωνιστικοί.Για τις τράπεζες, τα κίνητρα για να χρησιμοποιήσουν Big Data βρίσκονται κυρίως στην καλύτερη χαρτογράφηση των πελατών

της και την μείωση του επιχειρησιακού κινδύνου. Για τις εταιρίες ή τους ανθρώπους που ασχολούνται με το χρηματιστήριο ή τις επενδύσεις, τα αποτελέσματα που θέλουν χρειάζεται να είναι άμεσα, καθώς το κέρδος βρίσκεται στην λήψη γρήγορων αποφάσεων κατά την διάρκεια της ημέρας.

Οι ευκαιρίες που παρουσιάζονται από την επεξεργασία και την ανάλυση των δεδομένων είναι:

- Machine Learning αλγόριθμοι θα επεξεργάζονται Real-Time τα εισερχόμενα δεδομένα και θα είναι σε θέση να ενημερώνουν άμεσα για οικονομικές απάτες.

- Η επεξεργασία των δεδομένων θα προσφέρει στοχευμένες υπηρεσίες για τραπεζικές συναλλαγές ή δάνεια ανάλογα με το προφίλ του κάθε καταναλωτή.

- Θα παρέχουν σε επενδυτές μια ποιο ασφαλή εικόνα για τις μελλοντικές τους επενδύσεις. [17]

4.7 Δημόσια Διοίκηση

Δημόσια Διοίκηση: Οι περισσότερες κυβερνήσεις ανά τον κόσμο βρίσκονται υπό αυξανόμενη πίεση για να δοθεί ώθηση στην παραγωγικότητα, δηλαδή να προσφέρουν περισσότερα με το μικρότερο δυνατό κόστος. Αυτή η ανάγκη γίνεται ιδιαιτέρως επιτακτική ύστερα από τους κλυδωνισμούς που επέφερε η παγκόσμια οικονομική κρίση. Πολλές κυβερνήσεις προσπαθούν να παρέχουν όσο τον δυνατόν καλύτερες δημόσιες υπηρεσίες σε μια εποχή όπου υπάρχουν σημαντικοί δημοσιονομικοί περιορισμοί για να μειωθούν τα δημοσιονομικά ελλείμματα και το δημόσιο χρέος. Επιπλέον πολλές χώρες αντιμετωπίζουν και την γύρναση του πληθυσμού, που θα αυξήσει σημαντικά τη ζήτηση σε ιατρικές και κοινωνικές υπηρεσίες, άρα και τα έξοδα.

Για να καταφέρουν οι κυβερνήσεις να είναι μέσα στους δημοσιονομικούς τους στόχους, αλλά και για να δώσουν ώθηση στο δημόσιο τομέα θα πρέπει να αυξήσουν την παραγωγικότητά τους. Έρευνες έχουν δείξει ότι τα τελευταία χρόνια ο ιδιωτικός τομέας σε πολλάκρατη έχει ξεπεράσει τον δημόσιο. Πως όμως μπορούν τα Big Data να βοηθήσουν στην αύξηση της παραγωγικότητας της δημόσιας διοίκησης? Τα δεδομένα που παράγονται στον δημόσιο τομέα είναι κυρίως δεδομένα κειμένου ή αριθμητικά δεδομένα, και σε σχέση με τον τομέα της υγείας όπου τα δεδομένα είναι εικόνες υψηλής ευκρίνειας ή βίντεο από εγχειρήσεις, διαπιστώνουμε ότι στο δημόσιο τα δεδομένα είναι λιγότερα σε μέγεθος, αλλά εξίσου σημαντικά. Η σημαντική διαφορά όμως είναι ότι τα δεδομένα που παράγονται είναι κατά 90% ψηφιοποιημένα καθώς στις περισσότερες χώρες λειτουργούν μέσω ηλεκτρονικών υπηρεσιών για την διευκόλυνση των πολιτών. Τα σημαντικά πλεονεκτήματα που μπορούν να προσφέρουν τα Big Data αν ενσωματωθούν πλήρως στην δημόσια διοίκηση είναι:

- **Εξοικονόμηση χρόνου.** Τα στοιχεία των πολιτών είναι αποθηκευμένα σε datasets τα οποία δεν αξιοποιούνται και τους ζητούνται να ξαναγράφουν τα στοιχεία τους, ή τις φορολογικές τους ενημερότητες από την αρχή.

- **Καταπολέμηση φοροδιαφυγής.** Μέσω των Machine Learning αλγορίθμων θα μπορούν να ελέγχονται Real-Time τα στοιχεία που δηλώνουν οι πολίτες και να ενημερώνουν για πιθανές παραβατικές υποθέσεις ελέγχοντας παράλληλα ογκωδέστατα αρχεία που θα χρειαζόντουσαν δεκάδες εργατοώρες από το ανθρώπινο προσωπικό.

- **Παραγωγή πλούτου και μείωση του κόστους.** Νέα επιχειρηματικά μοντέλα, προϊόντα και υπηρεσίες γίνονται διαθέσιμα με τα Big Data, μειώνοντας το κόστος λειτουργίας του δημόσιου τομέα και βελτιώνοντας τις δημοσιονομικές επιδόσεις της οικονομίας του κάθε κράτους.

4.8 Marketing και Πωλήσεις

Marketing και Πωλήσεις: Αν και δεν μπορεί να υπάρξει άμεση σύγκριση μεταξύ κλάδων, θεωρείται από πολλούς ότι το marketing και οι πωλήσεις έχουν να παρουσιάσουν τα

μεγαλύτερα κέρδη από την χρήση των Big Data. Υπολογίζεται ότι με την πλήρη ενσωμάτωση των Big Data τεχνικών στους συγκεκριμένους κλάδους, θα υπάρχει αύξηση της παραγωγικότητας κατά 0.5% ανά έτος μέχρι το2020, χωρίς να υπολογίζεται σε αυτό η συνεχιζόμενη εμφάνιση τεχνικών και τεχνολογιών που δίνουν την δυνατότητα για ακόμη καλύτερη αξιοποίηση των Big Data.

Ο λόγος που υπάρχει αυτή η δυνατότητα άμεσου κέρδους στον συγκεκριμένο κλάδο βρίσκεται στην χρήση των ψηφιακών δεδομένων. Τα ψηφιακά δεδομένα αποκτούν καίριο ρόλο στον κλάδο καθώς οι καταναλωτές αναζητούν, ερευνούν, συγκρίνουν και τελικά αγοράζουν προϊόντα διαδικτυακά. Αυτό αφήνει ένα ψηφιακό μονοπάτι για τον κάθε καταναλωτή, το οποίο η κάθε εταιρία μπορεί να το αναλύσει προς όφελός της. Αλλά τα οφέλη θα είναι αρκετά και στους καταναλωτές καθώς προβλέπεται ότι με τις νέες τεχνολογίες που αναπτύσσονται στον χώρο των Big Data, οι τιμές θα πέσουν αισθητά. Μια εφαρμογή που ήδη κυκλοφορεί ονομάζεται RedLaser και επιτρέπει στους καταναλωτές να σκανάρουν το bar-code του προϊόντος που επιθυμούν να αγοράσουν σε ένα κατάστημα από τα κινητά τους, και αυτό τους επιστρέφει άμεσα συγκρίσεις με τιμές σε άλλα καταστήματα καθώς και σε παρεμφερή προϊόντα. Επιπλέον με την αύξηση των αγορών από το διαδίκτυο γίνεται ευκολότερο για τους καταναλωτές να επιλέξουν την καλύτερη τιμή και ποιότητα σε ένα προϊόν που επιθυμούν. Αυτοί οι νέοι τρόποι αγορών θα επιφέρουν τεράστια κέρδη στους καταναλωτές. Παρακάτω θα επισημάνω τα σημαντικότερα κέρδη που αποκομίζονται από την χρήση των Big Data τεχνικών στον κλάδο.

- **Προτεινόμενα προϊόντα.** Τα κέρδη από την ανάλυση των καταναλωτικών προτιμήσεων του πελάτη, το ιστορικό των αγορών του, την τοποθεσία του κλπ. μέσω των προτεινόμενων προϊόντων είναι τεράστια. Η Amazon ανέφερε ότι το 30% των πωλήσεων της είναι πλέον από τα προτεινόμενα προϊόντα και από προτεινόμενα παρεμφερή προϊόντα που μπορούν να φανούν χρήσιμα από την προηγούμενη αγορά.

- **Στοχευμένο marketing.** Τα Big Data δίνουν την δυνατότητα για στοχευμένο marketing πάνω στις ανάγκες των ανθρώπων ανα χώρα, πόλη ακόμα και βάση γεωγραφικών περιοχών μέσα στην πόλη, αναλύοντας τις οικονομικές δυνατότητες των καταναλωτών, τις προτιμήσεις τους, τον μέσο όρο ηλικίας, τα ήθη και έθιμα κλπ.

- **Real-time 'ανάλυση' πελατών.** Πολλές εταιρίες αναλύουν real-time την συμπεριφορά των καταναλωτών όταν βρίσκονται μέσα στο κατάστημα τους μέσα από σένσορες, από στοιχεία που βρίσκουν online, ακόμα και από το gps του κινητού που ενημερώνει τους ενδιαφερόμενους πόση ώρα αφιερώνουν ανά περιοχή του καταστήματος. Αυτό επιτρέπει στους πωλητές να έχουν μια καλύτερη εικόνα για τις προτιμήσεις των καταναλωτών και να υπάρχει η αντίστοιχη εξατομικευμένη προσφορά.

4.9 Τοποθεσία και μεταφορά

Τοποθεσία και μεταφορά: Μέσω του GPS που πλέον βρίσκεται ακόμα και στα κινητά μας ή στα αυτοκίνητα, είναι πολύ εύκολο να ξέρουν οι ενδιαφερόμενοι την ακριβή τοποθεσία του κάθε καταναλωτή. Τα προσωπικά στοιχεία χρησιμοποιούνται για να δημιουργήσουν ένα προφίλ του καταναλωτή σε σχέση με τα μέρη από τα οποία περνάει συνήθως, τα καταστήματα τα οποία επισκέπτεται κλπ. Επίσης εταιρίες που χρησιμοποιούν εφαρμογές για να παρέχουν βέλτιστες διαδρομές στους πελάτες της, μαζεύουν και αναλύουν ογκωδέστατα δεδομένα που προκύπτουν από την τοποθεσία των ανθρώπων κάθε χρονική στιγμή. Πολλές εφαρμογές ήδη χρησιμοποιούνται και άλλες βρίσκονται σε εξέλιξη και θα τις παραθέσω παρακάτω.

- **Βέλτιστη Διαδρομή.** Εφαρμογές σαν το γνωστό σε όλους μας google maps που αναλύοντας την κίνηση στους δρόμους καθορίζει την βέλτιστη διαδρομή από το σημείο Α στο Β. Η βελτίωση και η ακρίβεια που παρέχεται συνεχώς βελτιώνεται όσο οι real-time τεχνικές ανάλυσης των δεδομένων γίνονται ολοένα και αποτελεσματικότερες.

- **Έξυπνα αυτοκίνητα.** Υπάρχουν κάποια πρωτότυπα αυτοκίνητα (τα TESLA) τα οποία επιτρέπουν μέσα από μια σειρά από σένσορες στο αμάξι να κινείται αυτόνομα, αναλύοντας real-time σε microseconds τα δεδομένα από την άμεση περιοχή γύρο του. Αυτή η τεχνολογία αναπτύσσεται συνεχώς και βασίζεται στην πολύ γρήγορη ανάλυση των δεδομένων. Αναμένεται τα επόμενα χρόνια τα αυτοκίνητα στους δρόμους να είναι τελείως

αυτόνομα. **Μεγαλύτερη ασφάλεια.** Από τα δεδομένα τοποθεσίας είναι ευκολότερο να εντοπιστεί ένας άνθρωπος με πρόβλημα υγείας ή να εντοπιστεί από τα όργανα ασφαλείας σε περίπτωση που διατρέχει κίνδυνο.

Βιβλιογραφία

[1] Werner Hildenbrand (8 March 2015). Core and Equilibria of a Large Economy. (PSME-5). Princeton University Press.

[2] IBM Redbook: IMS Primer

[3] Barker, Richard (1990). CASE Method: Entity Relationship Modelling.

[4] Moore, Gordon E. (1965-04-19). "Cramming more components onto integrated circuits".. Retrieved 2016-07-01.

[5] McKinsey, 2011

[6] Juniper networks, 2012

[7] P.Chatterjee, 2013

[8] Chemawat, 2004

[9] Big Data technologies: A surveyAuthor links open overlay panelAhmedOussousaFatima-ZahraBenjellounaAyoubAit LahcenabSamirBelfkiha

[10] Rexer Analytics Data Miner Surveys (2007–2015)[41]

[11] Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

[12] Sarle, Warren (1994). "Neural Networks and statistical models". CiteSeerX 10.1.1.27.699.

[13] Kohavi, Ron; Longbotham, Roger (2017). "Online Controlled Experiments and A/B Tests". In Sammut, Claude; Webb, Geoff (eds.). Encyclopedia of Machine Learning and Data Mining

[14] ESRL, 2011

[15] Berkeley, 2012

[16] Adamson, 2016

[17] O'Dowd, 2016