ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΔΙΔΑΚΤΙΚΗ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ»

ΚΑΤΕΥΘΥΝΣΗ: ΔΙΚΤΥΟΚΕΝΤΡΙΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

¨Ανάπτυξη εφαρμογής για την ανάκτηση και κατηγοριοποίηση δεδομένων μέσω του κοινωνικού δικτύου Twitter, αναφορικά με τις διατροφικές συνήθειες πολιτών.¨

ΛΑΜΠΡΟΥ ΣΤΥΛΙΑΝΗ ΑΜ: ΜΕ 13045

ΕΠΙΒΛΕΠΟΥΣΑ: ΜΑΛΑΜΑΤΕΝΙΟΥ ΦΛΩΡΑ

ΠΕΙΡΑΙΑΣ 2016

Περιεχόμενα

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου στην κατεύθυνση «*Δικτυοκεντρικά Συστήματα*» του Μεταπτυχιακού προγράμματος «Διδακτική της Τεχνολογίας και Ψηφιακά Συστήματα» του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιά για όσα μου προσέφεραν.

Επιπλέον θα ήθελα να εκφράσω την ιδιαίτερη ευγνωμοσύνη μου στην επιβλέπουσα καθηγήτρια μου, Μαλαματένιου Φλώρα, για τις συμβουλές, την συνεχή επίβλεψή της και τη βοήθειά της.

Περίληψη

Ο όγκος των δεδομένων στον κλάδο της υγείας αυξάνεται συνεχώς και η κατάσταση αναμένεται να αλλάξει δραματικά τα επόμενα χρόνια. Για να επιτευχθεί η σωστή διαχείριση των υπάρχοντών δεδομένων, υπάρχουν συγκεκριμένοι τρόποι και τεχνικές. Ωστόσο, εντοπίζονται και τεχνικές αξιολόγησης οι οποίες δεν είναι τόσο αποδοτικές, όσο θα ήταν η τεχνική της αναλυτικής των δεδομένων, εφόσον αυτή εφαρμοζόταν. Αυτές οι τεχνικές ανάλυσης δεδομένων έχουν την ικανότητα να τα διαχωρίζουν και να διαχειρίζονται την ανάλυσή τους με τέτοιο τρόπο, ώστε να παρέχονται αξιόπιστες πληροφορίες. Με τη συγκεκριμένη διαδικασία, τεράστια ποσότητα δεδομένων του κλάδου της υγείας, αναλύεται με τέτοιο τρόπο, ώστε να υπάρξει βαθύτερη κατανόηση των αποτελεσμάτων. Αυτή η πιο ορθή διαχείριση δεδομένων καθιστά εφικτή την πρόοδο στον ευρύτερο τομέα της υγείας. Τέτοιες τεχνικές ανάλυσης μεγάλων δεδομένων είναι εξίσου αποτελεσματικές σε ποικίλους οργανισμούς υγειονομικής περίθαλψης. Συγκεκριμένα, τα πλεονεκτήματά αυτών των τεχνικών έχουν αντίκτυπο τόσο σε ατομικό επίπεδο, για τον εκάστοτε ιατρό, όσο και σε ευρύτερο επίπεδο, για μεγάλους οργανισμούς παροχής υγειονομικής περίθαλψης.

Η βιομηχανία παροχής υπηρεσιών υγείας αλλάζει με απίστευτη ταχύτητα. Ένας από τους σημαντικότερους παράγοντες για την αλλαγή αυτή είναι η δραματική έξαρση στην προβολή τέτοιων υπηρεσιών, που προέρχονται κυρίως από τα μέσα κοινωνικής δικτύωσης. Πλέον τα μέσα κοινωνικής δικτύωσης (Social media) αποτελούν για το ευρύτερο κοινό ένα βασικό εργαλείο αναζήτησης πληροφοριών σχετικών με την υγεία. Από τη φύση τους τα κοινωνικά μέσα ενημέρωσης και δικτύωσης επιτρέπουν την αμφίδρομη επικοινωνία στο κοινό τους. Με αυτόν τον τρόπο γίνεται αποδοτικότερη η αλληλεπίδραση μεταξύ ασθενών, ιατρών και των διαφόρων βιομηχανιών υγειονομικής περίθαλψης. Αξιοσημείωτο επίσης θεωρείται το γεγονός ότι οι συζητήσεις για διάφορα θέματα υγείας και η πληθώρα ποικίλων τέτοιων πληροφοριών είναι διαθέσιμη σε παγκόσμιο επίπεδο.

Η επιρροή της υγειονομικής περίθαλψης είναι έντονη και αμβλύνεται σε διάφορες ομάδες ατόμων, όπως των ηγετών μιας ομάδας, των ασθενών, των ιατρών, των διάφορων οργανισμών, καθώς και κυβερνητικών φορέων. Με αυτόν τον τρόπο καθημερινά δημιουργείται ένας μεγάλος όγκος πληροφοριών σχετικών με την υγεία.Οι δυνατότητες

που παρέχονται από τα διαθέσιμα δεδομένα των κοινωνικών δικτύων στην κλάδο της υγείας είναι πολύ σημαντικές. Ιδιαίτερα στην υγειονομική περίθαλψη η άντληση τέτοιων δεδομένων από τα κοινωνικά μέσα δικτύωσης γίνεται κυρίως με τη χρήση tweets. Η επιλογή του Twitter για αυτόν τον σκοπό δεν είναι τυχαία, καθώς συνεχώς αυξάνεται η άντληση τέτοιων πληροφοριών μέσω αυτού.

Ακόμα και σήμερα, το Twitter, όπως δημιουργήθηκε από τους ιδρυτές του, είναι πραγματικά το πιο απλό μέσο για να βρούμε πολλά σχόλια διαφόρων ατόμων και κοινωνικών ομάδων για ένα κοινό θέμα που μας ενδιαφέρει. Σε αυτό συμβάλλει το γεγονός ότι το Twitter έχει ενσωματωμένη μέθοδο κατηγοριοποίησης για όλα αυτά τα tweets. Σημαντικό επίσης είναι να τονιστεί ότι σημειώθηκε αύξηση από περίπου 400.000 tweets κατά το πρώτο τρίμηνο του 2007 σε 4.000.000.000 tweets κατά το πρώτο τρίμηνο του 2010. Επομένως καταλαβαίνουμε ότι μιλάμε για σχεδόν 45 εκατομμύρια tweets ανά ημέρα.

Στην παρούσα διπλωματική εργασία υλοποιήθηκε μια εφαρμογή με τη χρήση τεχνολογιών μεγάλων δεδομένων η οποία περιλαμβάνει δύο επιμέρους αλγορίθμους και έχει ως στόχο την ανάλυση, κατηγοριοποίηση και παρουσίαση των δεδομένων που αφορούν τις διατροφικές συνήθειες πολιτών που συλλέγονται μέσω του κοινωνικού δικτύου Twitter.

Ο αρχικός αλγόριθμος έχει ως στόχο τη συλλογή δεδομένων που σχετίζονται με τις διατροφικές συνήθειες των πολιτών τριών Ευρωπαϊκών χωρών, της Ελλάδας, της Αγγλίας και της Γαλλίας, επιλέγοντας τα κατάλληλα tweets μέσω του Twitter. Στη συνέχεια, ο δεύτερος αλγόριθμος αναλύει και κατηγοριοποιεί τα δεδομένα αυτά σε υγιεινά και ανθυγιεινά. Αυτό επιτυγχάνεται δίνοντας ένα ειδικό βάρος σε συγκεκριμένες λέξεις, έτσι ώστε να ελέγχονται και να αξιολογούνται μόνο τα Tweets που σχετίζονται με την υγεία και τη διατροφή. Τέλος, τα σημεία στα οποία εντοπίζονται οι αναφορές των Tweets προβάλλονται σε ένα χάρτη.

Λέξεις κλειδιά: Healthcare Big data, Healthcare Analytics, machine learning algorithms, social media, big data technologies

Abstract

The healthcare industry is changing with incredible speed, and one of the major contributors to this change is the dramatic upsurge in healthcare communication brought on by social media. Not only has social media become a place where the public goes to seek health information, but by nature these social media channels allow for two-way public communication between patients, providers and other third parties, effectively creating the largest source of health discussions available globally today. This vast network of healthcare influencers, thought-leaders, patients, providers, organizations, and governmental entities daily create rich healthcare content, messages and signals that provide incredible value if it is segmented, analyzed and curated in a meaningful way to answer your unique questions and needs.

The health data volume is increasing and the graph is expected to breed dramatically in the years ahead. However, the current data managing ways that are residing as prominent techniques of evaluation are not as capable as data analytics can prove to be. These data analysis techniques have the capacity to capture, process, distribute and manage the analysis in specific form that makes it easy to get reliable information. With this particular evaluation, vast amount of patient-related health data is analyzed in a better way to get a deeper understanding of outcomes, which may be applied at the point of care for better facilities.

Some of the most interesting on Twitter is that four and a half years after Jacks historic announcement, Twitter, as created by its founders, is really simple. So you try to find tweets that are all talking about a common subject that interests you. That's because Twitter has no built in method of categorizing all those tweets. And when I say "ALL those tweets", keep in mind that Twitter went from roughly 400,000 tweets in the first quarter of 2007 to 4,000,000,000 in the first quarter of 2010.

This thesis implemented an application using big data technologies including two individual algorithms and aims the analysis, classification and presentation of data on dietary habits citizens collected through the social network Twitter.

The original algorithm is designed to collect data relating to the eating habits of citizens three European countries, Greece, England and France, by selecting the appropriate tweets via Twitter. Then, the second algorithm analyzes and categorizes this data to healthy and unhealthy. This is achieved by giving a special weight in certain words, so be checked and evaluated only the Tweets related to health and nutrition. Finally, the points at which identify reports Tweets are displayed on a map.

Keywords: Big data, Healthcare Analytics, machine learning algorithms, social media, big data technologies

Λίστα Εικόνων

Λίστα Πινάκων

1. Μεγάλα Δεδομένα και Υγεία

1.1. Εισαγωγή

Η έννοια των μεγάλων δεδομένων πρωτοεμφανίστηκε στο 1970, όταν οι επιστήμονες συνειδητοποίησαν ότι στερούνταν τα εργαλεία για την ανάλυση συνόλων δεδομένων μεγάλου μεγέθους [1]. Μέχρι τότε, τα μεγάλα στοιχεία ήταν απλώς αρκετές εκατοντάδες megabytes σε αντίθεση με την σημερινή κατάσταση που ουσιαστικά αναφερόμαστε σε terabytes ποσότητας δεδομένων[2].Τα μεγάλα δεδομένα αναφέρονται σε πολλές ειδικότητες εφαρμόζονται από συσκευές ανίχνευσης πληροφοριών και λογισμικού [3].Τέτοια παραδείγματα αποτελούν τα αρχεία καταγραφής ιστού από ιστοσελίδες, όπως το Google ή το Facebook που καταγράφουν αυτόματα τις πληροφορίες των χρηστών σε κάθε επίσκεψη[4].

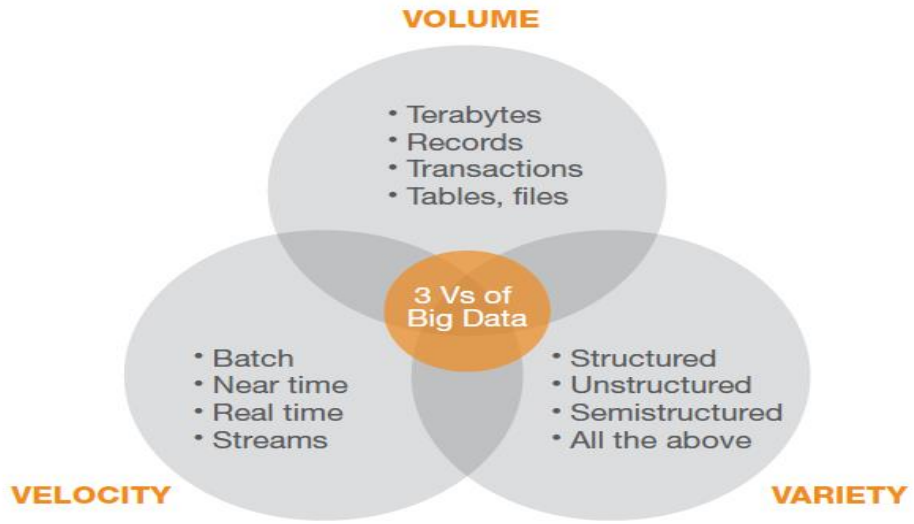
Το Κεφάλαιο 1 εστιάζει στην αναλυτική στον τομέα της υγείας. Στον συγκεκριμένο τομέα τα δεδομένα συνηθίζεται να είναι πολύ μεγάλα σε όγκο, [2] για αυτόν λοιπόν τον λόγο αρχικά γίνεται αναφορά στην έννοια των Μεγάλων Δεδομένων- Big Data[5]. Συγκεκριμένα δίνεται ο ορισμός τους και αναφέρονται όλες οι βασικές πληροφορίες αναφορικά με την υπόσταση τους και την ευρεία τους εφαρμογή και χρήση. Επίσης γίνεται μια συνολική αναφορά στα βασικά χαρακτηριστικά αναφορικά με τον όρο του Εξόρυξη γνώμης (Opinion Mining)[6].

1.2. Ορισμοί Μεγάλων δεδομένων

Το Gartner το 2012 έδωσε τον εξής ορισμό: «Τα big data είναι υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών»[7]. Στα «μεγάλα δεδομένα» συγκαταλέγονται όλες οι πληροφορίες των μέσων κοινωνικής δικτύωσης (social media) που είναι προσβάσιμες σε όλους μας και βρίσκονται στο Διαδίκτυο, δηλαδή φωτογραφίες, βίντεο και κείμενα, καθώς και όλα τα «κλειστά δεδομένα» των διαφόρων εταιριών αλλά και των κυβερνήσεων [8].
Μιλώντας για μεγάλα δεδομένα ουσιαστικά αναφερόμαστε σε πολύ μεγάλα σύνολα δεδομένων που χαρακτηρίζονται από πολύπλοκες δομές. Είναι εύκολο λοιπόν να

καταλάβουμε πως είναι δύσκολο να επεξεργαστούν χρησιμοποιώντας παραδοσιακές μεθόδους και εργαλεία επεξεργασία απλών δεδομένων [9].

Συγκεκριμένα αναφερόμαστε [10]:
Στη σύλληψη, στην αποθήκευση, στην μορφοποίηση, στην εξόρυξη, στην επιμέλεια, στην ολοκλήρωση, στην ανάλυση και τέλος στην οπτικο-ποίηση όλων αυτών των δεδομένων που χρίζουν επεξεργασία [11].



Εικόνα 1: Χαρακτηριστικά των μεγάλων δεδομένων.

Σύμφωνα με την Gartner Inc (Information Technology Research Company) μας παρέχεται ένας ιδιαίτερα δημοφιλής ορισμός αναφορικά με τα μεγάλα δεδομένα ο οποίος και είναι το μοντέλο "3V"[12]. Βασιζόμενοι σε αυτόν τον ορισμό αποδίδονται τρία θεμελιώδη χαρακτηριστικά στα μεγάλα δεδομένα: [13]

- υψηλό όγκο της μάζας των δεδομένων,
- υψηλή ταχύτητα της ροής των δεδομένων,
- και την υψηλή ποικιλία από διάφορους τύπους δεδομένων.

1.3. Εξόρυξη γνώμης

Μια πολύ χρήσιμη και σημαντική πληροφορία που μπορούμε να αντλήσουμε από τα μέσα κοινωνικής δικτύωσης είναι το τι σκέφτονται οι άλλοι και πως το σκέφτονται και αυτό γίνεται εφικτό με την Εξόρυξη γνώμης (Opinion Mining)[14]. Με την αυξανόμενη διαθεσιμότητα πόρων, νέες ευκαιρίες και προκλήσεις προκύπτουν για αυτούς που θέλουν να αντλήσουν δεδομένα χρησιμοποιώντας ενεργά τις τεχνολογίες των πληροφοριών. Η ξαφνική έκρηξη της δραστηριότητας στον τομέα της εξόρυξης γνώσης και ανάλυσης συναισθήματος, η οποία ασχολείται με την υπολογιστική επεξεργασία της γνωμοδότησης, του συναισθήματος και της θεματολογίας του κειμένου, έχει έτσι συνέβη, τουλάχιστον εν μέρει, ως άμεση απάντηση στην απότομη αύξηση του ενδιαφέροντος σε νέα opinion-driven συστήματα.[15]

Η αυτόματη εξόρυξη γνώμης έχει αναδειχτεί σε ένα από τα αντικείμενα που προσελκύουν ολοένα και αυξανόμενο ενδιαφέρον από εμπορικούς οργανισμούς, [15] καθώς η εξαγόμενη πληροφορία έχει εμπορική αξία και μπορεί να χρησιμοποιηθεί για διαφημιστικούς σκοπούς αλλά και για την βελτίωση των υπηρεσιών.

Σύμφωνα με μελέτες το 81% των χρηστών του διαδικτύου έχει αξιολογήσει τουλάχιστον μία φορά κάποιο προϊόν χρησιμοποιώντας το διαδίκτυο και το 73% με 87% όλων των χρηστών άλλαξε τη γνώμη τους για κάποιο εστιατόριο ή ξενοδοχείο διαβάζοντας μια γενική άποψη αυτών η οποία έχει αντληθεί από γνώμες άλλων.

Η ιστορία της εξόρυξης γνώμης ξεκινά το 2003 σε δημοσίευση στα πλαίσια του συνεδρίου WWW (WWW conference)[15]. Η δημοσίευση στο συγκεκριμένο συνέδριο μπορεί εν μέρει να εξηγήσει την δημοτικότητα του όρου της εξόρυξης γνώμης μεταξύ των κοινοτήτων που είναι προσανατολισμένες στην κατεύθυνση αναζήτησης στο διαδίκτυο (Web search) ή ανάκτησης πληροφορίας (information retrieval). Σύμφωνα με την εν λόγω δημοσίευση, η ιδανική εφαρμογή για εξόρυξη συναισθήματος «θα μπορούσε να επεξεργαστεί ένα σύνολο από δεδομένα αναζήτησης, δημιουργώντας μία λίστα των κύριων χαρακτηριστικών αυτών και συνοψίζοντας τις απόψεις που επικρατούν για κάθε ένα από αυτά τα χαρακτηριστικά σε θετικές, ουδέτερες και αρνητικές».

1.4. Στάδια εξόρυξης γνώμης

Εντοπισμός γνώμης ή συναισθήματος: Το πρώτο βήμα για την εξόρυξη γνώμης, είναι ο εντοπισμός των σχετικών σημείων στο κείμενο[6]. Η διαδικασία αυτή μπορεί να περιγραφεί ως μία ευρύτερη κατηγοριοποίηση του κειμένου σε αντικειμενικό ή υποκειμενικό και βασίζεται συνήθως στην εξέταση των υπαρχόντων επιθέτων αλλά και επιρρημάτων στις προτάσεις. [16]

Ανακάλυψη του αντικειμένου- στόχου: Η δεύτερη διαδικασία είναι η ανακάλυψη του αντικειμένου- στόχου στο οποίο αναφέρεται η γνώμη που εντοπίστηκε προηγουμένως και μπορεί να χρησιμοποιηθεί συμπληρωματικά με την κατηγοριοποίησης της πόλωσης. Η δυσκολία υλοποίησης της διαδικασίας αυτής εξαρτάται κατά πολύ από το είδος της ανάλυσης. Για παράδειγμα ανάλογα με το ουσιαστικό ή τα ουσιαστικά που υπάρχουν μέσα σε μια φράση, μπορούμε με σχετική ευκολία να προσδιορίσουμε την θεματολογία του συγκεκριμένου κειμένου. Βέβαια, είναι πιθανό σε μία πρόταση που περιέχει συναίσθημα/άποψη να αναφέρονται περισσότερα του ενός αντικειμένου, όπως συμβαίνει στην περίπτωση των συγκριτικών προτάσεων. Ένας τρόπος προσδιορισμού των συγκριτικών προτάσεων είναι η ανίχνευση για την ύπαρξη συγκριτικών επιθέτων και επιρρημάτων, υπερθετικών επιθέτων και κάποιων άλλων χαρακτηριστικών λέξεων όπως ίδιος, διαφέρω, προτιμώ και άλλων συναφών. [17-18]

Κατηγοριοποίηση: Με δεδομένο ένα τμήμα του κειμένου, το επόμενο βήμα είναι η κατηγοριοποίηση της άποψης που εκφράζεται. Η κατηγοριοποίηση μπορεί να γίνει μεταξύ των δύο αντίθετων Κατηγοριών (Θετικό ή αρνητικό) συναισθήματος ή σε περίπτωση σύγχυσης στην πλησιέστερη εκ των δύο αυτών. Αν η άποψη/συναίσθημα θεωρηθεί ως δυαδικό χαρακτηριστικό, η κατηγοριοποίηση πόλωσης είναι η δυαδική κατηγοριοποίηση συναισθήματος είτε σε συνολικά θετική ή σε συνολικά αρνητική άποψη. Επειδή στην συγκεκριμένη εργασία μας απασχολεί ένα microblogging site όπως το Twitter, όπου τα κείμενα που έχουμε είναι μικρά και μη πολύπλοκα, η διαδικασία της κατηγοριοποίησης είναι ευκολότερη σε σχέση με μεγάλα και σύνθετα κείμενα, όπως ένα ειδησεογραφικό άρθρο που μπορεί να περιέχει «κακά» νέα χωρίς ουσιαστικά να χρησιμοποιεί κανένα υποκειμενικό όρο. Επιπλέον, μεγάλο πλήθος κειμένων περιέχει τόσο θετικές όσο

αρνητικές απόψεις ώστε τελικά ο στόχος θα πρέπει να είναι ο προσδιορισμός του κυρίαρχου συναισθήματος εντός αυτού. [19]

Δυσκολίες: Η κατηγοριοποίηση της άποψης σε θετική και αρνητική είναι κάτι περίπλοκο, ακόμα και για τους ίδιους τους ανθρώπους. Πολλές φορές ο άνθρωπος δεν λέει ξεκάθαρα πως αισθάνεται για κάποιο θέμα, και είναι δύσκολο να καταλάβει κανείς αν η άποψή του είναι θετική ή αρνητική.

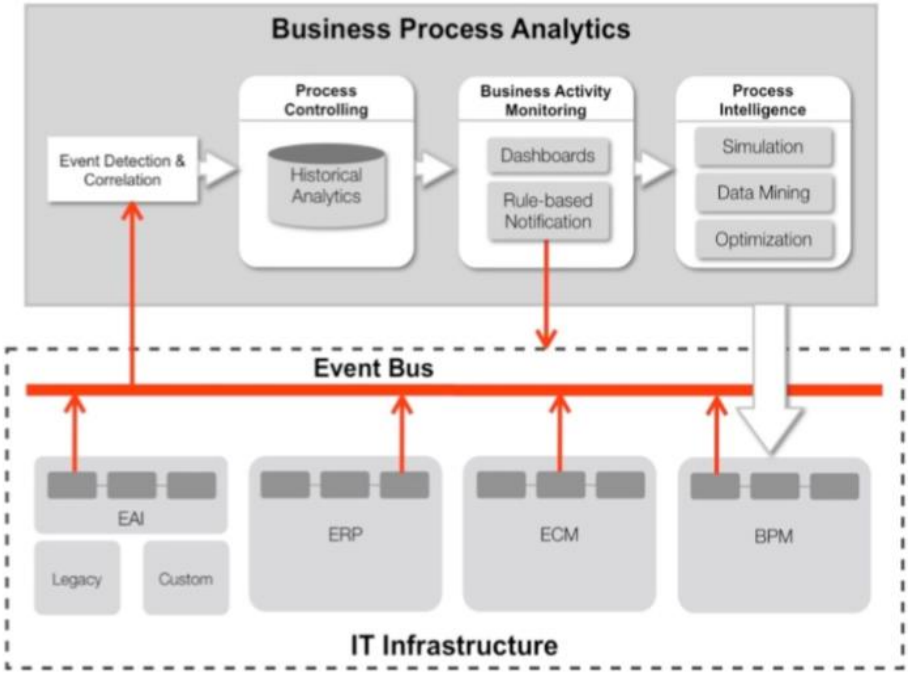
Ένα ακόμα στοιχείο που δυσχεραίνει την αυτόματη εξόρυξη συναισθήματος είναι η παράθεση αστείων, ιδιωματισμών, ειρωνείας και πολιτιστικών αναφορών καθώς απαιτεί ευρύτερη γνώση και εξοικείωση με τα σχετικά γεγονότα και τις αναφορές. Ο παράγοντας αυτός καθιστά την κατηγοριοποίηση κειμένου ιδιαίτερα πολύπλοκη, αν όχι αδύνατη για τους περισσότερους αλγόριθμους. [9-11]

1.5. Αναλυτική Μεγάλων Δεδομένων

Αρχικά για την ανάλυση των διαδικασιών, απαιτείται η καταγραφή των γεγονότων, ο συσχετισμός τους και η αξιολόγησή τους, κατά την διάρκεια ζωής μιας διαδικασίας. Τα γεγονότα αυτά υποδεικνύουν διάφορες αλλαγές στα διάφορα μέρη που αποτελούν την διαδικασία. Αυτά είναι δεδομένα, πληροφοριακά συστήματα, δράστες και διεργασίες. Τέτοιες πληροφορίες είναι η καταγραφή της εισόδου/εξόδου των χρηστών, και πότε και πιες αλλαγές έγιναν σε διάφορα δεδομένα. Όταν πραγματοποιείται μια τέτοια ανάλυση των διαδικασιών, αναλόγως του μεγέθους και του πλαισίου των διαδικασιών που μελετάμε, συγκεντρώνονται και τα δεδομένα. Αν επικεντρωθεί η έρευνα μόνο σε μια διαδικασία της επιχείρησης τότε ως πηγή θα είναι εκείνα μόνο τα γεγονότα που έχουν άμεση σχέση με την διαδικασία αυτή. Όταν το πλαίσιο της έρευνας μεγαλώνει, οι πηγές αυξάνονται, και υπάρχει περίπτωση να χρειαστεί να συμπεριληφθούν στοιχεία και γεγονότα από εξωτερικές πηγές. [20]

Όπως αναδεικνύεται και στο σχήμα τα στάδια της Ανάλυσης των Επιχειρησιακών Διαδικασιών, είναι ο Έλεγχος, η Παρακολούθηση και η Ευφυΐα. Στο κάτω επίπεδο είναι η βάση του πληροφοριακού συστήματος η οποία είναι ετερογενής και αποτελείται από

πολλά συστήματα (ERP, EAI, ECM,) [21,22]. Όλα αυτά τα συστήματα αλληλοεπιδρούν μεταξύ τους για την ολοκλήρωση των διεργασιών προκειμένου να φέρουν εις πέρας την επιχειρησιακή διαδικασία [23]. Έτσι κάθε ένα από αυτά αποτελεί και πηγή πληροφοριών και γεγονότων για κάθε διαδικασία που πραγματοποιεί αφού καταγράφονται ηλεκτρονικά τα δεδομένα για κάθε γεγονός [24]. Όλες αυτές οι πληροφορίες χρησιμοποιούνται για ιστορική ανάλυση παλαιών διαδικασιών, έλεγχο σε πραγματικό χρόνο και πρόβλεψη μελλοντικών γεγονότων. [25]

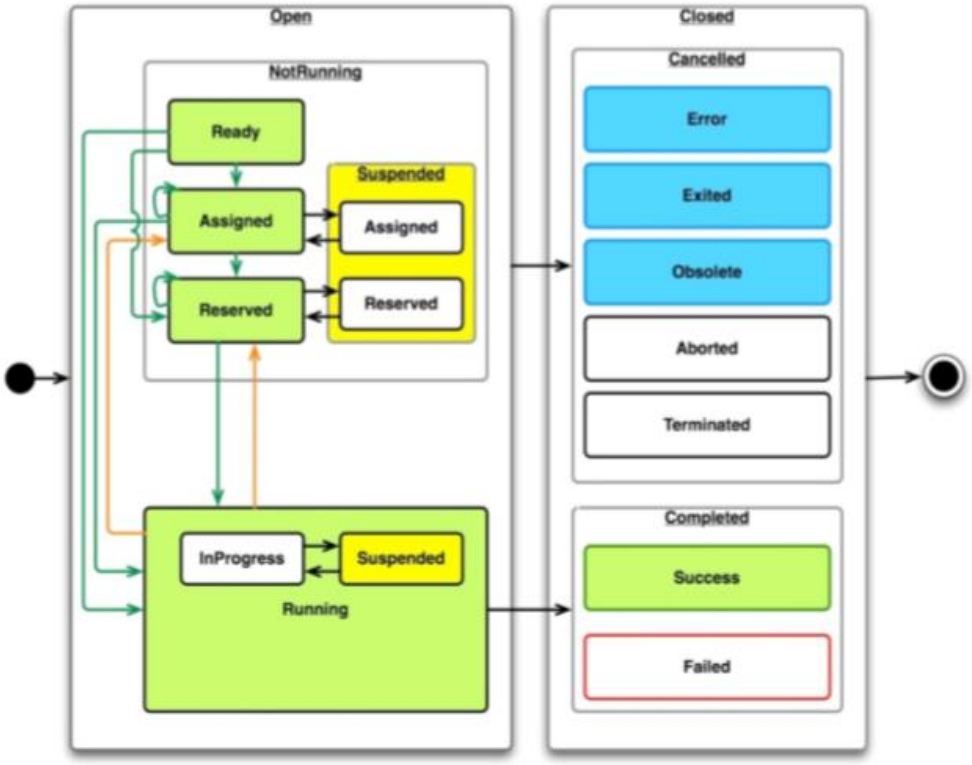


Εικόνα 2:Στάδια Ανάλυσης Επιχείρησης

1.6. Πηγαία Μορφή των Μεγάλων Δεδομένων

Για να γίνει εφικτή η ανάλυση των διαδικασιών μέσα από τα δεδομένα που παράγονται από τις διαδικασίες (γεγονότα) είναι απαραίτητο να έχουν μια μορφή η οποία δεν είναι απόλυτα συγκεκριμένη και στενά προσαρμοσμένη. Αυτό συμβαίνει καθώς τα συστήματα που επεξεργάζονται τέτοια δεδομένα είναι γενικής μορφής και δεν υποστηρίζουν σημασιολογικά την εκάστοτε επιχείρηση, αλλά είναι φτιαγμένα με κάποια γενική μορφή

που προσαρμόζεται στην επιχείρηση που τα χρησιμοποιεί. Έτσι έχει οριστεί ένα γενικό εύρος γεγονότων που αντιστοιχεί στις καταστάσεις που μπορεί να έχει μια διαδικασία [26] [standards: Wf-XML (WfMC, 2004), BPEL4People/WS-HumanTask (OASIS, 2008a, OASIS, 2008b)]. [96]

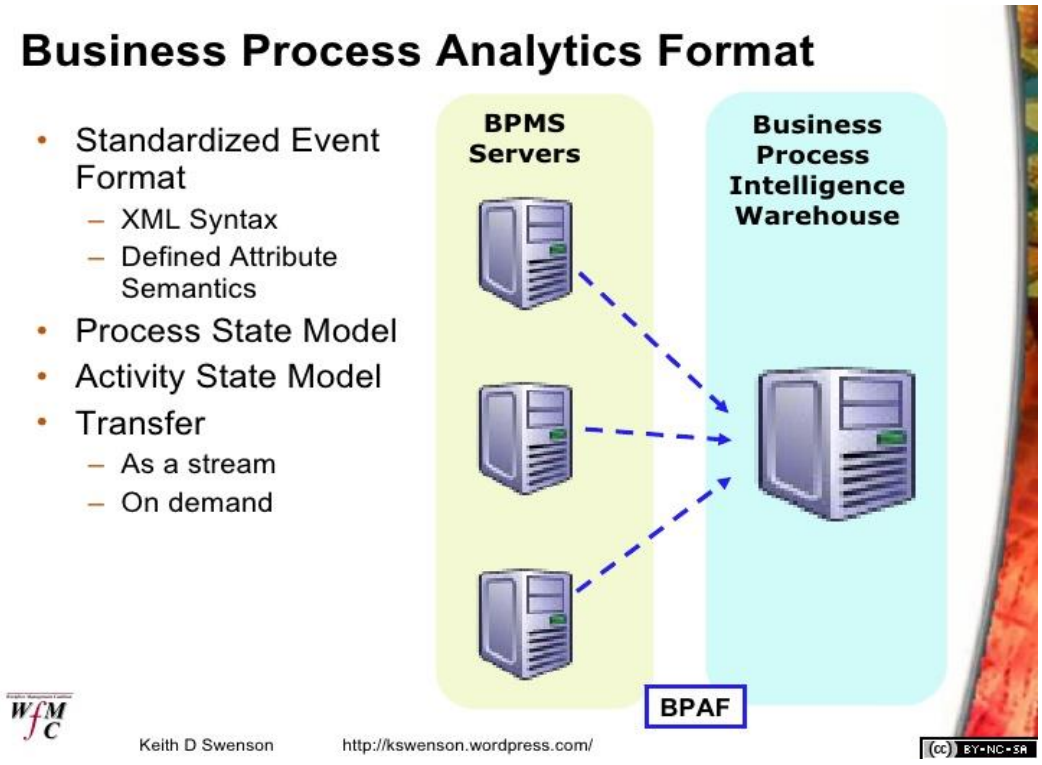


Εικόνα 3:Καταστάσεις μιας Διαδικασίας

Όπως φαίνεται και στο σχήμα, πιο αναλυτικά, οι γενικές καταστάσεις είναι ανοικτή (Open) και κλειστή (Closed). Όταν μια διαδικασία είναι ανοικτή μπορεί να μεταβεί σε διάφορες καταστάσεις, ενώ όταν κλείσει δεν μπορεί πλέον να αλλάξει καταστάσεις. Όταν μια διαδικασία είναι ανοιχτή, αυτή μπορεί να τρέχει (Open.Running) ή να είναι αδρανής (Open.NotRunning). Καθώς τρέχει μπορεί να σταματήσει και να ξανά ξεκινήσει. Εάν μια διαδικασία σταματήσει απότομα, από κάποιο λάθος, ή την ακυρώσει ο χρήστης, τότε μεταβαίνει στην κατάσταση ακυρωμένη (Closed.Cancelled). Μια διαδικασία όταν ολοκληρώνεται θεωρείται ολοκληρωμένη (Closed.Completed). Ωστόσο μπορεί να έχει επιτύχει ή όχι τον στόχο της. Όταν οι διαδικασίες τερματίζονται και δεν επιτυγχάνουν τον

σκοπό τους η κατάστασή τους είναι ανεπιτυχής (Closed.Completed.Failed). Από την άλλη αν επιτύχει θεωρείται επιτυχημένη (Closed.Completed.Success). [27]

Παρά τις γενικές καταστάσεις αυτές, που είναι πιο διαδεδομένες, μπορεί να υπάρξει ανάγκη σε ένα σύστημα να παράγει πληροφορίες και για άλλα γεγονότα που δεν είναι καταγεγραμμένα στο γενικό σύνολο. Για να επιτευχθεί αυτό πρέπει να προστεθούν και άλλες καταστάσεις που επιτρέπουν την καταγραφή νέων γεγονότων, προσαρμοσμένων στο περιβάλλον του πληροφοριακού συστήματος. Για παράδειγμα σε μια εταιρεία που παρήγαγε έργο σύμφωνα με κάποιες διορίες, το πληροφοριακό σύστημα των εργαζομένων της θα έπρεπε να μπορεί να ελέγχει την διορία των εργασιών που είναι αναθετημένες οι διαδικασίες της επιχείρησης. Έτσι θα έπρεπε να υπάρχει μια κατάσταση τύπου Closed.Completed.Timeout για να φανεί πως αυτή η διαδικασία δεν επιτεύχθηκε στο χρονικό διάστημα που ήταν ορισμένη.[28] Τα δεδομένα που παράγονται από τέτοια γεγονότα, που αναφέρονται στις επιχειρησιακές διαδικασίες, καταγράφονται με βάση το πρότυπο BPAF (Business Process Analytics Format για XML) Event Format [29] όπως φαίνεται στην Εικόνα 4:



Εικόνα 4: Πρότυπο BPAF γεγονότων (WfMC, 2009)

Το πρότυπο αυτό είναι ένα πρότυπο της γλώσσας XML και με βάση αυτό ένα γεγονός πρέπει να περιέχει τουλάχιστον ένα μοναδικό αναγνωριστικό (unique id / UID), το αναγνωριστικό της διαδικασίας, από που προήλθε (ποιός χρήστης/ποια διαδικασία την ξεκίνησε), χρόνο συμβάντων (χρονική σφραγίδα), και φυσικά την κατάσταση της διαδικασίας που αντιστοιχεί στο γεγονός. Με τα δεδομένα αυτά, ένα πληροφοριακό σύστημα ανάλυσης, μπορεί προσφέρει στο τμήμα λήψης αποφάσεων σημαντικές πληροφορίες για τις διαδικασίες, όπως ο χρόνος αναμονής, ο κύκλος ζωής, πόσες φορές έχει ολοκληρωθεί μια διαδικασία, κ.α.

Επιπλέον, είναι δυνατόν να προστεθούν και άλλες πέρα από τις βασικές πληροφορίες στα γεγονότα οι οποίες είναι προαιρετικές. Για παράδειγμα, σε περιβάλλον νέφους ή σε κατανεμημένα συστήματα είναι εύλογο υπάρχει η πληροφορία από ποιόν δια κομιστή προήλθε ένα γεγονός. Έτσι μπορεί να εντοπιστεί η τοποθεσία ενός προβλήματος που μπορεί να είναι καθαρά τεχνικό, και είναι ιδιαίτερα χρήσιμο ακόμη περισσότερο όταν αναφερόμαστε σε ένα περιβάλλον νέφους με πάρα πολύ μεγάλες διαστάσεις. Τέτοιου είδους επιπρόσθετα δεδομένα μπορούν να εμπεριέχονται στην καταγραφή γεγονότων προκειμένου να επιτευχθεί μεγαλύτερη σχέση ανάμεσα στην επιχείρηση και το πληροφοριακό της σύστημα. Έτσι τα δεδομένα είναι πιο συνδεδεμένα με την φύση της επιχείρησης, και παρέχουν πιο λεπτομερείς πληροφορίες για την ανάλυση των διαδικασιών. Η συνηθέστερη χρήση των πεδίων αυτών είναι για αναγνωριστικά, που αντιστοιχούν σε άλλες διαδικασίες, ρόλους ή μηχανήματα που έχουν άμεση αλληλεπίδραση με την κύρια διαδικασία του γεγονότος.

1.7. Ειδικά χαρακτηριστικά του Κύκλου Ζωής Δεδομένων

Φάση 1 : Δημιουργία της πληροφορίας

Διακυβέρνηση: Την ύπαρξη κάποιας δομής στην εταιρική διακυβέρνηση η οποία να εξασφαλίζει την προστασία και την διαχείριση των προσωπικών δεδομένων συμφώνα με τις πρότυπες πολιτικές του οργανισμού. Αν υπάρχει, η δομή αυτή είναι ικανή να διασφαλίσει την ροή ελέγχου για τα δεδομένα που μεταναστεύουν στο νέφος.

Ταξινόμηση: Άμεσο ενδιαφέρον έχει το πότε και το πώς τα προσωπικά δεδομένα ταξινομούνται.

Ιδιοκτησίας: Εξετάζουμε το ποιος από τον οργανισμό είναι ο κάτοχος αυτών των προσωπικών πληροφοριών και το πώς διατηρείται η ιδιοκτησία και τα δικαιώματά του σε αυτά τα δεδομένα εάν ο οργανισμός χρησιμοποιεί τεχνολογίες υπολογιστικών νεφών.[1,2]

Φάση 2 : Χρήση

Καταλληλόλητα: Γίνεται εξέταση της χρήση των δεδομένων, δηλαδή εάν χρησιμοποιούνται για τον σκοπό που αρχικά η συλλογή τους πραγματοποιήθηκε. Είναι η χρήση των δεδομένων ορθή όταν αυτά βρίσκονται σε περιβάλλον νέφους, και είναι σύμφωνη με τις νομικές δεσμεύσεις που έχει αναλάβει ο οργανισμός απέναντι στους πελάτες οι οποίοι παρέχουν τα δεδομένα αυτά.

Εξωτερική και εσωτερική χρήση: Ιδιαίτερη σημασία έχει το πού γίνεται η χρήση των δεδομένων προσωπικού χαρακτήρα, στο πλαίσιο του φορέα που αρχικά τα σύλλεξε ή χρησιμοποιούνται και εκτός του οργανισμού (π.χ. σε ένα δημόσιο νέφος).

Συμμόρφωση/Αποκάλυψη: Γίνεται η διαχείριση των πληροφοριών που βρίσκονται στο νέφος με τρόπο τέτοιο, ώστε να είναι δυνατή η συμμόρφωση του οργανισμού με τις νομικές απαιτήσεις σε περίπτωση δικαστικής διερεύνησης ή προστατευτικών μέτρων;

Τρίτοι: Επίσης εξετάζεται η περίπτωση του να είναι οι πληροφορίες αποθηκευμένες στον οργανισμό και να «μοιράζονται» από κοινού με τρίτους (π.χ., υπεργολάβους ενός προγράμματος πληροφορικής η τους παρόχους υπηρεσιών του νέφους).[1,9]

Φάση 3 : Μεταφορά

Έλεγχος πρόσβασης: Τέλος θα πρέπει οι μηχανισμοί ελέγχου πρόσβασης να εξεταστούν ως προς την επάρκειά τους. Πιο συγκεκριμένα αναφορικά με το ποιος μπορεί να έχει πρόσβαση σε ποιά δεδομένα προσωπικού χαρακτήρα.

Δημόσιο/ιδιωτικό νέφος: Όταν οι πληροφορίες μεταφέρονται σε ένα υπολογιστικό νέφος, αυτό μπορεί να είναι ιδιωτικό ή δημόσιο. Ωστόσο τα προσωπικά δεδομένα θα πρέπει να προστατεύονται πάντοτε, προκειμένου να αποφευχθούν διαρροές που αφορούν στοιχεία του τελικού πελάτη και τις νομικές συνέπειες που ακολουθούν.

Απαιτήσεις κρυπτογράφησης: Σε ένα νέφος τα προσωπικά δεδομένα καλό θα είναι να κρυπτογραφούνται. Θα πρέπει να εξεταστούν λοιπόν οι συνθήκες που ισχύουν, για την κρυπτογράφηση για τα δεδομένα ταξιδεύουν στο νέφος.[9,10]

Φάση 4: Μετασχηματισμός

Εξαγωγή ειδικών χαρακτηριστικών: Εξέταση των αρχικών δομών ελέγχου πρόσβασης και στο αν οι περιορισμοί χρήσης, διατηρούνται όταν τα προσωπικά δεδομένα μεταναστεύουν στο διαδίκτυο ή εάν αυτά μορφοποιούνται για περαιτέρω επεξεργασία.

Ακεραιότητα: Η ακεραιότητα των προσωπικών δεδομένων οφείλει να διατηρείται όταν αυτά πλέον υπάρχουν στο νέφος, και συνεπώς πρέπει να εξεταστεί.

Συγκέντρωση-συσχέτιση: Εδώ εξετάζεται η διατήρηση της σχέσης χρήστη και προσωπικών δεδομένων. Τα δεδομένα όταν μεταφερθούν σε πλατφόρμα νέφους συνεχίζουν είναι αναγκαίο να σχετίζονται με ένα αναγνωρίσιμο άτομο και άρα να διατηρούν το χαρακτήρα τους σαν προσωπικά δεδομένα.[9,10]

Φάση 5: Αποθήκευση

Έλεγχος πρόσβασης: Υπάρχουν δομές ελέγχου πρόσβασης για τα δεδομένα προσωπικού χαρακτήρα, που διασφαλίζουν ότι όταν πλέον αυτά είναι υποθηκευμένα στο νέφος, μόνο τα άτομα που είναι αναγκαίο και μόνο αυτά, θα μπορούν να έχουν την αντίστοιχη πρόσβαση.

Διαθεσιμότητα/Ακεραιότητα/Εμπιστευτικότητα: Απαραίτητη είναι και η ύπαρξη μηχανισμών που να διαφυλάττουν την ακεραιότητα την εμπιστευτικότητα και την διαθεσιμότητα των δεδομένων. Ποιοι είναι οι μηχανισμοί που εξασφαλίζουν την

ακεραιότητα των δεδομένων, ποια είναι η διαθεσιμότητα που επιτυγχάνεται και διατηρείται ορθά η εμπιστευτικότητα τους όταν αυτά αποθηκεύονται σε περιβάλλον νέφους; Πως λειτουργούν αυτοί οι μηχανισμοί, και κατά πόσο είμαστε σίγουροι για την αξιοπιστία τους;

Δομημένη/Αδόμητη Αποθήκευση: Ανακύπτει το ερώτημα στο αν υπάρχει κάποιος συγκεκριμένος τρόπος ή μεθοδολογία αποθήκευσης των πληροφοριών ώστε να διευκολύνεται η πρόσβαση αλλά και κάποιων αλγορίθμων που να λειτουργούν ως αρωγοί θέματα αναζήτησης και εξόρυξης δεδομένων.

Κρυπτογράφηση: Οι περισσότερες κανονιστικές και νομοθετικές διατάξεις, σε πολλές χώρες προβλέπουν ότι ορισμένοι τύποι προσωπικών δεδομένων, πρέπει να αποθηκεύονται σε κρυπτογραφημένη μορφή. Προκύπτει λοιπόν το ερώτημα του αν και σε ποιο βαθμό, ο πάροχος των υπηρεσιών υπολογιστικών νεφών είναι σε θέση να προσφέρει υπηρεσίες συμφώνα με τις παραπάνω νομικές απαιτήσεις.[9,11]

Φάση 6 : Αρχειοθέτηση

Τεχνικοί προβληματισμοί: Είναι σημαντικό να εξετάσουμε αν το αποθηκευτικό μέσο που θα αξιοποιηθεί για την αρχειοθέτηση των πληροφοριών, θα είναι προσπελάσιμο και στο μέλλον (π.χ. οι δισκέτες δεν μπορούν πλέον να διαβαστούν γιατί οι σχετικές συσκευές ανάγνωσης έχουν αποσυρθεί).

Νομικές Δεσμεύσεις: Τα προσωπικά δεδομένα υπόκεινται σε ρυθμίσεις που υπαγορεύουν το χρονικό διάστημα που θα πρέπει να αποθηκευτούν και να αρχειοθετηθούν. Είναι λοιπόν ζωτικής σημασίας η πλήρης συμμόρφωση του παρόχου προς της απαιτήσεις αυτές. Πρέπει να είμαστε σε θέση να διασφαλίσουμε την συμμόρφωση αυτή για να επιτύχουμε το επιθυμητό αποτέλεσμα.

Κατακράτησης: Επίσης πρέπει να ξέρουμε τον χρόνο που τα δεδομένα θα διατηρούνται από τον πάροχο της υπηρεσίας. Η περίοδος διατήρησης είναι σύμφωνη με την πολιτική του οργανισμού-πελάτη.[9]

Φάση 7 : Θάνατος της πληροφορίας

Ασφαλής καταστροφή: Σημαντικό κομμάτι στην αποθήκευση πληροφοριών είναι και η πολιτική αποδόμησης και καταστροφής τους. Όταν ο πελάτης επιθυμεί να εξαφανίσει κάποιο δεδομένο πρέπει να γίνεται το συντομότερο δυνατόν και χωρίς ίχνη αν θέλουμε να διαφυλάξουμε σωστά τον προσωπικό χαρακτήρα των δεδομένων αυτών. Οφείλουμε λοιπόν να εξετάσουμε πως η διαδικασία γίνεται με τον ενδεδειγμένο τρόπο ώστε να είναι αδύνατη η μη εξουσιοδοτημένη επανάκτηση της.

Αποτελεσματικότητα: Ως προς την αποτελεσματικότητα καταστροφής των πληροφοριών είναι αναγκαίο να γνωρίζουμε αν τα δεδομένα καταστρέφονται ολοσχερώς και με τρόπο που να κάνει αδύνατη την ανάκτηση τους από τον οποιοδήποτε.[100]

Οι επιπτώσεις διαφέρουν με βάση το μοντέλο νέφους που χρησιμοποιεί ο οργανισμός, τη φάση των προσωπικών πληροφοριών στο σύννεφο, καθώς και τη φύση της οργάνωσης. Η παρακάτω ανάλυση δίνει κάποιες από αυτές τις ανησυχίες. Ωστόσο, κάθε οργανισμός πρέπει να κάνει μια εκτίμηση των επιπτώσεων στην προστασία των προσωπικών δεδομένων πριν από την μετάβαση του σε κάποιο υπολογιστικό νέφος που περιλαμβάνει προσωπικές πληροφορίες.[30]

1.8. Μεγάλα Δεδομένα στην Υγεία

Οι αιτίες για τις οποίες η εξάπλωση των μεγάλων δεδομένων άργησε να εστιάσει στον κάδο της υγεία είναι αρχικά γιατί δεν υπήρχε ουσιαστική επένδυση στον συγκεκριμένο τομέα όσο αφορά τις τεχνολογίες πληροφοριών, καθώς υπήρχε επιφυλακτικότητα για την απόδοσή του. Έπειτα άλλη μία αιτία είναι ότι υπήρχε αντίσταση για την αλλαγή των μέχρι τότε δρώμενων. Στην συνέχεια άλλη αιτία είναι ότι υπήρχε επιφυλακτικότητα για την δηκτικότητα του απορρήτου των ασθενών[31].

1.8.1. Διαφορά μεγάλων δεδομένων και κλινικών μεγάλων δεδομένων

Τα μεγάλα δεδομένα που συλλέγονται για κλινικές δραστηριότητες στον γενικότερο χώρο της υγείας συλλέγονται βάση κάποιου συγκεκριμένου πρωτοκόλλου σε αντίθεση με τα

υπόλοιπα μεγάλα δεδομένα που αφορούν άλλους κλάδους. Αποτέλεσμα αυτής της διαφοροποίησης είναι ότι τα κλινικά- ιατρικά μεγάλα δεδομένα είναι μερικώς δομημένα.

Τα κλινικά μεγάλα δεδομένα μπορούν επίσης να χρησιμοποιηθούν για τον προσδιορισμό της αιτιότητας και των αποτελεσμάτων καθώς και την σχέση μεταξύ των παραγόντων κινδύνου αναφορικά με την ασθένεια που μας ενδιαφέρει.[32]

Με την πρόοδο της γονιδιακής τεχνολογίας όλο και πιο πολλές μελέτες που αναφέρονται στους παράγοντες εστιάζουν στον προσδιορισμό της σύνδεσης του κινδύνου εμφάνισης κάποια ασθένειας με την γονιδιακή κληρονομιά. Τέτοιες μελέτες γίνονται βάση υλικού που παρέχεται από δειγματοληψία διαφόρων ασθενών. Σημαντικό είναι να τονιστεί ότι ο όγκος αυτών των δειγμάτων αναφορικά με την γονιδιακή κληρονομιά ξεπερνά κάθε άλλο δείγμα δεδομένων που αναφέρονται σε άλλες θεματικές κλινικές έρευνες. Αυτό συμβαίνει καθώς βασιζόμενοι στην βιολογική και επιστημονική προσέγγιση της γονιδιακής έρευνας κατανοούμε με ευκολία ότι το σύνολο των διαφόρων συνδυασμών των γονιδίων ανέρχεται σε τεράστιο όγκο δεδομένων. [33] Ένα παράδειγμα αποτελεί η έρευνα Koefoed συμφωνά με την οποία φαίνεται πως οι πιθανοί συνδυασμοί που είναι δυνατό να παραχθούν από τρία γονότυπο είναι περίπου 2 δισεκατομμύρια[34].

Για να μπορέσουμε να επεξεργαστούμε, να αναλύσουμε και να μελετήσουμε αυτά τα τόσο μεγάλα κλινικά δεδομένα έγιναν ποικίλες έρευνες με σκοπό τον προσδιορισμό και την δημιουργία νέων τεχνικών και εργαλείων. [35]

1.8.2. Η ποικιλομορφία των δεδομένων στην κλινική ιατρική.

Οι ιατρικές έρευνες εκτελούνται χρησιμοποιώντας εξαιρετικά μεγάλα σύνολα δεδομένων και αυτό δείχνει το ευρύ φάσμα των πόρων που χρησιμοποιούνται. Παράλληλα όμως το είδος των δεδομένων που χρησιμοποιούνται εξαρτάται από την ερώτηση της έρευνας. Τα δεδομένα από διαφορετικές υποκατηγορίες της ιατρικής έχουν ευρεία ποικιλία από άποψη όγκου, είδους, διαστάσεων (ή μεγέθους) αλλά και δείγματος. [36]

Υπάρχουν δεδομένα αίματος, ιστών, γονιδίων, X-ray ή άλλου είδους εικόνων, βίντεο από επεμβάσεις και γενικά τεράστια ποικιλία δεδομένων στον τομέα της ιατρικής. Στα

δεδομένα, επίσης, υπάρχουν διαφορές σε μέγεθος: π.χ. σε σύνολα δεδομένων γονιδιακής έκφρασης, που προέρχονται από τεχνολογίες επόμενης γενιάς, όπως αυτές που αναλύουν τα SNPs (Single Nucleotide Polymorphism)[37], τείνουν να είναι εξαιρετικά μεγάλα, ενώ δεδομένα που αναφέρονται σε κλινικές δοκιμές είναι πολύ μικρότερα [38].

Για να υπάρχει ευκολότερη οργάνωση σε όλα αυτά τα δεδομένα, Ο Phan J. H. πρότεινε τα δεδομένα στον τομέα της ιατρικής να χωρίζονται σε τέσσερα διαφορετικά επίπεδα: Το βασικό επίπεδο βιοϊατρικής γνώσης, το κλινικό και ασθενών (clinical/patient), το κυτταρικό και το επίπεδο των ιστών (cellular/tissue) και το μοριακό επίπεδο (molecular π.χ., δεδομένα γονιδίων). Συγχρόνως, τα δεδομένα έχουν και διαφορετικά επίπεδα στις διαστάσεις τους (π.χ. διαφορετικές παραμέτρους) ή στα μεγέθη (π.χ. μεγάλος όγκος). Ωστόσο χάρη στις νέες τεχνολογίες, ολοένα περισσότεροι αλγόριθμοι μπορούν να αντιμετωπίσουν αυτά τα δεδομένα.[37]

1.8.3. Μέθοδοι Προσπέλασης Κλινικών Δεδομένων.

Τεχνολογίες Αποθήκευσης Κλινικών δεδομένων

Λόγω του μαζικού μεγέθους και της πολυπλοκότητας των στοιχείων, οι βάσεις δεδομένων που χρησιμοποιούνται είναι μη σχεσιακές και μαζικής παράλληλης επεξεργασίας, όπως η Apache Hadoop και η Google BigTable για την αποθήκευση των κλινικών δεδομένων. Πολλά λογισμικά βιοστατιστικής έχουν χρησιμοποιηθεί για τον χειρισμό αυτών των μεγάλων κλινικών δεδομένων, ορισμένες από τις οποίες επιτρέπουν και την χρήση του cloud ή κατανεμημένων συστημάτων όπως το SPSS για παράδειγμα.[39]

Τεχνολογίες προ επεξεργασίας Κλινικών δεδομένων

Τα κλινικά δεδομένα είναι πολυδιάστατα και δύσκολα να επεξεργαστούν. Επομένως όταν τα κλινικά δεδομένα είναι ακατέργαστα είναι ακόμη δυσκολότερη η διαδικασία αυτή. Επίσης πολλές φορές βασίζεται, μόνο στην εμπειρία ενός ειδικού. Μερικές άλλες μέθοδοι για την προ επεξεργασία των δεδομένων είναι η κατασκευή υπολογιστικών αλγορίθμων ή

στατιστικές προσεγγίσεις. Επίσης όταν τα δεδομένα είναι διάσπαρτα ή αποθηκεμένα σε διαφορετικές βάσεις, η ενοποίηση τους είναι σημαντική.[40]

Στατιστικές μέθοδοι στα μεγάλα δεδομένα υγείας:

- Πολλές δημοφιλείς στατιστικές μέθοδοι έχουν εφαρμοστεί στην ανάλυση κλινικών δεδομένων. Η πιο κοινές περιλαμβάνουν:
- την γραμμική παλινδρόμηση και λογιστική παλινδρόμηση,
- την ανάλυση λανθάνουσας τάξης,
- την ανάλυση κύριου συστατικού,
- την ταξινόμηση και τα δέντρα παλινδρόμησης.

Επιπλέον σε αυτές περιλαμβάνονται και ο μετασχηματισμός λογαριθμικής και τετραγωνικής ρίζας, οι αλγόριθμοι naive Bayes, τα δέντρα απόφασης, τα νευρωνικά δίκτυα καθώς και κρυμμένα μοντέλα Markov. Όλες αυτές οι τεχνικές χρησιμοποιούνται για τη μελέτη των προβλημάτων σε ιατρικά δεδομένα.

Όταν ένα σύνολο δεδομένων δεν είναι ιδιαίτερα περίπλοκο, μόνο με ένα πείραμα μπορεί κανείς να καταλήξει στην αποδοχή η στην απόρριψη μίας υπόθεσης του. Ωστόσο πολλές φορές όταν τα δεδομένα είναι περισσότερο περίπλοκα, είναι σημαντικό να διεξάγονται περισσότερα στατιστικά πειράματα και δοκιμές. Έτσι γίνεται η αναγνώριση των συσχετίσεων που χρίζουν αμεσότερης και ενδελεχέστερης έρευνας.[41]

Τα μεγάλα δεδομένα έχουν αρκετούς περιορισμούς. Έτσι η ακρίβεια, η επάρκεια, η πληρότητα αλλά και άλλα μέτρα της ποιότητας των δεδομένων αποτελούν τέτοιους περιορισμούς. Η μοντελοποίηση μπορεί συχνά να οδηγήσει σε μια βεβιασμένη στατιστική συσχέτιση, μερικές φορές γνωστή ως ψευδή ανακάλυψη. [42,43]

Η πληθώρα προκλήσεων που προκύπτουν από την χρήση των μεγάλων δεδομένων αναφέρεται ακολούθως περιγραφικά:

- Το μέγεθος του δείγματος: πολλές φορές το μέγεθος του δείγματος των κλινικών δεδομένων δεν είναι αρκετά μεγάλο. Αυτό έχει ως

αποτέλεσμα οι στατιστικές ιδιότητες να είναι ανόμοιες με την ανάλυση ενός μεγαλύτερου και ακριβέστερου δείγματος. Έτσι για να έχουμε μεγαλύτερα δείγματα βασιζόμαστε σε συγκεκριμένες μεθόδους οι οποίες γίνονται αυτόματα και εστιάζουν στην συλλογή μεγαλύτερου δείγματος. Ωστόσο συχνά δεν προτιμούνται καθώς χαρακτηρίζονται από ένα αρνητικό που δεν τις καθιστά απόλυτα έμπιστες, καθώς η ευστοχία της επιλογής των δειγμάτων είναι περιορισμένη.

- Η μεροληπτική επιλογή στοιχείων: τα διάφορα δεδομένα που χρησιμοποιούνται λαμβάνονται μεροληπτικά και δεν υφίστανται στον πραγματικό κόσμο.

- Τα προβλήματα ερμηνείας.

- Οι ελλιπής τιμές.

- Τέλος προβλήματα συσχετίσεων: κάποια δεδομένα εμφανίζουν συσχετίσεις στις ιδιότητες τους και έτσι κάποιες στατιστικές αρχές δεν υφίστανται. Και έτσι οι στατιστικές μέθοδοι δεν έχουν ασφάλεια αποτελέσματα και δεν καθίστανται έμπιστες. Παράδειγμα αποτελούν οι γονιδιακές εκφράσεις.[44]

2. Η Αναλυτική στην Υγεία

2.1. Εισαγωγή

Σήμερα, όπου επικρατεί η τεχνολογία, είναι σημαντική η εξαγωγή πληροφοριών για λήψη αποφάσεων. Το μεγαλύτερο πλεονέκτημα της αξιοποίησης της ανάλυσης δεδομένων είναι η μετατροπή του κλάδου της υγείας σε μία κουλτούρα με πραγματικό γνώμονα τις πληροφορίες. Μέσω της επεξεργασίας των δεδομένων που συγκεντρώνονται με σύγχρονες τεχνικές, εξάγεται γνώση που συντελεί και υποβοηθάει στην λήψη αποφάσεων. Ειδικότερα, στον κλάδο της υγείας η διαχείριση των δεδομένων αυτών εξελίσσεται μέσα από τρείς μεγάλες φάσεις. Η πρώτη φάση είναι η συλλογή των δεδομένων, ως προς την αξιοποίησή τους. Η δεύτερη φάση είναι ο διαμοιρασμός τους, ώστε να έχουν πρόσβαση τα κατάλληλα πρόσωπα στα κατάλληλα δεδομένα. Έως τώρα όμως δεν έχει υπάρξει ουσιαστική αλλαγή στα κόστη και στην ποιότητα των υπηρεσιών στην υγεία. Η τρίτη φάση έγκειται στην αναλυτική δεδομένων υγείας (healthcare analytics) όπου μέσα από αυτά εξάγεται γνώση για την λήψη αποφάσεων, με αποτέλεσμα καλύτερες υπηρεσίες και με μικρότερο κόστος. [45]

Με την χρήση κατάλληλων πληροφοριακών συστημάτων, την εκμετάλλευση της πληθώρας των δεδομένων και την ανάλυσή τους είναι δυνατόν να βελτιωθούν οι υπηρεσίες Υγείας που προσφέρουν οι Οργανισμοί και ταυτόχρονα να μειωθεί το κόστος. Τα πλεονεκτήματα που συγκαταλέγονται στο μοντέλο της αναλυτικής στην Υγεία, είναι μεταξύ άλλων, η μείωση των ουρών αναμονής για τους ασθενής, η ασφαλέστερη θεραπευτική αντιμετώπιση καθώς μειώνονται τα ανθρώπινα λάθη, μειώνονται τα κόστη και αυξάνεται η αποδοτικότητα και φυσικά η γενικότερη βελτίωση των υπηρεσιών Υγείας. [46]

Κυριότερα, το μοντέλο της αναλυτικής παρέχει[47]:

- Ένα κοινό πλαίσιο για την αξιολόγηση της αποδοχής των analytics στον κλάδο.
- Ένα χάρτη στις εταιρείες για να μπορέσουν να υιοθετήσουν τον νέο αυτό μοντέλο.

- Ένα μέσο για την αξιολόγηση των υπηρεσιών Υγείας.

Βασιζόμενοι στην βιομηχανία των Ηνωμένων Πολιτειών (ΗΠΑ) παρατηρούμε πως εντοπίζονται τρείς ουσιαστικές εξελίξεις στον τομέα της υγειονομικής περίθαλψη όσον αφορά την διαχείριση των δεδομένων και τη χρήση της τεχνολογίας των πληροφοριών[48]. Συγκεκριμένα:

- Η συλλογή δεδομένων, η οποία ουσιαστικά βασίζεται στην ουσιαστική χρήση των ηλεκτρονικών ιατρικών φακέλων από τους οποίους αντλούνται αυτά τα δεμένα.
- Η ανταλλαγή των δεδομένων, η οποία βασίζεται στην υιοθέτηση των ανταλλαγών πληροφοριών για την υγεία
- Η ανάλυση των δεδομένων, η οποία βασίζεται στην χρήση των αποθηκών δεδομένων που κατέχουν επιχειρήσεις και σε διάφορα άλλα αναλυτικά εργαλεία[49].

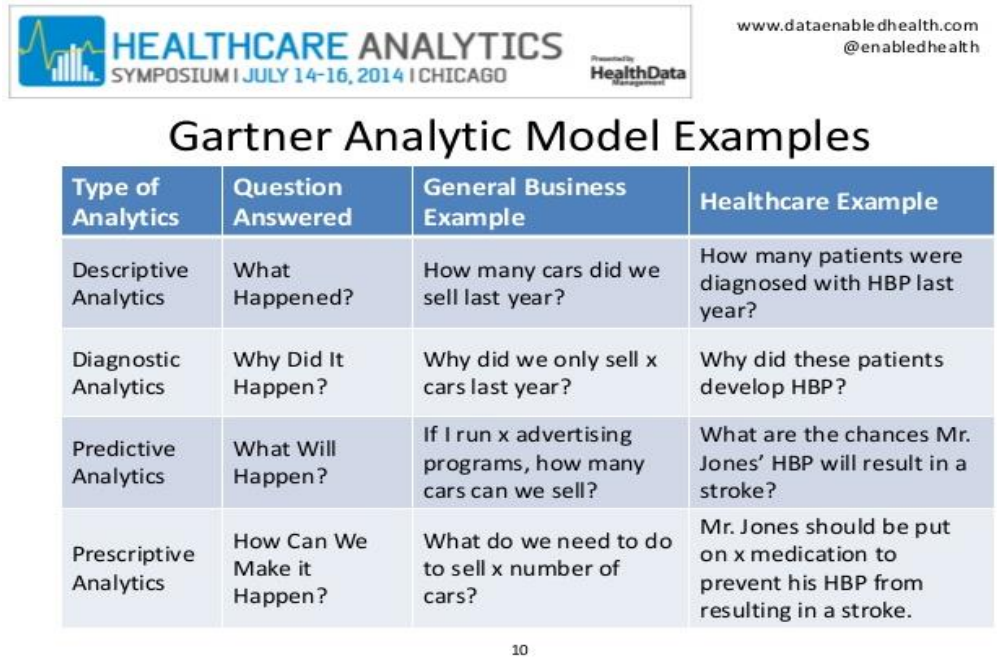
Η αναλυτική στον τομέα της υγείας εκτιμάται σε περίπου 10%, με σημαντική αύξηση αναμένεται στην επόμενη δεκαετία (Frost & Sullivan 2012)[49]. Ένα γενικά αποδεκτό πλαίσιο για την υιοθέτηση και την ουσιαστική χρήση των αποθηκών μεγάλων δεδομένων και αναλύσεων στον τομέα της υγείας θα μπορούσε είναι πολύ ευεργετική, κατά τρόπο παρόμοιο με το HIMSS [50] Emram [51-52].

2.2. Τύποι Ανάλυσης δεδομένων στην υγεία

Οι τύποι της αναλυτικής ουσιαστικά πηγάζουν από την κλάδο στον οποίο εφαρμόζονται. Ουσιαστικά λοιπόν πρόκειται για τους ευρύτερους τομείς στους οποίους έχει εφαρμοστεί η διερευνάτε η εφαρμογή της αναλυτικής. Έτσι εστιάζοντας στον τομέα της υγείας οι σημαντικότεροι τύποι αναλυτικής είναι οι ακόλουθοι:

- Αποφασιστική αναλυτική: υποστηρίζει τις ανθρώπινες αποφάσεις με οπτική αναλυτικής και με τα μοντέλα χρήστη να αντανακλούν τη συλλογιστική.
- Περιγραφική αναλυτική: υποστηρίζει την ενημέρωση από τα ιστορικά δεδομένα με την υποβολή εκθέσεων, scorecards, ομαδοποίησης κλπ.

- Προγνωστική αναλυτική: προγνωστική μοντελοποίηση με τη χρήση στατιστικών και τεχνικών μηχανικής μάθησης.
- Καθοδηγητική αναλυτική δεδομένων: είναι η σύνθεση μεγάλων δεδομένων μαθηματικών και υπολογιστικών επιστημών προκειμένου να ληφθούν αποφάσεις και να γίνουν προβλέψεις πάνω στις αποφάσεις αυτές.



Εικόνα 5: Τύποι ανάλυσης δεδομένων

2.3. Το μοντέλο υιοθέτησης της ανάλυσης δεδομένων στην υγεία

Η Κοινωνία Διαχείρισης Πληροφοριακών Συστημάτων Υγείας (ή HIMSS) [53] παρέχει ένα κεντρικό πλαίσιο κριτηρίων για την αξιολόγηση των οργανισμών όσον αφορά την χρήση αναλυτικής στην υγείας. Τα επίπεδα αυτού του μοντέλου που δίνονται ακολούθως αποτελούν έναν οδηγό για την υιοθέτηση και αξιοποίηση του μεγάλου όγκου πληροφοριών και δεδομένων στον κάδο της υγείας.

Αυτό το μοντέλο χαρακτηρίζεται από τα εξής επίπεδα:

Επίπεδο 0: Κατακερματισμένες λύσεις

Το επίπεδο 0 χαρακτηρίζεται από κατακερματισμένες λύσεις, που είναι πολύ συγκεκριμένες και με μικρή δυνατότητα εφαρμογής ανάλυσης. Αυτές οι λύσεις επικεντρώνονται σε ανάλυση δεδομένων μέσα σε ένα τμήμα του νοσοκομείου. Ότι γνώσεις παράγονται από την ανάλυση των δεδομένων στο επίπεδο αυτό έχει σχέση μόνο με τον τομέα που εξήχθη. Με την γνώση αυτή βελτιστοποιούνται περισσότερο οι υπό διαδικασίες μέσα σε ένα τμήμα έναντι όμως των γενικότερων διαδικασιών της επιχείρησης. Οι διάφορες εφαρμογές, στο επίπεδο αυτό, δεν υπάρχουν ούτε στην ίδια αποθήκη δεδομένων και ούτε μπορούν να αλληλοεπιδράσουν μεταξύ τους αφού δεν είναι αρχιτεκτονικά ολοκληρωμένες. [54]

Έτσι απαιτείται η αντιγραφή και η επαναχρησιμοποίηση των δεδομένων πολλές φορές για κάθε εφαρμογή. Αυτό συνεπάγεται πολλές εκδοχές της αλήθειας αφού δεν γίνεται να ενημερώνονται όλα τα δεδομένα ταυτόχρονα και μπορεί να οδηγήσει σε σύγχυση. Οι αναφορές που εξάγονται από τα δεδομένα επομένως είναι απαιτούν επιπλέον εργασία και μπορεί να είναι και ασυνεπής. Δεν υπάρχει ενοποιημένη διακυβέρνηση των δεδομένων ώστε να διευθύνει την ποιότητα και την αξία των δεδομένων στον οργανισμό. Οι λύσεις αυτού του τύπου είναι χρήσιμες για την δημιουργία εσωτερικών και εξωτερικών αναφορών και απαραίτητες για τα επίπεδα 3 και 4 αλλά δεν προσαρμόζονται στις πιο πολύπλοκες περιπτώσεις ανάλυσης που σχετίζονται με τα μεγαλύτερα επίπεδα. Συνοπτικά οι λύσεις αυτές απαιτούν πολύ περισσότερη επεξεργασία προκειμένου να συντηρούνται τα δεδομένα απ' ότι μια ενιαία βάση δεδομένων και επομένως και το κόστος της συντήρησής τους είναι μεγαλύτερο.

Επίπεδο 1: Ενσωματωμένη Επιχειρησιακή Βάση Δεδομένων

Για να πληρούνται οι προϋποθέσεις του επιπέδου 1, πρέπει τα βασικά δεδομένα που αλληλεπιδρούν στο πηγαίο σύστημα να είναι ενσωματωμένα σε μία ενιαία αποθήκη δεδομένων του οργανισμού (Enterprise Data Warehouse). Τέτοια δεδομένα είναι τα κλινικά δεδομένα του επιπέδου 3,τα οικονομικά δεδομένα της επιχείρησης, τα δεδομένα της αποθήκης και των υλικών ή πρώτων υλών, και τα δεδομένα των ασθενών.Αυτά τα

δεδομένα οφείλουν να ασφαλίζονται με τον καλύτερο δυνατό τρόπο. Στο επίπεδο αυτό πρέπει επίσης να είναι εφικτή η αναζήτηση δεδομένων μέσα από αποθήκες που κρατούν κάποια μετά-δεδομένα. Μια τέτοια αποθήκη προσφέρει και περιγραφές (σε φυσική γλώσσα) για το περιεχόμενό της. Είναι το πιο σημαντικό εργαλείο για την ενοποίηση των δεδομένων μέσα στον οργανισμό. Η βασική αυτή αποθήκη οφείλει να ανανεώνεται κάθε μήνα για τις διάφορες αλλαγές στο πηγαίο σύστημα. Η αρχή για την δημιουργία μιας ενιαίας Επιχειρησιακής Διακυβέρνησης των Δεδομένων είναι να διευρυνθεί η προσβασιμότητα και η ποιότητα των δεδομένων αυτών αλλά και με την προϋπόθεση πως τα δεδομένα αυτά έχουν τα απαραίτητα αναγνωριστικά και η πρόσβαση διατηρείται από τα κατάλληλα εξουσιοδοτημένα άτομα.

Επίπεδο 2: Τυποποιημένο Λεξικό για Μητρώα Ασθενών

Στο επίπεδο 2 ορίζονται οι μορφές των δεδομένων μέσα στην βάση με την χρήση κυρίαρχων λεξιλογίων και αναφορών. Περιέχουν όλα τα αναγνωριστικά ιατρών και ασθενών, τους κωδικούς των επιχειρησιακών διαδικασιών, αναγνωριστικά και κωδικούς διαγνώσεων, τμημάτων και εγκαταστάσεων. Επίσης τα ονόματα, οι ορισμοί και γενικά οι τύποι των δεδομένων προ τυποποιούνται για να υπάρχει κοινή αναγνωσιμότητα. Έτσι είναι εφικτή η διατύπωση προγραμματιστικών ερωτήσεων στο σύστημα. Στο σημείο αυτό υπάρχουν αρχεία ασθενών με τους λογαριασμούς τους και είναι απαραίτητη η ύπαρξη ομάδας για την υποστήριξη των βασικών αναλύσεων για τις πιο επικίνδυνες και ακριβές ασθένειες και τις διαδικασίες που τις συνοδεύουν. Είναι φανερή η διακυβέρνηση των δεδομένων για τα αρχεία των ασθενών.

Επίπεδο 3: Αυτοματοποιημένες Αναφορές

Το επίπεδο 3 χαρακτηρίζεται από την παραγωγή αυτοματοποιημένων αναφορών. Οι αναφορές αυτές είναι απαραίτητο να είναι συνεπής και αποδοτικές και είναι απαραίτητες για:

- την διαχείριση του οργανισμού σε εκτελεστικό και διοικητικό επίπεδο,

- την ανάλυση των επιδόσεων σε επίπεδο διεύθυνσης και διαχείρισης.

Η συνέπεια και η αποδοτικότητα των αναφορών είναι και το πιο σημαντικό χαρακτηριστικό σε αυτό εδώ το επίπεδο για να επιτευχθεί η μέγιστη αποτελεσματικότητα των δεδομένων. Το πλεονέκτημα στην χρήση τέτοιων αναφορών είναι πως, εάν χτιστεί το σύστημα σωστά, δεν απαιτεί σχεδόν καθόλου για την σύνταξη και την συντήρησή τους. Επίσης είναι αξιόπιστες και συνεχώς διαθέσιμες, ακριβείς και συνεπείς. Με τον τρόπο αυτό μειώνονται τα λάθη και ο φόρτος εργασίας.

Υπάρχει μια ομάδα αναλυτών υπηρεσιών που διευκολύνει την συνεργασία μεταξύ επιχειρηματικών μονάδων με την ευρύτερη εταιρική μονάδα. Η ουσιαστική της ευθύνη είναι να ορίζει τα πρότυπα ώστε τα δεδομένα να είναι συνεπή και αποτελεσματικά. Σε αυτό το επίπεδο επεκτείνεται και η διακυβέρνηση των δεδομένων ώστε να συμπεριλάβει τους ελέγχους ποιότητας.

Επίπεδο 4: Εξωτερικές Αναφορές

Το επίπεδο αυτό επικεντρώνεται στην δημιουργία αναφορών που είναι συνεπής και αποδοτικές για εξωτερικές ανάγκες. Τέτοιες ανάγκες είναι:

- Η συμμόρφωση σε κανονισμούς και οι πιστοποιήσεις από εξωτερικούς φορείς,
- Απαιτήσεις χρηματοδότησης
- Ειδικές δημόσιες βάσεις δεδομένων.

Για να μπορέσει ο οργανισμός να ανταπεξέλθει σε αυτές τις ανάγκες πρέπει τα δεδομένα του να είναι συμμορφωμένα με βάση τα πιο σύγχρονα βιομηχανικά πρότυπα. Στην ιατρική τέτοια πρότυπα είναι το LOINC, SNOMED, ICD, RxNorm κ.α.

Έτσι λοιπόν η επιχειρησιακή αποθήκη δεδομένων, πέραν του ότι πρέπει να είναι μικρού κόστους όσον αφορά την εργασία και την συντήρησή της αλλά και να είναι ευέλικτη (agile) καθώς αυτά τα πρότυπα αλλάζουν με την πρόοδο της τεχνολογίας και επομένως η επιχείρηση πρέπει να είναι σε θέση να παράγει εξωτερικές αναφορές με βάση τα εκάστοτε

πρότυπα. Στο σημείο αυτό εισέρχονται διαδικασίες που ελέγχουν τις αναφορές με βάση αυτά τα πρότυπα και τις αποστέλλουν στους εξωτερικούς φορείς. Η αποθήκη δεδομένων εδώ οφείλει να υποστηρίζει την εισαγωγή κειμένου για τις σημειώσεις των ιατρών καθώς και αναζήτηση με βάση λέξεις κλειδιά.

Επίπεδο 5: Μείωση Ιατρικών Αποβλήτων και Βελτίωση Θεραπευτικών Πρακτικών

Στο επίπεδο αυτό στόχος είναι η βελτίωση των υπηρεσιών και η μείωση του κόστους με την ενεργοποίηση της ανάλυσης δεδομένων. Τα δεδομένα του επιπέδου αυτού υπάρχουν καθαρά για την δημιουργία στρατηγικής και πολιτικής όσον αφορά την υγεία.

Το μοντέλο ανάλυσης επικεντρώνεται στην βελτιστοποίηση των κλινικών πρακτικών, στην μείωση των ιατρικών αποβλήτων και στην μείωση της μεταβλητότητας των θεραπειών χρησιμοποιώντας την διακύμανση αυτή ως δείκτη αντίστροφο της ποιότητας. Η διακυβέρνηση των δεδομένων επεκτείνεται και άλλο προκειμένου να υποβοηθήσει τις διεπιστημονικές ομάδες να αντιμετωπίσουν τα προβλήματα υγείας των ασθενών με τον καλύτερο τρόπο. Πληθυσμιακές αναλύσεις με βάση τα δεδομένα των ασθενών προτείνουν βέλτιστες θεραπευτικές πρακτικές για κάθε ασθενή.

Οι διεπιστημονικές ομάδες παρακολουθούν συνεχώς για ευκαιρίες που μειώνουν τον κίνδυνο και το κόστος αλλά βελτιώνουν την ποιότητα των υπηρεσιών υγείας για τις διαδικασίες των θεραπευτικών μεθόδων. Τα μητρώα των δεδομένων συνδέουν πλέον τα εργαστηριακά δεδομένα με τις κλινικές και φαρμακευτικές παρατηρήσεις και με τις περιπτώσεις των ασθενών. Εδώ η αποθήκη οφείλει να ενημερώνεται κάθε μια βδομάδα.

Επίπεδο 6: Διαχείριση Υγείας με μετρήσεις Πληθυσμού

Το επίπεδο 6 αποτελείται από οργανισμού που έχουν επιτύχει κουλτούρα με γνώμονα τα δεδομένα και έχουν εγκαθιδρύσει ένα ισχυρό περιβάλλον ανάλυσης δεδομένων για να κατανοεί καλύτερα τα κλινικά αποτελέσματα. Οργανισμοί τέτοιου επιπέδου συμμετέχουν ουσιαστικά στον οικονομικό κίνδυνο αλλά και στην ανταμοιβή που φέρει το αποτέλεσμα

μια κλινικής περίπτωσης. Τουλάχιστον το 50% των βαρέων περιστατικών υγείας προσφέρεται με πακέτα πληρωμών. Η ανάλυση των δεδομένων υποστηρίζει πλέον στο επίπεδο αυτό ενεργά τον τριπλό στόχο της βελτιστοποίησης των υπηρεσιών υγείας (Triple Aim, Institute for Healthcare improvement)[112].

Οι τρείς αυτοί στόχοι είναι:

- η βελτίωση της ποιότητας των υπηρεσιών υγείας,
- η βελτίωση της υγείας του πληθυσμού,
- και η μείωση του κόστους των υπηρεσιών.

Η επιχειρησιακή αποθήκη δεδομένων πλέον περιλαμβάνει δεδομένα από τα δωμάτια των ασθενών, οικιακές συσκευές παρακολούθησης, εξωτερικά φαρμακεία, και διαμόρφωση του κόστους με βάση την εκάστοτε δραστηριότητα.

Η διακυβέρνηση των δεδομένων παίζει σημαντικό ρόλο στην ακρίβεια και στην ποιότητα των μετρήσεων των δεδομένων. Η αποθήκη δημιουργεί αναφορές που αποστέλλονται στο ανάλογο στέλεχος με αρμοδιότητα την εξομάλυνση κόστους-ποιότητας των υπηρεσιών. Η αποθήκη οφείλει να ενημερώνεται κάθε μια ημέρα κατά μέσο όρο.

Επίπεδο 7: Μοντέλα Πρόβλεψης για Παρέμβαση Κινδύνου

Οι οργανισμοί αυτής της κατηγορίας είναι σε θέση να αξιοποιούν το μοντέλο πρόβλεψης από την ανάλυση των δεδομένων και έτσι είναι σε θέση να μειώσουν τα κόστη των υπηρεσιών τους. Επικεντρώνονται στην διαχείριση των περιπτώσεων των ασθενών μέσω της συνεργασίας κλινικών συνέταιρων και συμπεριλαμβάνουν την προγνωστική μοντελοποίηση, την πρόβλεψη και την διαστρωμάτωση κινδύνου. [113]

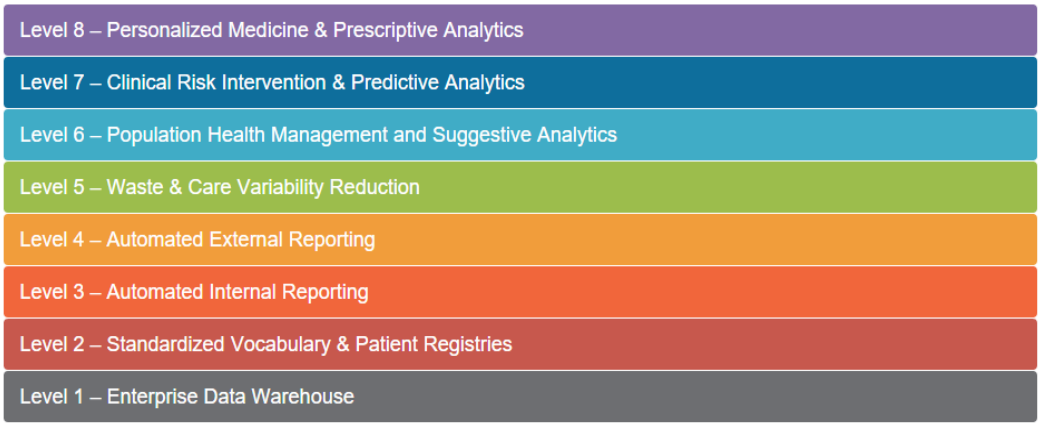
Τα μοτίβα των αναλύσεων εμπεριέχουν μοντέλα διάγνωσης και κοστολόγησης Επιπλέον οι ασθενής που δεν μπορούν να συμμετέχουν στα πρωτόκολλα φροντίδας είτε λόγο νοητικής ανεπάρκειας, ή θρησκευτικών αντιλήψεων, οικονομικής ανικανότητας ή γεωγραφικής πρόσβασης επισημαίνονται στα μητρώα τους.

Επίπεδο 8: Εξατομικευμένη Θεραπεία με βάση Πληθυσμιακά και Γενετικά Δεδομένα

Το μοτίβο των αναλύσεων στο επίπεδο αυτό είναι σε θέση να διαχειρίζεται τη φυσική και νοητική υγεία του ασθενούς. Το σύστημα παράγει μαζικά αλλά για κάθε ασθενή βελτιστοποιημένη φροντίδα. Οι οργανισμοί παροχής υπηρεσιών υγείας μετατρέπονται σε οργανισμούς βελτιστοποίησης υγείας και με μικρότερο κόστος [114]. Οι αναλύσεις είναι σε θέση να συμπεριλάβουν επεξεργασία της φυσικής γλώσσας από κείμενο, να ρυθμίσουν την φροντίδα και να παρέμβουν στην λήψη αποφάσεων. Με βάση πληθυσμιακές μετρήσεις κλινικών αποτελεσμάτων οι αναλύσεις είναι σε θέση να βελτιώσουν την φροντίδα κάθε ασθενούς. Η αποθήκη πλέον περιλαμβάνει βιομετρικά δεδομένα όλο το εικοσιτετράωρο, γενετικά δεδομένα και οικογενειακό ιστορικό.

Η αποθήκη ενημερώνεται μέσα σε λεπτά για τις αλλαγές. Ο σημαντικότερος ρόλος τον οργανισμών αυτών μεταβαίνει από την απλή παράδοση υπηρεσιών υγείας σε υποχρέωση πρόβλεψης κινδύνου και βελτίωσης των υπηρεσιών. Τα δεδομένα είναι σε θέση να συνδυάζονται με μελλοντικούς αλγορίθμους που θα βρίσκουν σχέσεις μεταξύ γενετικού υλικού, οικογενειακού ιστορικού και περιβάλλοντος του ασθενούς.

Φυσικά υπό τέτοιες συνθήκες η ιατρική θα οδηγηθεί στην δημιουργική καταστροφή της, και θα αλλάξει ριζικά ο τρόπος παράδοσης των υπηρεσιών υγείας. Από νωρίς στην ζωή του ασθενούς θα εισάγονται συσκευές μέτρησης προκειμένου όταν χρειαστεί περίθαλψη να μπορεί να λάβει την βέλτιστη και προσαρμοσμένη στις ανάγκες του φροντίδα.[40]



Εικόνα 6: Μοντέλο υιοθέτησης ανάλυσης δεδομένων

2.4. Κοινωνικά Δίκτυα στην Υγεία

Οι οργανισμοί υγείας συνεχίζουν να κατευθύνουν την συνεχή αλλαγή που υφίσταται ο κάδος της υγείας. Επιπλέον εντοπίζουν νέους τρόπους για να εμπλακούν με τους ευρύτερους καταναλωτές πέραν από τις εγκαταστάσεις ενός οργανισμού. Ο ρυθμός με τον οποίο οι πληροφορίες κυριαρχούνται στον τον κόσμο είναι πολύ γρήγορος, ωστόσο το να γίνει με επιτυχία αλληλεπίδραση των πληροφοριών αυτών με τους καταναλωτές μπορεί να φαίνεται σαν ένα δύσκολο έργο. [55] Όταν αναφερόμαστε στα κοινωνικά μέσα δικτύωσης και την χρήση τους στην υγειονομική περίθαλψη μπορεί να φαίνεται ότι πρόκειται για ένα εργαλείο το μάρκετινγκ, αλλά ουσιαστικά πρόκειται για πολύ πιο σημαντική διαδικασία από αυτό. Σίγουρα, τα κοινωνικά μέσα δικτύωσης μπορούν να χρησιμοποιηθούν για να προσελκύσουν και να διατηρήσουν τους καταναλωτές και το ενδιαφέρον τους, αλλά επιπλέον μπορεί ταυτόχρονα να είναι ένα ισχυρό εργαλείο για τη μείωση του κόστους της υγειονομικής περίθαλψης. σημαντικό επίσης είναι το γεγονός του ό,τι με τα κοινωνικά μέσα δικτύωσης συχνά παρέχεται βοήθεια στο ευρύ κοινό και σε άτομα με χρόνιες παθήσεις καθώς ο πληθυσμός ενημερώνεται αμεσότερα και σε ευρύτερο φάσμα για την ορθή διαχείρισης της υγείας. [56]

Τα κοινωνικά μέσα δικτύωσης έχουν χρησιμοποιηθεί στον τομέα της υγείας για ένα αρκετά μεγάλο χρονικό διάστημα. Το Ford Health System ήταν η πρώτη εμπειρία ενός ζωντανού tweet αναφορικά με μια χειρουργική επέμβαση το 2009, ενώ η κλινική Mayo Clinic πραγματοποίησε την πρώτη ετήσια Φροντίδα Υγείας Σύνοδο Κορυφής για τα κοινωνικά διαδίκτυά εκείνη τη χρονιά. Ωστόσο, στην πραγματικότητα το χρονικό διάστημα 2011-2012 αρχίσαμε να βλέπουμε υψηλά ποσοστά υιοθέτησης των κοινωνικών μέσων μαζικής ενημέρωσης στον τομέα της υγείας. Συχνά πολλά κέντρα και οργανισμοί υγείας έχουν, τα κοινωνικά μέσα δικτύωσης σε μεγάλο βαθμό ως εργαλείο δημοσίων σχέσεων Η δουλειά των κοινωνικών δικτύων είναι να εκμεταλλεύεται τα μέσα επικοινωνίας που χρησιμοποιούνται από τους καταναλωτές προκειμένου να συλλέγει πληροφορίες σχετικές με την υγεία. Τα κοινωνικά δίκτυα επίσης ασκούν μεγάλη επιρροή στην συμπεριφορά των καταναλωτών. Συγκριμένα, έρευνα σύμφωνα με το Webbed Feet, το 92% [57] ων καταναλωτών υποστηρίζει πως εμπιστεύονται τα δημοφιλέστερα κοινωνικά δίκτυα, τις φήμες που ακούνε στο άμεσο κοινωνικό περιβάλλον και τις

προτάσεις από συγγενείς και φίλους περισσότερο από κάθε άλλο μέσω διαφήμισης. Ταυτοχρόνως, στην υγεία πλέον μετακινούμαστε από το μοντέλο της φροντίδας των αρρώστων στο μοντέλο της διατήρησης και εξασφάλισης της υγείας ενός ατόμου. Οι συντελεστές αυτοί οδηγούν στην ανάγκη εύρεσης νέων τρόπων προσέγγισης του καταναλωτικού κοινού.[58]

Αυτή η κίνηση στο μοντέλο διασφάλισης της υγείας θα συντελέσει σε έναν μεγάλο βαθμό στην δημιουργία και στην υιοθέτηση τεχνολογιών, μέσω των οποίων θα συλλέγονται σημαντικές πληροφορίες και θα εμπλουτίζεται η ενημέρωση του κοινού προκειμένου να έχουν καλύτερες, πιο εξειδικευμένες υπηρεσίες αλλά και γνώσεις στο πως μπορούν να δράσουν οι ίδιοι για να εξασφαλίσουν την υγεία τους. Εν ολίγοις, θα έχουν μεγαλύτερη πρόσβαση στην ενημέρωση (και άρα στην πρόληψη) και ευκολότερη πρόσβαση στις υπηρεσίες.

Μερικά από τα θετικά που προσφέρουν τα κοινωνικά δίκτυα είναι τα παρακάτω:

Ταχύτατος Διαμοιρασμό Πληροφορίας: Μέσω των κοινωνικών δικτύων, τα γεγονότα και οι ειδήσεις γνωστοποιούνται πιο γρήγορα από ποτέ. Μπορεί να διαδραματίζεται ένα γεγονός παγκόσμιας σημασίας στην άλλη άκρη του κόσμου, και μόλις σε λίγες ώρες να μεταδοθεί σε όλον τον κόσμο. Στην περίπτωση της υγείας αυτό είναι χρήσιμο καθώς συμβάλλει στην ενημέρωση των ανθρώπων και στην ευαισθητοποίηση του κοινού σε θέματα υγείας.

Υπηρεσίες υγείας από απόσταση: Μαζί με τον διαμοιρασμό της πληροφορίας καθίσταται ευκολότερη και η ανταλλαγή απόψεων και επομένως και η ανταλλαγή γνώσεων πάνω σε θέματα υγείας. Αυτό επιτρέπει την πιο αξιόπιστη πρόληψη ασθενειών, την διάγνωση καθώς και την αντιμετώπισή τους. Οι πιο έμπειροι γνώστες του κλάδου μπορούν μέσω των κοινωνικών δικτύων να συμβάλουν να εκφράσουν απόψεις και να συνεργαστούν προκειμένου να λύσουν τέτοιου είδους θέματα. .[59]

Μηχανισμοί ελαχιστοποίησης κόστους: Τα κοινωνικά δίκτυα πέραν του ότι αποτελούν το ισχυρότερο εργαλείο στον τομέα του marketing – και όχι μόνο για τις υπηρεσίες Υγείας – προσφέρουν και την δυνατότητα για καλύτερη συνεργασία μέσα σε έναν οργανισμό που τα υιοθετεί αλλά και σε συνεργασίες μεταξύ οργανισμών. Αυτό είναι εφικτό να συμβεί με

τα πολλά εργαλεία που προσφέρονται πλέον από τα κοινωνικά δίκτυα. Ένα παράδειγμα θα μπορούσε να είναι το Google+ καθώς μέσω αυτού, κάθε εργαζόμενος π.χ. μιας εταιρείας είναι σε θέση να έχει το προφίλ του και ο υπεύθυνος ενός έργου αν έχει ένα κοινό φάκελο διαμοιρασμένο με όλους τους εργαζομένους ώστε να συνεργάζονται. Παράλληλα μέσω των κοινωνικών δικτύων παρέχεται ένας τεράστιος όγκος δεδομένων για επεξεργασία και συλλογή πληροφοριών.

Ένα από τα πιο σημαντικά οφέλη για τους οργανισμούς από τα κοινωνικά δίκτυα είναι η δυνατότητα που προσφέρουν για μετρήσεις απόψεις καταναλωτών και προβλέψεις στην κατεύθυνση της αγοράς. Ένας οργανισμός που κατέχει αυτά τα στοιχεία μπορεί να χτίσει την στρατηγική του ώστε να διατηρηθεί ανταγωνιστικός και κατέχει την πρώτη θέση στις προτιμήσεις των καταναλωτών. Οι οργανισμοί αυτοί που δεν θα κινηθούν σε αυτό το κομμάτι και δεν παρακολουθούν τις κινήσεις και τις επιλογές των καταναλωτών τους είναι λογικό να μείνουν πίσω στον ανταγωνισμό.

2.5. Σύστημα αναλυτικής - WANDA

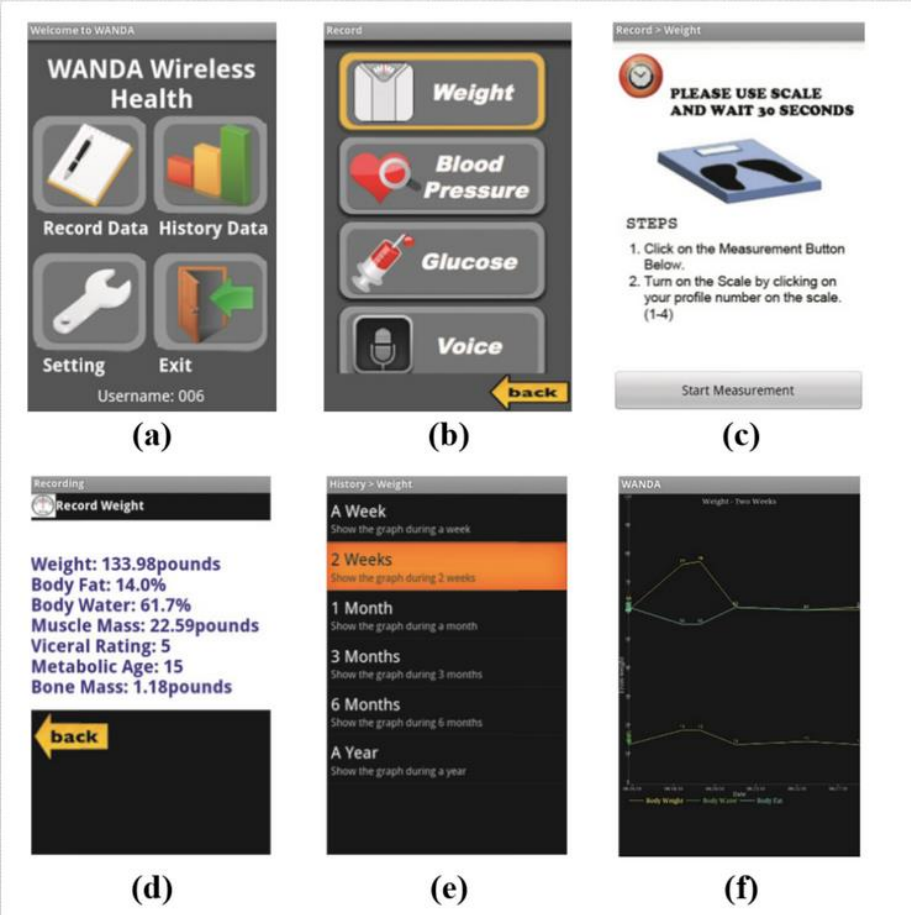
2.5.1.1. Περιγραφή πειράματος

Είναι ένα σύστημα ελέγχου και ανάλυσης δεδομένων, για ασθενείς με προβλήματα καρδιακής ανεπάρκειας. Το σύστημα υποστηρίζεται από κινητές συσκευές, και διαδικτυακό τόπο και τα δεδομένα αποθηκεύονται μέσω του διαδικτύου, όπως πραγματοποιούνται και οι αναζητήσεις. Επίσης υποστηρίζεται από μηχανή ανάλυσης δεδομένων, για την βελτιστοποίηση διαγνωστικών και προγνωστικών εξετάσεων.

Το σύστημα αυτό λειτουργεί από απόσταση και σκοπός του είναι η μείωση της θνησιμότητας ασθενών που πάσχουν από καρδιακή ανεπάρκεια και η μείωση του κόστους των νοσηλίων τους. Μελέτες έχουν δείξει πως, πολλές φορές οι ασθενείς που πάσχουν από τέτοια προβλήματα, νοσηλεύονται ξανά μέσα σε ένα διάστημα από 30 μέρες έως έξι μήνες και υπάρχει μεγάλος κίνδυνος θνησιμότητας. Ένα τέτοιο σύστημα χρησιμοποιεί αλγορίθμους για ανάλυση δεδομένων προκειμένου να διαγνώσει και να προβλέψει επιπλοκές σε τέτοιου είδους προβλήματα.

2.5.1.2. Αρχιτεκτονική του Wanda

Το Wanda είναι ένα σύστημα τριών επιπέδων, απομακρυσμένης παρακολούθησης με χρήση εξωτερικών υλικών και λογισμικού που έχουν σχεδιαστεί για να καλύπτουν ανάγκες τηλεϊατρικής και της εξ αποστάσεως παρακολούθησης.



Εικόνα 7: Η εφαρμογή του Wanda.

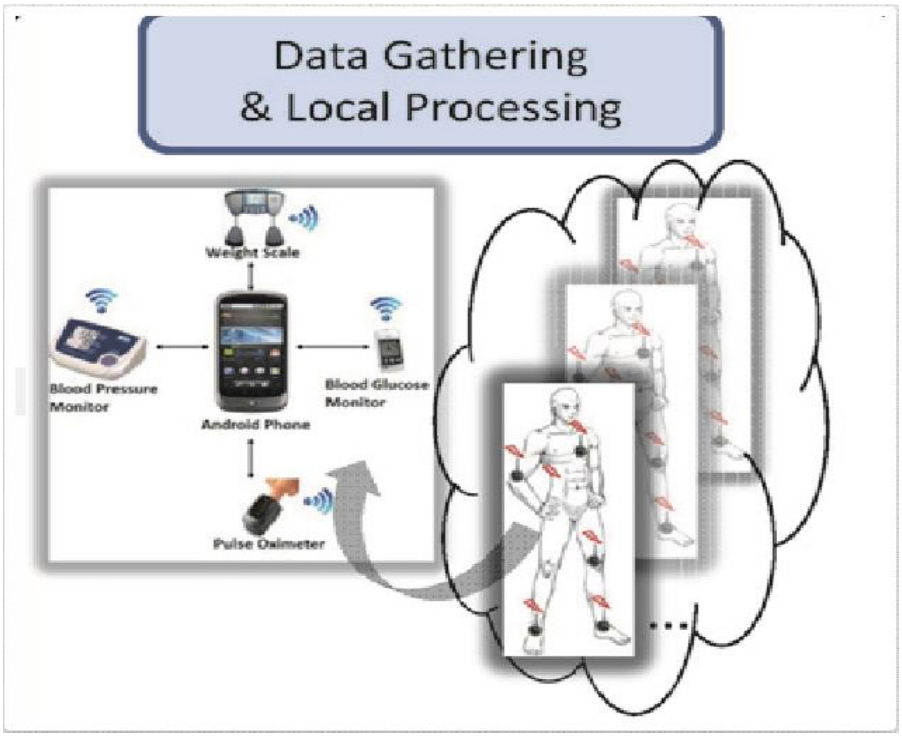
Η πρώτη βαθμίδα της αρχιτεκτονικής αποτελείται από μια συλλογή δεδομένων στο πλαίσιο, το οποίο σχηματίζεται από διάφορες αισθητικές συσκευές που μετρούν ποικίλα σωματικά στατιστικά στοιχεία, όπως το βάρος, το λίπος του σώματος, το νερό σώματος, η αρτηριακή πίεση, ο καρδιακός ρυθμός, η γλυκόζη στο αίμα, και τις κινήσεις του σώματος. Τα δεδομένα αυτά συλλέγονται, υποβάλλονται σε επεξεργασία και διαβιβάζονται μέσω

ενός κινητά τελευταίας γενιάς σε ένα υπολογιστικό νέφος που ουσιαστικά αποτελεί τη δεύτερη βαθμίδα της WANDA αρχιτεκτονικής.

Η τελευταία βαθμίδα της αρχιτεκτονικής του WANDA είναι μια μηχανή αναλυτικής ικανή συνεχώς να παράγει στατιστικά μοντέλα και πρόβλεψης των αποτελεσμάτων χρησιμοποιώντας διάφορους αλγορίθμους μάθησης και εξόρυξης δεδομένων.

2.5.1.3. Συλλογή δεδομένων

Το WANDA χρησιμοποίει έξυπνες συσκευές όπως κινητά προκειμένου να συλλέξει μετρήσεις με την βοήθεια εξωτερικών συσκευών και να τις στείλει στην συνέχεια στον κεντρικό Server. Οι εξωτερικές συσκευές επικοινωνούν με το την έξυπνη συσκευή μέσω Bluetooth. Πιο συγκεκριμένα συλλέγονται τα δεδομένα στην έξυπνη συσκευή και μέσω του WANDA οι χρήστες μπορούν να αναλύσουν τα δικά τους δεδομένα με την βοήθεια μιας γραφικής απεικόνισης που τους προσφέρεται από την εφαρμογή.



Εικόνα 8: Συλλογή δεδομένων.

2.5.1.4. Περιγραφή εφαρμογής

Ο χρήστης στην κεντρική διεπαφή μπορεί είτε να καταγράψει δεδομένα είτε να δει το ιστορικό του, δηλαδή να γίνεται κοινωνός εφόσον το επιθυμεί στα στοιχεία που έχει συλλέξει στο παρελθόν. Αν επιλέξει να καταγράψει νέα μέτρηση θα οδηγηθεί στις πέντε επιλογές βάρους, πίεσης, γλυκόζης, πυκνότητας οξυγόνου αίματος, φωνής, ερωτηματολογίου. Αφού επιλέξει ο χρήστης μία από τις επιλογές οδηγείται σε μία διεπαφή μαζί με σύντομες οδηγίες για την μέτρηση και επιπλέον έχει την δυνατότητα να επιλέξει την έναρξη της μέτρησης. Με την ολοκλήρωση των μετρήσεων η συσκευή αποθηκεύει τα δεδομένα.

2.5.1.5. Αποθήκευση και πρόσβαση δεδομένων

Το Wanda χρησιμοποιεί το υπολογιστικό νέφος του Amazon για να αποθηκεύει τα δεδομένα του. Το σύστημα απαιτεί εύκολα μεταβλητή αποθήκευση δεδομένων, ασφαλή μεταφορά, υψηλή διαθεσιμότητα καθώς και ιδιωτικότητα των δεδομένων τα οποία και καλύπτει πλήρως η Amazon.

Αρχικά η Amazon προσφέρει το Amazon SimpleDB, το οποίο ουσιαστικά είναι μια ψηλά διαθέσιμη και μεταβλητή μη σχεσιακή αποθήκη δεδομένων που στην ουσία δεν έχει κάποιον διαχειριστή. Οι χρήστες – προγραμματιστές αποθηκεύουν και ζητούν από την βάση δεδομένα μέσω Web Services. Αφού λοιπόν αναφέραμε συνοπτικά την λειτουργία του Amazon SimpleDB ας δούμε την λειτουργία της στην εφαρμογή του Wanda. Έτσι ουσιαστικά αποθηκεύει δομημένα δεδομένα όπως πληροφορίες ασθενών και των συσκευών που χρησιμοποιούν δυσμενή επεισόδια και σχολιασμούς δεδομένων.

Μια βάση δεδομένων SQL No SQL παρέχει ένα μηχανισμό για την αποθήκευση και την ανάκτηση των δεδομένων που διαμορφώνονται σχέσεις εντός και εκτός των πινάκων που χρησιμοποιούνται σε σχεσιακές βάσεις δεδομένων. Κίνητρα για την προσέγγιση αυτή είναι η απλότητα του σχεδιασμού, η οριζόντια κλιμάκωση και ο καλύτερος έλεγχος της διαθεσιμότητας. Η δομή των δεδομένων (π.χ. κλειδί-τιμή, γράφημα ή το έγγραφο) διαφέρει από τις RDBMS, και ως εκ τούτου ορισμένες εργασίες είναι πιο γρήγορες σε No SQL και

κάποια σε RDBMS. Υπάρχουν διαφορές αν και η συγκεκριμένη καταλληλότητα ενός δεδομένου No SQL DB εξαρτάται από το πρόβλημα που πρέπει να επιλυθεί. Οι βάσεις δεδομένων No SQL εμφανίζουν σημαντική και αυξανόμενη χρήση της βιομηχανίας σε μεγάλα δεδομένα και εφαρμογές web σε πραγματικό χρόνο. Τα συστήματα No SQL αναφέρονται επίσης ως "Όχι No SQL" για να τονιστεί ότι μπορεί στην πραγματικότητα να επιτρέψει SQL-όπως γλώσσες επερωτήσεων που πρέπει να χρησιμοποιούντα. Εμπόδια για την ευρεία υιοθέτηση της No SQL αποτελούν η χρήση του χαμηλού επιπέδου γλωσσών επερωτήσεων, η έλλειψη τυποποιημένων διεπαφών, καθώς και οι τεράστιες επενδύσεις που έχουν ήδη γίνει σε SQL από τις επιχειρήσεις.

2.5.1.6. Στοιχεία Analytics

Ένα από τα δυνατά σημεία του WANDA είναι συστήματα παρακολούθησης υγείας-τηλεϊατρική με την βοήθεια ειδικής μηχανής του analytics βάση του οποίου αναλύονται τα δεδομένα. Με βάση τα δεδομένα και τους σχολιασμούς που συλλέγονται, ο κινητήρας analytics μπορεί να δημιουργήσει πολλαπλά στατιστικά μοντέλα χρησιμοποιώντας αλγορίθμους μάθησης και εξόρυξης δεδομένων, συμπεριλαμβανομένης της ταξινόμησης, της ομαδοποίησης, των κανόνων συσχέτισης κλπ. Αυτά τα μοντέλα μπορούν στη συνέχεια να χρησιμοποιηθούν τόσο για διαγνωστικούς όσο και προγνωστικούς σκοπούς.

Η διαδικασία της ανάλυσης - analytics κατά κανόνα αποτελείται από δύο στάδια. Το offline στάδιο, και το online στάδιο. Στο offline στάδιο τα δεδομένα μεταφορτώνονται και αναλύονται σε offline βάση κάποιων υποθέσεων. Μόλις έχει δημιουργηθεί ένα ισχυρό μοντέλο και στην συνέχεια έχει επικυρωθεί, μπορεί έπειτα να φορτωθεί στον κεντρικό υπολογιστή για να χρησιμοποιηθεί για προβλέψεις στο δεύτερο στάδιο. Μία από τις προκλήσεις εδώ είναι η βελτιστοποίηση του αλγορίθμου, έτσι ώστε να μπορεί να εκτελεστεί σε ένα πραγματικό χρόνο. Αυτό συχνά απαιτεί τη χρήση της PubSub διεπαφής για πρόσβαση στα δεδομένα, σε συνδυασμό με μια προσωρινή αποθήκευση για μη επεξεργασμένα ιστορικά δεδομένα.

Όταν αναφερόμαστε στην έννοια του PubSub ουσιαστικά μιλάμε για την αρχιτεκτονική λογισμικού, publish που είναι ένα πρότυπο ανταλλαγής μηνυμάτων όπου οι αποστολείς

των μηνυμάτων, που ονομάζονται εκδότες. Είναι υπεύθυνα να δημοσιεύουν τα μηνύματα που χαρακτηρίζονται σε τάξεις, χωρίς γνώση του ποια, ενδεχομένως, οι συνδρομητές μπορούν να είναι εκεί. Ομοίως, οι συνδρομητές εκδηλώσουν το ενδιαφέρον σε ένα ή περισσότερα μαθήματα, και λαμβάνονται μόνο τα μηνύματα που παρουσιάζουν ενδιαφέρον, χωρίς τη γνώση του αν, ενδεχομένως, υπάρχουν εκδότες.

Τέλος, όταν ο αλγόριθμος ανιχνεύει ένα σχέδιο που είναι ισχυρά συσχετισμένο με ανεπιθύμητα αποτελέσματα, ένας συναγερμός ενεργοποιείται και το προσωπικό που είναι υπεύθυνο ειδοποιείται άμεσα αμέσως. Η μηχανή ανάλυσης προσαρμόζεται με το πλαίσιο Weka. Αυτό επιτρέπει στην μηχανή ανάλυσης να υποστηρίξει μια πληθώρα γνωστών αλγορίθμων μάθησης εξόρυξης δεδομένων.

2.5.1.7. Αποτελέσματα και Συζήτηση

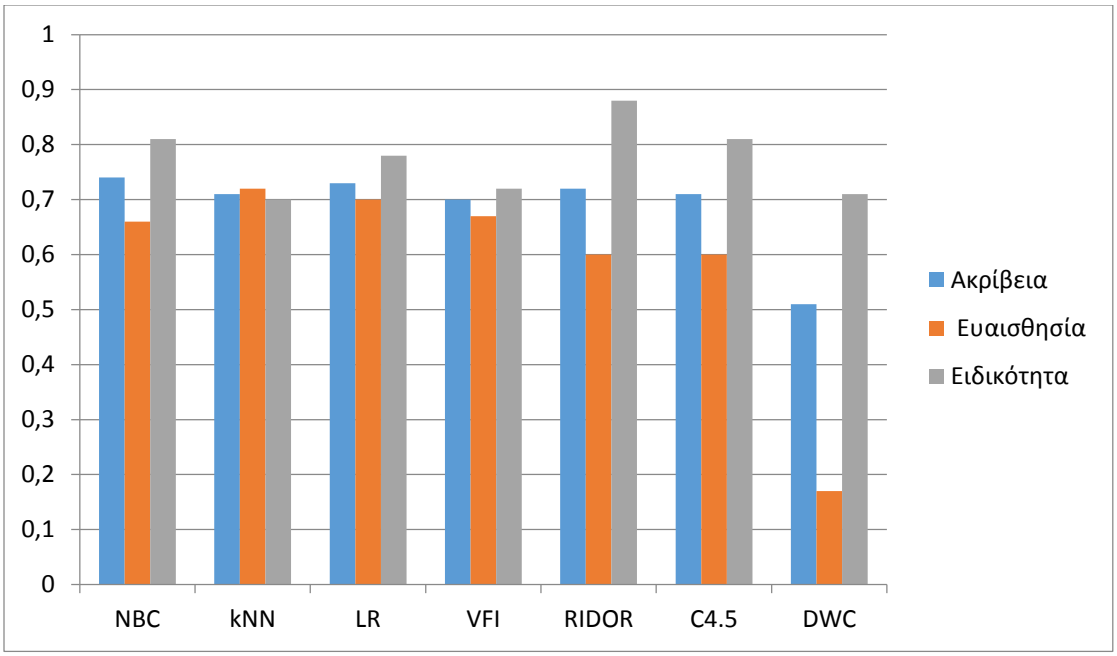
Οι αλγόριθμοι, αξιολογούνται ως προς την ακρίβεια τους, την ευαισθησία τους και την ειδικότητα τους. Έχουν υπολογιστεί στα αποτελέσματα της πρόβλεψης, με βάση τους αριθμούς ως:

- True Positive (TP),
- True Negative (TN),
- False Positive (FP)
- και False Negative (FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

Τα αποτελέσματα για κάθε αλγόριθμο παριστάνονται γραφικά στον Πίνακα 2.



Πίνακας 2-1: Ακρίβεια, Ευαισθησία, Ειδικότητα των αλγορίθμων NBC, kNN, LR, VFI, RIDO, C4.5, DWC.

Περιγραφή Χαρακτηριστικών Πίνακα 2-1.

Dcs: Daily change in systolic blood pressure

Dcd: Daily change in diastolic blood pressure

Dcw: Daily change in weight

sds.3d: Standard deviation of systolic blood pressure over the past 3 days

sdd.3d: Standard deviation of diastolic blood pressure over the past 3 days

sdw.3d: Standard deviation of weight over the past 3 days

sds.7d: Standard deviation of systolic blood pressure over the past 7 days

sdd.7d: Standard deviation of diastolic blood pressure over the past 7 days

sdw.7d: Standard deviation of weight over the past 7 days

Βασιζόμενοι στο γεγονός του ότι ο αριθμός των θετικών και των αρνητικών περιπτώσεις εξισορροπείται, η ελάχιστη ακρίβεια οποιουδήποτε αλγορίθμου την ευκαιρία θα πρέπει να είναι μεγαλύτερη από 0,5. Από τα αποτελέσματα των αλγορίθμων βλέπουμε ότι η πρόβλεψη και η επιδείνωση των συμπτωμάτων HF, με την χρήση του DWC φαίνεται να είναι οριακά πιο ακριβής περίπου στο 0,519. Αυτό δεν είναι έκπληξη, δεδομένου ότι μόνο το 15,9% των θετικών περιπτώσεων έχουν χαρακτηρισθεί ως τέτοια, ενώ ένα σημαντικό ποσό των αρνητικών περιπτώσεων (28,6%) καταλήγουν να έχουν μια DWC περισσότερες από 2 λίβρες.

Από την άλλη πλευρά, οι αλγόριθμοι των Analytics WANDA έχουν ως στόχο να επιτύχουν πολύ μεγαλύτερη ακρίβεια. Οι τρεις αλγόριθμοι, ο NBC, ο LR, και ο RIDOR, είναι σε θέση να προβλέψουν σωστά την επιδείνωση ή την σταθεροποίηση των συμπτωμάτων HF στο 74% του χρόνου. Μπορεί το VFI να έχει το χαμηλότερο ποσοστό 0,696, όμως εξακολουθεί να είναι σχεδόν 20% περισσότερο ακριβής από τον DWC. Επιπλέον, οι έξι αλγόριθμοι έχουν τιμές ευαισθησίας που είναι τουλάχιστον 45% μεγαλύτερες από εκείνες του DWC. Αυτό υποδηλώνει ότι τα χαρακτηριστικά που εντοπίζονται ακολούθως στην περιγραφή χαρακτηριστικών έχουν μια πολύ ισχυρότερη συσχέτιση με την επιδείνωση των συμπτωμάτων HF από την απλή καθημερινή αλλαγή βάρους, η οποία μπορεί να επιβεβαιωθεί περαιτέρω από το υψηλό ποσοστό που κυμαινόταν από 0.696 (για kNN) σε 0,87 (για RIDOR).

2.5.1.8. Σφάλματα Ανάλυσης

Είναι σημαντικό να εντοπίσουμε το κατώτερο όριο της ταξινόμησης σφάλματος, προκειμένου να τεθεί η ακρίβεια ενός αλγορίθμου σε μια προοπτική.

Μετά, ένας αλγόριθμος πρόβλεψης είναι τόσο καλός όσο και τα χαρακτηριστικά που δίνονται σε αυτόν. Το χαμηλότερο ποσοστό σφάλματος μπορεί να επιτευχθεί με οποιονδήποτε δυαδικό ταξινομητή που οριοθετείται από την Bayes Error Rate,

$$P(error) = \int_{R_2} p(x|w_1)P(w_1)dx + \int_{R_1} p(x|w_2)P(w_2)dx$$

όπου τα R1 και R2 είναι οι περιοχές όπου το x δίνεται για να ταξινομηθεί εσφαλμένα στις κατανομές των τάξεων w1 και w2.

Δυστυχώς, οι κατανομές είναι άγνωστες στην πραγματικότητα και ως εκ τούτου, μπορεί να εκτιμηθεί μόνο, χρησιμοποιώντας το απόδειξης, μπορούμε να υπολογίσουμε το ποσοστό σφάλματος Bayes ως το ήμισυ του ποσοστού λάθους με την μορφή 1-NN, η οποία μπορεί να ληφθεί εμπειρικώς από τα δεδομένα.

Ως αποτέλεσμα, το ποσοστό σφάλματος Bayes υπολογίζεται ότι είναι 0.174, το οποίο προϋποθέτει μια θεωρητική βέλτιστη ακρίβεια 0,826, η οποία είναι μόνο περίπου 9% υψηλότερη. Ενώ αυτό υποδηλώνει ότι οι αλγόριθμοι μπορούν να είναι καλύτερα συντονισμένοι, η βελτίωση μπορεί να έρθει σε βάρος της μειωμένης ευαισθησίας ή ειδικότητας. Συνεπώς, είναι γόνιμο να βρει χαρακτηριστικά που έχουν την ισχυρότερη προβλεπτική ικανότητα.

2.5.1.9. Αλγόριθμοι που εφαρμόστηκαν

Naïve Bayes Classifier (NBC): Ο NBC χρησιμοποιεί το θεώρημα του Bayes και θεωρεί πως κάθε χαρακτηριστικό είναι υπό όρους ανεξάρτητο. Η εκ των υστέρων πιθανότητα μιας ευρύτερης κατηγορίας C προσφέρει μια σειρά από χαρακτηριστικά τύπου F F1, F2, … Fn με αποτέλεσμα να έχουμε τον ακόλουθο τύπο:

$$p(C|F_1, F_2, \ldots, F_n) \propto p(C)\prod_{i=1}^{n} p(F_i|C)$$

Η προηγούμενη και η πιθανότητα μπορεί να υπολογιστεί από την εκπαίδευση που άμεσα και ισχύει για το σύνολο ελέγχου.

Nearest Neighbor (kNN): Μία μη σημειωμένη περίπτωση ταξινομείται με βάση τους κοντινότερους γείτονές του k από το σύνολο εκπαίδευσης με βάση την πλειοψηφία. Η Ευκλείδεια απόσταση χρησιμοποιείται για να προσδιοριστεί η εγγύτητα των δύο περιπτώσεων:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Με k = 5

όπου xi και yi οι αντίστοιχες λειτουργίες από τις δύο περιπτώσεις.

Logistic Regression (LR): Λογιστική παλινδρόμηση είναι ένα είδος παλινδρόμησης που να προβλέπει το αποτέλεσμα ενός δυαδικού εξαρτώνται μεταβλητή (τάξη) βασίζεται σε ένα σύνολο ανεξάρτητων μεταβλητών (χαρακτηριστικά) Xi χρησιμοποιώντας τη λογιστική συνάρτηση:

$$p(x) = \frac{1}{1 + e^{-f(x)}}, \quad f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

όπου p (x) είναι η πιθανότητα του x ον κατηγορίας 1, δεδομένου συντελεστές παλινδρόμησης β1 ... βi και το σημείο τομής β0.

Voting Feature Interval (VFI): Η VFI κατατάσσει και χτίζει άνω και κάτω φράγματα γύρω από κάθε κατηγορία για κάθε χαρακτηριστικό. Η κατάταξη βασίζεται στη λήψη αποφάσεων με πλειοψηφία, όπου η ψηφοφορία για την κατηγορία C γίνεται με βάση τον ακόλουθο τύπο:

$$v(a, C) = \frac{\text{interval\_class\_count}(a, i, C)}{\text{class\_count(C)}} \left(\frac{H(C|a)}{\text{max uncertainly}} \right)^{\text{bias}}$$

Με bias = 0.6

Ripple-Down Rule Learner ((RIDOR): Η RIDOR είναι μια έκδοση του Ripple Down κανόνα χρησιμοποιώντας το Indcut αλγόριθμο, όπου η προεπιλεγμένοι κανόνες πρώτα δημιουργούνται με βάση τουλάχιστον του ποσοστού σφάλματος. Ένα σύνολο κανόνων

εξαίρεσης που δημιουργούνται με σκοπό να προβλέψουμε τις κατηγορίες, εκτός εκείνων που καλύπτονται από τους προεπιλεγμένους κανόνες.

C4.5 Decision Tree (C4.5): Το C4.5 κατασκευάζει ένα δέντρο απόφασης με βάση τις πληροφορίες της εντροπίας. Κάθε χαρακτηριστικό σύνολο στην εκπαίδευση αξιολογείται για την αύξηση της πληροφόρησης

$$IG(T,a) = H(T) - H(T|a) \qquad H(T) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i)$$

Με confidence = 0.25

Το χαρακτηριστικό με τη μεγαλύτερη IG επιλέγεται για να χωρίσει το δέντρο απόφασης σε κάθε κόμβο έως ότου η εμπιστοσύνη πέφτει κάτω από ορισμένο όριο.

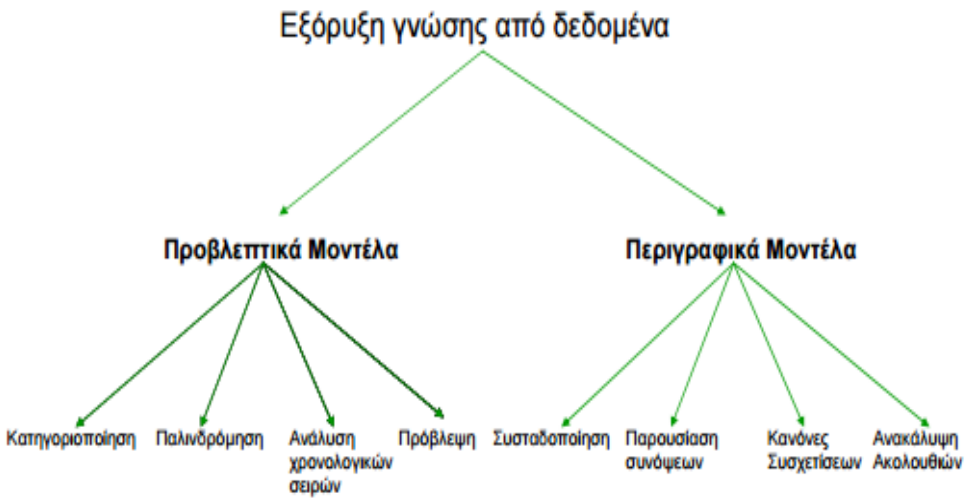
3. Αλγόριθμοι Μηχανικής Γνώσης

3.1. Εισαγωγή

Στην εξόρυξη γνώσης (data mining) υπάρχουν τρεις κύριες συνιστώσες: αναζήτηση μοντέλου, παράσταση μοντέλου και αξιολόγηση μοντέλου. Οι βασικές μέθοδοι αναζήτησης μοντέλου ψάχνουν και για παραμέτρους και για μοντέλα. Στην αναζήτηση παραμέτρων ο αλγόριθμος ψάχνει τις ελεύθερες παραμέτρους που βελτιστοποιούν την απόδοση ενός τελικού μοντέλου. Για απλά προβλήματα η αναζήτηση είναι εύκολη, αλλά για γενικά μοντέλα μία κλειστή λύση δεν είναι εφικτή, και χρησιμοποιούνται μέθοδοι όπως η συζυγής κατάβαση δυναμικού στον αλγόριθμο back-propagation για τα νευρωνικά δίκτυα.[47]

Η αναζήτηση μοντέλου από την άλλη ψάχνει για το κατάλληλο μοντέλο ή την οικογένεια μοντέλων και για κάθε μία τέτοια δομή μοντέλου που βρίσκει εφαρμόζει έπειτα την αναζήτηση για τις κατάλληλες παραμέτρους του. Αυτές οι δύο αναζητήσεις είναι χρονοβόρες όταν το μέγεθος του χώρου αναζήτησης είναι μεγάλο και οι υλοποιήσεις τους επωφελούνται ιδιαίτερα από τις τεχνικές παραλληλισμού. Ένα προβλεπτικό μοντέλο (predictive model)[66] κάνει μια πρόβλεψη για τις τιμές των δεδομένων, χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από άλλα δεδομένα. Η μοντελοποίηση πρόβλεψης μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων. Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός προβλεπτικού μοντέλου περιλαμβάνουν:

- Κατηγοριοποίηση ή ταξινόμηση (classification)
- Παλινδρόμηση (regression)
- Ανάλυση χρονολογικών σειρών (time series analysis)
- Πρόβλεψη (prediction)
- Συσταδοποίηση (clustering)
- Παρουσίαση συνόψεων (summarization)
- Κανόνες συσχετίσεων (association rules)
- Ανακάλυψη ακολουθιών (sequential pattern discovery)



Εικόνα 9: Εξόρυξη γνώσης από δεδομένα.

3.2. Εξόρυξη γνώσης

Η συμπεριφορά των διαδικασιών εξαρτάται από πολλούς παράγοντες:

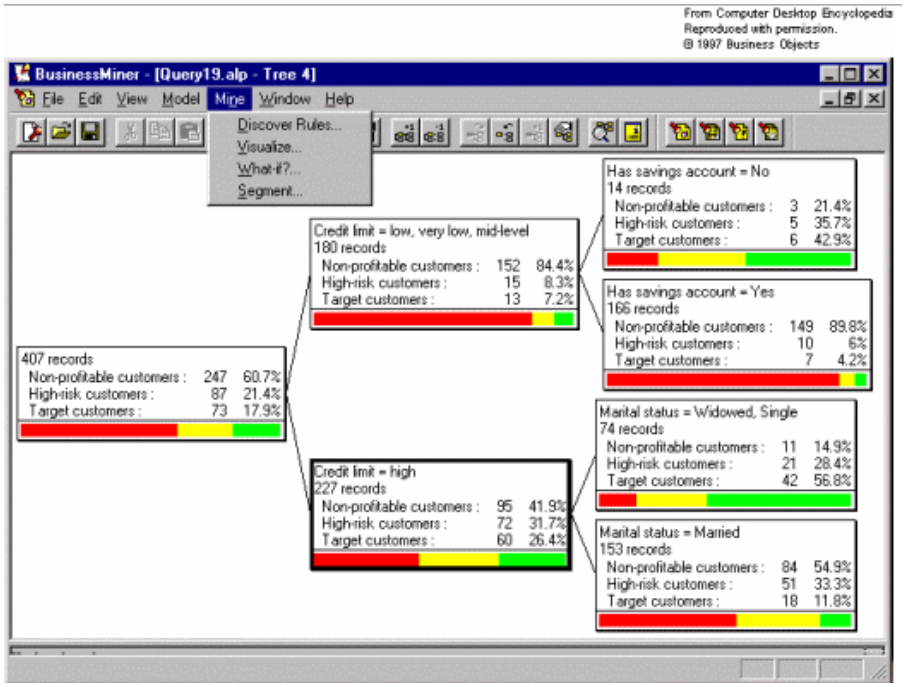
- Ο σχεδιασμός της διαδικασιών και των ενσωματωμένων κανόνων του,
- το μοτίβο άφιξη των νέων περιπτώσεων της διαδικασίας,
- η διαθεσιμότητα των πόρων και το επίπεδο δεξιοτήτων τους,
- οι ιδιότητες των υποθέσεων των επιχειρήσεων υποβάλλονται σε επεξεργασία σε κάθε περίπτωση διαδικασίας, καθώς και σε και άλλους εξωτερικούς παράγοντες των επιχειρήσεων.

Ιστορικά η διαδικασία τις ανάλυσης χρησιμοποιεί από τις επιχειρήσεις σχετικά δεδομένα για την ταξινόμηση και την πλοήγηση περιπτώσεων αναφορικά με την εκάστοτε διαδικασία και συναφείς πληροφορίες για τις επιδόσεις τους. Μοντέλα προσομοίωσης που χρησιμοποιούνται για την πρόβλεψη της μέσης συμπεριφοράς για ένα μεγάλο σύνολο των περιπτώσεων της διαδικασίας και μπορούν να χρησιμοποιήσουν την ροή εργασίας, και τα σχετικά δεδομένα για την προσομοίωση επιχειρηματικών κανόνων συνυφασμένες με τη δομή της διαδικασίας. Η εξόρυξη των δεδομένων για την analytics διαδικασία προσπαθεί να καθιερώσει συσχετίσεις μεταξύ των δεικτών Key Performance και της βασικής μας

διαδικασίας-από εξωτερικούς παράγοντες, όπως χαρακτηριστικά αποτελεί το στοιχείο της εργασίας, των χρονοδιαγραμμάτων των πόρων, ή τα μοτίβα άφιξης [60].

Εάν δεν μπορούν να καθοριστούν αυτές οι συσχετίσεις με επαρκή ακρίβεια, είναι δυνατό να προβλεφθεί η συμπεριφορά μίας ενιαίας instance διαδικασίας, με δεδομένη την τρέχουσα κατάσταση των υποδομών για την εκτέλεση της διαδικασίας και συγχρόνως τα χαρακτηριστικά της υπόθεσης των επιχειρήσεων να υποβληθούν σε επεξεργασία. Μια τυπική εφαρμογή ενός μοντέλου εξόρυξης θα είναι η ανάλυση της εισερχόμενης αίτησης του πελάτη με σκοπό να παρέχει σε ένα εκτιμώμενο χρόνο την απαραίτητη επεξεργασίας στον πελάτη. Προκειμένου να επιτευχθεί η πρόβλεψη αυτή ο αναλυτής θα εστιάσει σε περιπτώσεις αναφορικά με προηγούμενη διαδικασία έτσι τόσο ο καθορισμός της σχέσης μεταξύ των γνωρισμάτων σε κάθε περίπτωση των επιχειρήσεων όσο και του κύκλου χρόνου καταγράφεται για κάθε διαδικασία.

Με βάση τα γνωρίσματα μίας εισερχόμενης εφαρμογής ο αλγόριθμος εξόρυξης θα μπορούσε να προβλέψει το χρόνο επεξεργασίας για την συγκεκριμένη περίπτωση. Μια άλλη εφαρμογή ενός μοντέλου εξόρυξης θα είναι η βελτιστοποίηση της διακλάδωσης όσον αφορά τους κανόνες σε μια διαδικασία.



Εικόνα 10: BussinessMines

Βασιζόμενοι στο παράδειγμα που αναφέρθηκε πρωτίστως λαμβάνουμε ουσιαστικά υπόψη μας μόνο δύο παραμέτρους, σε αντίθεση με τις περισσότερες διαδικασίες οι οποίες επηρεάζονται από ένα σημαντικό αριθμό χαρακτηριστικών των επιχειρήσεων. Ο σχεδιασμός μιας κατάλληλης δομή εξόρυξης απαιτεί τόσο την ανάλυση των ιστορικών δεδομένων όσο και το θέμα της εμπειρογνωμοσύνη[61]. Υπάρχουν αρκετές διαφορές μεταξύ της προσομοίωσης και την προσέγγιση της εξόρυξης δεδομένων για την επεξεργασία των πεδίων της αναλυτικής:

- Ένα μοντέλο προσομοίωσης πρέπει να είναι επαρκώς ακριβής αναπαράσταση της συλλογής των διαδικασιών που εκτελούνται. Μπορεί να κάνει προβλέψεις για τις περιπτώσεις που δεν αντιμετώπιζε στο παρελθόν που οι σχετικές διεργασίες δεν έχουν αλλάξει.
- Οι προβλέψεις της εξόρυξη γνώσης [62] είναι βασισμένες σε μια στατιστική ανάλυση των περιπτώσεων της διαδικασίας που έχουν ήδη ολοκληρωθεί. [63]
- Οι Προσομοιώσεις που λαμβάνουν χώρα είναι υπολογιστικά εντατικές. Χρειάζεται πολύς χρόνος για να αποκτήσουν τις προβλέψεις, ιδίως αν ο κινητήρας χρησιμοποιεί προσομοίωση ροής εργασιών σχετικών δεδομένων με σκοπό να θεσπίσει κάθε Instance διαδικασίας στο σενάριο προσομοίωσης.

Στη διαδικασία της εξόρυξη γνώσης [31], η εκπαίδευση είναι υπολογιστικά εντατική, αλλά η εξόρυξη δεδομένων υλοποιείται σε συγκεκριμένο μοντέλο που έχει εκπαιδευτεί αποτελεσματικά για να είναι σε θέση να αντιμετωπίζει τις προβλέψεις που μπορούν να γίνουν σε ένα εξαιρετικά γρήγορο ρυθμό. Ωστόσο, η περιοδική επανεκπαίδευση μπορεί να απαιτείται για να κρατήσει το μοντέλο την επιθυμητή ακρίβεια που παρέχει.

3.3. Προγνωστική μοντελοποίηση

3.3.1. Κατηγοριοποίηση (classification)

Η κατηγοριοποίηση , (classification, Supervised learning, Pattern recognition) αντιστοιχίζει τα δεδομένα σε προκαθορισμένες κατηγορίες ή κλάσεις. Αναφέρεται συχνά σαν εποπτευομένη μάθηση , επειδή οι κατηγορίες– κλάσεις καθορίζονται πριν εξεταστούν

τα δεδομένα. Οι αλγόριθμου κατηγοριοποίησης, οι γνωστοί ταξινομητές, απαιτούν οι κατηγορίες να ορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων, και περιγράφουν αυτές τις κατηγορίες εξετάζοντας τα γνωρίσματα αυτά. Οι τεχνικές κατηγοριοποίησης χρησιμοποιούν δένδρα αποφάσεων, παλινδρόμηση, λογικούς κανόνες, νευρωνικά δίκτυα, στατιστικές μεθόδους διάκρισης ή κοντινότερους γείτονες ή τεχνικές Bayes. Με τη χρήση νευρωνικών δικτύων σαν ταξινομητές η κατηγοριοποίηση ανάγεται σε ένα πρόβλημα density estimation ή discrimination ή και regression . [64]

3.3.2. Παλινδρόμηση (regression) ή προσέγγιση συνάρτησης

Η Παλινδρόμηση (Regression) είναι μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίηση για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Έτσι η παλινδρόμηση μπορεί να απεικονίζει ένα στοιχειώδες δεδομένο x σε μια πραγματική μεταβλητή πρόβλεψης y. Η παλινδρόμηση περιλαμβάνει την εκμάθηση μιας συνάρτησης y=f(x) που κάνει αυτή την απεικόνιση. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης (γραμμική, μη-γραμμική ή πολυωνυμική κλπ.) και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα.

Η συνάρτηση παλινδρόμησης προβλέπει την συνάρτηση συμμετοχής του ανύσματος x στην κλάση με τιμή y. Η γραμμική παλινδρόμηση y=c0+c1x1+…+cnxn υποθέτει γραμμικές συσχετίσεις και μπορεί να βρει έτσι μία διαχωριστική συνάρτηση που διαχωρίζει έναν υπόχωρο σε δύο περιοχές κλάσεων. Και αυτή επίσης εμφανίζεται συχνά και στην κατηγοριοποίηση δεδομένων. Τα Τεχνητά νευρωνικά δίκτυα χρησιμοποιούνται ευρύτατα για εκτίμηση σημείων, ή εκτίμηση συνάρτησης, ή παλινδρόμηση, ή πρόβλεψη ή κατηγοριοποίηση. Στην αξιολόγηση μοντέλων υπάρχει το στάνταρ mean squared error και η cross entropy loss function για την παλινδρόμηση και κατηγοριοποίηση αντίστοιχα. Δένδρα παλινδρόμησης, κανόνες και regression splines χρησιμοποιούνται επίσης στην προβλεπτική μοντελοποίηση (predictive modelling) αν και μπορούν επίσης να εφαρμοστούν και στην περιγραφική μοντελοποίηση. [65]

3.3.3. Ανάλυση Χρονολογικών Σειρών (Time Series Analysis)

Με την ανάλυση χρονολογικών σειρών ή χρονοσειρών (time series analysis), μμελετάται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές λαμβάνονται σε ίσα χρονικά διαστήματα (ημερήσια, εβδομαδιαία, ωριαία, κοκ). Για να παρασταθούν οπτικά οι χρονοσειρές χρησιμοποιείται ένα διάγραμμα χρονοσειρών. Υπάρχουν τρεις βασικές λειτουργίες που πραγματοποιούνται στην ανάλυση χρονοσειρών. Στη μια περίπτωση, χρησιμοποιούνται μονάδες μέτρησης απόστασης για να καθορίσουν την ομοιότητα ανάμεσα σε διαφορετικές χρονοσειρές. Στη δεύτερη περίπτωση, εξετάζεται η δομή της χρονοσειράς για να καθορίσει (ίσως και να κατηγοριοποιήσει) τη συμπεριφορά της. Μια τρίτη περίπτωση είναι η χρήση διαγραμμάτων χρονοσειρών για την πρόβλεψη μελλοντικών τιμών. Μία άλλη πρόσφατη λειτουργία είναι η εύρεση των ίδιων των κατηγοριών των χρονοσειρών.[66]

3.3.4. Πρόβλεψη (prediction)

Πολλές από τις πρακτικές εφαρμογές εξόρυξης γνώσης μπορούν να θεωρηθούν σαν πρόβλεψη μελλοντικών καταστάσεων με γνώση των προηγούμενων και των σημερινών δεδομένων. Η πρόβλεψη (prediction - Forecasting) μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης. Αυτή η εργασία εξόρυξης γνώσης είναι διαφορετική από το μοντέλο πρόβλεψης, παρόλο που η διαδικασία πρόβλεψης αποτελεί ένα τύπο μοντέλου πρόβλεψης. Η διαφορά είναι ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μια μελλοντική κατάσταση παρά σε μια τρέχουσα. Έτσι εδώ αναφερόσαστε σε ένα είδος εφαρμογής παρά σε μια προσέγγιση μοντελοποίηση.[67*]

3.3.5. Συσταδοποίηση (clustering)

Η Συσταδοποίηση (clustering, Unsupervised learning, Segmentation, Partitioning) είναι η διαδικασία ομαδοποίησης αντικειμένων με όμοια χαρακτηριστικά και η κατάταξη σε κλάσεις ή συστάδες ή συμπλέγματα. Στην Συσταδοποίηση. Οι συστάδες δεν είναι προκαθορισμένες αλλά προσδιορίζονται από τα δεδομένα. Η Συσταδοποίηση αναφέρεται

εναλλακτικά και σαν μη εποπτευομένη μάθηση. Μπορεί να θεωρηθεί σαν μια διαίρεση ή τμηματοποίηση των δεδομένων σε ομάδες που μπορεί να είναι ή να μην είναι διακριτές μεταξύ τους. Η συσταδοποίηση συνήθως επιτυγχάνεται με τον καθορισμό της ομοιότητας, ως προς προκαθορισμένα γνωρίσματα, ανάμεσα στα δεδομένα. Τα πιο σχετικά δεδομένα ομαδοποιούνται σε ίδιες ομάδες.[68]

3.3.6. Παρουσίαση Συνόψεων (Summarization)

Η παρουσίαση συνόψεων (summarization, characterization) ή συνοπτικών μοντέλων αφορά μμεθόδους που βρίσκουν και απεικονίζουν τα δεδομένα σε υποσύνολα τους που συνοδεύονται με απλές και συνοπτικές περιγραφές. Η σύνοψη χαρακτηρίζει τα δεδομένα και παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδηγμένων. Αυτό είναι χρήσιμο και βοηθά στην κατανόηση της σημαντικότητας μερικών γνωρισμάτων έναντι άλλων. Βασικές έννοιες της στατιστικής όπως ο μέσος, η διακύμανση, η τυπική απόκλιση αποτελούν απλά μοντέλα ενός πληθυσμού. Το ταίριασμα ενός πληθυσμού σε μία κατανομή παρέχει ένα καλύτερο μοντέλο δεδομένων. Το ιστόγραμμα είναι επίσης μία πολύ κατατοπιστική τεχνική που δείχνει συνοπτικά πιθανά απλά μοντέλα για την κατανομή των δεδομένων. Το διάγραμμα διασποράς είναι μία άλλη τεχνική για την παρουσίαση δεδομένων.

3.3.7. Ανάλυση Κανόνων Συσχετίσεων (association rules)

Ένας κανόνας συσχέτισης (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ δεδομένων. Η ανάλυση κανόνων συσχετίσεων (association rules), αναφέρεται στην διαδικασία εκείνη της εξαγωγής γνώσης από βάσεις δεδομένων που αποκαλύπτει συγκεκριμένο τρόπο με τον οποίο τα δεδομένα είναι δυνατόν να συνδέονται. Το ευρύτερα γνωστό παράδειγμα αυτού του είδους εφαρμογής είναι ο προσδιοριστής κανόνων συσχέτισης από την ανάλυση δεδομένων συναλλαγών. Για παράδειγμα ο κανόνας $A \Rightarrow \{B,\Gamma\}$ δηλαδή οι πελάτες που αγοράζουν το Α, αγοράζουν ταυτόχρονα και τα Β και Γ είναι ένας κανόνας συσχέτισης.

Η ανάλυση κανόνων συσχέτισης είναι γνωστή και σαν Dependency Modelling, που γίνεται σε δύο επίπεδα: στο δομικό επίπεδο του μοντέλου (συνήθως σε γραφική μορφή) καθορίζεται ποιες μεταβλητές είναι εξαρτημένες τοπικά με ποιες, ενώ στο ποσοτικό επίπεδο του μοντέλου προσδιορίζεται η δύναμη αυτών των εξαρτήσεων με χρήση κάποιας αριθμητικής κλίμακας. Η ανάλυση κανόνων συσχετίσεων είναι ένα ενεργό πεδίο έρευνας και εκτός από το market basket analysis, χρησιμοποιείται σε πληθώρα άλλων εφαρμογών, από το Web usage mining, τη συσχέτιση υπόπτων εγκληματικών ενεργειών, την πρόβλεψη αποτυχίας της λειτουργίας τηλεπικοινωνιακών διακόπτων, το network intrusion detection καθώς και σε άλλες τεχνικές όπως graph mining, clustering with links, και bioinformatics.

3.3.8. Ανακάλυψη Προτύπων Ακολουθιών (Sequential Pattern Discovery)

Η ανακάλυψη προτύπων ακολουθιών (sequential pattern discovery) ή αλλιώς ακολουθιακή ανάλυση (sequential analysis) χρησιμοποιείται για να καθοριστούν σειριακά πρότυπα στα δεδομένα. Αυτά τα πρότυπα βασίζονται σε μια χρονική ακολουθία ενεργειών και είναι παρόμοια με τις συσχετίσεις στο ότι συσχετίζονται τα δεδομένα (ή τα γεγονότα) που εξάγονται, με την διαφορά ότι η συσχέτιση τους βασίζεται στον χρόνο. Οι παραπάνω μέθοδοι και αλγόριθμοι μπορούν και να συνδυάζονται.

3.3.9. Συστατικά αλγορίθμων εξόρυξης γνώσης

Τα παρακάτω συστατικά της ανάλυσης των αλγορίθμων εξόρυξης φαντάζουν στην αρχή θεωρητικά αλλά βοηθούν σαν χάρτης γνώσης για την γενικότερη μελέτη που παρουσιάζεται εδώ. Η πληθώρα των εκατοντάδων αλγορίθμων εξόρυξης γνώσης, εξαιτίας και των πολλών επιστημονικών πεδίων από τα οποία προέρχονται, είναι τέτοια που μπορεί να μπερδέψει ακόμη και έναν έμπειρο αναλυτή, και ο Fayyad και συνεργάτες τονίζουν ότι ενώ διαφωτίζονται πολλές μέθοδοι εξόρυξης γνώσης στη βιβλιογραφία υπάρχουν ουσιαστικά ορισμένες βασικές τεχνικές και συστατικά και οι υπόλοιπες τεχνικές αποτελούν παραλλαγές αυτών.

Οι 10 πιο γνωστοί αλγόριθμοι εξόρυξη γνώσης στην ψηφοφορία στο «International Conference in Data Mining» του 2006 είναι:

1): C4.5 (61 ψήφοι) – κατηγοριοποίηση (δέντρα απόφασης)

2): K-Means (60 ψήφοι) - συστηματοποίηση

3): SVM (58 ψήφοι) – κατηγοριοποίηση (support vector machine)

4): Apriori (52 ψήφοι) – κανόνες συσχέτισης

5): EM (48 ψήφοι) – στατιστική, συστηματοποίηση (expectation maximization)

6): PageRank (46 ψήφοι) – εξόρυξη γνώσης από ιστοσελίδες

7): AdaBoost (45 ψήφοι) – μετα-ταξινομητής

8): kNN (45 ψήφοι) – κατηγοριοποίηση-συσταδοποίηση (k-κοντινότεροι γείτονες)

9): Naive Bayes (45 ψήφοι) – στατιστική, κατηγοριοποίηση

10): CART (34 ψήφοι) – κατηγοριοποίηση (δέντρα απόφασης)

Οι βασικές συνιστώσες από τις οποίες αποτελείται ένας αλγόριθμος εξόρυξης γνώσης σύμφωνα με τους Hand και συνεργάτες είναι:

- Η δομή που θα αναζητηθεί
- Το κριτήριο αξιολόγησης
- Η μέθοδος βελτιστοποίησης και αναζήτησης
- Μια στρατηγική διαχείρισης δεδομένων

4. Τεχνολογίες αναλυτικής δεδομένων

4.1. Εισαγωγή

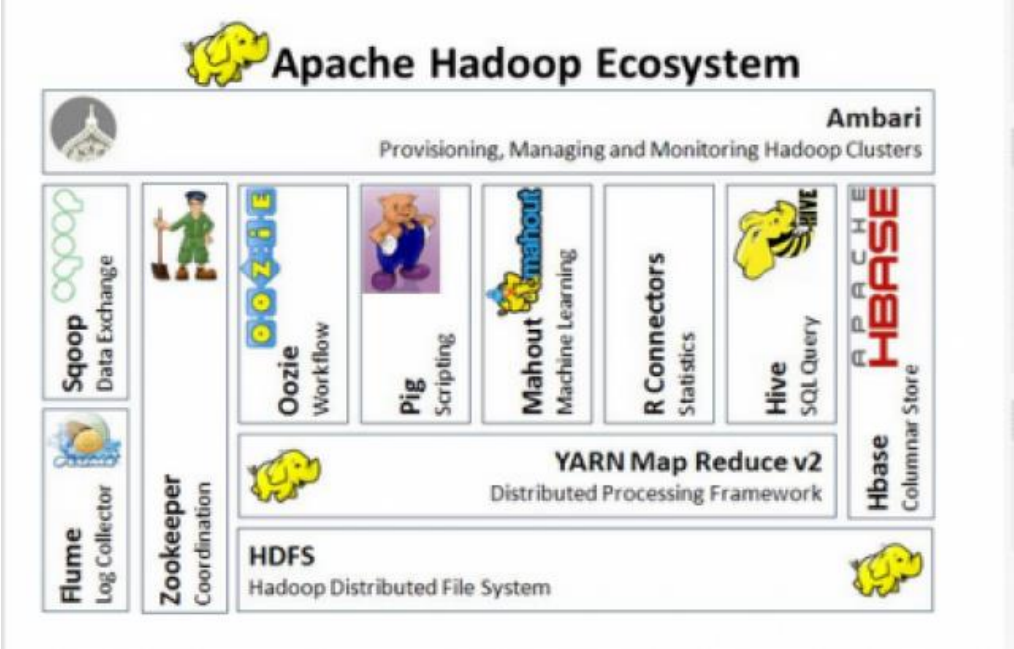
Το Apache Hadoop είναι λογισμικό ανοικτού πλαισίου που χρησιμοποιείται για αποθήκευση και επεξεργασία σετ δεδομένων σε υψηλή κλίμακα σε συστάδες για αξιοποιήσιμο hardware[69]. Το Hadoop αποτελεί ένα ύψιστο project του Apache που χρησιμοποιείται από μια μεγάλη παγκόσμια κοινότητα χορηγών και χρηστών. Είναι κατοχυρωμένα από το Apache License 2.0. Το Hadoop δημιουργήθηκε απο τους Doug Cutting και Mike Cafarella το 2005. [70]

4.2. Η τεχνολογία του Apache Hadoop

Το Apache Hadoop δομείται από τις ακόλουθες ενότητες:

- Hadoop Common: περιέχει βιβλιοθήκες και εργαλεία που χρειάζονται άλλες ενότητες Hadoop
- Hadoop Distributed File System (HDFS): ένα κατανεμημένο σύστημα αρχείων που αποθηκεύει δεδομένα σε μηχανές, παρέχοντας υψηλό εύρος ζώνης κατά μήκος της συστάδας. [69]
- Hadoop YARN: αποτελεί μια πλατφόρμα διαχείρισης πηγών που είναι υπεύθυνη για τη διαχείριση υπολογιστικών πηγών στις συστάδες και χρησιμοποιεί αυτές τις πηγές για οργάνωση των εφαρμογών των χρηστών[71]
- HadoopMapReduce: αποτελεί προγραμματιστικό μοντέλο για επεξεργασία μεγάλου σκέλους δεδομένων[72]
- Όλα τα τμήματα του Hadoop είναι σχεδιασμένα με το βασικό σκεπτικό πως οι αποτυχίες στα πλαίσια υλικού (ανεξάρτητων μηχανών ή αλληλουχίας μηχανών) είναι κοινότυπο φαινόμενο και πρέπει αυτόματα τον έλεγχο να τον παίρνει το λογισμικό.[73]

- Πέρα του HDFS, του YARN και του MapReduce, ολόκληρη η πλατφόρμα του Apache Hadoop " θεωρείται ότι δομείται από ένα σύνολο σχετιζόμενων αντικειμένων: το Apache Pig, το Apache Hive, το Apache.



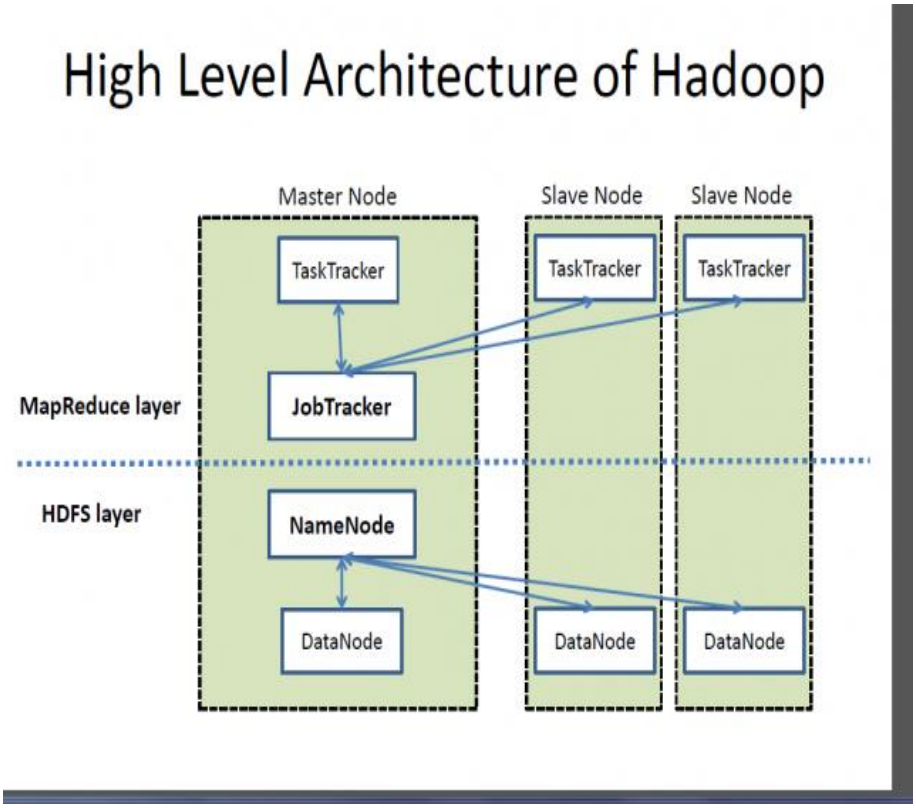
Εικόνα 11: Apache Hadoop

Για τους τελικούς χρήστες, αν η γλωσσά προγραμματισμού Java είναι πιο συνηθισμένη για το Map Reduce, οποιαδήποτε γλώσσα προγραμματισμού να χρησιμοποιηθεί με το "Hadoop Streaming"[74] για την υλοποίηση των τμημάτων "map" "reduce" στο πρόγραμμα του χρήστη. Τα Apache Pig και Apache Hive, μαζί με άλλα παρεμφερή projects, παρέχουν υψηλότερου επιπέδου διεπαφές όπως το Pig Latin και SQL παραλλαγές. Το Hadoop framework είναι κυρίως γραμμένο σε γλώσσα Java, με τμήματα βασικού κώδικα γραμμένο σε C και γραμμές εντολών (command lines) γραμμένα ως «σκριπτάκια»-φλοιού.[75]

4.3. HDFS και MapReduce

Υπάρχουν δύο κύρια συστατικά στοιχεία στον πυρήνα του Apache Hadoop 1.x: το Hadoop Distributed File System (HDFS) και το παράλληλης επεξεργασίας MapReduce framework.

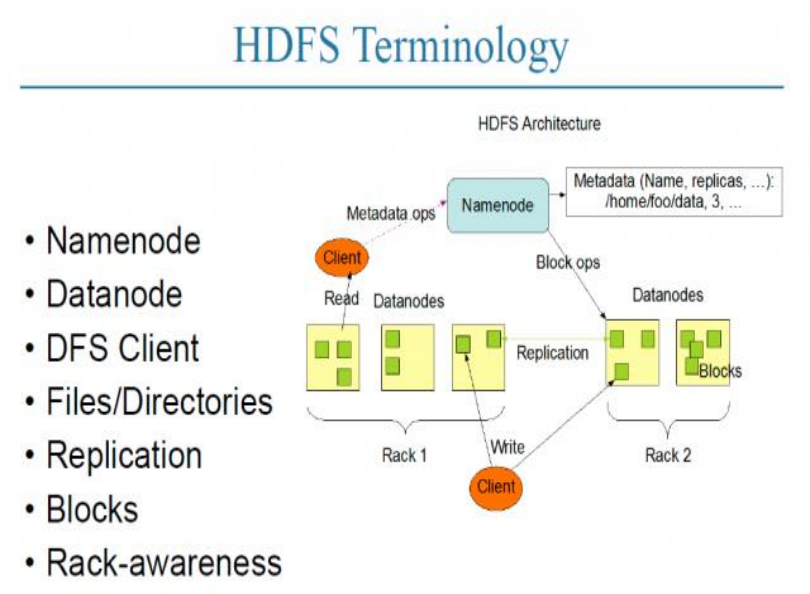
Και τα δύο αυτά είναι ανοιχτού λογισμικό και έχουν εμπνευστεί από τεχνολογίες που δημιουργήθηκαν στην Google.[76]



Εικόνα 12: High Level Architecture of Haddop

4.4. Hadoop distributed file system

Το Hadoop distributed file system (HDFS) [77] είναι ένα κατανεμημένο, επεκτάσιμο και φορητό σύστημα αρχείων γραμμένο σε Java για το Hadoop framework. Κάθε κόμβος σε στιγμιότυπο του Hadoop[78] έχει ένα κομβικό όνομα και μια συστάδα κόμβων δεδομένων που σχηματίζουν τη συστάδα του HDFS. Αυτή η κατάσταση είναι τυπική καθώς κάθε κόμβος δεν χρειάζεται κόμβο δεδομένων. Κάθε κόμβος δεδομένων εξυπηρετεί μπλοκ δεδομένων σ' όλο το δίκτυο χρησιμοποιώντας ένα πρωτόκολλο για μπλοκ το οποίο είναι συγκεκριμένο για HDFS[79]. Το σύστημα φακέλων χρησιμοποιεί το TCP/IP layer για επικοινωνία. Οι Clients επικοινωνούν μεταξύ τους μέσω κλήσεων απομακρυσμένης εμβέλειας (Remote procedure call (RPC)). [80]



Εικόνα 13: HDFS Terminology

Το HDFS αποθηκεύει μεγάλα αρχεία (κυρίως σε εμβέλεια από gigabytes μέχρι terabytes) σε εμβέλεια πολλαπλών μηχανών.[73] Εξασφαλίζει την αξιοπιστία του συστήματος αντιγράφοντας τα δεδομένα σε πολλαπλούς hosts,και ως εκ τούτου δεν απαιτείται χώρος του RAID για τους hosts. Με τη προκαθορισμένη replication value, 3, τα δεδομένα αποθηκεύονται σε 3 κόμβους: δύο του ίδιου σκέλους, και ένα διαφορετικού. Οι κόμβοι δεδομένων μπορούν να επικοινωνούν μεταξύ τους για να εξισορροπούν , να μεταφέρουν αντίγραφα και να διατηρούν την αντιγραφή δεδομένων σε υψηλά στάνταρ. Το HDFS δεν είναι πλήρως συμβατό με το POSIX καθώς οι απαιτήσεις για ένα POSIX σύστημα αρχείων διαφέρουν απ' τους στόχους μιας εφαρμογής τύπου Hadoop. Η απορία της ύπαρξης μη-πλήρους συμβατού συστήματος αρχείων με το POSIX είναι η αυξημένη απόδοση στη ροή των δεδομένων και η συμβατότητα με μη POSIX λειτουργίες όπως είναι το Append.[81]

Το HDFS έχει προσθέσει νέες δυνατότητες για την έκδοση 2.x που είναι άμεσα διαθέσιμες, επιτρέποντας έτσι στον κύριο metadata εξυπηρετητή (τον NameNode) να κάνει αυτόματα se περίπτωση σφάλματος του συστήματος.

Το σύστημα αρχείων του HDFS περιλαμβάνει και ένα δευτερεύον namenode, που κάνει πολύ κόσμο να πιστεύει πως σε περίπτωση που ο πρωτεύων namenode αποσυνδέεται από το σύστημα τότε τη θέση του αναλαμβάνει ο δευτέρων. Ουσιαστικά, ο δευτερεύων namenode συχνά συνδέεται με τον πρωτεύων namenode και δημιουργεί στιγμιότυπα του πρωτεύοντος των πληροφοριών του δευτερεύοντος namenode's, που έπειτα το σύστημα αποθηκεύει σε τοπικά ή απομακρυσμένα directories. Αυτά τα σημεία μπορούν να χρησιμοποιηθούν για επανεκκίνηση του πρωτεύοντος namenode χωρίς να χρειάζεται ο επανέλεγχος όλων των δραστηριοτήτων. Καθώς το namenode αποτελεί ένα απλό σημείο για αποθήκευση και διαχείριση των metadata, μπορεί να αναλάβει την υποστήριξη μεγάλου αριθμού αρχείων ειδικά μεγάλο αριθμό μικρών αρχείων. Το HDFS Federation, αποτελεί μια νέα προσθήκη η οποία στοχεύει στην αντιμετώπιση του προβλήματος αυτού μέχρι ένα σημείο επιτρέποντας πολλαπλά να λειτουργούν ως ανεξάρτητα namenodes.[82]

Ένα από τα προτερήματα της χρήσης του HDFS είναι η επίγνωση δεδομένων μεταξύ του job tracker και του task tracker. Ο job tracker δημιουργεί χάρτες ή ελαττώνει τις εργασίες στους task trackers έχοντας επίγνωση της τοποθεσία των δεδομένων. Π.χ, αν ένας κόμβος Α περιέχει τα δεδομένα (x, y, z) και ο κόμβος Β περιέχει τα δεδομένα (a, b, c), ο job tracker «προγραμματίζει» τον κόμβο Β να σχηματίσει χάρτη ή να περιορίσει τις διεργασίες στα (a,b,c) και ο κόμβος Α θα «προγραμματιστεί» να σχηματίσει χάρτη ή να ελαττώσει τις διεργασίες στα (x,y,z). Έτσι ελαττώνεται η κίνηση στο δίκτυο και εμποδίζεται η άσκοπη μεταφορά δεδομένων. Όταν το Hadoop χρησιμοποιείται με άλλα συστήματα αρχείων, αυτό το προνόμιο δεν είναι πάντα διαθέσιμο. Αυτή η κατάσταση έχει μεγάλο αντίκτυπο στο χρόνο εκτέλεσης διαδικασιών, κάτι που είναι εμφανές σε διαδικασίες που τρέχουν δεδομένα. Το HDFS έχει δημιουργηθεί κυρίως για αμετάβλητα αρχεία και μπορεί να είναι κατάλληλο για συστήματα που απαιτούν ταυτόχρονες λειτουργίες εγγραφής.[83]

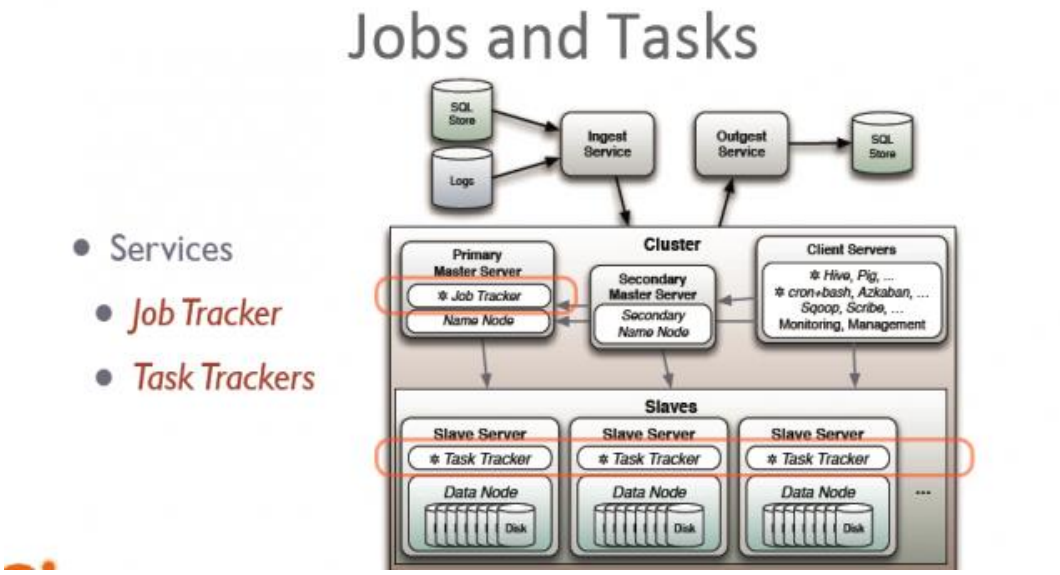
Άλλος περιορισμός του HDFS αποτελεί και το γεγονός ότι δεν μπορεί να χρησιμοποιηθεί απευθείας από ένα ήδη υπάρχων υπολογιστικό σύστημα. Η μεταφορά δεδομένων εντός και έκτος του συστήματος αρχείων του HDFS , είναι μια διαδικασία που συχνά χρειάζεται να υλοποιηθεί πριν και μετά την εκτέλεση κάποιας «δουλειάς» και μπορεί να καταλήξει μη-βολική. Ένα «filesystem in Userspace (FUSE)» [84] εικονικό σύστημα αρχείων έχει

αναπτυχθεί για την αντιμετώπιση αυτού του προβλήματος, τουλάχιστον για τα Linux και άλλα συστήματα Unix.[85]

Η πρόσβαση σε φακέλους μπορεί να επιτευχθεί μέσω του Java API, το Thrift API, για να δημιουργήσει client στη γλώσσα της επιλογής του χρήστη (C++, Java, Python, PHP, Ruby, Erlang, Perl [86], Haskell, C#, Cocoa, Smalltalk, ή OCaml), του interface της γραμμής εντολών ή να αναζητηθεί μέσω της εφαρμογής HDFS-UI στο HTTP. [87]

4.5. Job Tracker and Task Tracker: The Map Reduce engine

Πάνω του συστήματος αρχείων έρχεται η μηχανή Map Reduce, η οποία δομείται από το Job Tracker, στον οποίο οι εφαρμογές τύπου client υποβάλουν εργασίες του Map Reduce. Ο Job Tracker μειώνει την απαραίτητη δουλειά προκειμένου να καθιστούν διαθέσιμοι οι Task Tracker κόμβοι στη συστάδα, προσπαθώντας να κρατήσει τη δουλεία όσο το δυνατόν πιο κοντά στα δεδομένα.

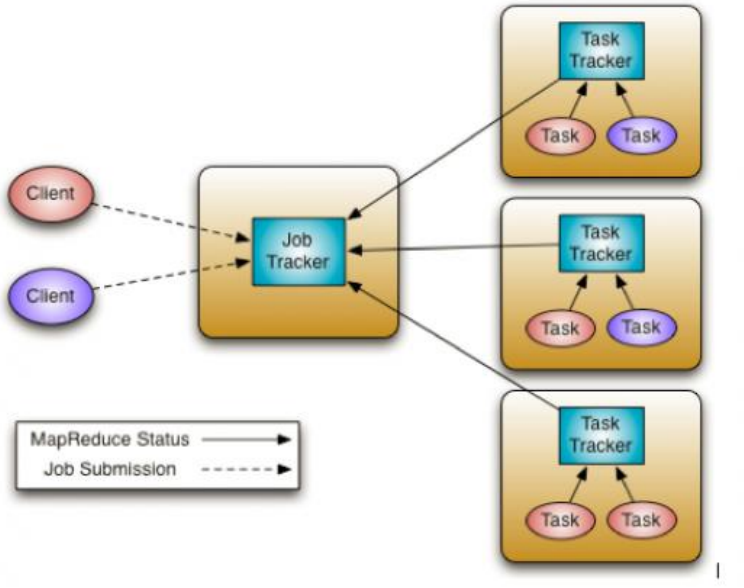


Εικόνα 14: JobTracker and TaskTracker

Με ένα «rack-aware» σύστημα αρχείων, ο Job Tracker γνωρίζει ποιος κόμβος περιέχει δεδομένα, και ποιες από τις μηχανές είναι κοντά σε αυτόν. Αν τη δουλειά δεν μπορεί να

την αναλάβει ένας κόμβος στον οποίο βρίσκονται τα δεδομένα, η προτεραιότητα δίνεται σε κόμβους ίδιου επιπέδου. Με αυτόν τον τρόπο ελαττώνεται η κίνηση στο κυρίως δίκτυο.

Αν ένας Task Tracker αποτύχει ή ξεμείνει από χρόνο, το τμήμα εργασίας που έχει αναλάβει δέχεται επανόρθωση. Το Task Tracker σε κάθε κόμβο δημιουργεί μια ξεχωριστή διεργασία σε Java Virtual Machine για να εμποδίσει το Task Tracker από το να «αποτύχει» αν η εργασία που υλοποιείται υπερφορτώσει το JVM. Ο Task Tracker στέλνει στο Job Tracker κάθε λίγα λεπτά προκειμένου να ελέγξει την κατάσταση του. Η κατάσταση και οι πληροφορίες των Job Tracker και Task Tracker εκθέτονται από το Jetty και είναι προσβάσιμα από web browser.



Εικόνα 15: Task Tracker

Αν ο Tracker αποτύχει στο Hadoop 0.20 ή σε προηγούμενη έκδοση, όλη η δουλεία χάνεται. Η έκδοση 0.21 πρόσθεσε λειτουργία για checkpoint στην περίπτωση αυτή. Το Job Tracker κρατάει αρχείο για την κατάσταση του συστήματος αρχείων. Όταν ο Job Tracker ξεκινάει, κάνει έλεγχο για αυτά τα δεδομένα ώστε να συνεχίσει η δουλεία από το σημείο που σταμάτησε.

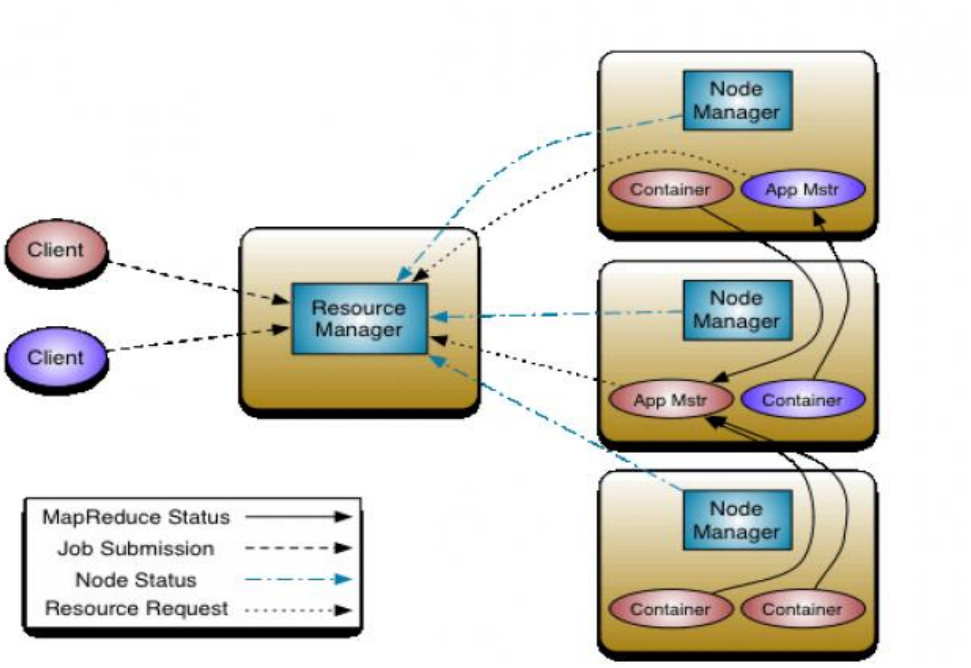
4.6. Γνωστοί περιορισμοί αυτής της προσέγγισης στο Hadoop 1.x

Η κατανομή της δουλειάς στους Task Trackers είναι πολύ απλή διαδικασία. Κάθε Task Tracker περιέχει ένα σύνολο διαθέσιμων θέσεων(slots) (πχ "4 slots"). Κάθε ενεργός χάρτης ή διαδικασία ελάττωσης λαμβάνει χώρο σε ένα slot. Ο Job Tracker αναθέτει τη δουλειά στον κόμβο που είναι κοντινότερος στα δεδομένα που υπάρχει διαθέσιμος χώρος. Δεν λαμβάνεται υπόψη ο φόρτος του συστήματος της μηχανής που έχει ανατεθεί η δουλειά και συνεπώς ούτε η διαθεσιμότητά της. Αν ένας Task Tracker είναι πολύ αργός, μπορεί να καθυστερήσει όλη η δουλειά του Map Reduce — ειδικά προς το τέλος όπου μπορούν να καταλήξουν να περιμένουν την πιο αργή διεργασία. Αν τεθεί σε εφαρμογή η θεωρητική εκτέλεση τότε μία διεργασία μπορεί να ανατεθεί σε πολλαπλούς κόμβους-σκλάβους.[88]

4.7. Apache Hadoop NextGenMapReduce (YARN)

Το Map Reduce έχει υποστεί πλήρη αναδόμηση στο Hadoop-0.23 και αυτό είχε σαν αποτέλεσμα τη δημιουργία του αποκαλούμενου, Map Reduce 2.0 (MRv2) ή αλλιώς YARN.

Το Apache™ Hadoop® YARN είναι ένα υπό-project του Hadoop της Apache Software Foundation που εισήχθη στο Hadoop 2.0 που διαχωρίζει τη διαχείριση πηγών και τα εξαρτήματα επεξεργασίας. Το YARN δημιουργήθηκε από ανάγκη να γίνει εφικτό ένα ευρύτερο σύνολο προτύπων αλληλεπίδρασης για αποθηκευμένα δεδομένα στο HDFS πέρα του Map Reduce. Η YARN-based αρχιτεκτονική του Hadoop 2.0 μια ευρύτερη επεξεργαστική πλατφόρμα η οποία δεν περιορίζεται μόνο στο Map Reduce.[89-91]

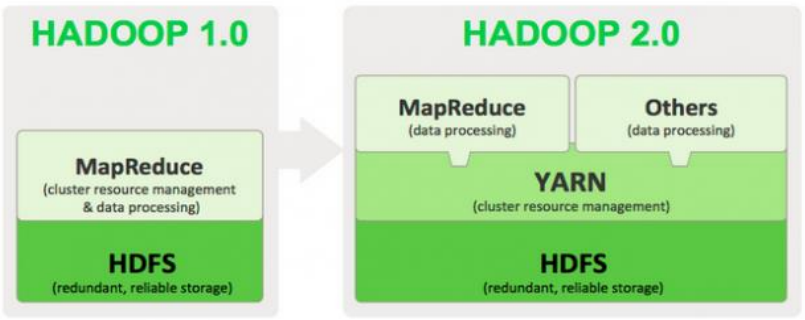


Εικόνα 16: Map Reduce

Η βασική ιδέα για το MRv2 είναι να μπορεί να διαχωρίζει τις δύο βασικές λειτουργίες του Job Tracker, (τη διαχείριση πηγών και επίβλεψη διεργασιών), σε διαφορετικούς τομείς. Η ιδέα είναι να υπάρχουν ένας Διαχειριστής Πηγών(Resource Manager (RM)) και ανα εφαρμογή ένας Application Master (AM). Μια εφαρμογή είναι είτε μία απλή διεργασία στα πλαίσια του ή ένα DAG διεργασιών.

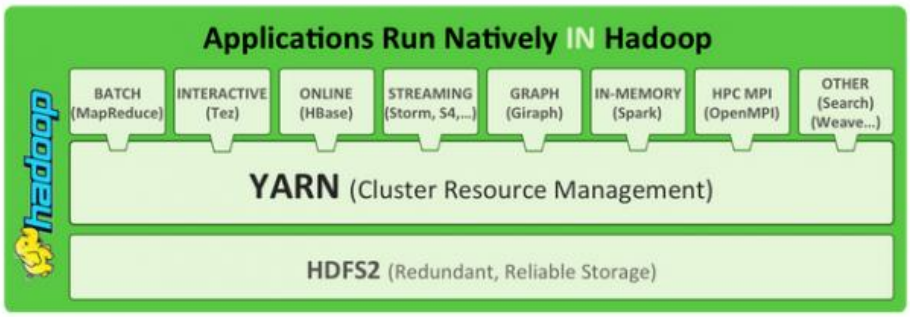
Το Resource Manager και ο per-node slave, ο Node Manager (NM), δημιουργούν το framework για επεξεργασίας δεδομένων. Ο Resource Manager είναι η απόλυτη αρχή που κατανέμει τις πηγές σόλες τις εφαρμογές του συστήματος.[92]

Ο ανά-εφαρμογή Application Master αποτελεί ουσιαστικά μια συγκεκριμένα του framework βιβλιοθήκη και έχει ως στόχο τη δραπέτευση πηγών από το Resource Manager και τη συνεργασία με τον/τους Node Manager(s) για την υλοποίησης και επίβλεψη διεργασιών.



Εικόνα 17: Hadoop 2.0

Όταν μέρος του Hadoop 2.0, το YARN παίρνει τις δυνατότητες διαχείρισης πηγών του Map Reduce και τις καθιστά προς χρήστη νέων μηχανών. Αυτό επίσης συμβάλλει στο το Map Reduce να κάνει αυτό που μπορεί να κάνει καλύτερα, να επεξεργάζεται δεδομένα. Με το YARN, γίνεται εφικτό να τρέχουν πολλαπλές εφαρμογές στο Hadoop, οι οποίες μοιράζονται μια κοινή διαχείριση πηγών. Πολλοί οργανισμοί έχει ήδη ξεκινήσει να φτιάχνουν εφαρμογές στο YARN προκειμένου να τις μεταφέρουν IN στο Hadoop.



Εικόνα 18: YARN

Ως τμήμα του Hadoop 2.0, το YARN παίρνει τις δυνατότητες διαχείρισης πηγών του Map Reduce ώστε να μπορούν να χρησιμοποιηθούν από νέες μηχανές. Χάρη στο YARN, επιτυγχάνεται το τρέξιμο πολλαπλών εφαρμογών στο Hadoop, οι οποίες εφαρμογές ανήκουν στο ίδιο σύστημα διαχείρισης πηγών. Πολλοί οργανισμοί έχουν αρχίσει τη δημιουργία εφαρμογών στο YARN για να τις μεταφέρουν στο Hadoop. Όταν τα επιχειρησιακά δεδομένα καθίστανται διαθέσιμα στο HDFS, καθιστάται σημαντικό να

υπάρχουν πολλαπλοί τρόποι επεξεργασίας δεδομένων. Με το Hadoop 2.0 και το YARN οι οργανισμοί μπορούν να χρησιμοποιούν το Hadoop για streaming, [74] διαδραστικές λειτουργίες και λοιπές εφαρμογές που καθίστανται δυνατές μέσω του Hadoop. [75]

4.7.1. YARN

Το YARN ενισχύει την ισχύ μιας υπολογιστικής συστάδας Hadoop με τους ακόλουθους τρόπους:

- Επεκτασιμότητα: Η επεξεργαστική δύναμη δεδομένων συνεχώς αυξάνεται. Λόγω του ότι το YARN Resource Manager εστιάζει αποκλειστικά στο σχεδιασμό, μπορεί να διαχειριστεί συστάδες δεδομένων πιο εύκολα.
- Συμβατότητα με το Map Reduce: Υπάρχουσες εφαρμογές και χρήστες του Map Reduce μπορούν να λειτουργήσουν στο YARN χωρίς να υπάρξει καμία διαταραχή στις υπάρχουσες διεργασίες.
- Βελτιστοποιημένη αξιοποίηση συστάδων: The Resource Manager είναι ουσιαστικά ένας οργανωτής που βελτιστοποιεί την αξιοποίηση συστάδων ανάλογα των δοθέντων κριτηρίων όπως η εγγύηση χώρου, η σωστή κατανομή και τα SLAs. Επίσης, σε αντίθεση με προηγουμένως δεν υπάρχουν ονομαστικοί χάρτες και reduce slots,πράγμα που βοηθάει στην καλύτερη αξιοποίηση των πηγών που παρέχουν οι συστάδες.
- Υποστήριξη για φόρτο εργασίας πέρα του Map Reduce: Επιπρόσθετα προγραμματιστικά μοντέλα όπως η επεξεργασία γραφημάτων και η διαδραστική μοντελοποίηση γίνονται πλέον δυνατά για την επεξεργασία δεδομένων. Αυτά τα μοντέλα επιτρέπουν στις επιχειρήσεις να επιτυγχάνουν επεξεργασία σε σχεδόν real-time καταστάσεις και αυξημένο στα πλαίσια της επένδυσης τους στο Hadoop.
- Ευλυγισία: Με το Map Reduce να μετατρέπεται σε μια προσιτή για το χρήση βιβλιοθήκη library, μπορεί να εξελιχθεί πέρα του πλαισίου διαχείρισης δεδομένων.
- Πως λειτουργεί το YARN

- Η βασική ιδέα για το YARN είναι ο διαχωρισμός των Job Tracker/Task Tracker σε διαφορετικές οντότητες:
- Ένας γενικός Resource Manager
- Ένας ανα-εφαρμογή Application Master
- Ένας ανά-κόμβο σκλάβος Node Manager και
- Ένας ανά-εφαρμογή container ο οποίος τρέχει σε έναν Node Manager
- Ο Resource Manager και ο Node Manager σχηματίζουν το νέο και απλοποιημένο σύστημα για διαχείριση διεργασιών σε κατανεμημένου χαρακτήρα. Ο Resource Manager είναι η απόλυτη οντότητα που διαχειρίζεται τις πηγές κατά μήκος του συστήματος. Ο ανά-εφαρμογή Application Master αποτελεί μια framework-specific οντότητα και έχει σαν στόχο του με τη διαπραγμάτευση πηγών από τον Resource Manager και τη συνεργασία με τον/τους Node Manager(s) για την εκτέλεση και το έλεγχο διεργασιών. Ο Resource Manager έχει ένας διοργανωτή, ο οποίος είναι υπεύθυνος για την κατανομή πηγών και τρέχουσες εφαρμογές σύμφωνα με περιορισμούς όπως τη χωρητικότητα των ουρών, όρια χρήσης κτλ. Ο σχεδιαστής πραγματοποιεί τη λειτουργία του σύμφωνα με απαιτήσεις για πηγές των εφαρμογών. Ο Node Manager είναι ένα ανά-μηχανή σκλάβος, ο οποίος είναι υπεύθυνος να τρέχει containers εφαρμογών, να επιβλέπει την χρησιμοποίηση πηγών (cpu, μνήμη, σκληρό δίσκο, δίκτυο) και να δίνει αναφορά στον Resource Manager. Κάθε Application Master έχει την ευθύνη για την διαπραγμάτευση πηγών από τον σχεδιαστή ,να επιβλέπει την κατάσταση του και να ελέγχει την πρόοδο τους. Από την πλευρά του συστήματος ο Application Master λειτουργεί σαν απλός container.[91]

5. Μελέτη Περίπτωσης

5.1. Εισαγωγή

Η εφαρμογή που υλοποιήθηκε στην παρούσα διπλωματική εργασία αποτελείται από δύο επιμέρους τμήματα (αλγορίθμους) και έχει ως σκοπό την ανάλυση και την κατηγοριοποίηση των δεδομένων που συλλέγονται μέσω του κοινωνικού δικτύου Twitter. Τα δεδομένα αυτά συλλέγονται από κατοίκους τριών ευρωπαϊκών χωρών. Οι 3 χώρες που επιλέχτηκαν να αναλυθούν είναι η Ελλάδα, η Γαλλία και η Αγγλία.

Συγκεκριμένα με ειδικό αλγόριθμο που υλοποιήσαμε συλλέγουμε τα δεδομένα μας (Tweets) μέσω του Twitter api. Στην συνέχεια αναλύουμε αυτά τα δεδομένα με την βοήθεια λεξικών στα οποία έχει δοθεί ειδικό βάρος σε συγκεκριμένες λέξεις έτσι ώστε να ελέγχονται και να αξιολογούνται τα δεδομένα (Tweets) που μας ενδιαφέρουν. Εφόσον ολοκληρωθεί η εν λόγω αξιολόγηση τα δεδομένα (Tweets) κατηγοριοποιούνται.

Η εφαρμογή που υλοποιήθηκε αποτελείται από δύο επιμέρους τμήματα που το κάθε ένα έχει διαφορετικό ρόλο στην λειτουργία της. Το πρώτο τμήμα απότελείται από ένα Windows Console Application όπου μέσου αυτού γίνεται η διαδικασία ανάκτησης, αξιολόγησης και αποθήκευσης των tweets. Αυτή απευθύνεται στον κάτοχο της εφαρμογής ώστε να «ταΐσει» τον big data cluster με δεδομένα. Η δεύτερη είναι ένα ASP.NET MVC[94] Web Application που απευθύνεται στους End-Users της εφαρμογής.

Μετά το πέρας του αλγορίθμου ουσιαστικά θα είναι δυνατή η αξιολόγηση σε αυτές τις 3 χώρες. Συγκεκριμένα θα επιτευχθεί η διαπίστωση συμπερασμάτων αναφορικά με τις διατροφικές συνήθειες των κατοίκων τους.

5.2. Λειτουργικότητα

Αρχικά, το πρώτο στάδιο λειτουργίας της εφαρμογής μας είναι η ανάλυση των tweets και η αποθήκευση των αποτελεσμάτων.

Αυτό επιτυγχάνεται μέσω των παρακάτω ενεργειών:

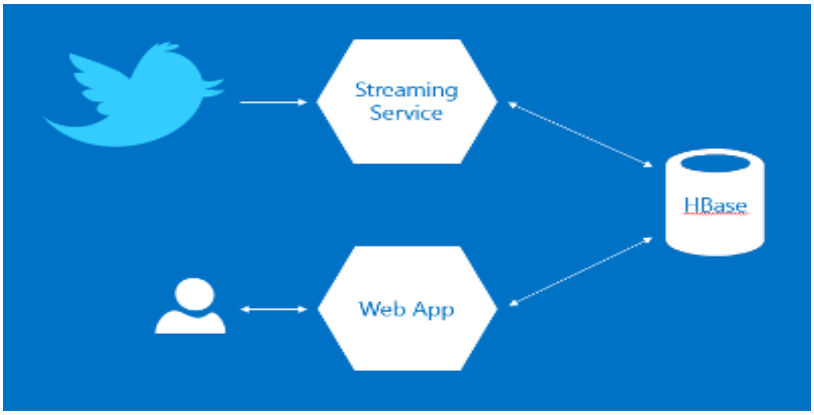
- Ανάκτηση των tweets των κατοίκων, συνεχώς και σε πραγματικό χρόνο
- Φιλτράρισμα αυτών που έχουν σχέση με τον τρόπο ζωής τους μέσω lexicons που έχουν σχέση με υγιεινή και ανθυγιεινή ζωή.
- Αξιολόγηση των συναισθημάτων τους δίνοντας βαρύτητα σε κάθε μία λέξη μέσα στο tweet χρησιμοποιώντας sentiment lexicon.
- Υπολογισμός της βαθμολογίας του tweet
- Αποθήκευση αποτελεσμάτων σε μια αποθήκη δεδομένων.

Έπειτα, όσο τα δεδομένα συλλέγονται, το web application παίρνει τα αποτελέσματα κατηγοριοποιημένα ανά χώρα και ανά τρόπου ζωής από την αποθήκη δεδομένων και τα οπτικοποιεί τοποθετώντας τα πάνω σε χάρτη και εμφανίζοντας χρήσιμες πληροφορίες που θα χρησιμοποιηθούν στην παρακάτω έρευνα της διπλωματικής εργασίας.

Εφόσον έχει ήδη γίνει η εύρεση και επεξεργασία των tweets, εδώ οι διαδικασίες που γίνονται είναι λιγότερες και απαιτούν λιγότερη υπολογιστική ισχύ ώστε να μπορεί να τρέχει σε οποιοδήποτε μηχάνημα ακόμα και όταν τα tweets για οπτικοποίηση είναι πολλά.

Οι κυριότερες από αυτές τις διαδικασίες είναι:

- Ανάκτηση των δεδομένων από την βάση
- Υπολογισμός ποσοστού υγιεινών ανά ανθυγιεινών tweets
- Εμφάνιση των tweets σε διαφορετικά layers πάνω στον χάρτη και δεδομένων



Εικόνα 19: Ανάλυση του κύκλου ζωής του Twitter

5.3. Τεχνολογίες

5.3.1. Τεχνολογίες που χρησιμοποιήθηκαν

Both: .NET Framework & C#

Server-Side

- Apache Hadoop
- Apache HBase
- Microsoft Azure HDInsight
- Microsoft Azure Storage
- Twitter Streaming API[74]

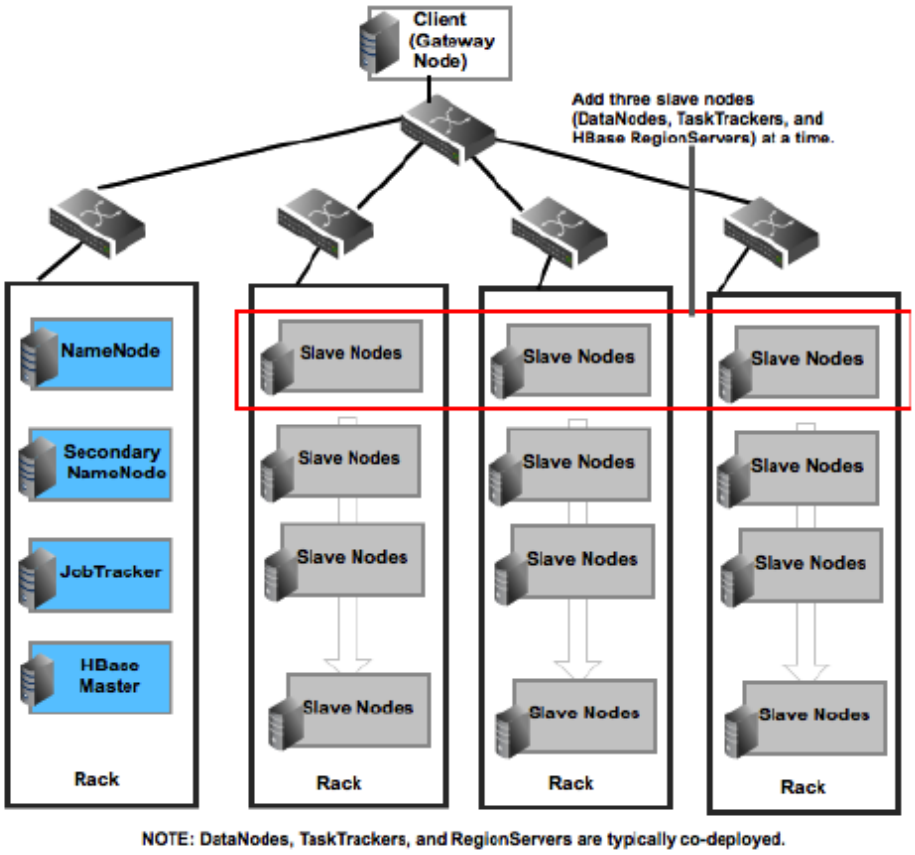
Client-Side

- ASP.NET MVC[94]
- Bing Heat Maps
- Bootstrap Framework [11]
- HTML5 & CSS3
- JQuery & JavaScript

5.3.2. Επεξήγηση των κυριότερων Τεχνολογιών στην εφαρμογή

Apache Hadoop

Το Apache Hadoop είναι ένα open-source framework που χρησιμοποιείται από κατανεμημένα συστήματα για αποθήκευση και επεξεργασία πολύ μεγάλου όγκου δεδομένων μέσω του Hadoop Distributed File System και του Map Reduce Engine (το οποίο δεν χρησιμοποιούμε διότι παίρνουμε τα δεδομένα μέσω του Apache HBase)[90]. Ο ελάχιστος αριθμός κόμβων που απαιτεί το Hadoop είναι 4 (1 Master και 3 Slaves).



Εικόνα 20: Hadoop Cluster 1 - Typical Roles and Tasks

Από την σχεδίαση του, εάν υπάρξουν αποτυχίες υλικού κάποιου κόμβου μέσα στον cluster, φροντίζει να μπορεί να δουλέψει και χωρίς αυτόν, παρέχοντας έτσι σχεδόν 100% διαθεσιμότητα.

Το κύριο μειονέκτημα του είναι ότι είναι αρκετά δαπανηρό, αν σκεφτεί κανείς ότι χρειάζεται τουλάχιστον 4 servers για να στηθεί. Βέβαια, δεν έχει καθόλου μεγάλες απαιτήσεις σε υπολογιστική ισχύ, άρα, μπορεί να χρησιμοποιηθεί εξίσου καλά σε low-cost μηχανήματα.

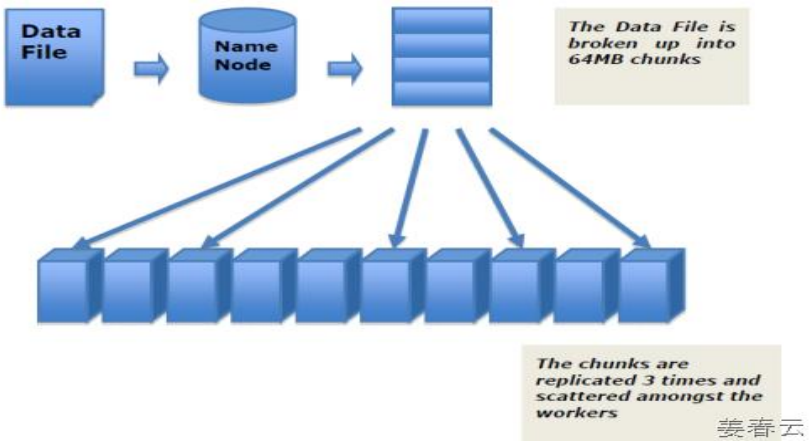
Hadoop Distributed File System

Το HDFS είναι το σύστημα αρχείων που χρησιμοποιεί το Hadoop ώστε να αποθηκεύσει τα δεδομένα. Επιτρέπει να συνδέονται οι κόμβοι που υπάρχουν μέσα στο cluster και να

λειτουργεί ως ένα ομοιογενές σύστημα αρχείων. Είναι fault-tolerant και παρέχει πρόσβαση υψηλής απόδοσης για μεγάλα σύνολα δεδομένων.

Οι στόχοι λειτουργίας του είναι:

- Ανίχνευση σφαλμάτων και εφαρμογή γρήγορης και αυτόματης ανάκτησης
- Επεξεργαστική λογική κοντά στα δεδομένα ανά slave Server, αντί τα δεδομένα κοντά στον κεντρικό Server
- Φορητότητα σε ετερογενείς hardware και λειτουργικά συστήματα
- Επεκτασιμότητα να αποθηκεύουν και να επεξεργάζονται αξιόπιστα μεγάλες ποσότητες δεδομένων
- Οικονομία στο bandwidth
- Αποδοτικότητα με τα δεδομένα και την επεξεργαστική λογική στον ίδιο κόμβο
- Αξιοπιστία με την αυτόματη διατήρηση πολλαπλών αντιγράφων των δεδομένων σε διαφορετικούς κόμβους και ανάκτηση αυτών σε περίπτωση βλάβης



Εικόνα 21: Hadoop Cluster 2 - HDFS Example

Εφαρμογή του Hadoop Cluster 1 - HDFS στην εργασία.

Το Hadoop είναι το κύριο framework που τρέχει στα Windows Servers και έχει πρόσβαση στο Azure.

Η επιλογή του Hadoop ως την κύρια πλατφόρμα για την αποθήκευση των δεδομένων της εφαρμογής έγινε, καθώς μακροπρόθεσμα και μιλώντας πλέον για μια real-world application, τα δεδομένα που αντλούνται από το Twitter και πρέπει να αναλυθούν και να αποθηκευτούν είναι μεγάλα σε όγκο. Επιπλέον εντοπίζονται και άλλα πλεονεκτήματα όπως:

- Η μορφή των δεδομένων είναι unstructured, το οποίο έχει περισσότερα πλεονεκτήματα όταν δουλεύουμε με big data.
- Η ανάκτηση των δεδομένων είναι γρηγορότερη σε σχέση με μια SQL Database.
- Είναι εύκολα επεκτάσιμο.
- Η εγκατάσταση και η συντήρηση του είναι πολύ εύκολη μέσω του Microsoft Azure HDInsight, το οποίο θα εξετάσουμε στην συνέχεια.

Apache HBase

Το Apache HBase είναι μια μη σχεσιακή, κατανεμημένη βάση δεδομένων (NoSQL) ανοιχτού κώδικα που τρέχει πάνω από το Hadoop και προσφέρει read/write πρόσβαση στα big data μας σε πραγματικό χρόνο. Πιο συγκεκριμένα, το HBase μας παρέχει δυνατότητες βάσης δεδομένων, κάτι το οποίο δεν προσφέρεται απευθείας από το Hadoop[90].

Μερικά πλεονεκτήματα του Apache HBase είναι:

- Εύκολα επεκτάσιμο
- Συνεπής read/writes ώστε να μην υπάρχουν deadlocks.
- Αυτόματοι και παραμετροποιήσιμοι πίνακες από τα μεγάλα δεδομένα
- Αυτόματη υποστήριξη failover μεταξύ των Servers.
- Ευέλικτη και εύκολη χρήση του API
- Μπλοκ κρυφής μνήμης για ερωτήματα σε πραγματικό χρόνο.
- Αυτόματη δημιουργία RESTful Web service ερωτημάτων που υποστηρίζει XML και Json

Εφαρμογή του Apache HBase στην εργασία

Εφόσον το Hadoop δεν υποστηρίζει από μόνο του SQL-like ερωτήματα σε πραγματικό χρόνο, χρησιμοποιήθηκε το HBase για να επιτευχθεί πολύ εύκολα η εισαγωγή των δεδομένων από το Twitter στον cluster και ταυτόχρονα να μπορούν αυτά να αναπαρασταθούν στον end-user μέσω του Web Application. [26]

Στην συνέχεια η εφαρμογή μπορεί να αναπτυχθεί περαιτέρω ώστε μέσω του HBase να γίνονται και άλλα ερωτήματα από τον τελικό χρήστη στην βάση και να παρέχονται σε αυτόν και άλλα χρήσιμα δεδομένα.

Microsoft Azure HDInsight

Το Azure HDInsight είναι ένα Service στο cloud platform Azure της Microsoft, οπού μπορείς πολύ εύκολα να δημιουργήσεις Hadoop clusters που τρέχουν είτε Linux είτε Windows στο cloud, είναι εύκολα παραμετροποιήσιμο ορίζοντας τον αριθμό των κόμβων και πληρώνοντας μόνο για την υπολογιστική ισχύ και τον χώρο που χρησιμοποιείτε.

Μας παρέχει πολλά εργαλεία παραμετροποίησης, στατιστικών και ανάλυσης του Hadoop cluster μας. Επίσης, μας επιτρέπει να δουλέψουμε με .NET και C# κάτι το οποίο δεν μας προσφέρει το Hadoop εξ αρχής, το οποίο είναι γραμμένο σε Java.

Η επιλογή του έγινε Azure HDInsight καθώς η δημιουργία του cluster και η εγκατάσταση προεκτάσεων (όπως του HBase) είναι πολύ εύκολη και μπορεί να γίνει μέσα σε λίγα λεπτά. Έτσι, έγινε εστίαση περισσότερο στην ανάπτυξη της εφαρμογής και όχι στο στήσιμο των servers και στην παραμετροποίηση τους.

Το αρνητικό είναι το κόστος που βέβαια θα υπήρχε ακόμα και με έναν server έστω και μικρότερο. Σύμφωνα με το Azure το κόστος ενοικίασης και προφανώς ο χρόνος εγκατάστασης και ρύθμισης του HDInsight σε ξεχωριστούς servers αντί του ολοκληρωμένου HDInsight cluster θα ήταν αρκετά μεγαλύτερα.

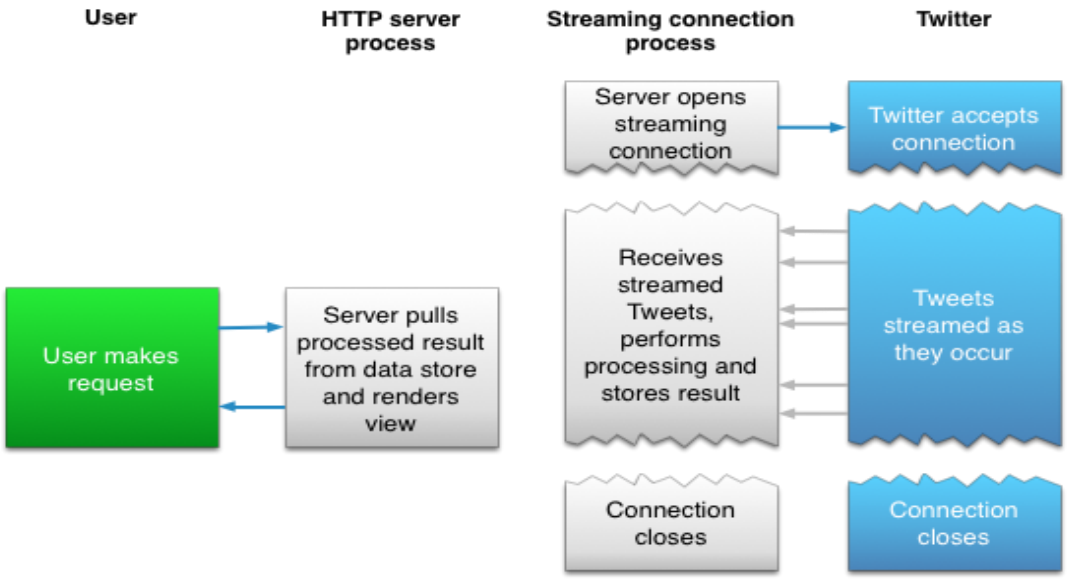
Microsoft Azure Storage

Το Microsoft Azure Storage είναι το cloud storage solution που μας προσφέρει το Azure και μπορεί εύκολα να χρησιμοποιηθεί ως το Hadoop Distributed File System που

αναφέραμε παραπάνω ώστε να αποθηκεύουμε τα δεδομένα μας σε blob μορφή. Ο χώρος που μας προσφέρει είναι τεράστιος (μπορεί να φτάσει σε μέγεθος Petabytes).

Twitter Streaming API

Το Twitter μας προσφέρει δωρεάν 2 ειδών APIs, το REST API και το Streaming API. Το πρώτο μας δίνει την δυνατότητα να «τραβάμε» ιστορικά tweets, δηλαδή, αυτά που έχουν ήδη γραφτεί στο Twitter και είναι δημόσια, μέσω RESTful calls σε ένα web service του Twitter. Το δεύτερο μας επιτρέπει να ανοίγουμε μια ροή δεδομένων με ένα endpoint του Twitter και να παίρνουμε τα tweets την στιγμή που γράφονται. Μέσω παραμέτρων που ορίζουμε κατά την έναρξη της σύνδεσης όπως την γλώσσα του tweet ή την τοποθεσία που γράφτηκε μας επιτρέπεται να τραβάμε tweets από ένα συγκεκριμένο μέρος και κατ' επέκταση από μία συγκεκριμένη χώρα.[74]



Εικόνα 22: Twitter Streaming API 1

Εφαρμογή του Twitter Streaming API στην εργασία

Η χρήση Streaming API[74] του Twitter έγινε ώστε για να ανακτάμε τα tweets σε πραγματικό χρόνο και να γίνεται η ανάλυση τους.

Ορίζοντας μια συνθήκη ώστε τα coordinates του κάθε tweet να είναι μέσα στα bounding boxes για την κάθε μια από τις 3 χώρες που είχαν οριστεί, είναι δυνατό με αυτόν τον τρόπο να φιλτράρονται τα tweets πολύ γρήγορα έτσι ώστε να λαμβάνονται εξ' αρχής αυτά που είναι επιθυμητά και απαραίτητα στα πλαίσια της εργασίας.

Κατά μέσο όρο τα tweets που λαμβάνονταν και από τις 3 χώρες ήταν 17/sec ενώ αυτά που είχαν σχέση με το lifestyle ήταν περίπου 0.5/sec

ASP.NET MVC

Στο μέρος της διεπαφής με τον χρήστη το κύριο framework που χρησιμοποιήθηκε ήταν το ASP.NET MVC[94], ένα open source server-side web application framework που χρησιμοποιείτε για την δημιουργία δυναμικών ιστοσελίδων. Η αρχιτεκτονική η οποία βασίζεται είναι η MVC (Model-View-Controller) όπου διαχωρίζονται σε ξεχωριστά αρχεία και φακέλους οι Model Classes, δηλαδή τον πυρήνα της εφαρμογής (Business Layer), και το View, δηλαδή το User Interface(Display Layer). Ενδιάμεσα από αυτά υπάρχει άλλο ένα είδος αρχείων, οι Controllers (Input/Output Control) που παίρνουν τα δεδομένα από το Model, τα επεξεργάζονται αν χρειαστεί και τα εμφανίζει στο User Interface του χρήστη. Οι Controllers παίρνουν μέρος επίσης στην αντίστροφη διαδικασία η οποία είναι να παίρνουν το input του χρήστη και δημιουργούν αντικείμενα Model κλάσεων και να τα χρησιμοποιούν κατάλληλα ή ακόμα τον κατευθύνουν στην σωστή σελίδα την οποία ο ίδιος ζήτησε με το Input του. Τρέχει πάνω στο IIS και στο.NET Framework και χρησιμοποιεί κάποια .NET γλώσσα προγραμματισμού στο back-end (server-side) και HTML5,CSS3,JavaScript που γίνονται render στον browser του χρήστη (front-end).

Η MVC αρχιτεκτονική είναι σχεδιασμένη να παρέχει ευελιξία στον developer χωρίζοντας το project (κυρίως web) σε επιμέρους κομμάτια. Έτσι τον βοηθάει να αναπτύξει αλλά και να συντηρήσει πολύ εύκολα και γρήγορα την εφαρμογή εφόσον ξέρει ανά πάσα στιγμή που να ανατρέξει ώστε να διορθώσει κάτι για παράδειγμα. Επιπλέον, είναι πιο ελαφριά από τις κλασσικές μεθόδους ανάπτυξης ιστοσελίδων.

Εφαρμογή του ASP.NET MVC στην εργασία

Όλο το Web Application αναπτύχθηκε χρησιμοποιώντας το ASP.NET [94] MVC Framework.

Αρχικά, μόλις ο χρήστης της εφαρμογής ζητήσει τα δεδομένα (υγιεινά ή ανθυγιεινά tweets), ο server που τρέχει το application στέλνει αίτηση στον cluster, μέσω ενός query στην HBase του, όλα τα tweets του συγκεκριμένου είδους. Μόλις ο server παραλάβει τα δεδομένα και υπολογίσει τα στατιστικά τους, τα εμφανίζει στον browser του χρήστη.

Bing Heat Maps

Οι χάρτες του Bing και ειδικότερα το Heat Maps Module είναι μια επέκταση του Bing Maps[95] γραμμένο σε JavaScript που μας επιτρέπει να εμφανίσουμε σημεία αιχμής πάνω στον χάρτη της Bing.

5.4. Οθόνες Συστήματος

5.4.1. Back-End

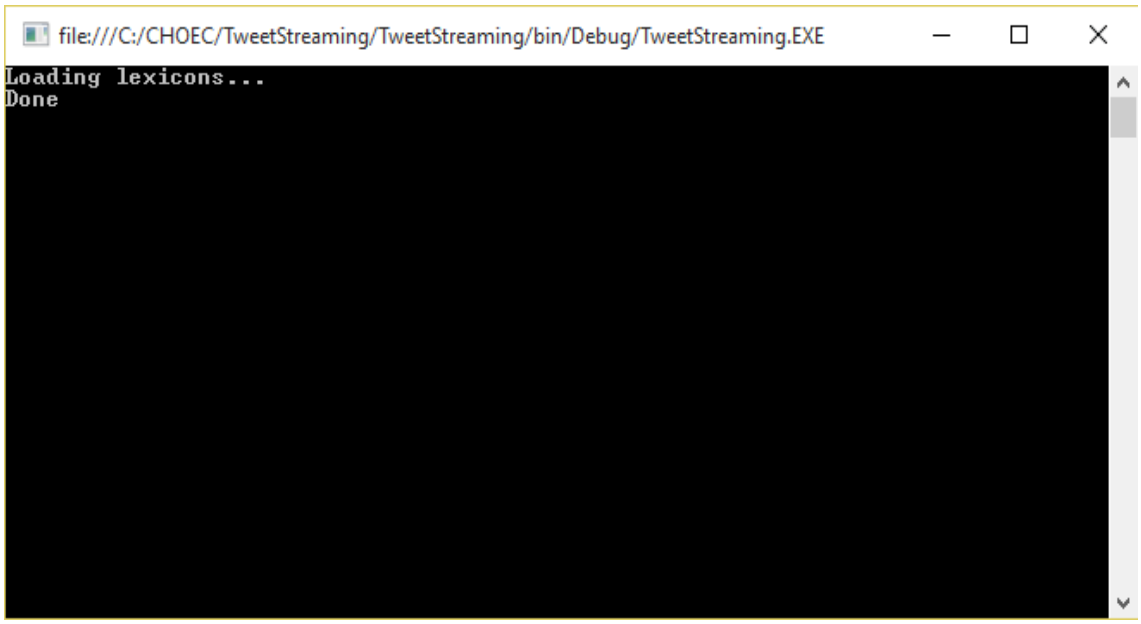
Αρχικά, εφόσον ο Hadoop Cluster μας μέσω του Azure λειτουργεί, αρχίζουμε την διαδικασία ανάκτησης, αξιολόγησης και αποθήκευσης των tweets μέσω του Windows Console Application που εξηγήσαμε παραπάνω.

1. Διάβασμα των lexicons για κάθε γλώσσα

   Τα lexicons που χρησιμοποιούνται και για τις τρεις γλώσσες είναι:

   - Healthy Dictionary – Greek, French, English – 200 Λέξεις
   - Unhealthy Dictionary – Greek, French, English – 200 Λέξεις
   - Sentiment Dictionary – Greek, French, English – 8000 Λέξεις

   Μόλις το διάβασμα των lexicons από την εφαρμογή ολοκληρωθεί εμφανίζεται το παρακάτω μήνυμα.



Έπειτα, αρχίζει το διάβασμα των tweets από τις χώρες που επιλέξαμε και εμφανίζει πόσα tweets διαβάζει το δευτερόλεπτο, πόσα από αυτά έγραψε στην HBase μας, δηλαδή πόσα Healthy/Unhealthy tweets υπάρχουν μέσα σε αυτά που διάβασε και εμφανίζει το κείμενο αυτών που βρήκε.

Η διαδικασία της κατηγοριοποίησης των tweets εφαρμόζεται με τον παρακάτω αλγόριθμο (απλοποιημένος):

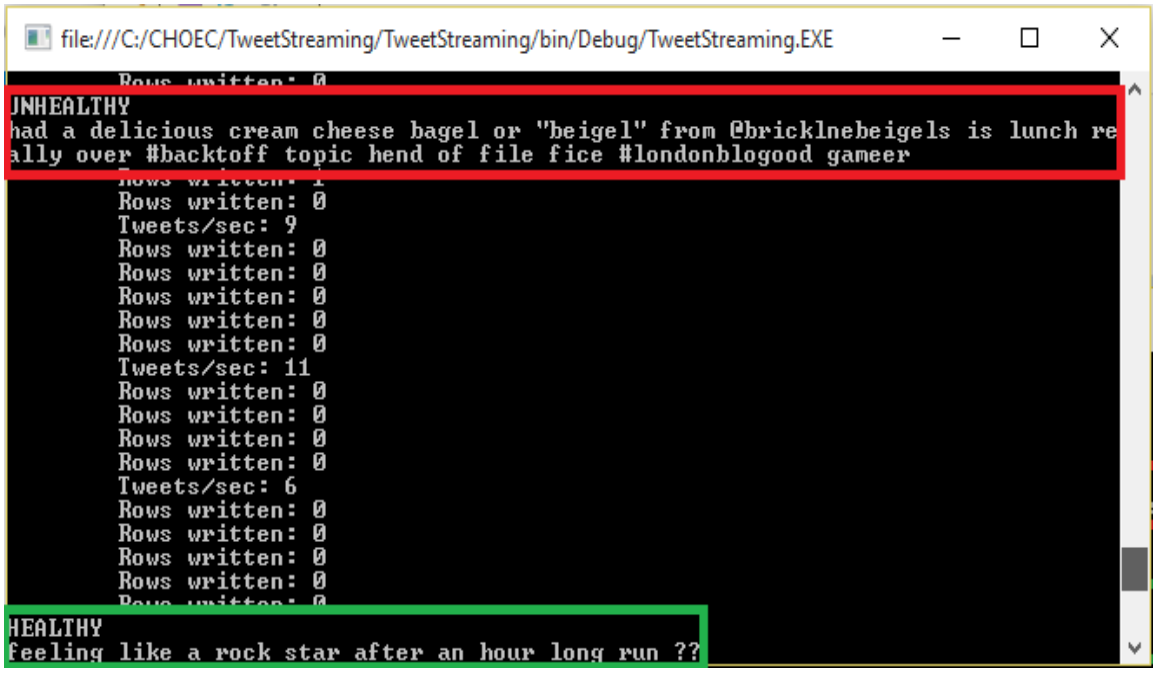
```
If (tweet is healthy)
{
        If (sentiment_score is good)
                addToHealthy(tweet);
        else
                addToUnhealthy(tweet);
}
Else if (tweet is unhealthy)
{
        If (sentiment_score is good)
                addToUnhealthy(tweet);
        else
                addToHealthy(tweet);

}
Else
{
        //ignore
}
```

Έτσι, με απλά λόγια αν το tweet αναφέρεται σε κάτι υγιεινό και τα συναισθήματα του είναι θετικά ή ουδέτερα τότε αυτό εισάγεται στην βάση ως HEALTHY. Αν αναφέρεται σε κάτι υγιεινό αλλά γράφει με αρνητικό ύφος, τότε προστίθεται ως UNHEALTHY.

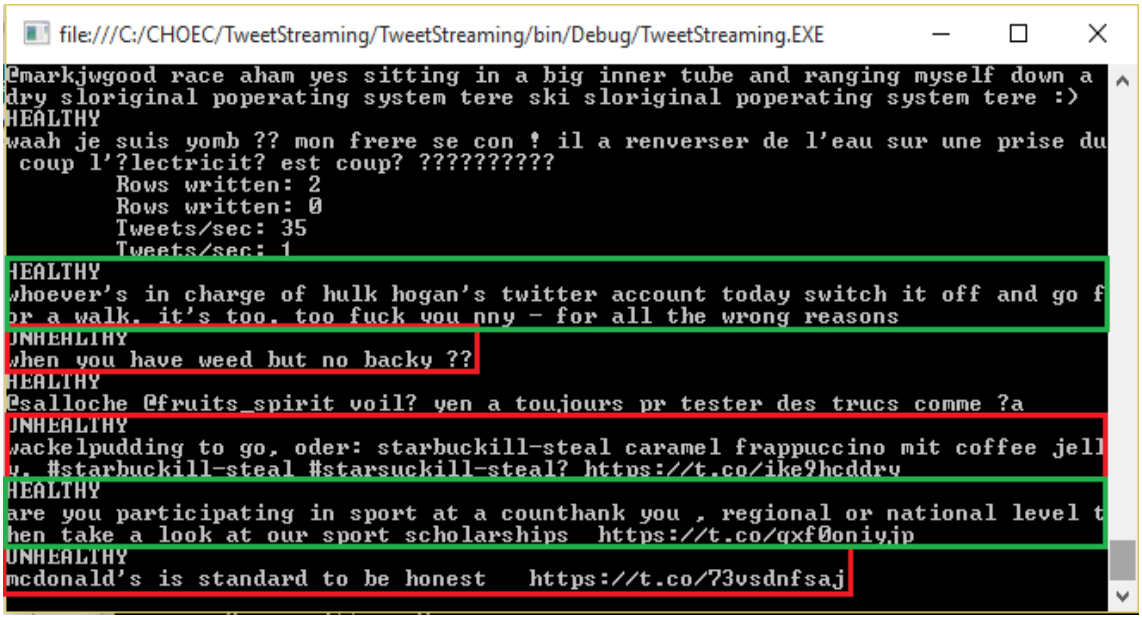
Αντιθέτως, αν το tweet αναφέρεται σε κάτι ανθυγιεινό και τα συναισθήματα είναι θετικά ή ουδέτερα τότε εισάγεται ως UNHEALTHY. Αλλιώς, αν τα συναισθήματα είναι αρνητικά τότε εισάγεται ως HEALTHY. Το ποσοστό ευστοχίας σύμφωνα με μερικές παρακολουθήσεις μας είναι περίπου στο 70-80%. Ποσοστά αρκετά ενθαρρυντικά για την έρευνα μας.

Ακολουθούν μερικά παραδείγματα:



Στο πρώτο παράδειγμα βλέπουμε 2 tweets που προέρχονται από την Αγγλία. Το πρώτο, που έχει σαν ετικέτα UNHEALTHY, μας περιγράφει για ένα νόστιμο Bagel (το οποίο είναι σίγουρα ανθυγιεινό) που έφαγε ο συγγραφέας από ένα μαγαζί BrickLneBeigels στο Λονδίνο. Μέσω του sentiment lexicon υπολογίζουμε ότι τα συναισθήματα του συγγραφέα είναι περισσότερο θετικά παρά αρνητικά μέσω συντελεστών βαρύτητας που παίρνει η κάθε λέξη μέσα στην πρόταση. Έτσι κατηγοριοποιούμε αυτό το tweet ως Unhealthy και το προσθέτουμε στην βάση δεδομένων μας.

Παρακάτω βλέπουμε ένα tweet με την ένδειξη ως HEALTHY. Αυτό μας περιγράφει το πως νιώθει ο συγγραφέας μετά από μια ώρα τρέξιμο. Επειδή τα συναισθήματα του είναι θετικά το προσθέτουμε στην βάση ως Healthy.

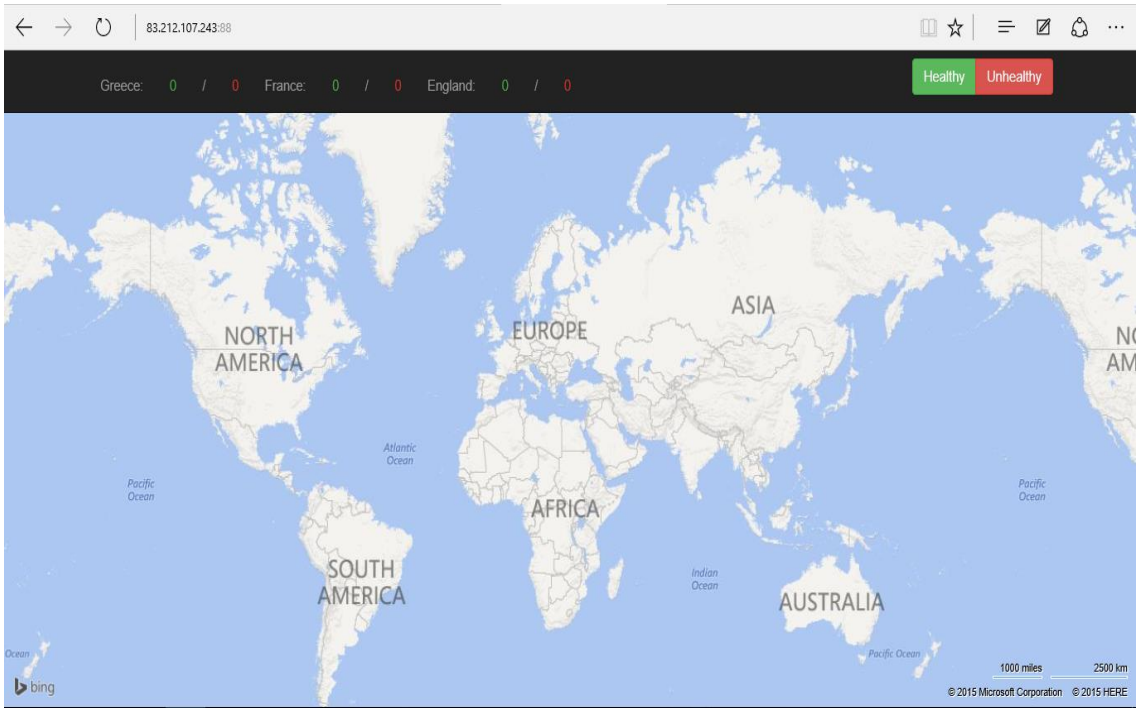


Σε αυτό το παράδειγμα έχουμε περισσότερα tweets που έχουν σχέση με την υγεία, όχι μόνο από την Αγγλία, αλλά και από την Γαλλία. Μπορείτε να διακρίνετε την κατηγοριοποίηση του καθενός ανάλογα με τα περιεχόμενα του.

5.4.2. Front-End

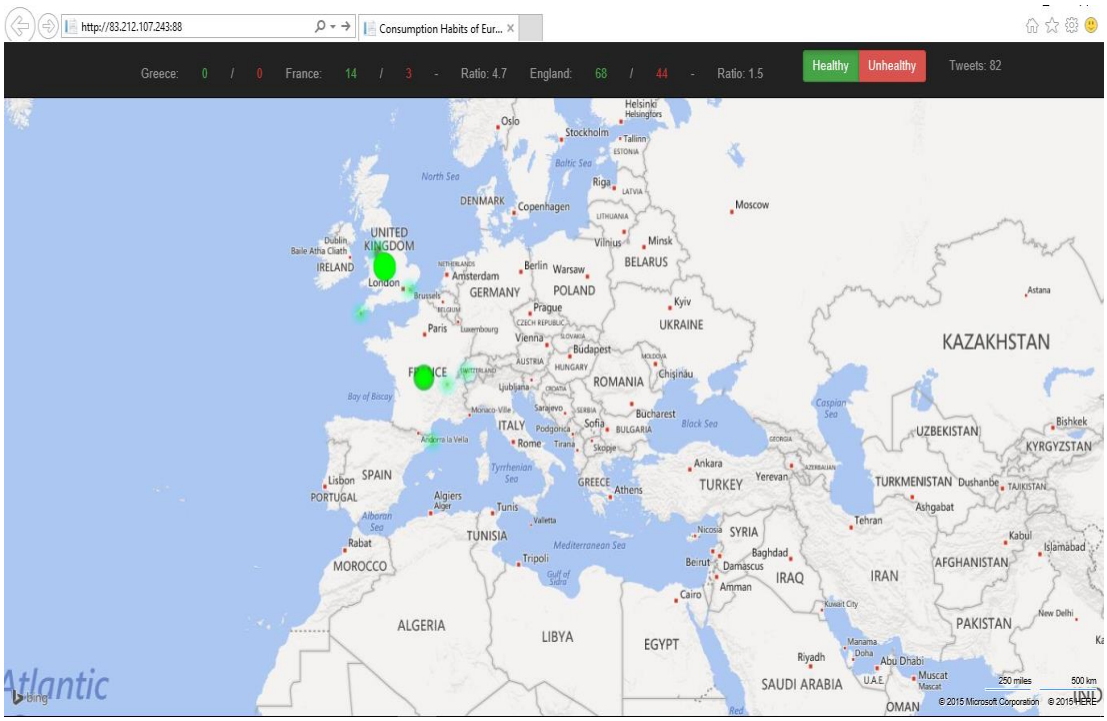
Το Web Application όπως αναφέρεται και παραπάνω, αναπτύχθηκε έτσι ώστε να επικεντρώνεται στις λειτουργίες της εφαρμογής, κάνοντας την διεπαφή του χρήστη εύχρηστη και απλή. Το Web Application στήθηκε σε ένα VM μέσω του Okeanos το οποίο είναι ένα cloud service που μας προσφέρει το GRNET δωρεάν.

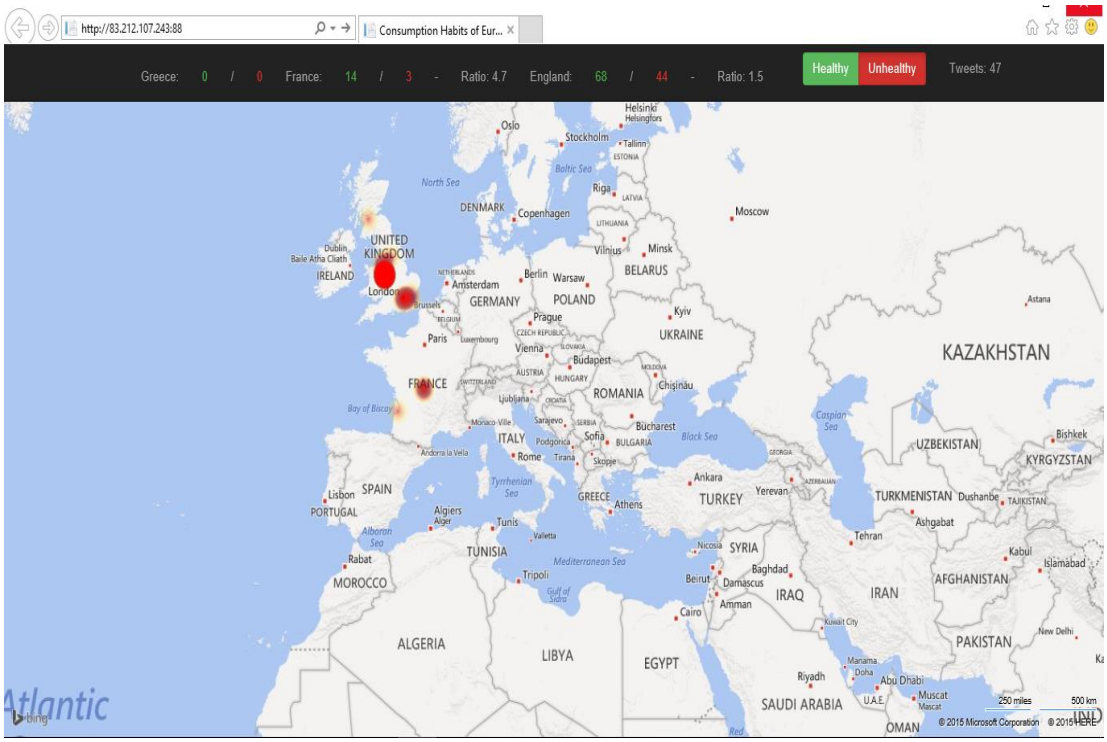
Αρχικά μπαίνοντας στην διεύθυνση του Web Application βλέπουμε έναν άδειο από σημεία χάρτη, και μία μπάρα στο πάνω μέρος στην οποία υπάρχουν πληροφορίες για τα Tweets για την κάθε χώρα (τα οποία είναι 0 σε πρώτη φάση) και δύο κουμπιά που γράφουν HEALTHY και UNHEALTHY όπως φαίνονται παρακάτω.



Πατώντας ένα από τα δύο κουμπιά, μέσω JavaScript γίνεται αίτημα στο web service του ASP.NET MVC application που αναφέραμε και πριν, απαιτώντας δεδομένα από την HBase στο Azure HDInsight.

Αυτό, στέλνει το αίτημα και το HBase μας επιστρέφει τα δεδομένα που έχουν συγκεντρωθεί μέχρι τώρα σε αυτήν. Έτσι, πατώντας το κουμπί του Healthy ή του Unhealthy θα εμφανιστούν τα παρακάτω δεδομένα.



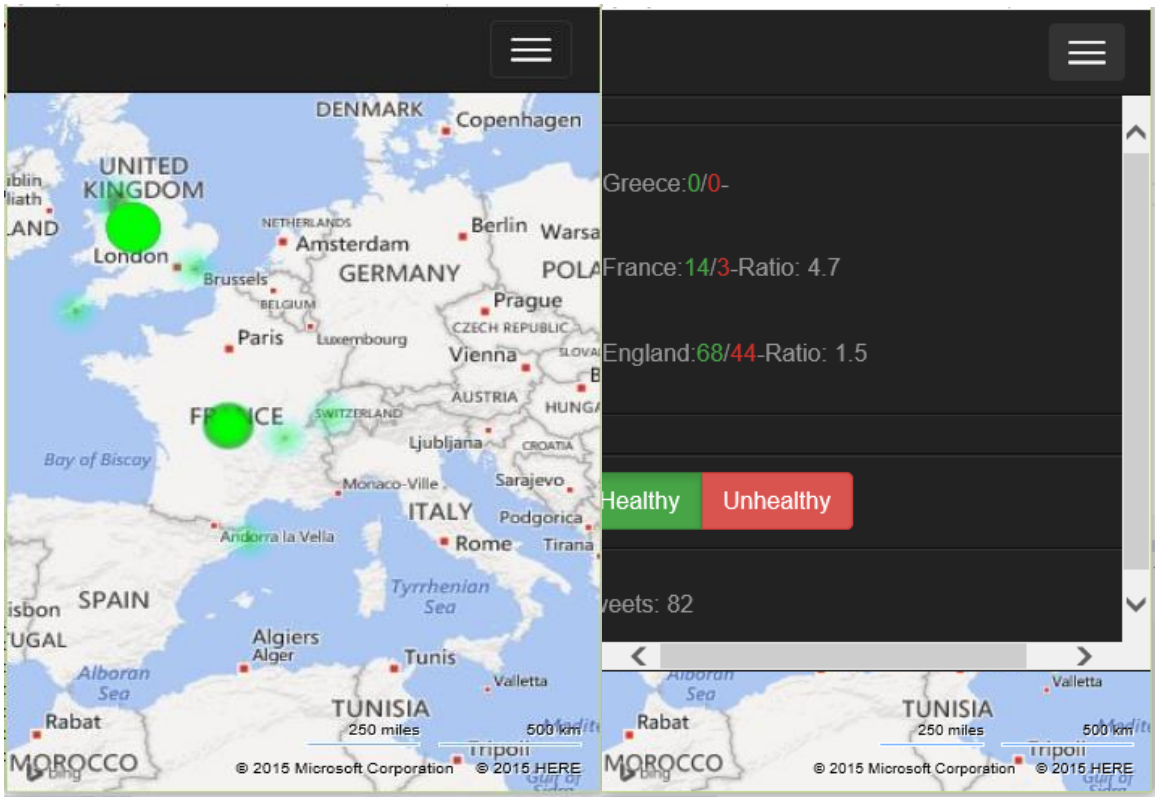


Και στις δύο περιπτώσεις, στον χάρτη εμφανίζονται τα σημεία σε μορφή Heat δηλαδή, όπου είναι συγκεντρωμένα πολλά tweets τα ομαδοποιεί και η κουκίδα είναι μεγαλύτερη

και πιο έντονη. Πάνω στην μπάρα, φαίνονται ο συνολικός αριθμός των tweets στον χάρτη, ο αριθμός των υγιεινών και ανθυγιεινών tweets, καθώς και το ποσοστό υγιεινών ανά ανθυγιεινών tweets, που θα χρειαστούν για την περεταίρω έρευνα μας.

Να σημειωθεί, ότι τα δεδομένα του παραδείγματος αυτού ήταν αποτέλεσμα την εκτέλεσης του προγράμματος ανάκτησης των tweets για 10 λεπτά μόνο. Δυστυχώς, δεδομένα για την Ελλάδα δεν υπάρχουν σε αυτό το παράδειγμα.

Η web εφαρμογή θα φαινόταν όπως φαίνεται παρακάτω σε κινητή συσκευή κάνοντάς την εύχρηστη.



6. Ερωτηματολόγιο

6.1. Εισαγωγή

Για την καλύτερη διασταύρωση των αποτελεσμάτων αναφορικά με τις διατροφικές συνήθειες των ατόμων δημιουργήσουμε ένα ερωτηματολόγιο το οποίο ζητήσαμε από άτομα διαφόρων ηλικιών να συμπληρώσουν. Πιο κάτω παραθέτουμε το ερωτηματολόγιο.

Το εν λόγο ερωτηματολόγιο γίνεται στα πλαίσια διπλωματικής εργασίας αναφορικά με τις διατροφικές συνήθειες των πολιτών στην Ελλάδα. Η εφαρμογή που υλοποιήθηκε στην διπλωματική αυτή εργασία αποτελείται από δύο επιμέρους εφαρμογές και έχει ως σκοπό την ανάλυση και οπτικοποίηση του τρόπου ζωής των κατοίκων τριών ευρωπαϊκών χωρών μέσω του κοινωνικού δικτύου Twitter.

Ουσιαστικά με το συγκεκριμένο ερωτηματολόγιο συλλέγουμε δεδομένα από ένα συγκεκριμένο τμήμα πληθυσμού με σκοπό να διασταυρώσουμε την εγκυρότητα των αποτελεσμάτων που θα διεξαχθούν από τα δεδομένα που θα αντλήσουμε από τον αλγόριθμο της εφαρμογής που υλοποιήθηκε στην παρούσα διπλωματική εργασία.

Οι ερωτήσεις του ερωτηματολογίου έχουν βασιστεί στις επιλεγόμενες λέξεις που έχουμε βάλει στον αλγόριθμο ώστε να κατηγοριοποιεί τα tweets που λαμβάνει από το κοινωνικό διαδίκτυο ως «υγιεινά» ή «ανθυγιεινά». Ουσιαστικά λοιπόν μέσω τον ερωτηματολογίου ελέγχουμε την εγκυρότητα των αποτελεσμάτων του αλγορίθμου αναφορικά με τον τρόπο ζωής των Ελλήνων.

Ακολούθως δίνεται το link του ερωτηματολογίου:

http://goo.gl/forms/6C64mhPPHf

6.2. Παράθεση ερωτηματολογίου

Προσδιορίστε το φύλλο σας.

- ◉ Άνδρας
- ○ Γυναίκα

Προσδιορίστε την ηλικία σας.

- ○ <15
- ○ 15-20
- ○ 20-25
- ○ 25-30
- ○ 30-40
- ○ >40

Θεωρείτε ότι είναι σημαντικό να λαμβάνεται πρωινό.

- ○ Ναι
- ○ Όχι

Πόσες ημέρες την εβδομάδα λαμβάνετε πρωινό;

- ○ Ποτέ
- ○ Σπάνια (1-2)
- ○ Συχνά (2-4)
- ○ Πολύ συχνά (4-6)
- ○ Κάθε μέρα (7)

Σας αρέσει το junk food;

- ○ Ναι

- Όχι

Πόσο συχνά την εβδομάδα τρώτε junk food;

- Ποτέ
- Σπάνια (1-2)
- Συχνά (2-4)
- Πολύ συχνά (4-6)
- Κάθε μέρα (7)

Καταναλώνετε αλκοολ;

- Ναι
- Όχι

5. Γνωρίζετε τις επιπτώσεις του αλκοόλ στον οργανισμό σας;

- Γνωρίζω
- Δεν γνωρίζω

Ποια είναι η επιθυμία σας σε εβδομαδιάια βάση για την κατανάλωση αλκοόλ;

Βάση των ημερών της εβδομάδας (π.χ η επιλογή 1 αντιστοιχεί σε 1 μέρα στην εβδομάδα).

1	2	3	4	5	6	7

Πόσες φόρες την εβδομάδα καταναλώνεται αλκοόλ;

- Ποτέ

- Σπάνια (1-2)
- Συχνά (2-4)
- Πολύ συχνά (4-6)
- Κάθε μέρα (7)

1. Σας αρέσουν τα γλυκά;

- Ναι
- Όχι

5. Πόσες φόρες την εβδομάδα καταναλώνεται γλυκά;

- Ποτέ
- Σπάνια (1-2)
- Συχνά (2-4)
- Πολύ συχνά (4-6)
- Κάθε μέρα (7)

Καπνίζετε;

- Ναι
- Όχι

Γνωρίζετε τις επιπτώσεις του καπνίσματος στον οργανισμό σας;

- Γνωρίζω
- Δεν γνωρίζω

Αν γνωρίζετε, τότε αναφέρετε κάποιες από αυτές.

Πόσο καπνίζετε σε ημερίσια βάση;

- 0-10 τσιγάρα
- 10-20 τσιγάρα
- 1-3 πακέτα
- >3 πακέτα

Πόσο συχνά την εβδομάδα αθλήστε;

	1	2	3	4	5	
Ποτέ						Κάθε μέρα

Πιστεύετε ότι κοιμάστε καλά;

	1	2	3	4	5	
Καθόλου καλά						Πολύ καλά

Πόσες φορές την εβδομάδα ξενυχτάτε;

- Ποτέ
- Σπάνια (1-2)
- Συχνά (2-4)
- Πολύ συχνά (4-6)
- Κάθε μέρα (7)

6.3. Ανάλυση Ερωτηματολόγιου

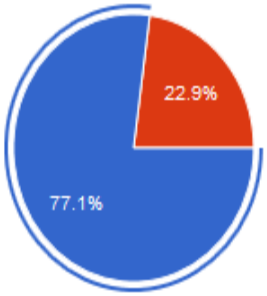
Το συγκεκριμένο ερωτηματολόγιο απαντήθηκε από 36 άτομα και τα αποτελέσματα που λάβαμε δίνονται ακολούθως. Η στατιστική ανάλυση του εν λόγο ερωτηματολόγιου έγινε με την χρήση του εργαλείου SPSS.

6.4. Αποτελέσματα ερωτηματολογίου

Consumption Habits
of Greek Citizens bas

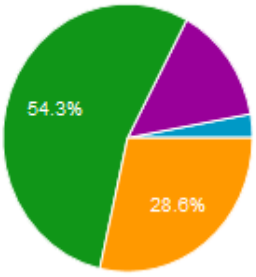
6.4.1. Ποσοτική ανάλυση

Προσδιορίστε το φύλλο σας.



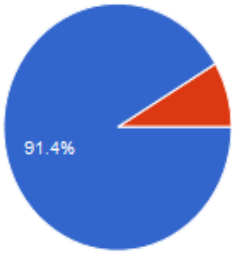
Άνδρας	27	77.1%
Γυναίκα	8	22.9%

Προσδιορίστε την ηλικία σας.



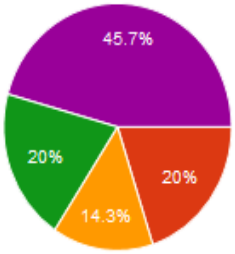
<15	**0**	0%
15-20	**0**	0%
20-25	**10**	28.6%
25-30	**19**	54.3%
30-40	**5**	14.3%
>40	**1**	2.9%

Θεωρείτε ότι είναι σημαντικό να λαμβάνεται πρωινό.



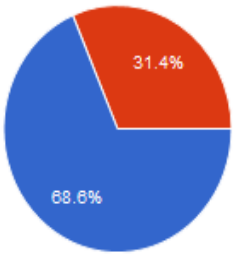
Ναι	**32**	91.4%
Όχι	**3**	8.6%

Πόσες ημέρες την εβδομάδα λαμβάνετε πρωινό;



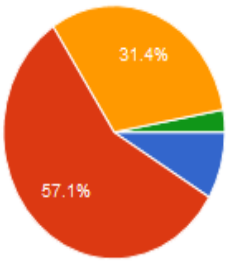
Ποτέ	**0**	0%
Σπάνια (1-2)	**7**	20%
Συχνά (2-4)	**5**	14.3%
Πολύ συχνά (4-6)	**7**	20%
Κάθε μέρα (7)	**16**	45.7%

Σας αρέσει το junk food;



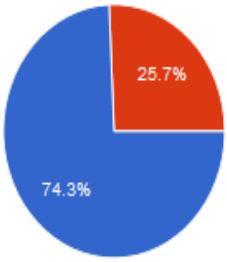
Ναι	**24**	68.6%
Όχι	**11**	31.4%

Πόσο συχνά την εβδομάδα τρώτε junk food;



Ποτέ	3	8.6%
Σπάνια (1-2)	20	57.1%
Συχνά (2-4)	11	31.4%
Πολύ συχνά (4-6)	1	2.9%
Κάθε μέρα (7)	0	0%

Καταναλώνετε αλκοολ;

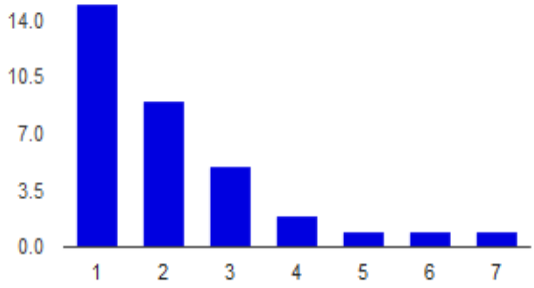


Ναι	26	74.3%
Όχι	9	25.7%

5. Γνωρίζετε τις επιπτώσεις του αλκοόλ στον οργανισμό σας;

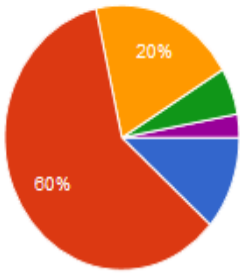
Γνωρίζω	35	100%
Δεν γνωρίζω	0	0%

Ποια είναι η επιθυμία σας σε εβδομαδιάια βάση για την κατανάλωση αλκοόλ;



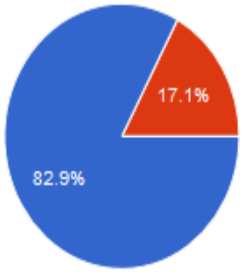
1	**15**	44.1%
2	**9**	26.5%
3	**5**	14.7%
4	**2**	5.9%
5	**1**	2.9%
6	**1**	2.9%
7	**1**	2.9%

Πόσες φόρες την εβδομάδα καταναλώνεται αλκοόλ;



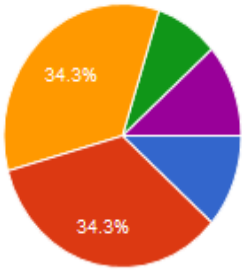
Ποτέ	**4**	11.4%
Σπάνια (1-2)	**21**	60%
Συχνά (2-4)	**7**	20%
Πολύ συχνά (4-6)	**2**	5.7%
Κάθε μέρα (7)	**1**	2.9%

1. Σας αρέσουν τα γλυκά;



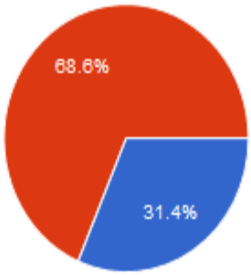
Ναι	**29**	82.9%
Όχι	**6**	17.1%

Πόσες φόρες την εβδομάδα καταναλώνεται γλυκά;



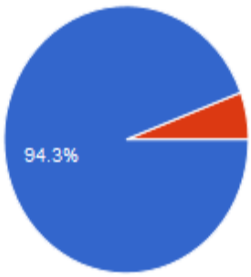
Ποτέ	**4**	11.4%
Σπάνια (1-2)	**12**	34.3%
Συχνά (2-4)	**12**	34.3%
Πολύ συχνά (4-6)	**3**	8.6%
Κάθε μέρα (7)	**4**	11.4%

Καπνίζετε;



Ναι	**11**	31.4%
Όχι	**24**	68.6%

Γνωρίζετε τις επιπτώσεις του καπνίσματος στον οργανισμό σας;



Γνωρίζω	**33**	94.3%
Δεν γνωρίζω	**2**	5.7%

Αν γνωρίζετε, τότε αναφέρετε κάποιες από αυτές.

Καρκίνος του πνεύμονα

ΑΥΞΗΣΗ ΠΙΕΣΗΣ, ΑΣΘΜΑ, ΒΛΑΒΗ ΠΝΕΥΜΟΝΩΝ

καρδιά, πνευμονία, συκώτι

καρκίνος , γήρανση του δέρματος ,αρτηριοσκλήρυνση

Καρκίνος

Πρόκληση καρκίνου του λάρυγγα και των πνευμόνων. Πρόκληση προβλημάτων στην εγκυμοσύνη. Υπογεννητικότητα.

Καρκίνος Ανικανότητα Έλλειψη δυνάμεων και αντοχών

Καρκίνος πνεύμονα

καρκίνος, εξασθένηση αισθήσεων, υποτονικότητα.

καρκίνος πνεύμονα, αλλοίωση δέρματος, κιτρίνισμα δοντιών, περιορισμένη αντοχή στην αναπνοή

Δυσλειτουργία πνευμόνων Καρδιακές παθήσεις Κόπωση

Καρκίνος, πνευμονικό οίδημα, φράξιμο αρτηριών

καρκίνος πνεύμονα, βουλωμένες αρτηρίες, άσθμα

Πτώση φυσικής κατάστασης. Δύσπνοια λόγω επιβάρυνσης πνευμόνων. Πρόωρη γήρανση κυττάρων. Σοβαρές πιθανότητες εμφάνισης Καρκίνου.

χρόνια προβλήματα υγείας, μείωση γονιμότητας, άσθμα, επίδραση στην εμφάνιση του ανθρώπου,

-Στεφανιαία νόσος -Μείωση μέσου όρου ζωής -Καρκίνο στον πνεύμονα -Κιτρίνισμα δοντιών -Αναπνευστικά προβλήματα

Βλάβη στους πνεύμονες, καρδιά, αγγεία κλπ

καρκίνος

Πρόωρη γήρανση Δέρματος

Πρόωρη γήρανση. Αυξημένες πιθανότητες καρκίνου. Απώλεια καλής φυσικής κατάστασης. Επιβάρυνση άλλων προβλημάτων υγείας.

προβλήματα στυτικής δυσλειτουργίας

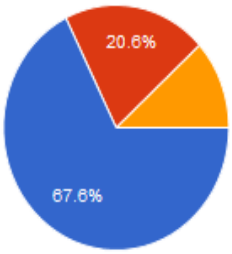
καρκίνος σε διάφορα μέρη του σώματος,

καρκίνος πνευμόνων

Αύξηση της πιθανότητας εμφάνισης όλων των μορφών καρκίνου και κυρίως καρκίνου των πνευμόνων.

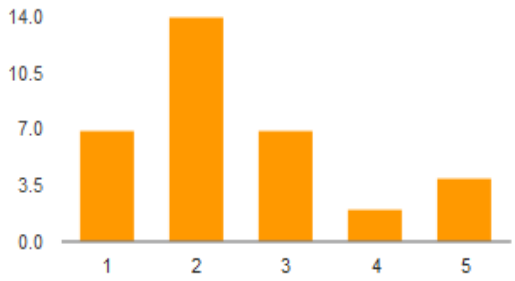
καρκίνος του πνεύμονα & γλώσσας απώλεια όσφρησης & γεύσης κιτρίνισμα δαχτύλων

Πόσο καπνίζετε σε ημερίσια βάση;



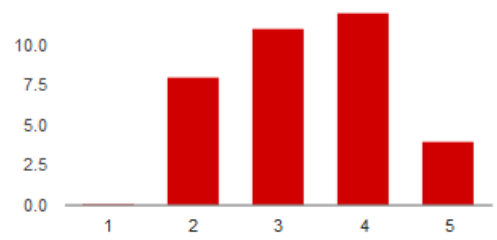
0	**23**	65.7%
1-10 τσιγάρα	**7**	20%
10-20 τσιγάρα	**4**	11.4%
1-3 πακέτα	**0**	0%
>3 πακέτα	**0**	0%

Πόσο συχνά την εβδομάδα αθλήστε;



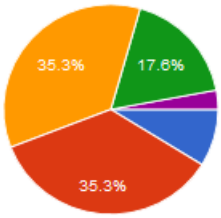
Ποτέ: 1	**7**	20.6%
2	**14**	41.2%
3	**7**	20.6%
4	**2**	5.9%
Κάθε μέρα: 5	**4**	11.8%

Πιστεύετε ότι κοιμάστε καλά;



Καθόλου καλά: 1	**0**	0%
2	**8**	22.9%
3	**11**	31.4%
4	**12**	34.3%
Πολύ καλά: 5	**4**	11.4%

Πόσες φορές την εβδομάδα ξενυχτάτε;



Ποτέ	**3**	8.6%
Σπάνια	**12**	34.3%
Συχνά	**12**	34.3%
Πολύ συχνά	**6**	17.1%
Κάθε μέρα	**1**	2.9%

Number of daily responses

7. Συμπεράσματα

Η ανάλυση δεδομένων στην υγεία είναι μία σύνθεση κλινικών καινοτομιών και τεχνολογίας. Μεγάλες ποσότητες δεδομένων έχουν γίνει διαθέσιμες σε πολλούς τομείς στον κάδο της υγείας. Είναι σχεδόν αδύνατο να επεξεργαστούμε και να εκμεταλλευτούμε αυτά τα δεδομένα χωρίς την χρήση νέων τεχνολογιών και εργαλείων. Με την χρήση τεχνικών ανάλυσης δεδομένων είναι δυνατό να εξάγουμε σημαντικές πληροφορίες από terabytes η ακόμα και petabytes δεδομένων πολύ πιο εύκολα. Χάρη σε αυτή την πληροφορία οι οργανισμοί της υγείας μπορούν να έρθουν σε θέση να προσφέρουν καλύτερες υπηρεσίες, πιο εύστοχες λύσεις και απαντήσεις σε ερευνητικά θέματα. Συγκεκριμένα μπορούν να αντιμετωπιστούν διάφορες ασθένειες, να προληφθούν επιδημίες και να προβλεφθούν ανάγκες. Επιπλέον παρέχεται η δυνατότητα εξειδικευμένων συνταγών για ασθενείς με βάση το προφίλ τους, τα γονίδια τους, και το ιστορικό τους.

Ταυτόχρονα όμως εξίσου σημαντικά είναι και τα οφέλη των κοινωνικών δικτύων στον κλάδο της υγείας. Συγκεκριμένα σε στιγμές κρίσης είναι εύκολη η ενημέρωση ατόμων που αναμιγνύονται σε ένα γεγονός. Επίσης είναι δυνατή η εκπαίδευση του ιατρικού προσωπικού παρακολουθώντας επεμβάσεις μέσω του κοινωνικού διαδικτύου του YouTube και παρατηρώντας τα σχόλια των γιατρών στο Tweeter. Επίσης είναι πολύ πιο εύκολο να ευαισθητοποιηθεί και να ενημερωθεί το κοινό πολύ πιο άμεσα και εύκολα μέσω των κοινωνικών διαδικτύων για σημαντικά θέματα υγείας. Τα κοινωνικά δίκτυα βοηθούν καταλυτικά τον κλάδο της υγείας καθώς παρέχουν μια μεγάλη συλλογή πληροφοριών αναφορικά με τις συμπεριφορές και τις επιλογές τον πολιτών, έτσι βοηθούν πολλές επιστημονικές έρευνες όπως για παράδειγμα στην επιστήμη της ψυχολογίας.

Στην συγκεκριμένη περίπτωση χρησιμοποιούμε μεγάλα δεδομένα πραγματικού χρόνου από το Tweeter προκειμένου να αναλύσουμε πόσα από αυτά έχουν σχέση με την υγεία και κατά πόσο αξιολογούνται ως θετικά ή αρνητικά. Αρχικά, το πρώτο στάδιο λειτουργίας της εφαρμογής είναι η ανάλυση των tweets και η αποθήκευση των αποτελεσμάτων. Αυτό επιτυγχάνεται μέσω τις ανάκτησης των tweets των κατοίκων συνεχώς και σε πραγματικό χρόνο σε φιλτράρισμα αυτών που έχουν σχέση με τον τρόπο ζωής τους μέσω lexicons που

έχουν σχέση με υγιεινή και μη υγιεινή ζωή με βάση την αξιολόγηση των διατροφικών συνηθειών τους δίνοντας βαρύτητα σε κάθε μία λέξη μέσα στο tweet χρησιμοποιώντας sentiment lexicon. Στο επόμενο βήμα, υπολογίζεται η βαθμολογία του tweet. Τελευταίο βήμα της διαδικασίας είναι η αποθήκευση αποτελεσμάτων σε μια αποθήκη δεδομένων.

Συμπεράνουμε λοιπόν ότι αν το tweet αναφέρεται σε κάτι υγιεινό και τα συναισθήματα του είναι θετικά ή ουδέτερα τότε αυτό εισάγεται στην βάση ως HEALTHY. Αν αναφέρεται σε κάτι υγιεινό αλλά γράφεται με αρνητικό ύφος, τότε προστίθεται ως UNHEALTHY.Για την διαπίστωση της εγκυρότητας του αλγορίθμου κάναμε πολλά παραδείγματα βάση των οποίων συμπεράνουμε ότι τα tweets που έχουν σχέση με την υγεία είναι αρκετά, όχι μόνο για την Αγγλία αλλά και για την Γαλλία. Ωστόσο είναι σημαντικό να αναφερθεί ότι διαπιστώσαμε πως τα tweets υγείας για την Ελλάδα ήταν ελάχιστα. Συγκεκριμένα για την Αγγλία συνήθως εντοπίζουμε τον τετραπλάσιο αριθμό δεδομένων συγκριτικά με εκείνα της Γαλλίας, ενώ στα αντίστοιχα ποσά η Ελλάδα εμφανίζει σχεδόν μηδενικά νούμερα.

Βασιζόμενοι λοιπόν στα αποτελέσματα της εφαρμογής που υλοποιήσαμε παρατηρήσαμε ότι το μεγαλύτερο τμήμα του πληθυσμού γνωρίζει τι είναι και τι όχι επιβλαβές για την διατήρηση της σωματικής υγείας του ανθρώπου. Ωστόσο συχνά υπάρχουν καταχρήσεις που ξεπερνούν το μέτρο. Στην έρευνα επιλέχτηκαν να αναλυθούν 3 χώρες: η Ελλάδα, η Γαλλία και η Αγγλία.Το επίπεδο ζωής αυτών των χωρών σύμφωνα με των αλγόριθμο και των δεδομένων που αντλήσαμε από αυτόν διαφέρει σημαντικά. Επίσης ο αριθμός δεδομένων αναφορικά με τις ανάρτησης στην κοινωνικό διαδίκτυο ποικίλει αισθητά.

Συγκεκριμένα για την Ελλάδα αντλήσαμε ένα πολύ μικρό αριθμό δεδομένων τα οποία μάλιστα θελήσαμε να τα προσεγγίσουμε και σε ρεαλιστική προσομοίωση με την εκπόνηση ειδικού ερωτηματολόγιου που αφορά τον τρόπο ζωής των Ελλήνων. Τα δεδομένα που αντλήσαμε από το ερωτηματολόγιο, έχουν δοθεί και αναλυθεί σε προηγούμενο κεφάλαιο και ουσιαστικά δηλώνουν ότι μεγαλύτερο ποσοστό τον Ελλήνων γνωρίζουν τι είναι ανθυγιεινό και τι όχι, ωστόσο συχνά κάνουν καταχρήσεις αλκοόλ ειδικά στις ηλικίες 25-35 ενώ το μεγαλύτερο ποσοστό αυτών είναι καπνιστές. Όμως οι περισσότεροι αθλούνται τακτικά και αποφεύγουν την κατάχρηση γλυκών και junk food.Όπως μπορούμε να δούμε και τα γενικά αποτελέσματα του αλγορίθμου και για τις 3 χώρες, δείχνουν πως τα ποσοστά των tweets που θεωρούνται υγιεινά είναι μεγαλύτερα από αυτά τον μη υγιεινών.

8. Βιβλιογραφία

1. Big Data, Urbandictionary. [Internet]. [Τελευταία πρόσβαση 2015 Ιούνιος 15]. Διαθέσιμο στο URL: http://www.urbandictionary.com/define.php?term=Big%20Data

2. Elliott, T. 7 Definitions of big data you should know about. [Internet]. 2015 Ιούλιος 5. [Τελευταία πρόσβαση 2015 Νοέμβριος 22]. Διαθέσιμο στο URL: http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html

3. Gasper, T. Big Data Right Now: Five Trendy Open Source Technologies. TechCrunch. [Internet]. 2012 Οκτώβριος. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://techcrunch.com/2012/10/27/big-data-right-now-five-trendy-open-source-technologies/

4. Hopkins B, Evelson B. Big Data Brewer and a Couple of Webinars. Forrester. [Internet]. 2015 Ιούνιος 29. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://blogs.forrester.com/brian_hopkins/11-08-29 big_data_brewer_and_a_couple_of_webinars

5. Mauritz, P. Mauritz: Pivotal Platform Will Sidestep Amazon Tax For Big Data. Αναφορά σε άρθρο του Burke S, στο CRN. [Internet].[Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL:http://www.crn.com/news/applications-os/240152130/maritz-pivotal-platform-will-sidestep-amazon-tax-for-big-data-apps.htm?pgno=2

6. Γλωσσολογικές πηγές για τεχνικές εξόρυξης γνώμης (opinion mining) προσαρμοσμένες στις ιδιαιτερότητες της Νέας Ελληνικής [Τελευταία πρόσβαση 2015 Ιουνίου 26]. Διαθέσιμο στο URL: http://nemertes.lis.upatras.gr/jspui/handle/10889/8156

7. Chen H, Chiang R, Storey V. From Big Data to Big Impact. [Internet]. 2012 Δεκέμβριος. MIS Quarterly Vol. 36 No. 4, Page 1171. [Τελευταία πρόσβαση 2014 Αύγουστος 12]. Διαθέσιμο στο URL: http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-Chiang-Storey%20December%202012.pdf

8. Η θεαματική ανάδυση των «μεγάλων δεδομένων» [Τελευταία πρόσβαση 2015 Ιουνίου 16]. Διαθέσιμο στο URL: http://www.saith.gr/2012-10-25-21-03-11/2012-10-30-16-09-22/250-2013-12-28-16-13-23

9. Johnston, L. Defining the big in big data. Library of Congress website. [Internet]. 2012 Μάιος 17. [Τελευταία πρόσβαση 2015 Ιούνιος 12]. Διαθέσιμο στο URL:http://blogs.loc.gov/digitalpreservation/2012/05/defining-the-big-in-big-data/

10. Gualtieri, M. The pragmatic definition of big data. Forrester Research Blog. [Internet]. 2012 Δεκέμβριος 5. [Τελευταία πρόσβαση 2015 Ιούνιος 12]. Διαθέσιμο στο URL: http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data

11. Anssen, C. Big Data. Techopedia. [Internet]. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www.techopedia.com/definition/27745/big-data

12. Howie,T. The Big Bang: How the Big Data Explosion Is Changing the World. TheMicrosoft Enterprise Insight Blog. [Internet]. 2013 Απρίλιος 15. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world.aspx

13. Lohr, S. The Age of Big Data. The New York Times. [Internet]. 2012 Φεβρουάριος 11. [Τελευταία πρόσβαση 2015 Ιούνιος 31]. Διαθέσιμο στο URL: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0

14. Opinion Mining, Sentiment Analysis, and Opinion Spam Detection [Τελευταία πρόσβαση 2015 Ιουνίου 23]. Διαθέσιμο στο URL: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

15. Opinion Mining and Sentiment Analysis [Τελευταία πρόσβαση 2015 Ιουνίου 24]. Διαθέσιμο στο URL: https://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf

16. Goebel, Michael; Gruenwald, Le (1999); A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, Vol. 1, Issue 1, pp. 20–33

17. Dumbill, E. What is big data? O'Reilly Radar. [Internet]. 2012 Ιανουάριος 11. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://radar.oreilly.com/2012/01/what-is-big-data.html

18. Laney, D. 3-D Data Management: Controlling Data Volume, Velocity and Variety. [Internet]. ADS 6 Feb 01.949 Addendum. [Τελευταία πρόσβαση 2015 Ιούνιος 25]. Διαθέσιμο στο URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

19. Big Data at the Speed of Business. IBM. [Internet]. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www-01.ibm.com/software/data/bigdata/

20. Big Data Now: 2012 Edition [Τελευταία πρόσβαση 2015 Ιουνίου 06]. Διαθέσιμο στο URL: http://www.merlot.org/merlot/viewMaterial.htm?id=801537&hitlist=keywords%3DBig%2BData%2BNow&fromUnified=true

21. ECM [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/ECM

22. Enterprise resource planning [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Enterprise_resource_planning

23. EAI (enterprise application integration) definition [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://searchsoa.techtarget.com/definition/EAI

24. EAI [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.webopedia.com/TERM/E/EAI.html

25. Big Data analytics [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.techopedia.com/definition/28659/big-data-analytics/

26. "Continuuity Raises $10 Million Series A Round to Ignite Big Data Application Development Within the Hadoop Ecosystem". finance.yahoo.com. Marketwired. 2012-11-14. Retrieved 2015 -10-30.

27. Russom, Ph. Big Data Analytics. [Internet]. TDWI Research, Executive Summary; Fourth Quarter 2011. [Τελευταία πρόσβαση 2015 Ιούλιος 3]. BIG DATA ANALYTICS - TDWI

28. O' Reilly. Big Data Now. Kindle Books Edition. [Internet]. 2012. [Τελευταία πρόσβαση 2015 Ιούνιος 24]. Διαθέσιμο στο URL: http://www.merlot.org/merlot/viewMaterial.htm?id=801537&hitlist=keywords%3DBig%2BData%2BNow&fromUnified=true

29. Hendrix, M. Big Data in the Time of Cholera. U.S. Chamber of Commerce Foundation. [Internet]. 2015 Σεπτέμβριος 26. [Τελευταία πρόσβαση 2014 Νοέμβριος Διαθέσιμο στο URL: http://www.builtinla.com/blog/significant-benefits-big-data-analytics-healthcare-industry

30. Definition of: Big Data [Τελευταία πρόσβαση 2015 Ιουνίου 16]. Διαθέσιμο στο URL:http://www.pcmag.com/encyclopedia/term/62849/big-data

31. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

32. Hamilton, B. Big Data is the Future of Healthcare. Cognizant. [Internet]. [Τελευταία πρόσβαση 2014 Αύγουστος 20]. Διαθέσιμο στο URL: http://www.cognizant.com/InsightsWhitepapers/Big-Data-is-the-Future-of-Healthcare.pdf

33. Stiglich P, Rajagopal H. Big Data in Healthcare. Perficient. [An Interview on the Internet]. 2013 Ιανουάριος 3. [Τελευταία πρόσβαση 2014 Αύγουστος 8]. Διαθέσιμο στο URL:http://www.perficient.com/Thought-Leadership/Perficient-Perspectives/2013/Big-Data-in-Healthcare

34. Big Data is the Future of Healthcare [Τελευταία πρόσβαση 2015 Ιουλίου 13]. Διαθέσιμο στο URL: http://www.cognizant.com/InsightsWhitepapers/Big-Data-is-the-Future-of-Healthcare.pdf

35. Big Data In Healthcare [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.slideshare.net/sanderklous/big-data-in-healthcare

36. Τσαγκαράκης, Π. Η μεγάλη αξία των μεγάλων δεδομένων. Reporter. [Internet]. 2013 Ιανουάριος 9. [Τελευταία πρόσβαση 2015 Ιούνιος 20]. Διαθέσιμο στο URL: http://www.reporter.gr/Apopseis/MarketingBrowser/Panos-Tsagkarakhs/item/215774-H-megalh-axia-twn-megalwn-dedomenwn-(Big-Data)

37. Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); A Data Mining & Knowledge Discovery Process Model. In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438–453, February 2009, I-Tech, Vienna, Austria.

38. Yan, J. Big Data, Bigger Opportunities. [Internet]. 2013 Απρίλιος 9. [Τελευταία πρόσβαση 2015 Ιούνιος 10]. Διαθέσιμο στο URL: http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

39. Hutchins, R. Perspective: Looking Forward to Life with Big Data. Emcien. [Internet]. 2015 Ιανουάριος 2. [Τελευταία πρόσβαση 2015 Ιούνιος 22]. Διαθέσιμο στο URL: http://emcien.com/perspective-looking-forward-life-big-data/

40. Allouche, G. Can big data save healthcare? Techopedia. [Internet]. 2013 Δεκέμβριος 13. [Τελευταία πρόσβαση 2014 Νοέμβριος 20]. Διαθέσιμο στο URL: http://www.techopedia.com/2/29792/trends/big-data/can-big-data-save-health-care

41. Murdoch T, Detsky A. The inevitable application of big data to health care. JAMA. [Internet]. April 3, 2013, Vol 309, No. 13. [Τελευταία πρόσβαση 2015 Αύγουστος 20]. Διαθέσιμο στο URL: http://jama.jamanetwork.com/article.aspx?articleid=1674245

42. Gartner. Roundup of Big Data Forecasts and Market Estimates. Αναφορά σε άρθρο του Columbus L. στο Forbes. [Internet]. 2012 Ιούνιος 26. [Τελευταία πρόσβαση 2015 Ιούνιος 20]. Διαθέσιμο στο URL: http://www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/

43. Wu Z, Chin O.B, From Big Data to Data Science: A Multi-disciplinary Perspective. [Internet]. Big Data Research, Volume 1, 2012 Ιούνιος; doi:10.1016/j.bdr. 2015.08.002. [Τελευταία πρόσβαση 2015 Ιούνιος 01]. Διαθέσιμο στο URL: http://www.sciencedirect.com/science/article/pii/S2214579614000082

44. Groves P, Kayyali B, Knott D, Van Kuiken S. The 'big data' revolution in healthcare. McKinsey&Company. [Internet]. 2015 January. [Τελευταία πρόσβαση 2015 Ιουνίου 10]. Διαθέσιμο στο URL: http://www.search.ask.com/web?o=APN10386&gct=SB&q=The%20big%20data%20revolution%20in%20healthcare

45. HIMSS Health Information Exchange Wiki [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://himsshie.pbworks.com/w/page/537838/FrontPage

46. Big data save health care [Τελευταία πρόσβαση 2015 Ιουλίου 15]. Διαθέσιμο στο URL: http://www.techopedia.com/2/29792/trends/big-data/can-big-data-save-health-care

47. HIMSS Part of NHIT Collaborative for Underserved Challenge to Develop Precision Medicine Tools [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.himss.org/

48. HIMS [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.himss.eu/

49. HIMS [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://ww2.frost.com/

50. HIMSS Legacy Workgroup (January 2013), History of the Healthcare Information and Management Systems Society (Formerly Hospital Management Systems Society) (PDF)

51. Jodock, Pam (17 March 2015). "Sun Sets on Record-setting HIMSS14". HIMSS.

52. "HIMSS Events (Local chapter)". HIMSS. Retrieved 22 May 2015.

53. Health information management [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Health_information_management

54. Electronic Medical Record Adoption Model (EMRAM) [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.himssanalytics.eu/emram

55. Healthcare Social Media Analytics. [Internet]. 2013 Σεπτέμβριος 26. [Τελευταία πρόσβαση 2015 Ιανουάριο Διαθέσιμο στο URL: http://www.symplur.com/healthcare-social-media-analytics/

56. Adler-Milstein J, Jha A, K. JAMA Forum: A Strong Start for Electronic Health Records in the United States. [Internet]. 2013 Ιούνιος 2. [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: http://newsatjama.jama.com/2013/08/02/jama-forum-a-strong-start-for-electronic-health-records-in-the-united-states/

57. Social Media Use in the United States: Implications for Health Communication [Internet]. 2012 Μάρτιος 10. [Τελευταία πρόσβαση 2016 Ιανιυάριος]. Διαθέσιμο στο URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2802563/

58. The Evolution of Social Media in Healthcare. [Internet]. 2013 Σεπτέμβριος 26. [Τελευταία πρόσβαση 2016 Ιανουάριο Διαθέσιμο στο URL: http://www.perficient.com/Thought-Leadership/Perficient-Perspectives/2015/The-Evolution-of-Social-Media-in-Healthcare

59. Healthcare-hashtags-social project. [Internet]. 2013 Σεπτέμβριος 26. [Τελευταία πρόσβαση 2016 Ιανουάριο Διαθέσιμο στο URL: http://www.symplur.com/blog/healthcare-hashtags-social-project/

60. Han, Jiawei; Kamber, Micheline (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long

61. Goebel, Michael; Gruenwald, Le (1999); A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, Vol. 1, Issue 1, pp. 20–33

62. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

63. Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 Ed.). Elsevier. ISBN 978-0-12-374856-0.

64. Classification [Τελευταία πρόσβαση 2015 Ιουνίου 03]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Classification

65. Regression_analysis [Τελευταία πρόσβαση 2015 Ιουνίου 06]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Regression_analysis

66. Time_series [Τελευταία πρόσβαση 2015 Ιουνίου 06]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Time_series

67. Prediction [Τελευταία πρόσβαση 2015 Ιουνίου 06]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Prediction

68. Cluster analysis [Τελευταία πρόσβαση 2015 Ιουνίου 06]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Cluster_analysis

69. Chapter 3. Typical Hadoop Cluster [Τελευταία πρόσβαση 2015 Ιουνίου 25]. Διαθέσιμο στο URL: http://docs.hortonworks.com/HDPDocuments/HDP1/HDP-1.3.2/bk_getting-started-guide/content/ch_hdp1_getting_started_chp3.html

70. Welcome to Apache™ Hadoop®! [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://hadoop.apache.org/

71. Hadoop and Big Data. Cloudera. [Τελευταία πρόσβαση 2015 Ιούνιος 12]. Διαθέσιμο στο URL: http://cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html

72. Reliable, economical cloud storage for data big and small Managed Apache Hadoop, Spark, HBase, and Storm made easy [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: http://azure.microsoft.com/en-us/services/hdinsight/

73. An introduction to the Hadoop Distributed File System [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: http://www.ibm.com/developerworks/library/wa-introhdfs/

74. The Streaming APIs Τελευταία πρόσβαση 2015 Φεβρουάριος 30. Διαθέσιμο στο URL: https://dev.twitter.com/streaming/overview

75. "Hadoop-related projects at". Hadoop.apache.org. Retrieved 2013-10-17.

76. The Hadoop Distributed File System [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://www.aosabook.org/en/hdfs.html

77. "What is the Hadoop Distributed File System (HDFS)?" Ibm.com. IBM. Retrieved 2015 -10-30.

78. "Hadoop Releases". Apache.org. Apache Software Foundation. Retrieved 2015 -12-06.

79. "Hadoop Releases". Hadoop.apache.org. Retrieved 2015-07-29.

80. Apache Hadoop [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html

81. Malak, Michael (2015 -09-19). "Data Locality: HPC vs. Hadoop vs. Spark". datascienceassn.org. Data Science Association. Retrieved 2015 -10-30.

82. "Resource (Apache Hadoop Main 2.5.1 API)". apache.org. Apache Software Foundation. 2015 -09-12. Retrieved 2015 -09-30.

83. Hadoop Toolbox: When to Use what [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: http://www.smartdatacollective.com/mtariq/120791/hadoop-toolbox-when-use-what

84. "Michael J. Cafarella". Web.eecs.umich.edu. Retrieved 2013-04-05.

85. Vance, Ashlee (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times. Archived from the original on 11 February 2010. Retrieved 2010-01-20.

86. Adventures with Hadoop and Perl". Mail-archive.com. 2010-05-02. Retrieved 2013-04-05.

87. Hadoop: Open Insight Anywhere [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL:] http://www-01.ibm.com/software/data/infosphere/hadoop/

88. Bhattacherjee A, Hikmet N. Physician's resistance toward healthcare information technology: a theoretical model and empirical test. European Journal of Information Systems. [Internet]. 2007; 16:725-737. [Τελευταία πρόσβαση 2015 Ιουλίου 30]. Διαθέσιμο στο URL: http://www.researchgate.net/publication/211382563_Physicians%27_resistance_toward_healthcare_information_technology_a_theoretical_model_and_empirical_test

89. YARN [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

90. Apache HBase [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Apache_HBase

91. Murthy, Arun (2012-08-15). "Apache Hadoop YARN – Concepts and Applications". hortonworks.com. Hortonworks. Retrieved 2015 -09-30.

92. ASP.NET MVC Overview [Τελευταία πρόσβαση 2015 Ιούνιος 10]. Διαθέσιμο στο URL: https://msdn.microsoft.com/en-us/library/dd381412%28v=vs.108%29.aspx

93. Bing Maps V7 Modules [Τελευταία πρόσβαση 20115 Ιουνίου 10]. Διαθέσιμο στο URL: http://bingmapsv7modules.codeplex.com/wikipage?title=Client%20Side%20Heatmap

94. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons. 2015 -12-19. p. 300. ISBN 9781118876220. Retrieved 2015-01-29.

95. Ovadia, St. 2013, The Role of Big Data in the Social Sciences. *Behavioral & Social Sciences Librarian*. [Internet]. 2013. [Τελευταία πρόσβαση 2015 Ιούλιος 18].Διαθέσιμο στο URL: http://eric.ed.gov/?q=Educate+physiacians+in+the+use+of+big+data&id=EJ1005014

96. Yan, J. Big Data, Bigger Opportunities. [Internet]. 2013 April 9. [Τελευταία πρόσβαση 2015 Ιούλιος 4]. Διαθέσιμο στο URL: http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

97. Pcmag encyclopedia. [Internet]. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www.pcmag.com/encyclopedia/term/62849/big-data

98. Rauser, J. Defining big data depends on who's doing the defining. Network World. [Internet]. 2012 Μάρτιος 10. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www.networkworld.com/article/2188435/data-center/defining--big-data--depends-on-who-s-doing-the-defining.html

99. Loukides, M. What is data science. O'Reilly Radar. [Internet]. 2010 Ιούνιος 2. [Τελευταία πρόσβαση 2015 Μάρτιος 25]. Διαθέσιμο στο URL: http://radar.oreilly.com/2010/06/what-is-data-science.html

100. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. [Internet]. 2011 Μάιος. [Τελευταία πρόσβαση 2015 Μάρτιος 12]. Διαθέσιμο στο URL: http://www.academia.edu/5740927/McKinsey_Global_Institute_The_McKinsey_Global_Institute

101. Brust, A. Big Data: Defining it definition. ZDNet Blog. [Internet].[Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www.zdnet.com/article/big-data-defining-its-definition/

102. Defining Big Data. FCW Blog. [Internet].. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL:http://fcw.com/blogs/conversation/2013/04/defining-big-data.aspx

103. Ebbert, J. Define It – What Is Big Data? Adexchanger. [Internet].[Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL:http://www.adexchanger.com/online-advertising/big-data/

104. Gourley, B. Demystifying Big Data Industry Leaders Release Comprehensive Report on Big Data in Government. The TechAmerica Foundation's Federal Big Data Commission Comprehensive Guide to Best Practices for Big Data. [Internet]. 2015 Οκτώβριος. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://www.techamericafoundation.org/demystifying-big-data-industry-leaders-release-comprehensive-report-on-big-data-in-government

105. Johnston, L. Data is The New Black. Library of Congress. [Internet]. 2015 Οκτώβριος 14. [Τελευταία πρόσβαση 2015 Ιούλιος 25]. Διαθέσιμο στο URL: http://blogs.loc.gov/digitalpreservation/2011/10/data-is-the-new-black/

106. Weathington, J. Big Data Defined. TechRepublic Blog. [Internet]. 2012 Σεπτέμβριος 3. [Τελευταία πρόσβαση 2015 Ιούνιος 25]. Διαθέσιμο στο URL:http://www.techrepublic.com/blog/big-data-analytics/big-data-defined/

107. Why should I use Hadoop instead of Microsoft SQL? [Τελευταία πρόσβαση 2015 Ιανουάριος 30]. Διαθέσιμο στο URL: http://www.idganswers.com/question/2388/why-should-i-use-hadoop-instead-of-microsoft-sql

108. Blobs, Tables, Queues, and Files [Τελευταία πρόσβαση 2015 Φεβρουάριος 30]. Διαθέσιμο στο URL: http://azure.microsoft.com/en-us/services/storage/

109. Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web. [Τελευταία πρόσβαση 2015 Ιουνίου 23]. Διαθέσιμο στο URL: http://getbootstrap.com/

110. Snowflake [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://snowflake.net/

111. Business activity monitoring (BAM) definition [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: http://searchcio.techtarget.com/definition/business-activity-monitoring-BAM

112. Healthcare Information and Management Systems Society [Τελευταία πρόσβαση 2015 Ιουλίου 23]. Διαθέσιμο στο URL: https://en.wikipedia.org/wiki/Healthcare_Information_and_Management_Systems_Society

113. Haughton, Dominique; Deichmann, Joel; Eshghi, Abdolreza; Sayek, Selin; Teebagy, Nicholas; and Topi, Heikki (2003); A Review of Software Packages for Data Mining, The American Statistician, Vol. 57, No. 4, pp. 290–309

114. "National Health IT Week". Healthcare Information and Management Systems Society. Retrieved 22 May 2015.