ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Βιοιατρικής

ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ BIG DATA ΣΤΗΝ ΨΗΦΙΑΚΗ ΕΠΙΔΗΜΙΟΛΟΓΙΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΑΝΑΣΙΟΣ Ζ. ΓΡΙΒΑΣ

Επιβλέπων: Δημήτριος Κουτσούρης, Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου, ΕΔΙΠ Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Βιοιατρικής

ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ BIG DATA ΣΤΗΝ ΨΗΦΙΑΚΗ ΕΠΙΔΗΜΙΟΛΟΓΙΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΑΝΑΣΙΟΣ Ζ. ΓΡΙΒΑΣ

Επιβλέπων: Δημήτριος Κουτσούρης, Καθηγητής Ε.Μ.Π.

Συνεπιβλέπουσα: Ουρανία Πετροπούλου, ΕΔΙΠ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2η Σεπτεμβρίου 2020

.....	.....	.....
Δ. Κουτσούρης	Γ. Ματσόπουλος	Π. Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

.....

Αθανάσιος Ζ. Γρίβας Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Επιδημιολογία ονομάζεται η επιστήμη που μελετά τη φύση και τους μηχανισμούς εξάπλωσης των μεταδοτικών ασθενειών. Ένας αρκετά νέος τομέας της επιδημιολογίας είναι η ψηφιακή επιδημιολογία. Στην ψηφιακή επιδημιολογία χρησιμοποιούνται τεχνικές επεξεργασίας και ανάλυσης διαδικτυακών δεδομένων με στόχο τον εντοπισμό και την παρακολούθηση ξεσπασμάτων μεταδοτικών ασθενειών. Τα διαδικτυακά δεδομένα, που χρησιμοποιούνται στην ψηφιακή επιδημιολογία, ονομάζονται και ‘’μεγάλα’’ δεδομένα και διακρίνονται για το πολύ μεγάλο μέγεθος, πολυπλοκότητα και ανομοιομορφία που τα χαρακτηρίζει. Η επεξεργασία των ‘’μεγάλων’’ δεδομένων δεν είναι εύκολη διαδικασία και κατά κύριο λόγο για την ανάλυσή τους χρησιμοποιούνται τεχνολογίες όπως η μηχανική μάθηση, η επεξεργασία φυσικής γλώσσας και η τεχνητή νοημοσύνη. Επίσης, για την επεξεργασία και ανάλυση των μεγάλων δεδομένων απαιτούνται ισχυρά συστήματα με μεγάλη ταχύτητα επεξεργασίας, μνήμη και χώρο αποθήκευσης. Καθώς το μέγεθος των ‘’μεγάλων’’ δεδομένων αυξάνεται τα συστήματα αυτά πρέπει να εξελίσσονται συνεχώς, όμως ο σχεδιασμός και η δημιουργία τέτοιου είδους συστημάτων έχει ιδιαίτερα υψηλό κόστος. Γι’ αυτό το λόγο έχουν εφευρεθεί τεχνικές κλιμάκωσης, με χρήση των οποίων οι δυνατότητες ενός ήδη υπάρχοντος συστήματος επεκτείνονται. Η κλιμάκωση διακρίνεται στην οριζόντια κλιμάκωση και στη κάθετη κλιμάκωση, αλλά για τον σκοπό της επεξεργασίας ‘’μεγάλων’’ δεδομένων χρησιμοποιείται κατά κύριο λόγο η οριζόντια κλιμάκωση. Κατά την οριζόντια κλιμάκωση μεμονωμένα συστήματα διασυνδέονται μεταξύ τους και έτσι η συνολική επεξεργαστική ισχύς αυξάνεται. Υπάρχουν αρκετές πλατφόρμες οριζόντιας κλιμάκωσης συστημάτων οι οποίες αναλαμβάνουν τη διασύνδεση των συστημάτων και, στη συνέχεια, την επεξεργασία των ‘’μεγάλων’’ δεδομένων, κάποιες από τις οποίες είναι το Hadoop, το Storm και το Spark της Apache. Μέχρι σήμερα έχουν δημιουργηθεί αρκετά συστήματα στον τομέα της ψηφιακής επιδημιολογίας τα οποία συμβάλουν στον εντοπισμό νέων ξεσπασμάτων ασθενειών και στη μελέτη της πορείας εξάπλωσής τους. Κάποια από αυτά τα συστήματα ανήκουν σε οργανισμούς υγείας, ενώ κάποια άλλα έχουν δημιουργηθεί από εταιρείες. Μερικά πολύ δημοφιλή τέτοιου είδους συστήματα αποτελούν το HealthMap, το GPHIN, το ProMED-mail και το σύστημα AI της BlueDot. Επίσης, η πανδημία COVID-19, που πλήττει σήμερα το μεγαλύτερο μέρος του κόσμου, ήταν αφορμή για τη δημιουργία αρκετών λογισμικών για την απεικόνιση των κρουσμάτων και των θανάτων εξ’ αιτίας του κορονοίού. Τα λογισμικά αυτά έχουν καθαρά στόχο την ενημέρωση του πληθυσμού και περιλαμβάνουν διαδραστικούς χάρτες και γραφικές παραστάσεις για την πιο λεπτομερή απεικόνιση των δεδομένων που παρέχουν. Το πρώτο από αυτά τα συστήματα που δημιουργήθηκε είναι το COVID-19 Dashboard του Παγκοσμίου Οργανισμού Υγείας (WHO), ενώ ακολούθησαν και άλλα λογισμικά όπως το COVID-19 Dashboard του Πανεπιστημίου Johns Hopkins (JHU). Η ψηφιακή επιδημιολογία είναι ένα πεδίο που συνεχώς εξελίσσεται παράλληλα με την εξέλιξη της τεχνολογίας και όσο βελτιώνεται αυξάνονται και τα ευεργετικά αποτελέσματα στην ποιότητα ζωής των ανθρώπων. Απαραίτητη, όμως, είναι η προσοχή κατά τη λήψη πληροφοριών από ηλεκτρονικά δεδομένα, καθώς η δημοσιοποίηση εσφαλμένων πληροφοριών και η παραβίαση του προσωπικού απορρήτου μπορεί να έχει πολύ δυσάρεστες συνέπειες.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Ψηφιακή Επιδημιολογία, Επεξεργασία Δεδομένων, ‘’Μεγάλα’’ Δεδομένα, Πλατφόρμες Επεξεργασίας, Πλατφόρμες Κλιμάκωσης, Ηλεκτρονικές Εφαρμογές, Τεχνητή Νοημοσύνη, Μεταδοτικές Ασθένειες

ABSTRACT

Epidemiology is the science of studying the nature and the spread mechanisms of communicable diseases. A fairly new field of epidemiology is digital epidemiology. In digital epidemiology, methods of internet data processing are applied for the purpose of detecting and tracking communicable disease outbreaks. The internet data that are used in digital epidemiology are also called big data and are distinguished by the great volume, complexity and the lack of uniformity that describes them. Big data processing is not a simple task and, for that purpose, technologies like machine learning, natural language processing and artificial intelligence are used. Also, for the purpose of big data processing, powerful computers with high processing speed, memory and storage are needed. As the size of big data is increasing, the processing power of the computing systems must increase too, but designing and building such systems has a high cost. In order to increase the systems power and keep the cost low, scaling techniques are used, and by using these techniques the processing power of an existing system can be increased. There are two types of scaling, horizontal scaling and vertical scaling. For the purposes of digital epidemiology horizontal scaling is used, in which individual systems can be interconnected and so the overall processing power is increased. There are several horizontal scaling platforms that are used both for the interconnection of the systems and for the big data processing, some of which are the Apache's Haddop, Storm and Spark. In the last few years several systems have been designed in the field of digital epidemiology, which contribute in the detection of new disease outbreaks and in the study of the way they spread. A number of this systems belong to health organizations and others are designed by private companies. A few popular systems in this field are HealthMap, GPHIN, ProMED-mail and BlueDot's AI system. Also, the COVID-19 pandemic, which today has spread in almost every part of the world, triggered the design of new software platforms which provide information on the cases and deaths due to coronavirus. These software platforms are built with the sole purpose of informing the population, and for the more detailed projection of information they feature interactive maps and graphs. The first of these systems that came online was World Health Organization's COVID-19 Dashboard, while other software followed later like Johns Hopkins University's COVID-19 Dashboard. Digital epidemiology is a field that continues to evolve and this evolution comes with positive effects in people's quality of life. Though, in the process of extracting information from big data, it is necessary to proceed with care, because the publication of false information and the violation of citizens' privacy may have very disturbing consequences.

KEYWORDS

Digital Epidemiology, Data Processing, Big Data, Processing Platforms, Scaling Platforms, Digital Applications, Artificial Intelligence , Communicable Diseases

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

1.1 Ορισμός επιδημιολογίας

Επιδημιολογία ονομάζεται η επιστήμη που μελετά και αναλύει την κατανομή, την πορεία εξάπλωσης και τη συχνότητα εμφάνισης ασθενειών στις διάφορες ομάδες του ανθρώπινου πληθυσμού καθώς και των παραγόντων που τις διαμορφώνουν ή μπορούν να τις επηρεάσουν. Στόχος της επιδημιολογίας , μέσω της διερεύνησης των αιτιών και των επιπτώσεων της εξάπλωσης των διαφόρων νοσημάτων, είναι να δημιουργηθούν πιο αποτελεσματικές διαγνωστικές μέθοδοι, πιο κατάλληλες θεραπείες για τους ήδη νοσούντες και πιο αξιόπιστες προγνώσεις για τη διάδοση των ασθενειών. [1]

1.2 Ιστορική αναδρομή

Ο όρος επιδημία, από ιατρικής πλευράς, διατυπώθηκε για πρώτη φορά από τον αρχαίο Έλληνα ιατρό Ιπποκράτη (460 - 375 πΧ) ο οποίος όρισε την έννοια και τη σημασία του, και αποτέλεσε έναν από τους πρώτους κλάδους της ιατρικής. Ο Ιπποκράτης μετά από παρατηρήσεις πολλών ετών χώρισε τις ασθένειες σε τέσσερις κατηγορίες, τις οξείες, τις χρόνιες, τις ενδημικές και τις επιδημικές. Ενδημικές ονόμασε τις ασθένειες που παρουσιάζονται μόνιμα ή συχνά σε έναν συγκεκριμένο τόπο ή ασθενή, ενώ επιδημικές τις νόσους που εμφανίζονται σπάνια, σε διαφορετικές τοποθεσίες και σε βαθμό μεγαλύτερο του αναμενόμενου.

Σχεδόν 2000 χρόνια αργότερα, στα μέσα του 16^{ου} αιώνα, ο Ιταλός ιατρός, ποιητής, αστρονόμος και γεωλόγος από τη Βερόνα Girolamo Fracastoro (1748-1553) ήταν ο πρώτος που πρότεινε μια θεωρία ότι υπάρχουν κάποια απειροελάχιστα μικρά αόρατα έμβια σωματίδια τα οποία αποτελούν την αιτία όλων των ασθενειών. Όπως υποστήριξε τα σωματίδια αυτά μεταδίδονται μέσω του αέρα, πολλαπλασιάζονται μόνα τους και μπορούν να καταστραφούν με τη χρήση φωτιάς. Η θεωρία αυτή επιβεβαιώθηκε το 1675 από τον Antoine van Leeuwenhoek (1632-1723), έναν Ολλανδό έμπορο και επιστήμονα, ο οποίος κατασκεύασε το πρώτο μικροσκόπιο με αρκετά μεγάλη μεγέθυνση ώστε να μπορεί κάποιος να παρατηρήσει μικροοργανισμούς και κύτταρα.

Μία πολύ μεγάλη συνεισφορά στην επιδημιολογία ήταν αυτή του John Snow (1813-1858) που θεωρείται και πατέρας της σύγχρονης επιδημιολογίας. Ο Snow ζούσε τον 19^{ο} αιώνα στο Λονδίνο την εποχή που η πόλη μαστίζονταν από μία επιδημία χολέρας. Μετά από έρευνα παρατήρησε ότι τα κρούσματα ήταν αυξημένα στις περιοχές που παρείχε νερό η εταιρία Southwark Company, στη συνέχεια εντόπισε την μολυσμένη αντλία νερού και απολυμαίνοντας το νερό με χλωρίνη έδωσε τέλος στην επιδημία.

Εικόνα 1.1: Μία παραλλαγή του αρχικού χάρτη μιας περιοχής του Λονδίνου, που σχεδιάστηκε από τον John Snow, που δείχνει τα αυξημένα κρούσματα χολέρας γύρω από τα σημεία παροχής νερού. [2]

Από τις αρχές του 20ου αιώνα άρχισε η εφαρμογή μαθηματικών μεθόδων στο πεδίο της επιδημιολογίας από επιστήμονες όπως ο Ronald Ross (1857-1932) και ο Anderson Gray McKendrick (1876-1943) . Η νέα αυτή προσέγγιση οδήγησε σε ριζικές αλλαγές στον τρόπο που διεξαγόταν μια επιδημιολογική έρευνα μέχρι τότε και στη δημιουργία νέων επαναστατικών για την εποχή μεθόδων που, κάποιες από αυτές, εφαρμόζονται ακόμα και σήμερα. [3][4][2]

1.3 Η επιδημιολογία σήμερα

Στη σημερινή εποχή η επιδημιολογία, που καλύτερα να διαχωριστεί ως κλινική επιδημιολογία, αποτελεί μια ιατρική εξειδίκευση και βασικότατο κομμάτι στην ανάλυση ιατρικής γνώσης. Ως θεμέλιος λίθος της δημόσιας υγείας διαμορφώνει πολιτικές αποφάσεις, επισημαίνοντας τον υψηλό κίνδυνο εξάπλωσης μιας ασθένειας καθώς και τα προληπτικά μέτρα που θα πρέπει να ληφθούν για την προστασία των πολιτών, και κυρίως εκείνων που ανήκουν σε ευπαθείς κοινωνικές ομάδες. Στις μέρες μας όμως οι ασθένειες που οδηγούν σε θάνατο στις ανεπτυγμένες χώρε είναι ως επί το πλείστον μη μεταδοτικές, ενώ ακόμα και στις χαμηλότερου βιοτικού επιπέδου χώρες, οι μεταδοτικές ασθένειες έχουν μειωθεί σε ποσοστό μεγαλύτερο από το ήμισυ. Νέα είδη ασθενειών έχουν προκύψει τα οποία οφείλονται στη συμπεριφορά, τις συνήθειες και την καθημερινή ζωή των ανθρώπων. Τέτοιες ασθένειες είναι είτε ψυχικές, όπως η κατάθλιψη, είτε προκαλούνται από καταχρήσεις, όπως ο τύπου ΙΙ διαβήτης που προέρχεται από υπερβολικό σωματικό λίπος και έλλειψη σωματικής άσκησης και ο καρκίνος του πνεύμονα ο οποίος προέρχεται σε πολλές περιπτώσεις από το κάπνισμα.

Όλα τα παραπάνω έχουν περιορίσει το πεδίο μελέτης της κλινικής επιδημιολογίας και πλέον ο βασικός του στόχος είναι η πρόληψη και η πολύ γρήγορη αντιμετώπιση νέων ασθενειών ή γνωστών ασθενειών με πολύ μεγάλη μεταδοτικότητα που δεν μπορούν να θεραπευθούν. Το βασικό πρόβλημα, όμως, είναι πως για την αποφυγή μελλοντικών ξεσπασμάτων ασθενειών χρειάζεται

αρχικά να γίνει λεπτομερέστερη μελέτη των αιτιών που τις προκαλούν και του τρόπου εξάπλωσης τους , ώστε να γίνει δυνατή η προσομοίωση της μετάδοσης των νοσημάτων και η δημιουργία μοντέλων πρόβλεψης. Επίσης, με την εμφάνιση μιας νέας επιδημίας πρέπει να έχουμε τη δυνατότητα να εντοπίσουμε στο λιγότερο δυνατό χρόνο την προέλευσή της, καθώς και τους παράγοντες που επηρεάζουν τη μετάδοσή της για να την περιορίσουμε άμεσα. Επιπλέον, η κατασκευή ενός εμβολίου ή κάποιου άλλου χρήσιμου φαρμάκου είναι πολύ χρονοβόρες διαδικασίες στις οποίες έστω και ένα μικρό χρονικό προβάδισμα θα είχε σημαντικά αποτελέσματα στην πορεία εξάπλωσης μιας επιδημίας.

Η κλινική επιδημιολογία, σήμερα, βασίζεται στη συλλογή δεδομένων από φορείς δημόσιας υγείας (κυρίως από νοσοκομεία και ιδιωτικά ιατρεία) και από παρατηρήσεις που γίνονται από τους ίδιους τους επιδημιολόγους σε διάφορες περιοχές. Όμως τα δεδομένα αυτά και η ταχύτητα επεξεργασίας αυτών από τους επιδημιολόγους, με χρήση συμβατικών υπολογιστικών συστημάτων, δεν αρκούν για να καλυφθούν όλες οι παραπάνω απαιτήσεις. Η παγκοσμιοποίηση και η μεγάλη αύξηση της ανθρώπινης κοινωνικής και χωρικής κινητικότητας τα τελευταία χρόνια, όπως τα επαγγελματικά ταξίδια και τα ταξίδια αναψυχής, έχουν συνδράμει στην ιδιαίτερα μεγαλύτερη εξάπλωση των επιδημιών. Αυτή η μεγαλύτερη εξάπλωση έχει δημιουργήσει έναν επιπλέον, ιδιαίτερα μεγάλο, όγκο πληροφοριών που πρέπει να διαχειριστούν επίσης οι επιδημιολόγοι. Όλα τα παραπάνω έχουν δημιουργήσει κάποια σημαντικότατα προβλήματα τα οποία στοχεύει να λύσει η ψηφιακή επιδημιολογία. [5][6][7][8]

2 ΨΗΦΙΑΚΗ ΕΠΙΔΗΜΙΟΛΟΓΙΑ

2.1 Εισαγωγή

Η αύξηση του αριθμού των ηλεκτρονικών μέσων μαζικής ενημέρωσης, η μαζική χρήση του διαδικτύου και των ηλεκτρονικών μέσων κοινωνικής δικτύωσης από το μεγαλύτερο μέρος του παγκόσμιου πληθυσμού, καθώς και η ευρεία χρήση των smartphones τα τελευταία χρόνια οδήγησε στη δημιουργία νέων πηγών δεδομένων που δεν είχαμε στη διάθεσή μας παλιότερα. Επίσης νέες αλγοριθμικές τεχνικές επέτρεψαν τη δημιουργία καινούριων εργαλείων για την επεξεργασία δεδομένων όπως η τεχνητή νοημοσύνη, η μηχανική μάθηση και η επεξεργασία φυσικής γλώσσας. Όλα τα παραπάνω συνέβαλλαν στην εμφάνιση ενός νέου κλάδου της επιδημιολογίας που ονομάζεται ψηφιακή επιδημιολογία.

Η ψηφιακή επιδημιολογία ενστερνίζεται τους στόχους της κλινικής επιδημιολογίας αλλά ακολουθεί διαφορετική προσέγγιση ως προς την υλοποίησή τους. Αντί να στηρίζεται μόνο σε δεδομένα που προέρχονται από τον χώρο της υγείας κάνει χρήση αυτών των νέων πηγών δεδομένων. Οι νέες αυτές πηγές δεδομένων, όπως τα ηλεκτρονικά μέσα κοινωνικής δικτύωσης, ονομάζονται και πηγές ‘’μεγάλων’’ δεδομένων (Big data) και χαρακτηρίζονται από πολύ μεγάλους όγκους δεδομένων τα οποία έχουν πολύπλοκη δομή και παρουσιάζουν μεγάλη ετερογένεια. Η μεγάλη πρόκληση που έρχεται να αντιμετωπίσει η ψηφιακή επιδημιολογία είναι η εύρεση των κατάλληλων εργαλείων για την επεξεργασία και την ανάλυση αυτών των δεδομένων, η αποφυγή εσφαλμένων συμπερασμάτων και η εκμαίευση των σωστών πληροφοριών καθώς ένα λάθος αποτέλεσμα μπορεί να έχει ανεπιθύμητες συνέπειες. [8][9]

2.2 Πηγές ‘’μεγάλων’’ δεδομένων (Big data)

Στις μέρες μας ο όρος ‘’μεγάλα’’ δεδομένα έχει διττή σημασία που άλλοτε αφορά τα ίδια τα δεδομένα και τις πηγές από τις οποίες προέρχονται, ενώ κάποιες φορές αναφέρεται στις μεθόδους επεξεργασίας αυτών των δεδομένων. Στην παρούσα ανάλυση όταν χρησιμοποιούμε τον όρο ‘’μεγάλα’’ δεδομένα θα αναφερόμαστε στα ίδια τα δεδομένα που παράγονται από την καθημερινή ζωή των ανθρώπων που χρησιμοποιούν το διαδίκτυο για διάφορους λόγους. Αυτοί οι λόγοι περιλαμβάνουν αγορές ή πωλήσεις, μεταφορά χρημάτων προς ή από κάποιον τραπεζικό λογαριασμό, χρήση ηλεκτρονικών υπηρεσιών κτλ.

Η ιδέα της αξιοποίησης αυτού του μεγάλου όγκου αχρησιμοποίητων δεδομένων σε διάφορους επιστημονικούς τομείς υπήρχε από τα τέλη της δεκαετίας του 90’ και ο όρος ‘’μεγάλα’’ δεδομένα είχε χρησιμοποιηθεί από τους Michael Cox και David Ellsworth, δύο επιστήμονες της NASA, το 1997, χωρίς όμως να ορισθεί η σημασία του. Ένα χρόνο αργότερα, το 1998, ο επιστημονικός διευθυντής της εταιρίας Silicon Graphics, John Mashey (1946-), όρισε τη σημασία του όρου και των μεγεθών που τα χαρακτηρίζουν.

Σήμερα υπάρχουν οκτώ χαρακτηριστικές λέξεις που χρησιμοποιούνται για την περιγραφή των μεγάλων δεδομένων που γνωστές ως ‘’*The 8 V’s*’’. Τα *V’s* μπορούν να χωριστούν σε δύο ομάδες από

τις οποίες η πρώτη περιέχει τρία τα οποία είναι τα γενικά χαρακτηριστικά της φύσης των ‘’μεγάλων’’ δεδομένων, ενώ η δεύτερη περιέχει τα χαρακτηριστικά που αποκτούν τα ‘’μεγάλα’’ δεδομένα από τη στιγμή που θα εισέλθουν σε ένα σύστημα. Στη συνέχεια θα αναφέρουμε τα οκτώ *V's* και θα περιγράψουμε τη σημασία τους:

- ***Βασικά χαρακτηριστικά:***

  - *Volume (Όγκος)*: Αναφέρεται στον πολύ μεγάλο όγκο που έχουν τα ‘’μεγάλα’’ δεδομένα. Μια μέτρηση του όγκου αυτού από την IBM το 2016 έδειξε πως ανερχόταν στα 2.5 Exabytes (1 Exabyte = 10^{18} bytes). Έτσι, γίνεται εύκολα κατανοητή η μεγάλη δυσκολία συλλογής και επεξεργασίας ενός τόσο μεγάλου όγκου δεδομένων.

  - *Velocity (Ταχύτητα)*: Αναφέρεται στην ταχύτητα με την οποία παράγονται τα δεδομένα που ανήκουν στην κατηγορία των ‘’μεγάλων’’ δεδομένων από τις διάφορες ηλεκτρονικές πηγές.

  - *Variety (Ποικιλία)*: Αναφέρεται στους διαφορετικούς τύπους των δεδομένων που ανήκουν στην κατηγορία των ‘’μεγάλων δεδομένων’’ (π.χ. εικόνα, ήχος, ηλεκτρονικές αποδείξεις) και στη διαφορετική δομή τους.

- ***Χαρακτηριστικά που αποκτούν μετά τη είσοδό τους σε ένα σύστημα:***

  - *Value (Αξία)*: Αναφέρεται στην αξία που μπορεί να προσφέρει σε διάφορους τομείς η αξιοποίησή τον ‘’μεγάλων’’ δεδομένων.

  - *Veracity (Εγκυρότητα)*: Αναφέρεται στην εγκυρότητα των αποτελεσμάτων που παράγονται από την αξιοποίηση των ‘’μεγάλων’’ δεδομένων και στην αξιοπιστία των δεδομένων αυτών.

  - *Variability (Μεταβλητότητα)*: Αναφέρεται στις διάφορες μορφές που μπορούν να μετατραπούν, διάφορα μοντέλα με τα οποία μπορούν να επεξεργαστούν και τις διάφορες συσχετίσεις που μπορούν να γίνουν μετά την είσοδό τους σε ένα σύστημα.

  - *Virality (Εξαπλωσιμότητα)*: Αναφέρεται στο πόσο γρήγορα μπορούν να εξαπλωθούν μέσω ενός δικτύου σε διάφορους χρήστες.

  - *Viscosity (Ιξώδες)*: Αναφέρεται στο πόση αντίσταση-καθυστέρηση μπορεί να παρατηρηθεί στη ροή-μετάδοση ενός συγκεκριμένου όγκου μεγάλων δεδομένων

Τα ‘’μεγάλα’’ δεδομένα μπορούν επίσης να χαρακτηριστούν από την συνεχώς αυξανόμενη ταχύτητα συλλογής τους και χρησιμοποίησής τους στο μεγαλύτερο μέρος των επιστημονικών ερευνών σήμερα. Είναι πολύ σημαντικό, σε αυτό το σημείο, να τονίσουμε πως το κέρδος της αξιοποίησης των ‘’μεγάλων’’ δεδομένων εξαρτάται απόλυτα από το χρονικό διάστημα που μεσολαβεί από τη δημιουργία τους μέχρι τη στιγμή που η χρησιμοποίησή τους θα έχει δώσει τα επιθυμητά αποτελέσματα. Όσο μικρότερο είναι αυτό το διάστημα τόσο περισσότερο αυξάνεται η

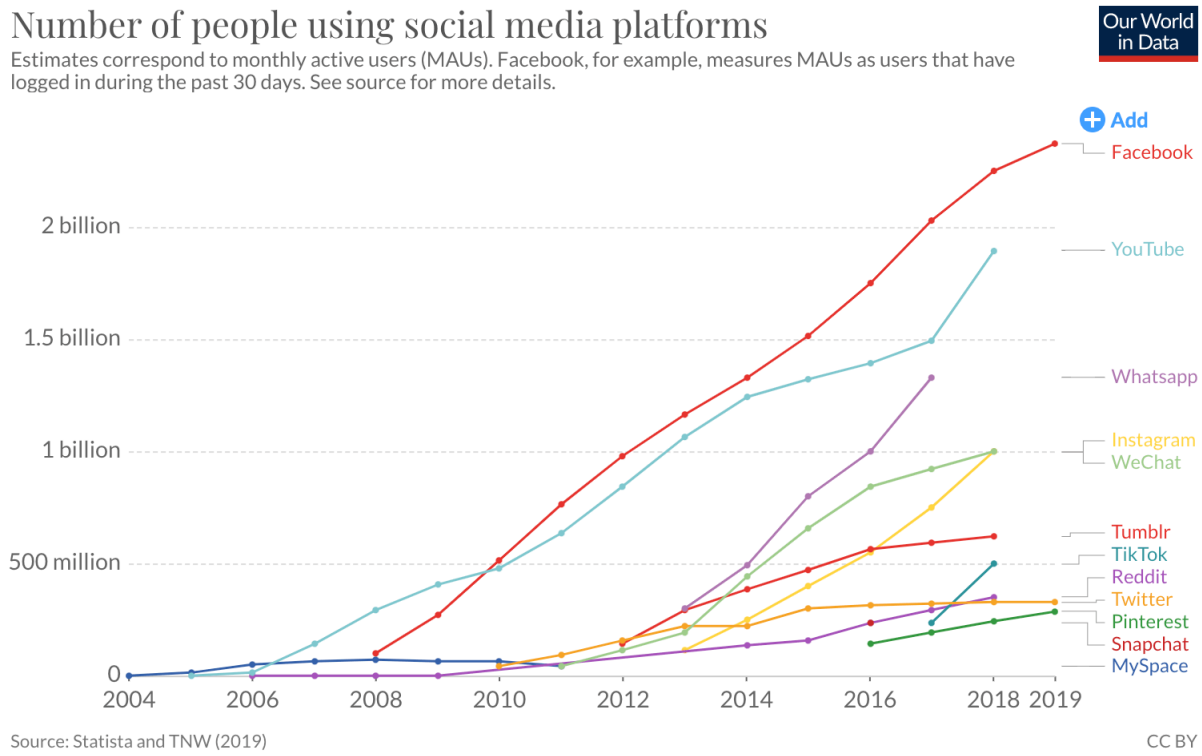
αξία των πληροφοριών που εκμαιεύονται από αυτά. Αυτό ισχύει για όλες σχεδόν τις χρήσεις των ‘’μεγάλων’’ δεδομένων και ακόμα περισσότερο για τον τομέα της ψηφιακής επιδημιολογίας, αφού ο χρόνος από το ξέσπασμα μιας επιδημίας μέχρι τη στιγμή που θα ληφθούν τα πρώτα μέτρα για την αντιμετώπισή της είναι σημαντικότατος στον περιορισμό της. Για την αποτελεσματική μείωση αυτού του χρονικού διαστήματος τα συστήματα που αναλαμβάνουν τη συλλογή και συσχέτιση αυτών των δεδομένων πρέπει να είναι κατάλληλα σχεδιασμένα ώστε να επιτυγχάνουν τις υψηλότερες δυνατές ταχύτητες.

Αν και ο χώρος της υγείας ήταν για πολύ καιρό αποστασιοποιημένος από τη χρήση των ‘’μεγάλων’’ δεδομένων, οι νέες δυνατότητες που προκύπτουν με τη σωστή εκμετάλλευσή τους είναι τεράστιες. Στην ψηφιακή επιδημιολογία, δεν είναι όλες οι πηγές ‘’μεγάλων’’ δεδομένων εξίσου χρήσιμες, για το λόγο αυτό στη συνέχεια θα περιγράψουμε τις σημαντικότερες από αυτές ως προς τη χρησιμότητά τους στην παρακολούθηση και πρόβλεψη ξεσπασμάτων ασθενειών, και ως προς των δυσκολιών που υπεισέρχονται στην αξιοποίηση δεδομένων από αυτές. [7][10] [11][12][13]

2.2.1 Ηλεκτρονικά μέσα κοινωνικής δικτύωσης

Τα τελευταία χρόνια οι περισσότεροι κάτοικοι των ανεπτυγμένων και πολλών υπό ανάπτυξη χωρών είναι εγγεγραμμένοι και χρησιμοποιούν ένα ή περισσότερα ηλεκτρονικά μέσα κοινωνικής δικτύωσης, ενώ ο αριθμός των χρηστών αυξάνεται συνεχώς και με επιταχυνόμενο ρυθμό. Τα δεδομένα που παράγονται από τα μέσα αυτά, τα οποία ανήκουν στην κατηγορία των ‘’μεγάλων δεδομένων’’, περιέχουν πληροφορίες που μπορούν να οδηγήσουν σε συμπεράσματα για την υγεία, την ψυχολογική κατάσταση και την τοποθεσία τον χρηστών τους. Παραδείγματος χάριν αν κάποιος χρήστης είναι φορέας κάποιας ασθένειας, οι συζητήσεις που κάνει και τα άτομα με τα οποία επικοινωνεί συχνότερα μπορούν να μας δώσουν μια εικόνα για τις κινήσεις του τις τελευταίες μέρες και με ποια άτομα ήρθε σε επαφή, και έτσι να αναγνωρίσουμε νέα μοτίβα εξάπλωσης της ασθένειας, νέους πιθανούς φορείς που θα πρέπει να εξεταστούν καθώς και να ενορχηστρώσουμε νέα προληπτικά μέτρα αντιμετώπισης μιας επιδημίας. Τα δεδομένα που προέρχονται από τα ηλεκτρονικά μέσα κοινωνικής δικτύωσης έχουν μεγάλη ανομοιογένεια και χρειάζεται ιδιαίτερη προσοχή στην επεξεργασία και την ανάλυσή τους, καθώς μπορούν να οδηγήσουν πολύ εύκολα σε λάθος συμπεράσματα.

Τα πιο ευρέως χρησιμοποιούμενα μέσα κοινωνικής δικτύωσης καθώς και η αυξητική τάση του αριθμού των χρηστών τους τα τελευταία δεκαπέντε χρόνια παρουσιάζονται στο παρακάτω διάγραμμα. [14][15]



Διάγραμμα 2.1: Η αύξηση του αριθμού των χρηστών των χρηστών ηλεκτρονικών μέσων κοινωνικής δικτύωσης τα τελευταία 15 χρόνια. [16]

2.2.2 Κινητά τηλέφωνα και ‘’Έξυπνες’’ φορητές συσκευές

Μία άλλη πηγή ‘’μεγάλων’’ δεδομένων αποτελούν τα έξυπνα κινητά τηλέφωνα και γενικότερα οι έξυπνες φορητές συσκευές των οποίων η χρήση εδραιώνεται όλο και περισσότερο στην καθημερινότητά μας. Μέσω των δεδομένων που συλλέγονται από αυτές τις συσκευές μπορούμε να παρακολουθήσουμε τις κινήσεις ενός ατόμου χρησιμοποιώντας τις κεραίες του τηλεφωνικού δικτύου μιας περιοχής ή και δορυφορικά σήματα από το Παγκόσμιο Σύστημα Θεσιθεσίας (GPS). Έτσι, παρακολουθώντας τις κινήσεις διαφόρων χρηστών τέτοιων συσκευών κατά τη διάρκεια μιας επιδημίας μπορεί να μελετηθεί ο τρόπος εξάπλωσής της και να δημιουργηθούν μοντέλα τα οποία προβλέπουν την πορεία της ασθένειας χωρικά και χρονικά. Με τη χρήση, λοιπόν, αυτών των μεθόδων γίνεται εφικτή η λήψη των κατάλληλων προληπτικών μέτρων για την καταπολέμηση της επιδημίας έγκαιρα, καθώς και η αποφυγή μιας αντίστοιχης επιδημίας στο μέλλον. Επίσης μέσω των δεδομένων των τηλεφωνικών κλήσεων μπορούν να ληφθούν πληροφορίες για άτομα με τα οποία ένας φορέας μπορεί να ήρθε σε επαφή και έτσι να εντοπιστούν νέες περιοχές στις οποίες μπορούν να εμφανισθούν κρούσματα της ασθένειας.

Σε αυτό το σημείο αξίζει να σημειωθεί πως κάποιες από τις παραπάνω λειτουργίες, όπως ο εντοπισμός μέσω του τηλεφωνικού δικτύου και η καταγραφή τηλεφωνημάτων, μπορούν να εφαρμοστούν και στα συμβατικά κινητά τηλέφωνα, οπότε και από αυτά μπορούν να αντληθούν χρήσιμες πληροφορίες σε περίπτωση ανάγκης. Πάντα όμως η χρήση δεδομένων που αντλήθηκαν με τους τρόπους που αναφέρθηκαν πρέπει να γίνεται με ιδιαίτερη προσοχή και με τη συγκατάθεση των χρηστών, αν είναι δυνατό, διότι η παραβίαση και δημοσίευση προσωπικών δεδομένων αποτελεί αδίκημα.

Τα δεδομένα που λαμβάνονται από τις ‘’έξυπνες’’ φορητές συσκευές δεν έχουν τόσο μεγάλο βαθμό ανομοιομορφίας όσο τα δεδομένα που προέρχονται από ηλεκτρονικά μέσα κοινωνικής δικτύωσης, έχουν όμως μεγάλο όγκο και για το λόγο αυτό χρειάζεται καλή υποστήριξη από πλευράς υλικού και λογισμικού για την επεξεργασία τους. [5][8]

2.2.3 Ηλεκτρονικά ΜΜΕ

Με την εξέλιξη της τεχνολογίας και την ραγδαία αύξηση της χρήσης του διαδικτύου, τα τελευταία τριάντα περίπου χρόνια, η εμφάνιση αμέτρητων ιστοσελίδων μέσων μαζικής ενημέρωσης ήταν αναπόφευκτη. Σήμερα σχεδόν όλα τα ειδησεογραφικά πρακτορεία διαθέτουν και μία αντίστοιχη ηλεκτρονική ιστοσελίδα για τη διευκόλυνση της ενημέρωσης των πολιτών, ενώ ταυτόχρονα έχουν προκύψει και χιλιάδες άλλοι καθαρά διαδικτυακοί ιστότοποι με τον ίδιο στόχο. Έτσι πλέον η ενημέρωση είναι μία πολύ εύκολη και γρήγορη διαδικασία και το γεγονός ότι, ως επί το πλείστον, είναι δωρεάν έχει οδηγήσει στη δραστική μείωση στην πώληση των αντίστοιχων έντυπων μέσων. Δυστυχώς αυτή η αύξηση στις διαθέσιμες πηγές ενημέρωσης έχει οδηγήσει σε μία αντίστοιχη, ολοένα αυξανόμενη, διεύρυνση της παραπληροφόρησης. Πλέον, υπάρχει πληθώρα ιστοσελίδων που υποστηρίζουν πως έχουν ως στόχο την ενημέρωση των ανθρώπων που τις επισκέπτονται, ενώ ταυτόχρονα προφέρουν ψευδείς ειδήσεις με σκοπό την παραπλάνησή τους. Εξαιτίας αυτού ιδιαίτερη προσοχή απαιτείται στην ενημέρωση μέσω του διαδικτύου και καλή θα ήταν η διασταύρωση των πληροφοριών που λαμβάνονται από τέτοιες πηγές πριν ληφθούν ως βάσιμες.

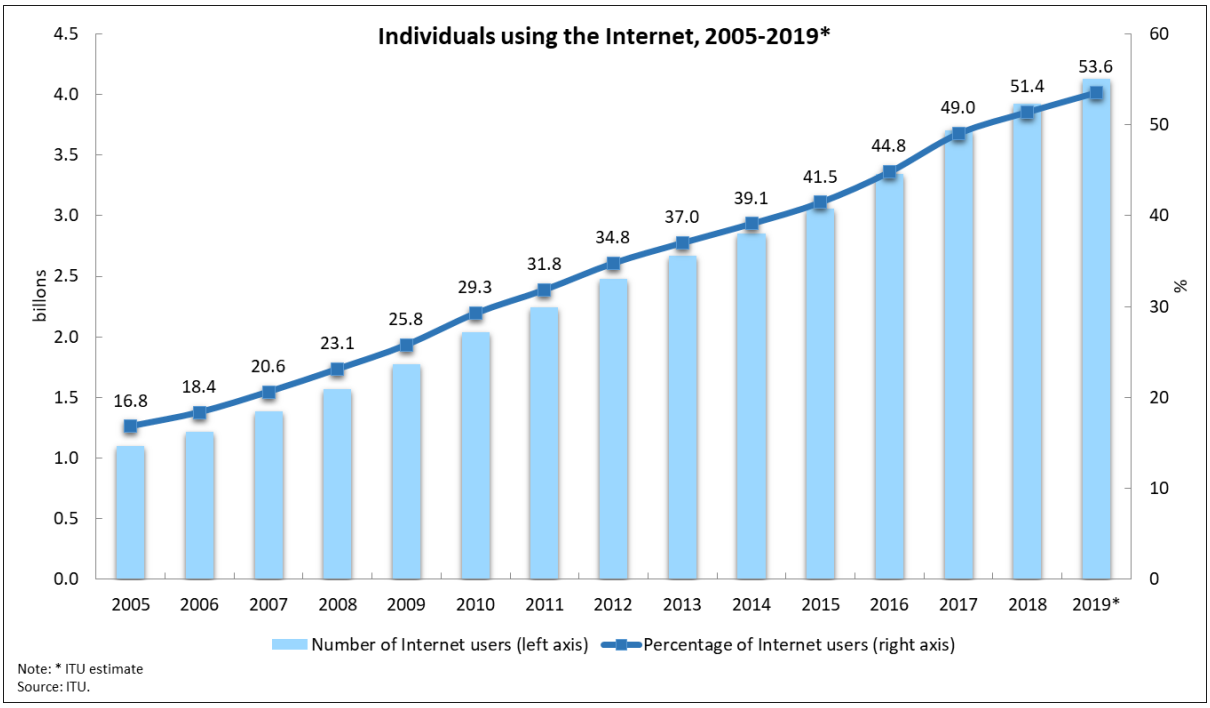
Για την ψηφιακή επιδημιολογία τα ηλεκτρονικά μέσα μαζικής ενημέρωσης αποτελούν μια πολύ χρήσιμη πηγή δεδομένων. Η συλλογή και επεξεργασία των δεδομένων από αυτές τις πηγές είναι αρκετά πιο εύκολη από τα δεδομένα των άλλων πηγών που αναφέραμε. Πάντα όμως χρειάζεται προσοχή και ενδελεχής προκαταρκτική έρευνα στη χρήση των δεδομένων από τέτοιες πηγές για τους λόγους που αναφέραμε προηγουμένως. Σε κάθε περίπτωση, όμως, η σωστή αξιοποίησή τους μόνο ευεργετική μπορεί να είναι για την ψηφιακή επιδημιολογία.

Αναλύοντας τη χρησιμότητα των δεδομένων των ηλεκτρονικών μέσων μαζικής ενημέρωσης στα διάφορα πεδία της ψηφιακής επιδημιολογίας, ένα βασικό μειονέκτημα τους είναι πως δεν εμφανίζονται σε έγκαιρο χρόνο και πάντα ακολουθούν το ξέσπασμα μιας επιδημίας. Έτσι, η δημιουργία συστημάτων πρόβλεψης και πρόληψης με τη χρήση αυτών των δεδομένων είναι αδύνατη, τουλάχιστον για τις πάσχουσες περιοχές. Μπορούν όμως να χρησιμοποιηθούν για τη δημιουργία διαδραστικών χαρτών για την ενημέρωση του πληθυσμού και για τη μελέτη της ‘’μηχανικής’’ με βάση την οποία εξαπλώνεται μια επιδημία που μπορεί να βοηθήσει στη δημιουργία μιας προσομοίωσής της. Με αυτόν τον τρόπο μπορούν να εντοπισθούν και να ειδοποιηθούν έγκαιρα περιοχές και χώρες στις οποίες υπάρχει πιθανότητα να υπάρξει μεταγενέστερη εξάπλωση της ασθένειας. Τέλος, υπάρχουν κάποιες περιπτώσεις που μία είδηση σχετικά με την υγεία σε έναν τόπο, που μπορεί να δημοσιευτεί από ένα μικρό τοπικό ειδησεογραφικό πρακτορείο, να μην γίνει γρήγορα ευρέως γνωστή. Σε αυτή την περίπτωση ένα σύστημα που συλλέγει τις ειδήσεις από όλο τον κόσμο μπορεί να ειδοποιήσει άμεσα τις αρμόδιες αρχές για έναν κίνδυνο για τη δημόσια υγεία, με αποτέλεσμα να ληφθούν έγκαιρα τα απαραίτητα μέτρα και να μη χαθεί πολύτιμος χρόνος. [8][11][17][18]

2.2.4 Μηχανές ηλεκτρονικής αναζήτησης

Πληροφορίες σχετικά με την ευημερία ενός τόπου μπορούν να ληφθούν πολλές φορές από τις αναζητήσεις που κάνουν οι κάτοικοι της περιοχής σε πλατφόρμες ηλεκτρονικής αναζήτησης. Συνήθως όταν υπάρχουν προβλήματα υγείας οι άνθρωποι προσπαθούν να βρουν τον τρόπο να τα κατανοήσουν και να τα αντιμετωπίσουν.

Στη σημερινή εποχή η βασική πηγή πληροφοριών είναι το διαδίκτυο και χρησιμοποιείται από το μεγαλύτερο μέρος του παγκόσμιου πληθυσμού. Έτσι, μπορεί κανείς να αναζητήσει πληροφορίες που μπορεί να σχετίζονται με κάποια ασθένεια ή να προσπαθήσει να εντοπίσει κάποιον ιατρό στην περιοχή του. Συνεπώς, τα δεδομένα των ηλεκτρονικών αναζητήσεων μπορούν να δώσουν στοιχεία σχετικά με την εξάπλωση μιας ασθένειας σε έναν τόπο όταν οι ηλεκτρονικές αναζητήσεις από τους κατοίκους σχετίζονται σε μεγάλο βαθμό με την υγεία, τις ασθένειες και το ιατρικό προσωπικό της περιοχής. Ακόμη ανάλογα με το αντικείμενο των ηλεκτρονικών αναζητήσεων μπορούν να εξαχθούν δεδομένα σχετικά με το είδος της ασθένειας που πλήττει την συγκεκριμένη περιοχή.



Διάγραμμα 2.2: Η αύξηση του αριθμού των χρηστών του διαδικτύου στο διάστημα 2005-2019, σε αριθμό και ποσοστό, παγκοσμίως. [19]

Κλείνοντας, επισημαίνουμε πως απαιτείται προσεκτική ανάλυση στην συσχέτιση και την επαλήθευση των πληροφοριών που λαμβάνονται, καθώς οι ηλεκτρονικές αναζητήσεις μπορεί να μην αντικατοπτρίζουν πάντα και σε κάθε χρονική στιγμή το τι επικρατεί σε έναν τόπο. Σε κάθε περίπτωση όμως μία ένδειξη ότι κάτι ίσως συμβαίνει μπορεί να οδηγήσει σε περισσότερη έρευνα για την επαλήθευση αυτών των πληροφοριών, κάτι το οποίο δεν μπορεί να έχει αρνητικά αποτελέσματα αν εκμηδενίσουμε την σπατάλη περεταίρω πόρων. [20][21]

2.3 Μέθοδοι και εργαλεία αξιοποίησης ‘’μεγάλων’’ δεδομένων (Big data)

Για την ανάλυση των δεδομένων από τις πηγές που αναφέραμε, στην ηλεκτρονική επιδημιολογία, υπάρχουν κάποια βασικά ηλεκτρονικά εργαλεία και τεχνικές τα οποία συνεχώς εξελίσσονται παράλληλα με την εξέλιξη της τεχνολογίας των υπολογιστών τόσο στο υλικό όσο και στο λογισμικό. Αυτά τα εργαλεία, τα οποία αποτελούνται από τη μηχανική μάθηση, την επεξεργασία φυσικής γλώσσας, τη ‘’βαθιά’’ μάθηση και τα νευρωνικά δίκτυα γενικότερα, ανήκουν στην κατηγορία των συστημάτων τεχνητής νοημοσύνης και ο τρόπος λειτουργίας τους θα αναλυθεί στη συνέχεια.[22][23][21]

2.3.1 Η τεχνητή νοημοσύνη (artificial intelligence - AI)

Η τεχνητή νοημοσύνη (artificial intelligence – AI) αποτελεί το βασικότερο εργαλείο για την επεξεργασία, την ανάλυση και την εξαγωγή συμπερασμάτων από τα ‘’μεγάλα’’ δεδομένα. Αποτελεί μια τεχνολογία που ήδη έχει μεγάλη επίδραση στο πώς οι χρήστες του διαδικτύου αλληλεπιδρούν και επηρεάζονται από αυτό. Ένα πολύ απλό παράδειγμα είναι ο διαχωρισμός των ηλεκτρονικών μηνυμάτων σε φακέλους ανεπιθύμητης αλληλογραφίας που γίνεται από τις υπηρεσίας ηλεκτρονικής αλληλογραφίας ή η εξατομίκευση που γίνεται στις ηλεκτρονικές εφαρμογές παρακολούθησης ταινιών και τηλεοπτικών σειρών όπου ανάλογα με τις προτιμήσεις και τις επιλογές που κάνει ένας χρήστης οι προτάσεις, ως προς το περιεχόμενο που γίνονται από την εφαρμογή, αλλάζουν και προσαρμόζονται στις προτιμήσεις του. Η τεχνολογία της τεχνητής νοημοσύνης υπήρχε για δεκαετίες όμως μέχρι προσφάτως δεν μπορούσε να εφαρμοστεί σε μεγάλη κλίμακα καθώς δεν υπήρχαν τα απαιτούμενα τεχνολογικά μέσα. Η ραγδαία όμως εξέλιξη των υπολογιστών τα τελευταία χρόνια έχει ανοίξει το δρόμο για μεγάλες εξελίξεις για τον τρόπο που ζούμε και αλληλεπιδρούμε με την τεχνολογία. Ο όρος τεχνητή νοημοσύνη διατυπώθηκε για πρώτη φορά σε ένα εργαστήριο του κολλεγίου Ντάρτμουθ, το οποίο βρίσκεται στο Ανόβερο της Πολιτείας του Νέου Χαμσάιρ των Ηνωμένων Πολιτειών Αμερικής, το 1956, από τον τότε επίκουρο καθηγητή και επιστήμονα Τζον Μακάρθι με σκοπό να διαχωρίσει αυτό το πεδίο αυτό από το πεδίο της ‘’κυβερνητικής’’ (cybernetics). Έκτοτε έχουν γίνει αμέτρητες έρευνες στο πεδίο της τεχνητής νοημοσύνης και έχουν οδηγήσει σε μεγάλες εξελίξεις με αποτέλεσμα να χρησιμοποιείται σε πάρα πολλά σύγχρονα επιστημονικά πεδία όπως η πληροφορική, η γλωσσολογία και πολλά ιατρικά πεδία.

Ουσιαστικά ο όρος τεχνητή νοημοσύνη περιγράφει μια κατηγορία υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς, τα οποία υπονοούν έστω και μια στοιχειώδη ευφυΐα όπως μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση μέσω συμφραζόμενων και επίλυση προβλημάτων. Τα υπολογιστικά συστήματα αυτά χρησιμοποιούν ένα συνδυασμό αλγοριθμικών τεχνικών όπως η μηχανική μάθηση, η ‘’βαθιά μάθηση’’ και τα νευρωνικά δίκτυα με στόχο την επίλυση προβλημάτων τα οποία απαιτούν μια νοημοσύνη παρόμοια με την ανθρώπινη για να επιλυθούν, αλλά περιέχουν τόσο πολλές πληροφορίες ή έχουν τόσο μεγάλη πολυπλοκότητα που ο ανθρώπινος εγκέφαλος δεν μπορεί να επεξεργαστεί.

Τα βασικά χαρακτηριστικά – ικανότητες που αντιπροσωπεύουν τα συστήματα τεχνητής νοημοσύνης είναι τα εξής:

- Η ικανότητα πρόβλεψης
- Η προσαρμοστικότητά τους
- Η ικανότητά τους για λήψη αποφάσεων
- Η συνεχής τους μάθηση και η δυνατότητα εξέλιξης
- Η ικανότητά τους για διαφορετικές αντιδράσεις (προσεγγίσεις) σε κάθε νέο πρόβλημα
- Η ικανότητά τους για κίνηση και αντίληψη

Για τη δημιουργία τους χρειάζεται μία περίοδος εκπαίδευσης του συστήματος με συγκεκριμένα δεδομένα που σχετίζονται με τη λειτουργία που επιθυμείται, από τον κατασκευαστή τους, να επιτελέσουν. Ιδιαίτερη προσοχή όμως χρειάζεται κατά την εκπαίδευση του συστήματος διότι κακής ποιότητας δεδομένα μπορεί να οδηγήσουν σε ένα δυσλειτουργικό σύστημα. Άλλωστε υπάρχει και ένα ρητό ανάμεσα στους ειδήμονες που ασχολούνται με τα συστήματα τεχνητής νοημοσύνης: ‘’Ένα σύστημα τεχνητής νοημοσύνης είναι τόσο καλής ποιότητας όσο τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή του’’. Μεγαλύτερη ανάλυση για τη λειτουργία και την εκπαίδευση των συστημάτων τεχνητής νοημοσύνης ακολουθεί στη συνέχεια. [21][22][23][24]

2.3.2 Οι αλγόριθμοι

Ένας αλγόριθμος, με τη γενική του έννοια, ορίζεται ως μια πεπερασμένη σειρά ενεργειών, οι οποίες πρέπει να εκτελεστούν με καθορισμένη σειρά και σε πεπερασμένο χρόνο, που έχουν ως στόχο την επίλυση ενός προβλήματος. Μπορούμε πιο απλά να πούμε πως οι αλγόριθμοι αποτελούν έναν συνδυασμό των μαθηματικών με τη λογική. Ένας υπολογιστικός αλγόριθμος είναι μια σειρά από ρητές εντολές προς το υπολογιστικό σύστημα οι οποίες καθορίζουν την συμπεριφορά του συστήματος ως προς τα δεδομένα που του εισάγονται με σκοπό την εκτέλεση μιας συγκεκριμένης λειτουργίας από το σύστημα, η οποία μπορεί να σχετίζεται με την επεξεργασία, την αποθήκευση ή την προβολή των δεδομένων εισόδου.

Μία ομάδα αλγορίθμων που επιτελούν μια συγκεκριμένη λειτουργία ονομάζεται συνάρτηση. Τα υπολογιστικά προγράμματα αποτελούνται από ένα συνδυασμό συναρτήσεων οι οποίες αλληλεπιδρούν με σκοπό την επίλυση ενός ή περισσοτέρων πιο σύνθετων προβλημάτων, ή την εκτέλεση μιας πιο σύνθετης λειτουργίας από το σύστημα. Έτσι, ένα ολοκληρωμένο υπολογιστικό σύστημα αποτελείται από πληθώρα προγραμμάτων, δηλαδή στην ουσία αλγορίθμων.

Επειδή στα υπολογιστικά συστήματα μία συγκεκριμένη λειτουργιά είναι δυνατόν να επιτελεσθεί από παραπάνω από έναν αλγόριθμο, ένας αλγόριθμος θεωρείται βέλτιστος όταν οδηγεί στην εκτέλεση της συγκεκριμένης λειτουργίας που επιτελεί στον ελάχιστο δυνατό χρόνο και σπαταλώντας τους ελάχιστους υπολογιστικούς πόρους από το σύστημα. Με την πάροδο των ετών οι αλγόριθμοι που χρησιμοποιούνται στα υπολογιστικά συστήματα έχουν βελτιωθεί σε πολύ μεγάλο βαθμό καθώς και νέες αλγοριθμικές τεχνικές έχουν προκύψει. Τέτοιες τεχνικές, όπως προαναφέραμε, αποτελούν η τεχνητή νοημοσύνη, η μηχανική μάθηση, η ‘’βαθιά μάθηση’’ και τα νευρωνικά δίκτυα, οι οποίες θα αναλυθούν στη συνέχεια. [25][26]

2.3.3 Η μηχανική μάθηση (Machine Learning - ML)

Η λέξη *μηχανική* στη μηχανική μάθηση αναφέρεται ουσιαστικά σε έναν αλγόριθμο ή μια μέθοδο υπολογισμού. Η μηχανική μάθηση αποτελεί μια σημαντικότατη εξέλιξη στην τεχνολογία αλγορίθμων η οποία όμως υπάρχει για δεκαετίες, αν και μέχρι πρόσφατα δεν μπορούσαμε να εκμεταλλευτούμε πλήρως όλες τις νέες δυνατότητες που προσφέρει. Τα τελευταία χρόνια χρησιμοποιείται στους περισσότερους τομείς όλων των επιστημών, ως μια μέθοδος ανάλυσης δεδομένων, και χρησιμοποιείται για τη δημιουργία συστημάτων τα οποία μπορούν να αναγνωρίζουν αυτόματα συγκεκριμένα μοτίβα, να τα επεξεργάζονται και να τα διαχειρίζονται κατάλληλα ανάλογα με τον σκοπό για τον οποίο κατασκευάστηκαν. Βασίζεται στην ιδέα της κατασκευής υπολογιστικών συστημάτων που μπορούν να μαθαίνουν από τα δεδομένα που επεξεργάζονται, χωρίς να έχουν προγραμματιστεί ρητά για τη συγκεκριμένη λειτουργία, και όσο επεξεργάζονται καινούρια δεδομένα να γίνονται πιο ''έξυπνα'', δηλαδή πιο ακριβή και με λιγότερα σφάλματα στην αναγνώριση συγκεκριμένων προτύπων.

Ένα σύστημα μηχανικής μάθησης πριν να μπορέσει να λειτουργήσει αυτόνομα πρέπει να εκπαιδευτεί. Η εκπαίδευση αυτή γίνεται μέσω ενός αλγορίθμου που ονομάζεται αλγόριθμος μάθησης. Κατά τη διάρκεια της εκπαίδευσης εισάγονται στο σύστημα δεδομένα ανάλογα με την διεργασία που πρέπει τελικά να επιτελεί, για παράδειγμα αν το σύστημα πρέπει να διαχωρίζει εικόνες από ζώα, συγκεκριμένα τις γάτες από τους σκύλους, εισάγονται δεδομένα που περιλαμβάνουν πολλαπλές εικόνες από γάτες και σκύλους με αποτέλεσμα να μπορεί στο μέλλον να διακρίνει οποιαδήποτε εικόνα γάτας από εικόνα σκύλου και όχι μόνο αυτές που είχε λάβει ως αρχική είσοδο. Συγκεκριμένα ο αλγόριθμος επεξεργάζεται τα δεδομένα εισόδου και βγάζει συμπεράσματα μέσω ενός συστήματος ανάδρασης. Στη συνέχεια με βάση τα συμπεράσματα αυτά δημιουργεί μία νέα ομάδα κανόνων που ουσιαστικά αλλάζουν τον αρχικό αλγόριθμο και το σύστημα μαζί, και έτσι γεννάται ένας νέος τροποποιημένος αλγόριθμος. Αξίζει να σημειωθεί πως χρησιμοποιώντας διαφορετικά αρχικά δεδομένα στον ίδιο αλγόριθμο μπορεί να δώσει έναν τελείως διαφορετικό τελικό σύστημα το οποίο μπορεί να χρησιμοποιηθεί για άλλες λειτουργίες, όπως για την πρόβλεψη των μεταβολών του χρηματιστηρίου ή τη μετάφραση μιας γλώσσα σε μία άλλη. Υπάρχουν τρείς βασικές στρατηγικές που χρησιμοποιούνται στην αρχική εκπαίδευση ενός συστήματος μηχανικής μάθησης και αυτές είναι η μάθηση υπό επίβλεψη (supervised learning), η μάθηση χωρίς επίβλεψη (unsupervised learning) και η ενισχυμένη μάθηση (reinforcement learning), που θα αναλυθούν στη συνέχεια.

- *Μάθηση υπό επίβλεψη (supervised learning)*:

  Στην τεχνική της μάθησης υπό επίβλεψη τροφοδοτούνται στον αλγόριθμο μάθησης δεδομένα με επεξήγηση του τι αντιπροσωπεύουν καθώς και η επιθυμητή έξοδος που θα πρέπει να έχει το τελικό σύστημα. Για παράδειγμα τροφοδοτούνται στον αλγόριθμο εικόνες με σκύλους οι οποίες περιλαμβάνουν την επιγραφή σκύλος καθώς και το επιθυμητό αποτέλεσμα που είναι να αναγνωρίζονται από το πρόγραμμα εικόνες σκύλων ανάμεσα σε άλλες εικόνες. Έτσι, το σύστημα αρχίζει να αναγνωρίζει τους κανόνες που περιγράφουν έναν σκύλο και στη συνέχεια με βάση αυτούς τους κανόνες μπορεί να αναγνωρίσει σε οποιαδήποτε εικόνα έναν σκύλο και όχι μόνο τις εικόνες και τους σκύλους που αρχικά χρησιμοποιήθηκαν στην εκπαίδευσή του.

- *Μάθηση χωρίς επίβλεψη (unsupervised learning)*:

Στην τεχνική της μάθησης χωρίς επίβλεψη τροφοδοτούνται στον αλγόριθμο μάθησης δεδομένα χωρίς κάποια επεξήγηση του τι αντιπροσωπεύουν και ζητείται από τον αλγόριθμο να αναγνωρίσει μοτίβα στα δεδομένα εισόδου που τα συνδέουν. Για παράδειγμα τέτοιου είδους συστήματα συναντώνται σε ιστοσελίδες ηλεκτρονικών καταστημάτων που προτείνουν στους πελάτες προϊόντα που μπορεί να τους ενδιαφέρουν που συνήθως αγοράζονται μαζί με τα προϊόντα που έχουν ήδη στο ηλεκτρονικό τους καλάθι.

- *Η ενισχυμένη μάθηση (reinforcement learning)*:

  Στην τεχνική της ενισχυμένης μάθησης ο αλγόριθμος μάθησης αλληλεπιδρά με ένα δυναμικό περιβάλλον το οποίο ανάλογα με τις διεργασίες που επιτελεί ο αλγόριθμος τον επιβραβεύει ή τον τιμωρεί. Για παράδειγμα αυτή η τεχνική μάθησης χρησιμοποιείται στα αυτοοδηγούμενα αυτοκίνητα, που όταν παραμένουν μέσα στη σωστή λωρίδα κυκλοφορίας δέχονται επιβράβευση ενώ όταν τείνουν να βγουν εκτός του δρόμου τιμωρούνται. [27][23][28]

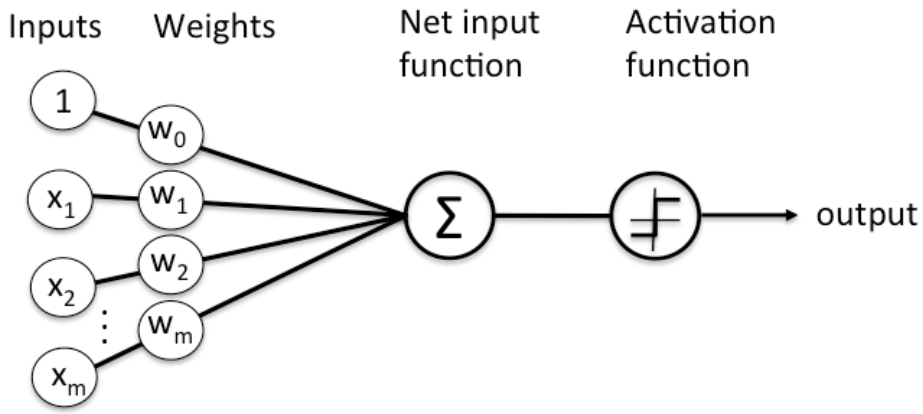
2.3.4 ‘’Βαθιά’’ μάθηση (deep learning) και νευρωνικά δίκτυα (neural networks)

Η ‘’βαθιά’’ μάθηση αποτελεί μια εξελιγμένη υποκατηγορία της μηχανικής μάθησης και ενώ η μηχανική μάθηση χρησιμοποιείται σε ένα ευρύ πεδίο εφαρμογών η ‘’βαθιά’’ μάθηση συναντάται κυρίως ως ακρογωνιαίος λίθος των συστημάτων τεχνητής νοημοσύνης. Η ‘’βαθιά’’ μάθηση είναι άρρηκτα συνδεδεμένη με τα νευρωνικά δίκτυα τα οποία, όπως υποδηλώνει και η ονομασία τους, είναι υπολογιστικά συστήματα τα οποία μιμούνται τη δομή και τη λειτουργία των νευρώνων ενός εγκεφάλου.

Ένα νευρωνικό δίκτυο βασίζεται σε μια ομάδα αλγορίθμων που επικοινωνούν μεταξύ τους και αποτελούν τους νευρώνες του νευρωνικού δικτύου με αντίστοιχη λογική όπως σε έναν φυσικό εγκέφαλο, και, όπως και στη μηχανική μάθηση, η λειτουργία του είναι η αναγνώριση μοτίβων. Στην οσία ένα νευρωνικό δίκτυο είναι ένα σύνολο διασυνδεδεμένων επεξεργαστών που επικοινωνούν μεταξύ τους και εκτελούν πράξεις. Όταν ένας νευρώνας λάβει ένα σήμα το επεξεργάζεται (εκτελείται ο αλγόριθμος) και στη συνέχεια μπορεί να μεταβιβάσει το ίδιο το σήμα, ή κάποιο αποτέλεσμα της ανάλυσης που έκανε, στους νευρώνες που είναι συνδεδεμένοι με αυτόν. Τα σήματα που εισέρχονται ή εξέρχονται από έναν νευρώνα και περνούν από τις συνδέσεις του δικτύου, που ονομάζονται και ‘’άκρες’’ (edges), είναι πραγματικοί αριθμοί οι οποίοι παράγονται από τους ίδιους τους νευρώνες. Αντίστοιχα οι νευρώνες επεξεργάζονται το άθροισμα των δεδομένων που εισέρχονται σε αυτούς με τη χρήση κάποιας καθορισμένης μη γραμμικής συνάρτησης και έτσι παράγονται νέοι πραγματικοί αριθμοί που μεταβιβάζονται μέσω των συνδέσμων στους γειτονικούς νευρώνες του νευρωνικού δικτύου. Με αυτόν τον τρόπο τα νευρωνικά δίκτυα επεξεργάζονται τα δεδομένα και σιγά-σιγά εξελίσσονται -‘’μαθαίνουν’’ ώστε να εκτελούν ακριβέστερα τις λειτουργίες για τις οποίες έχουν κατασκευασθεί.

Σε ένα νευρωνικό δίκτυο οι πληροφορίες που περιέχονται σε έναν νευρώνα κάθε στιγμή αποτελούνται από τα δεδομένα εισόδου καθώς και μια ομάδα συντελεστών. Οι συντελεστές αυτοί, που ονομάζονται και βάρη (weights), ανάλογα με τις τιμές τους, μπορούν να αυξήσουν ή να μειώσουν τη βαρύτητα που έχει η συγκεκριμένη είσοδος στους υπολογισμούς του νευρώνα. Με

αυτόν τον τρόπο, τα βάρη, επιτρέπουν στα δεδομένα εισόδου που έχουν μεγαλύτερη σημασία, σε σχέση με τη λειτουργία που θέλουμε να διδάξουμε στο σύστημα, να επηρεάσουν πιο πολύ το σύστημα σε σχέση με άλλα όχι τόσο χρήσιμα δεδομένα. Μέσα στο σύστημα γίνονται συνδυασμοί της εισόδων των νευρώνων με τους διάφορους συντελεστές και στη συνέχεια όλα αυτά τα δεδομένα αθροίζονται και τροφοδοτούνται σε μια συνάρτηση που περιέχεται στο σύστημα και ονομάζεται συνάρτηση ενεργοποίησης. Η συνάρτηση ενεργοποίησης είναι αυτή που αποφασίζει το αν και κατά πόσο τα δεδομένα θα επεξεργαστούν ή όχι από τον νευρώνα και θα συνεχίσουν τη μετάδοσή τους στο υπόλοιπο σύστημα. Αν αποφασιστεί τελικά η διέλευση των δεδομένων, από τη συνάρτηση ενεργοποίησης, τότε αυτά συνεχίζουν στο επεξεργαστικό τμήμα του νευρώνα και στη συνέχεια το αποτέλεσμα της επεξεργασίας τους μεταδίδεται στους γειτονικούς νευρώνες. Όταν ένας νευρώνας επεξεργάζεται δεδομένα ονομάζεται ενεργοποιημένος. Τέλος, υπάρχει ένας ακόμα συντελεστής που επηρεάζει τη λειτουργία του νευρωνικού δικτύου ο οποίος ονομάζεται συντελεστής αντιστάθμισης (offset). Ο συντελεστής αυτός μπορεί να επηρεάσει την τελική έξοδο του δικτύου έτσι ώστε να πλησιάσει την επιθυμητή έξοδο κατά τον κατασκευαστή του συστήματος.



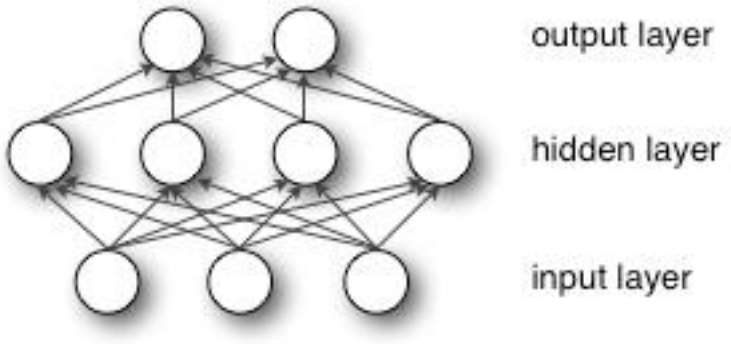
Σχήμα 2.1: Οι είσοδοι, τα βάρη, ο αθροιστής και η συνάρτηση ενεργοποίησης ενός νευρώνα που ανήκει σε ένα νευρωνικό δίκτυο. [29]

Η ονομασία ‘’βαθιά’’ μάθηση αντιστοιχεί σε μία συγκεκριμένη κατηγορία νευρωνικών δικτύων που ονομάζονται ‘’στοιβαγμένα’’ ή ‘’βαθιά’’ νευρωνικά δίκτυα (stacked-deep neural networks). Αυτή η κατηγορία των νευρωνικών δικτύων αποτελείται από πολλά ξεχωριστά επίπεδα (layers), όπου το κάθε επίπεδο αποτελείται από ένα συμβατικό νευρωνικό δίκτυο. Τα δεδομένα αναλύονται ξεχωριστά σε όλα τα επίπεδα με τη σειρά και η συνολική ανάλυση γίνεται εις βάθος και στον μεγαλύτερο βαθμό με αποτέλεσμα να είναι δυνατή η αναγνώριση προτύπων μέσα σε αδόμητα και μεγάλης ετερογένειας δεδομένα όπως τα ‘’μεγάλα’’ δεδομένα (big data) που αποτελούν τις βασικές πηγές δεδομένων της ψηφιακής επιδημιολογίας.

Εμβαθύνοντας, η λειτουργία ενός συστήματος ‘’βαθιάς’’ μάθησης είναι η επεξεργασία και η συσχέτιση των δεδομένων εισόδου και η απεικόνιση των συσχετίσεων που έγιναν στην έξοδο μέσω κάποιων συναρτήσεων των δεδομένων εισόδου. Επειδή τα συστήματα αυτά μπορούν να δημιουργήσουν συσχετίσεις μέσα σε κάθε σύνολο δεδομένων που επεξεργάζονται, εκτός από την περίπτωση όπου τα δεδομένα εισόδου είναι παντελώς ασυσχέτιστα, περιγράφονται μεταφορικά ως ‘’καθολικοί εκτιμητές’’ (‘’universal approximators’’).

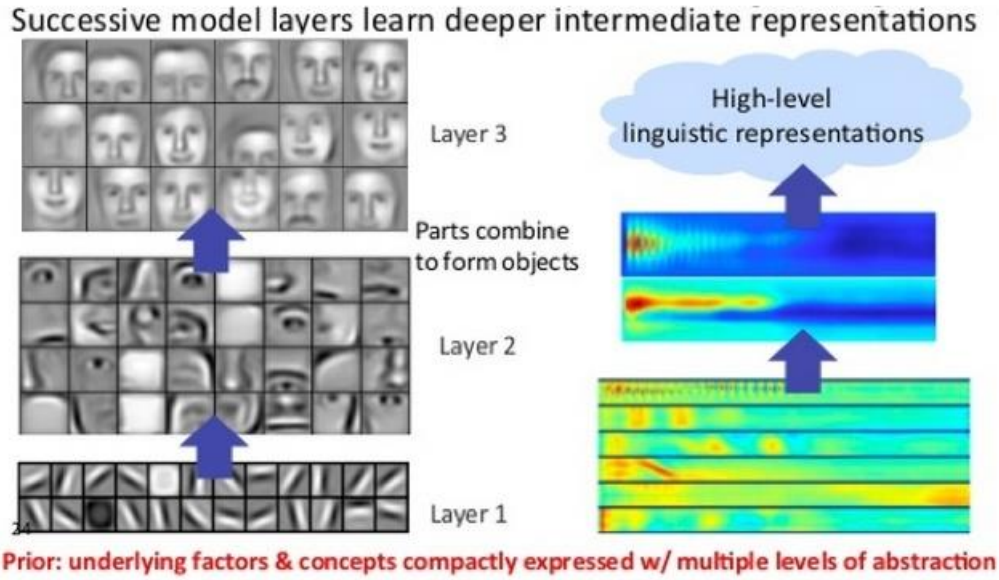
Όπως προαναφέραμε, τα ‘’βαθιά’’ νευρωνικά δίκτυα αποτελούνται από κάποια επίπεδα νευρώνων. Το εξωτερικό σήμα εισόδου το οποίο παρέχεται από τον διαχειριστή του συστήματος

εισέρχεται πρώτα στο πρώτο επίπεδο, που ονομάζεται επίπεδο εισόδου (input layer), επεξεργάζεται από τους νευρώνες αυτού του επιπέδου με την λογική που αναλύσαμε και στη συνέχεια τροφοδοτείται στα επόμενα εσωτερικά επίπεδα, που ονομάζονται κρυφά επίπεδα (hidden layers), με τη σειρά και χωρίς αναδρομές. Τελικά τα δεδομένα εισόδου εισέρχονται στο τελικό επίπεδο, που ονομάζεται επίπεδο εξόδου (output layer), στο οποίο παράγεται η τελική έξοδος του συστήματος. Πρέπει να επισημανθεί πως τα βαθιά νευρωνικά δίκτυα περιέχουν τουλάχιστον δύο κρυφά επίπεδα σε αντίθεση με άλλους τύπους στοιβαγμένων νευρωνικών δικτύων που μπορεί να αποτελούνται μόνο από ένα επίπεδο εισόδου και ένα επίπεδο εξόδου ή να περιέχουν και ένα κρυφό επίπεδο. Σε ένα ‘’βαθύ’’ νευρωνικό δίκτυο, όσο αυξάνεται ο αριθμός των κρυφών επιπέδων τόσο βαθύτερη, ακριβέστερη και λεπτομερέστερη ανάλυση επιτελεσθεί από το δίκτυο.



Σχήμα 2.2: Τα επίπεδα ενός νευρωνικού δικτύου (πάνω: το επίπεδο εξόδου, στη μέση: το κρυφό επίπεδο, κάτω: το επίπεδο εισόδου). [29]

Ένα άλλο σημαντικό χαρακτηριστικό αυτής της κατηγορίας των νευρωνικών δικτύων είναι η ανάλυση των δεδομένων με λειτουργία ιεραρχίας. Σε αυτή τη λειτουργία το κάθε επίπεδο αναλύει διαφορετικής πολυπλοκότητας μοτίβα με την πολυπλοκότητα να αυξάνεται σε κάθε επόμενο επίπεδο. Έτσι, ουσιαστικά επιτελείται ένας καταμερισμός της συνολικής εργασίας στα διάφορα επίπεδα με τα πρώτα να επεξεργάζονται απλά μοτίβα και μέρη σχημάτων, ενώ τα τελευταία να μπορούν να αναλύσουν ολόκληρες εικόνες. Με αυτόν τον τρόπο γίνεται δυνατό για τα ‘’βαθιά’’ νευρωνικά δίκτυα να επεξεργάζονται πολύ μεγάλες και μεγάλων διαστάσεων ομάδες δεδομένων κάνοντας δισεκατομμύρια συσχετίσεις σε μικρά χρονικά διαστήματα. Τέτοια πολυεπίπεδα συστήματα μπορούν να χρησιμοποιηθούν για προηγμένες λειτουργίες όπως η αναγνώριση προσώπου ή ομιλίας και η επεξεργασία φυσικής γλώσσας που η χρησιμότητά της στην ψηφιακή επιδημιολογία θα εξηγηθεί στη συνέχεια.



Σχήμα 2.3: Ανάλυση με λειτουργία ιεραρχίας. Οι λεπτομέρειες αυξάνονται όλο και περισσότερο σε κάθε επόμενο επίπεδο (από κάτω προς τα πάνω). [29]

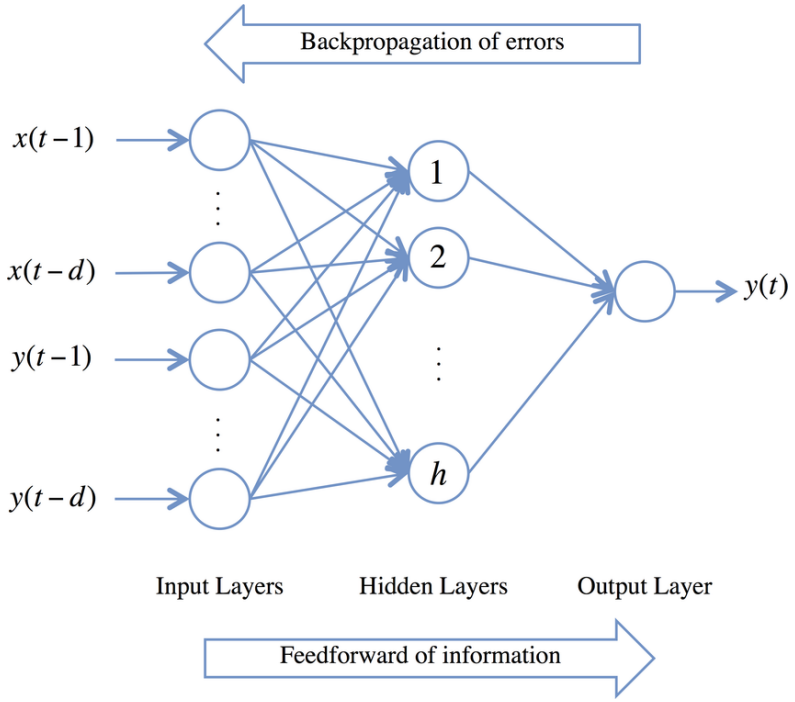
Τα νευρωνικά δίκτυα χωρίζονται σε δύο βασικές κατηγορίες ανάλογα με την κίνηση των δεδομένων εισόδου μέσα στο δίκτυο και τη μέθοδο παραγωγής της τελικής εξόδου. Οι κατηγορίες αυτές συναντώνται σε όλους τους τύπους νευρωνικών δικτύων και σχετίζονται με την ακρίβεια της εξόδου του δικτύου σε σχέση με την επιθυμητή για τον κατασκευαστή έξοδο. Οι δύο αυτές κατηγορίες είναι οι εξής:

- *Χωρίς ανάδραση (Forward Propagation – FP)*:

  Στα νευρωνικά δίκτυα χωρίς ανάδραση το σήμα εισόδου περνά αρχικά από το επίπεδο εισόδου, στη συνέχεια από τα κρυφά επίπεδα και τέλος από το επίπεδο εξόδου γραμμικά και χωρίς επαναλήψεις. Κατά τη διαδικασία αυτή η τιμή του συντελεστή αντιστάθμισης και των βαρών του δικτύου παραμένει σταθερή. Κατά την χωρίς ανάδραση λειτουργία ενός νευρωνικού δικτύου οι συντελεστές ενός επιπέδου μπορούν να επηρεαστούν μόνο ως αποτέλεσμα της επεξεργασίας των δεδομένων από το αμέσως προηγούμενο επίπεδο, από τη φυσιολογική λειτουργία του επιπέδου.

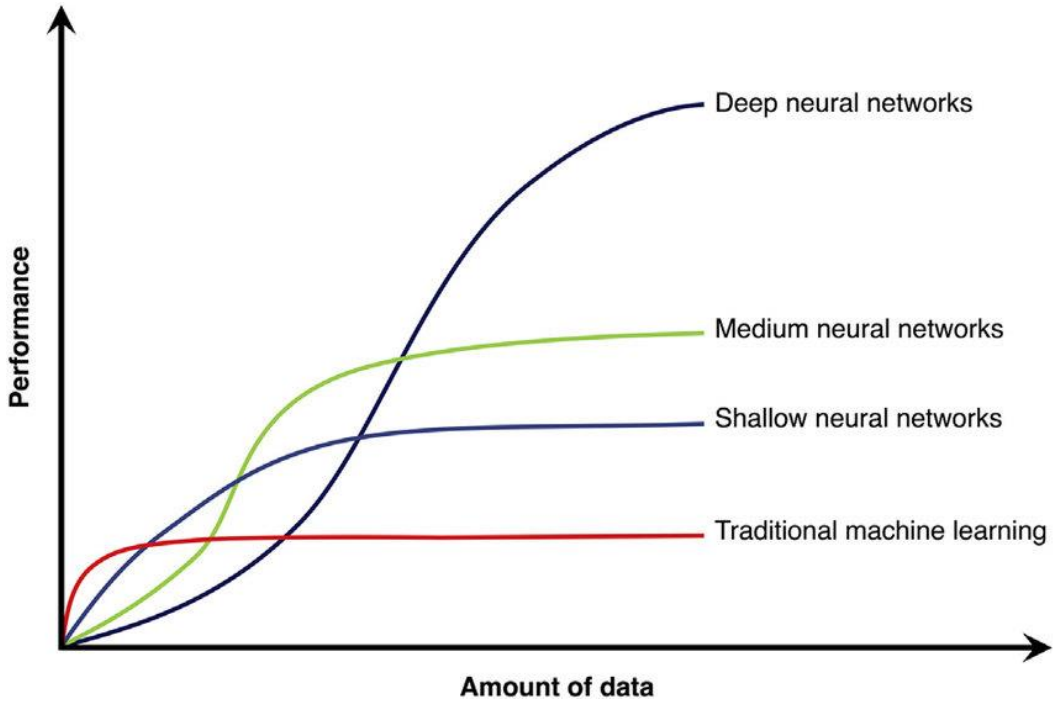
- *Με ανάδραση (Back Propagation – BP)*:

  Στα νευρωνικά δίκτυα με ανάδραση το σήμα εισόδου ακολουθεί την ίδια πορεία όπως και στα δίκτυα χωρίς ανάδραση. Σε αυτή την κατηγορία όμως γίνεται μία σύγκριση ανάμεσα στην έξοδο του συστήματος και σε μία επιθυμητή για τον κατασκευαστή έξοδο. Η απόκλιση των δύο εξόδων μεταφράζεται σε ένα σήμα σφάλματος το οποίο τροφοδοτείται από το επίπεδο εξόδου στο επίπεδο εισόδου περνώντας μέσα από τα ενδιάμεσα επίπεδα διαδοχικά, μεταβάλλοντας τις τιμές των βαρών τους. Κατά τη λειτουργία του συστήματος γίνονται συνεχείς τροποποιήσεις στις τιμές των βαρών και του συντελεστή αντιστάθμισης μέχρι η έξοδος του συστήματος να γίνει ίση με την επιθυμητή.



Σχήμα 2.4: Ένα νευρωνικό δίκτυο με ανάδραση του σήματος σφάλματος (πάνω βέλος). Όταν δεν υπάρχει ανάδραση το πάνω βέλος δεν υπάρχει. [30]

Σε αυτό το σημείο είναι πολύ σημαντικό να αναφερθεί πως η ανάδραση είναι η λειτουργία που επιτρέπει στα συστήματα να μαθαίνουν από τα σφάλματά τους, δηλαδή από την απόκλισή τους από την σωστή κατά τους ανθρώπους έξοδο. Η μεγάλη διαφορά στην απόδοση διαφόρων τύπων νευρωνικών δικτύων ανάλογα με το βάθος τους και μια σύγκριση με συστήματα συμβατικής μηχανικής μάθησης παρουσιάζεται στο παρακάτω διάγραμμα.



Διάγραμμα 2.3: Μία σύγκριση της απόδοσης των ''βαθιών'' (πάνω), μετρίου βάθους (2ο από πάνω) και ρηχών (3ο από πάνω) νευρωνικών δικτύων και των συστημάτων συμβατικής μηχανικής μάθησης (κάτω) ανάλογα με τον όγκο των δεδομένων που τους τροφοδοτείται. [31]

Εν κατακλείδι, αναφέρουμε πως συστήματα ‘’βαθιάς’’ μάθησης δεν περιορίζονται απλά στην αναγνώριση προτύπων αλλά μπορούν να μοντελοποιήσουν περίπλοκες μη γραμμικές συσχετίσεις μεταξύ των δεδομένων και να κατασκευάσουν πολύπλοκα μοντέλα όπου το αντικείμενο της ανάλυσης να μπορεί να περιγραφεί πολύπλευρα. Επομένως, τα συστήματα αυτά όχι μόνο μπορούν να λύσουν πρόβλημα της επεξεργασία των ‘’μεγάλων’’ δεδομένων αλλά μπορούν να μας βοηθήσουν να ανακαλύψουμε πως εξαπλώνεται μια επιδημία και να κατασκευάσουμε συστήματα πρόβλεψης για μελλοντικά ξεσπάσματα ασθενειών. [29][32][33][34]

2.3.5 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP)

Η ανθρώπινη επικοινωνία, είτε γραπτώς είτε προφορικώς, περιλαμβάνει έναν πολύ μεγάλο όγκο δεδομένων. Ακόμη, ο τόνος της φωνής, οι λέξεις που χρησιμοποιούμε ακόμα και ο συλλαβισμός των λέξεων ή ο τρόπος γραφής περιέχουν κρυφές πληροφορίες που μέχρι πρόσφατα μόνο ένας άνθρωπος μπορούσε να αντιληφθεί και να αποκωδικοποιήσει. Η εξέλιξη των υπολογιστών τα τελευταία χρόνια και η εμφάνιση νέων τεχνολογιών, και κυρίως της τεχνητής νοημοσύνης, έχουν ανοίξει το δρόμο σε νέες δυνατότητες όπως η δημιουργία συστημάτων που μπορούν να αποκωδικοποιήσουν και να κατανοήσουν τον ανθρώπινο λόγο.

Η επεξεργασία φυσικής γλώσσας αποτελεί μία νέα αναπτυσσόμενη τεχνολογία της οποίας στόχος είναι η κατανόηση των ανθρωπίνων γλωσσών από τους υπολογιστές. Η επεξεργασία φυσικής γλώσσας είναι ιδιαίτερα σημαντική για τα συστήματα τεχνητής νοημοσύνης, που απαιτούν για την αποδοτική τους λειτουργία όσο το δυνατόν περισσότερες διαφορετικές πηγές δεδομένων. Η άμεση επικοινωνία των συστημάτων τεχνητής νοημοσύνης με τους ανθρώπους, μέσω του λόγου, κάνει πλέον δυνατή την περεταίρω πολύπλευρη εξέλιξη αυτών των συστημάτων και την εξάλειψη της απόστασης μεταξύ ανθρώπου και μηχανής. Αυτό μπορεί να οδηγήσει σε μία μεγάλη βελτίωση στην ποιότητας ζωής του ανθρώπινου πληθυσμού, έναν στόχο κοινό για όλες τις νέες τεχνολογίες.

Η επεξεργασία φυσικής γλώσσας γίνεται δυνατή με τη χρήση συστημάτων ‘’βαθιάς’’ μάθησης και ο βασικός στόχος αυτής της τεχνολογίας, όπως αναφέραμε, είναι η εξαγωγή πληροφοριών από τον προφορικό και τον γραπτό λόγο χρησιμοποιώντας αλγορίθμους. Η εφαρμογή βασίζεται στην υλοποίηση δύο υποσυστημάτων με διαφορετικούς στόχους τα οποία είναι:

- *Παραγωγή Φυσικής Γλώσσας (Natural Language Generation – NLG)*:

  Η Παραγωγή Φυσικής Γλώσσας αποτελεί το πρώτο υποσύστημα και αφορά τη δυνατότητα των υπολογιστικών συστημάτων να σχηματίσουν φράσεις που θα μπορούσαν να διατυπωθούν από άνθρωπο.

- *Κατανόηση Φυσικής Γλώσσας (Natural Language Understanding – NLU)*:

  Η Κατανόηση Φυσικής Γλώσσας αποτελεί το δεύτερο υποσύστημα και αφορά τη δυνατότητα των υπολογιστών να κατανοήσουν το νόημα μίας φράσης, δηλαδή σε τι αναφέρεται ο συνδυασμός των λέξεων που περιέχει η φράση και τον σκοπό για τον οποίο η συγκεκριμένη φράση διατυπώθηκε.

Σε ένα ολοκληρωμένο σύστημα επεξεργασίας φυσικής γλώσσας η λειτουργία των δύο υποσυστημάτων εναλλάσσεται καθώς οι αλγόριθμοι κατανοούν μια φράση και δημιουργούν μια κατάλληλη απάντηση προς αυτήν. Τα συστήματα αυτά είναι άκρως σημαντικά στην ανάπτυξη του πεδίου της ψηφιακής επιδημιολογίας. Όπως προαναφέραμε τα ‘’μεγάλα’’ δεδομένα προέρχονται κυρίως από ηλεκτρονικά μέσα κοινωνικής δικτύωσης και ηλεκτρονικά μέσα μαζικής ενημέρωσης, και έτσι πολύ μέρος τους αποτελείται από ηλεκτρονικά μηνύματα και άλλα δεδομένα ανθρώπινου λόγου. Συνεπώς, για την επεξεργασία και την κατανόησή τους από τα υπολογιστικά συστήματα η επεξεργασία φυσικής γλώσσας είναι απαραίτητη. [35][36][37]

2.4 Υποστήριξη υλικού (hardware) για την αξιοποίηση των ‘’μεγάλων’’ δεδομένων

Σήμερα τα ‘’μεγάλα’’ δεδομένα έχουν εισχωρήσει και αξιοποιούνται από σχεδόν όλους τους τομείς της βιομηχανίας και των κυβερνήσεων για διάφορους σκοπούς. Η σωστή εκμετάλλευση των δεδομένων αυτών μπορεί να έχει μεγάλο κέρδος το οποίο μπορεί να είναι οικονομικό, πολιτικό ή η βελτίωση της ποιότητας ζωής. Φυσικά, για την απόκτηση αυτού του κέρδους πρέπει πρώτα να γίνει δυνατή η εκμαίευσή του από τους διάφορους τύπους ‘’μεγάλων’’ δεδομένων και ενώ το απαραίτητο λογισμικό υπάρχει, εξελίσσεται συνεχώς και η χρήση του δεν είναι ιδιαίτερα ακριβή, δεν μπορούμε να πούμε το ίδιο για την υποστήριξη υλικού που απαιτείται.

Η αξιοποίηση των ‘’μεγάλων’’ δεδομένων απαιτεί ισχυρά υπολογιστικά συστήματα με μεγάλες ταχύτητες επεξεργασίας και πολύ μεγάλους χώρους αποθήκευσης με υψηλές ταχύτητες μεταφοράς δεδομένων. Υπάρχουν δυο βασικές προσεγγίσεις στην ανάλυση ‘’μεγάλων’’ δεδομένων. Η πρώτη κάνει χρήση εξειδικευμένων αλγορίθμων για την ανάλυση ροών δεδομένων ενώ η δεύτερη βασίζεται στην ταυτόχρονη επεξεργασία μεγάλων όγκων ομαδοποιημένων δεδομένων. Και οι δύο αυτές προσεγγίσεις εκμεταλλεύονται τις δυνατότητες που προσφέρουν τα διανεμημένα συστήματα υπολογιστών.

Τα συστήματα που χρησιμοποιούνται για αυτή την ανάλυση πρέπει συνεχώς να εξελίσσονται παράλληλα με την αύξηση των ‘’μεγάλων’’ δεδομένων, έτσι ώστε να μπορούν να διατηρούν σταθερή υψηλή απόδοση κατά την επεξεργασία τους. Για τη λύση αυτού του προβλήματος χρησιμοποιούνται κάποιες διαφορετικές προσεγγίσεις που θα αναλυθούν στη συνέχεια. [10][38]

2.4.1 Κλιμάκωση (Scaling)

Επειδή τα ‘’μεγάλα’’ δεδομένα που είναι διαθέσιμα για συλλογή και επεξεργασία αυξάνονται συνεχώς με πολύ μεγάλη ταχύτητα, αντίστοιχα αυξάνονται και οι απαιτήσεις για τα συστήματα που χρησιμοποιούνται για την επεξεργασία και αποθήκευσή τους. Η κατασκευή όμως τέτοιου είδους συστημάτων είναι μια ιδιαίτερα δαπανηρή διαδικασία οπότε οι φορείς και οι επιχειρήσεις που χρησιμοποιούν ‘’μεγάλα’’ δεδομένα δεν έχουν τη δυνατότητα να αλλάζουν τα συστήματά τους ανά τα τακτά χρονικά διαστήματα, κάτι που απαιτεί η αξιοποίηση των ‘’μεγάλων’’ δεδομένων. Αυτό το πρόβλημα λύνεται με τη χρήση της κλιμάκωσης.

Η κλιμάκωση ενός συστήματος είναι μία διαδικασία η οποία αυξάνει την επεξεργαστική ισχύ ενός συστήματος και/ή τους χώρους αποθήκευσης του προσθέτοντας νέα υποσυστήματα σε ένα

ήδη υπάρχον σύστημα, με την προϋπόθεση πως αυτό έχει τη δυνατότητα να κλιμακωθεί. Τα περισσότερα υψηλού κόστους συστήματα σήμερα κατασκευάζονται με τέτοιο τρόπο ώστε να έχουν τη δυνατότητα μελλοντικής κλιμάκωσης. Υπάρχουν δύο διαφορετικά είδη κλιμάκωσης που μπορούν να γίνουν σε ένα σύστημα που αναλύονται στη συνέχεια.

- *Κάθετη Κλιμάκωση (Vertical Scaling)*:

  Η κάθετη κλιμάκωση, η οποία πολλές φορές αναφέρεται ως ‘’scale up’’ (κλιμάκωση προς τα πάνω), περιλαμβάνει την εγκατάσταση περισσότερων επεξεργαστών, μνήμης (RAM) και καρτών γραφικών (GPU) σε μία υπολογιστική μονάδα (π.χ. έναν server) καθώς και περισσότερου χώρου αποθήκευσης. Έτσι επιτυγχάνεται η αύξηση της υπολογιστικής δύναμής και ταχύτητας της μονάδας.

- *Οριζόντια κλιμάκωση (Horizontal Scaling)*:

  Η οριζόντια κλιμάκωση περιλαμβάνει την εγκατάσταση περισσότερων υπολογιστικών μονάδων που συνδέονται με την προϋπάρχουσα μονάδα. Η διασύνδεση των συστημάτων αυτών μπορεί να γίνει είτε μέσω καλωδίων (π.χ. δίκτυα LAN) είτε μέσω του διαδικτύου. Έτσι μοιράζεται ο συνολικός φόρτος εργασίας ανάμεσα στα συστήματα που λειτουργούν παράλληλα και επιτυγχάνεται ταχύτερη επεξεργασία των δεδομένων. Εξ ου τα συστήματα αυτά ανήκουν στην κατηγορία των διανεμημένων συστημάτων.

Και οι δύο παραπάνω μέθοδοι βοηθούν στην επίλυση του ίδιου προβλήματος με διαφορετικές στρατηγικές. Στην πράξη όμως το ποια είναι καλύτερο να χρησιμοποιηθεί έχει να κάνει με τις λειτουργίες που επιτελεί το συγκεκριμένο σύστημα και τον σκοπό για τον οποίο θα χρησιμοποιηθεί. Στον επόμενο, λοιπόν, πίνακα παρουσιάζουμε τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου. [13][39][40][41]

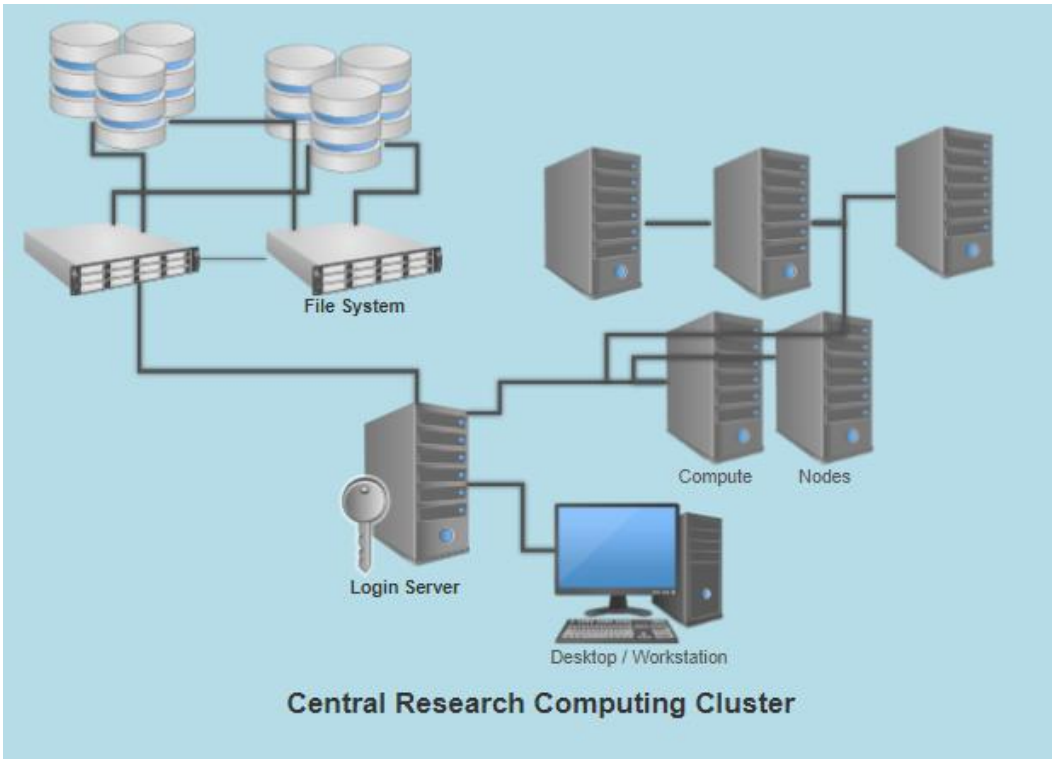
Κλιμάκωση	**Πλεονεκτήματα**	**Μειονεκτήματα**
Οριζόντια κλιμάκωση	→ Αυξάνει την επίδοση του συστήματος με μικρά βήματα και όσο χρειάζεται	→ Απαιτείται κατάλληλο λογισμικό για τη διαχείριση του διαμοιρασμού των δεδομένων ανάμεσα στα συστήματα και την παράλληλη επεξεργασία
	→ Το οικονομικό κόστος της κλιμάκωσης είναι σχετικά μικρότερο	→ Υπάρχει μικρός αριθμός λογισμικών που μπορούν να εκμεταλλευτούν πλήρως τις δυνατότητες της οριζόντιας κλιμάκωσης
	→ Η κλιμάκωση του συστήματος μπορεί να γίνει σε οποιοδήποτε βαθμό επιθυμείται χωρίς περιορισμό	

Κάθετη κλιμάκωση	→ Τα περισσότερα λογισμικά μπορούν να εκμεταλλευτούν εύκολα τους νέους πόρους του συστήματος → Η εγκατάσταση και η διαχείριση του νέου υλικού στο σύστημα είναι ιδιαίτερα εύκολη	→ Απαιτεί σημαντική οικονομική επένδυση → Η κλιμάκωση του συστήματος πρέπει είναι μεγαλύτερη από την απαιτούμενη για να μπορεί το σύστημα να επεξεργάζεται και μελλοντικό αυξημένο φόρτο εργασίας, με αποτέλεσμα αρχικά οι επιπλέον δυνατότητες να μένουν αχρησιμοποίητες → Δεν είναι δυνατή η κλιμάκωση του συστήματος πάνω από ένα συγκεκριμένο όριο

Πίνακας 2.1: Τα πλεονεκτήματα και τα μειονεκτήματα της κάθετης και της οριζόντιας κλιμάκωσης. [40]

2.4.2 Πλατφόρμες και εργαλεία οριζόντιας κλιμάκωσης

Όπως ήδη αναφέραμε η οριζόντια κλιμάκωση ενός υπολογιστικού συστήματος είναι μία διαδικασία η οποία συνδέει το αρχικό σύστημα με άλλα συστήματα και διαμοιράζει το φόρτο εργασίας μεταξύ τους. Έτσι, η συνολική απόδοση και ταχύτητα επεξεργασίας των δεδομένων αυξάνεται όσο επιθυμούμε. Λόγω του γεγονότος ότι η οριζόντια κλιμάκωση μπορεί να γίνει σε όποιο βαθμό θέλουμε, υπάρχουν αρκετές διαφορετικές πλατφόρμες-τεχνικές οριζόντιας κλιμάκωσης. Οι πλατφόρμες αυτές αφορούν, κυρίως, διασυνδεμένα συστήματα που η απόσταση μεταξύ τους είναι πεπερασμένη και ονομάζονται Clusters. Τα δίκτυα αυτά έχουν κάποιες βασικές διαφορές με τα δίκτυα Cloud, που και αυτά αποτελούν μια μέθοδο οριζόντιας κλιμάκωσης, η λειτουργία και οι πλατφόρμες των οποίων θα περιγραφούν σε επόμενη ενότητα. Η επεξεργασία δεδομένων με χρήση δικτύων συστημάτων Cluster ονομάζεται Cluster Computing. Μερικές από τις πιο βασικές πλατφόρμες οριζόντιας κλιμάκωσης για τη δημιουργία δικτύων επεξεργασίας Cluster αναλύονται στη συνέχεια. [40]



Σχήμα 2.5: Ένα δίκτυο Cluster. Στο δίκτυο περιλαμβάνονται τα συστήματα αρχείων και αποθήκευσης (File System) οι μονάδες επεξεργασίας (Compute Nodes), ο κεντρικός κόμβος του δικτύου (Login Server) καθώς και το σύστημα διαχείρισης του δικτύου από τον χρήστη (Desktop / Workstation). Η αποστάσεις μεταξύ των διαφόρων συσκευών του δικτύου είναι πεπερασμένες. [42]

2.4.2.1 Peer to Peer Networks

Αυτού το είδους τα δίκτυα περιλαμβάνουν χιλιάδες συστήματα συνδεδεμένα στο ίδιο δίκτυο. Η αρχιτεκτονική δικτύου σε αυτή την περίπτωση βασίζεται στην αποκέντρωση και τον διαμοιρασμό των δεδομένων και τη χρήση του κάθε συστήματος για την επεξεργασία ενός μέρους από τα συνολικά δεδομένα. Η πλατφόρμα αυτή είναι ιδιαίτερα παλιά αλλά επιτρέπει τη σύνδεση ακόμα και εκατομμυρίων συστημάτων ταυτόχρονα, με το μόνο πρόβλημα την επικοινωνία μεταξύ των συστημάτων του δικτύου. Η γλώσσα παράλληλου προγραμματισμού που χρησιμοποιείται σε αυτά τα δίκτυα, για το διαμοιρασμό των δεδομένων στα διάφορα συστήματα και την επικοινωνία των συστημάτων, είναι η MPI. Στον παρακάτω πίνακα παρουσιάζονται συνοπτικά τα βασικά πλεονεκτήματα και μειονεκτήματα της χρήσης της MPI στην ανάλυση ''μεγάλων'' δεδομένων. Σημειώνουμε πως η MPI δεν χρησιμοποιείται σήμερα σε τέτοιου είδους συστήματα. [40][43]

	Πλεονεκτήματα	**Μειονεκτήματα**
MPI	→ Δεν υπάρχει ανάγκη μελλοντικής ανάγνωσης ήδη αναγνωσμένων δεδομένων. Έτσι μπορούν να χρησιμοποιηθούν τεχνικές επαναληπτικής επεξεργασίας → Μπορεί να χρησιμοποιηθεί λειτουργία ιεραρχίας αφέντη-σκλάβου όπου κάποιοι κόμβοι αναλαμβάνουν το ρόλο του αφέντη και οι υπόλοιποι του σκλάβου για ταχύτερη επεξεργασία δεδομένων. Αυτό είναι ιδιαίτερα χρήσιμο στο δυναμικό καταμερισμό πόρων όταν οι κόμβοι-σκλάβοι χρειάζεται να επεξεργαστούν μεγάλους όγκους δεδομένων → Είναι διαθέσιμη σε πολλές γλώσσες προγραμματισμού → Διαθέτει χρήσιμες λειτουργίες για την αποστολή δεδομένων και τον συγχρονισμό των αποστολών ανάμεσα στα διάφορα συστήματα του δικτύου	→ Δεν έχει κάποιο μηχανισμό για τη διαχείριση των σφαλμάτων. Όταν χρησιμοποιείται σε Peer to Peer δίκτυα, τα οποία θεωρούνται ιδιαίτερα αναξιόπιστα, ένα σφάλμα σε ένα σύστημα του δικτύου μπορεί να οδηγήσει στην απενεργοποίηση όλου του δικτύου

Πίνακας 2.2: Πλεονεκτήματα και μειονεκτήματα χρήσης της MPI στην ανάλυση μεγάλων δεδομένων. [40]

2.4.2.2 Apache Hadoop

Το Apache Hadoop είναι ένα ελεύθερο για χρήση πλαίσιο για την αποθήκευση και την επεξεργασία μεγάλων όγκων δεδομένων χρησιμοποιώντας ομάδες καταναλωτικών συστημάτων,

όπως προσωπικούς υπολογιστές και servers. Είναι ειδικά σχεδιασμένο για να λειτουργεί με χρήση των προτύπων HDFS και MapReduce που θα αναλυθούν στη συνέχεια.

Ένα βασικό χαρακτηριστικό του Hadoop είναι πως προσφέρει στους χρήστες τη δυνατότητα υψηλής κλιμάκωσης αυξάνοντας τον συνολικό αποθηκευτικό χώρο, το εύρος του διαύλου εισόδου-εξόδου και την συνολική επεξεργαστική ισχύ του δικτύου προσθέτοντας σε αυτό νέες μονάδες, όποτε αυτό είναι απαραίτητο. Έτσι οι χρήστες του μπορούν να εκμεταλλευτούν τους πόρους από χιλιάδες μεμονωμένα συστήματα ταυτόχρονα. Επίσης, έχει την δυνατότητα της διαίρεσης των διεργασιών που πρόκειται να εκτελεστούν σε τμήματα και τον διαμοιρασμό των τμημάτων αυτών στα διάφορα διασυνδεμένα συστήματα για παράλληλη εκτέλεση, καθώς επίσης εκτελεί πολύ αποτελεσματικά τεχνικές επεξεργασίας ομαδοποιημένων δεδομένων πολύ μεγάλου όγκου. Επιπλέον ένα χαρακτηριστικό του Hadoop που το κάνει ιδιαίτερα δημοφιλές είναι η παράλληλη εκτέλεση των διαφόρων διεργασιών κοντά στα δεδομένα τους. Αυτό σημαίνει πως οι διεργασίες εκτελούνται στα ίδια συστήματα που περιλαμβάνουν και τα δεδομένα προς επεξεργασία και έτσι δεν υπάρχει λόγος μεταφοράς των δεδομένων σε κάποιο άλλο σύστημα, το οποίο αποτελεί μια χρονοβόρα διαδικασία. Τέλος, μία άλλη ιδιότητα του Hadoop είναι η πολύ μεγάλη αντοχή της πλατφόρμας σε σφάλματα, που πηγάζει από τη χρήση των συστημάτων HDFS και MapReduce.

Το Hadoop περιέχει δύο βασικά δομικά στοιχεία:

- *Distributed File System (HDFS)*: Το HDFS είναι επίπεδο αποθήκευσης βασισμένο σε UNIX που αποτελεί το σύστημα αποθήκευσης δεδομένων του Hadoop. Ουσιαστικά είναι ένα σύστημα διαμοιρασμού αρχείων που χρησιμοποιείται για την αποθήκευση αρχείων στα διασυνδεδεμένα συστήματα μιας ομάδας συστημάτων, ενώ ταυτόχρονα επιτρέπει ταχύτατη διαθεσιμότητα αυτών των αρχείων για χρήση.

  Ένα πολύ βασικό χαρακτηριστικό του συστήματος HDFS που το κάνει ιδιαίτερα χρήσιμο είναι ότι μπορεί να λειτουργήσει σε καταναλωτικά συστήματα τα οποία δεν χρειάζεται να είναι ειδικά σχεδιασμένα. Επίσης, διαθέτει υψηλή ανθεκτικότητα σε σφάλματα, έχει τη δυνατότητα να διαχειρίζεται πολύ μεγάλους όγκους δεδομένων και μπορεί να εγγράφει αρχεία μόνο με μία και μόνο πρόσβαση στο αποθηκευτικό μέσο, κάτι που εξηγεί την υψηλή ταχύτητα λειτουργίας του. Τέλος, μπορεί να λειτουργήσει και με βάση την ιεραρχία αφέντη-σκλάβου που είναι πολύ χρήσιμο σε ορισμένες περιπτώσεις.

- *Hadoop YARN*: Το YARN είναι ένα επίπεδο που διαχειρίζεται τους συνολικούς υπολογιστικούς πόρους των διασυνδεμένων συστημάτων, ενώ επίσης αναλαμβάνει τη χρονοδρομολόγηση των διαφόρων διεργασιών που εκτελούνται από αυτά τα συστήματα.

Τα πιο βασικά πλεονεκτήματα και μειονεκτήματα του Hadoop είναι:

■ ***Πλεονεκτήματα:***

- <u>*Το εύρος των πηγών δεδομένων*</u>: Γενικά το εύρος των διαφόρων πηγών δεδομένων είναι πάρα πολύ μεγάλο και τα δεδομένα που προέρχονται από αυτές τις πηγές μπορεί να έχουν διάφορες μορφές δομημένες ή και όχι. Το Hadoop μπορεί να επεξεργαστεί

δεδομένα κάθε μορφής χωρίς να απαιτείται αρχικά η μετατροπή τους σε έναν συγκεκριμένο τύπο, κάτι που συνήθως είναι αρκετά χρονοβόρο.

- *Οικονομικότητα*: Οι διάφορες εταιρίες συχνά επενδύουν χρήματα για την κατασκευή εγκαταστάσεων για την αποθήκευση των δεδομένων που χρειάζονται. Η συνεχής επέκταση τέτοιων εγκαταστάσεων, όμως, δεν είναι οικονομική. Πλέον είναι αρκετά πιο οικονομικό να αποθηκεύουν τα δεδομένα τους ή παλιά δεδομένα που θέλουν να διατηρήσουν στο Hadoop, στο οποίο δεν υπάρχει επίσης ο κίνδυνος απώλειάς τους. Έτσι ακόμα και δεδομένα τα οποία θα διαγράφονταν σε άλλη περίπτωση, για τη δημιουργία αποθηκευτικού χώρου, μπορούν πλέον να διατηρηθούν και να ανακτηθούν όποτε αυτό είναι επιθυμητό.

- *Ταχύτητα*: Στόχος κάθε οργανισμού ή εταιρίας σήμερα είναι η διεκπεραίωση των καθηκόντων τους στο λιγότερο δυνατό χρόνο. Η επεξεργασία των απαραίτητων δεδομένων όμως είναι συχνά μια χρονοβόρα διαδικασία. Με την αποθήκευση των δεδομένων τους στο Hadoop, οι διάφοροι φορείς, μπορούν να τα επεξεργαστούν με αρκετά αυξημένη ταχύτητα. Αυτό συμβαίνει διότι, συνήθως, όταν τα δεδομένα προς επεξεργασία βρίσκονται στο ίδιο σύστημα με τα εργαλεία που χρησιμοποιούνται για την επεξεργασία τους, η ταχύτητα επεξεργασίας των δεδομένων αυξάνεται.

- *Πολλαπλά αντίγραφα*: Το Hadoop παρέχει προστασία ως προς τα δεδομένα που είναι αποθηκευμένα σε αυτό. Ένας όγκος δεδομένων δεν μπορεί να χαθεί εκτός αν διαγραφεί μετά από επίσημη εντολή του χρήστη. Επίσης, για περαιτέρω αύξηση της ασφάλειας των δεδομένων, όταν αυτά βρίσκονται αποθηκευμένα στο Hadoop, μπορούν να δημιουργηθούν πολλαπλά αντίγραφα αυτών και να παραμείνουν και αυτά αποθηκευμένα στην πλατφόρμα.

■ ***Μειονεκτήματα:***

- *Δεν υπάρχει προεπιλεγμένη ασφάλεια των δεδομένων*: Πολλά από τα δεδομένα που αποθηκεύονται στο Hadoop αποτελούν πνευματική ιδιοκτησία μιας εταιρίας ή/και περιέχουν ευαίσθητο περιεχόμενο που πρέπει να προστατευτεί. Το Hadoop δε διαθέτει κάποια προεπιλεγμένη ασφάλεια για τα δεδομένα που αποθηκεύονται στην πλατφόρμα και είναι καθαρά επιλογή του προγραμματιστή που αποθηκεύει τα δεδομένα να ορίσει κάποια σχετική ασφάλεια.

- *Δεν είναι αποδοτικό στην επεξεργασία μικρών όγκων δεδομένων*: Αν και οι περισσότερες πλατφόρμες που χρησιμοποιούνται για την επεξεργασία ‘’μεγάλων’’ δεδομένων μπορούν να επεξεργαστούν εξίσου αποδοτικά οποιαδήποτε όγκο δεδομένων, το Hadoop δεν είναι ιδιαίτερα αποδοτικό στην επεξεργασία μικρών όγκων δεδομένων. Εξαιτίας αυτού χρησιμοποιείται κυρίως από μεγάλες εταιρίες και οργανισμούς που διαθέτουν πολύ μεγάλους όγκους δεδομένων προς επεξεργασία.

- *Επικινδυνότητα για παραβιάσεις*: Μια από τις πιο συχνά χρησιμοποιούμενες προγραμματιστικές γλώσσες είναι η Java η οποία συναντάται και στο Hadoop. Το

Hadoop είναι εξολοκλήρου κατασκευασμένο σε Java. Η Java είναι μία γλώσσα ιδιαίτερα ευάλωτη σε επιθέσεις και παραβιάσεις κάτι που κάνει την πλατφόρμα του Hadoop όχι ιδιαίτερα ασφαλή. [10][40][44][45]

2.4.2.3 MapReduce

Το MapReduce είναι ένα προγραμματιστικό μοντέλο που χρησιμοποιείται στο Hadoop και σε άλλες εφαρμογές ενώ παράλληλα αναπτύσσονται και νέες υπορουτίνες που το χρησιμοποιούν. Ο βασικός σκοπός της χρήσης του MapReduce είναι για την επεξεργασία ''μεγάλων'' δεδομένων.

Το MapReduce ουσιαστικά είναι μια στρατηγική επεξεργασίας δεδομένων οι οποία χωρίζει όλες τις διεργασίες του συστήματος σε δύο μέρη που ονομάζονται mappers και reducers, και στις περισσότερες εφαρμογές του χρησιμοποιείται σε συνδυασμό με το σύστημα HDFS. Οι mappers, που λειτουργούν σε υψηλό επίπεδο, διαβάζουν τα δεδομένα από το σύστημα HDFS, τα επεξεργάζονται και παράγουν κάποια ενδιάμεσα αποτελέσματα που τροφοδοτούν στους reducers. Οι reducers με τη σειρά τους συσσωρεύουν αυτά τα ενδιάμεσα αποτελέσματα και τα ενοποιούν παράγοντας έτσι την τελική έξοδο η οποία αποθηκεύεται ξανά στο σύστημα HDFS. Αυτός ο τρόπος επεξεργασίας εφαρμόζεται παράλληλα στα διάφορα διασυνδεμένα συστήματα του δικτύου και έτσι επιτυγχάνεται υψηλή ταχύτητα επεξεργασίας μεγάλου όγκου δεδομένων. Αξίζει να αναφέρουμε πως για μία συνολική εργασία του Hadoop δημιουργούνται αρκετοί mappers και reducers στα διάφορα διασυνδεμένα συστήματα.

Κάποιες υπορουτίνες του MapReduce που χρησιμοποιούνται ευρέως είναι το Apache Pig το οποίο είναι ένα περιβάλλον παρόμοιο με SQL που κατασκευάστηκε από τη Yahoo και το Hive που κατασκευάστηκε από την Facebook. Οι δύο αυτές υπορουτίνες παρέχουν ένα καλύτερο προγραμματιστικό περιβάλλον κατά τη δημιουργία προγραμμάτων καθώς οι προγραμματιστές δεν χρειάζεται να ασχοληθούν άμεσα με την πολύπλοκη δομή του MapReduce.

Το MapReduce φυσικά έχει και κάποιους περιορισμούς ο πιο σημαντικός από τους οποίους είναι η ανικανότητά του να χρησιμοποιεί αποδοτικά επαναληπτικούς αλγορίθμους και γενικώς επαναληπτικές διαδικασίες, που αποτελούν μια πολύ χρήσιμη λειτουργία στην επεξεργασία ''μεγάλων'' δεδομένων. Έτσι αναγκαστικά οι mappers κάνουν ανάγνωση των ίδιων δεδομένων ξανά και ξανά και σε κάθε επανάληψη μιας επαναλαμβανόμενης διαδικασίας και τα δεδομένα πρέπει να αποθηκευτούν στον δίσκο έτσι ώστε να είναι διαθέσιμα στην επόμενη επανάληψη. Αυτή η επανάληψη αναγνώσεων και εγγραφών στο δίσκο δημιουργεί σημαντικές καθυστερήσεις και επηρεάζει σε μεγάλο βαθμό τη συνολική απόδοση. Τέλος, ένα άλλο σημαντικό πρόβλημα του MapReduce είναι ότι δημιουργούνται mappers και reducers σε κάθε νέα εργασία του συστήματος. Αυτό συμβαίνει ακόμα και σε εργασίες που έχουν πολύ μικρή διάρκεια με αποτέλεσμα να σπαταλιέται περισσότερος χρόνος όταν για την εργασία χρησιμοποιούνται mappers και reducers απ' ότι θα χρειαζόταν αν δεν χρησιμοποιούνταν αυτή η τεχνική. [40][46][47]

2.4.2.4 Microsoft Dryad

Το Dryad ήταν ένα πλαίσιο μεγάλης κλίμακας κατασκευασμένο για εντατική επεξεργασία δεδομένων. Δημιουργήθηκε από τη Microsoft αλλά η λειτουργία του διακόπηκε το 2011. Αναπτύχθηκε για ένα ευρύ φάσμα εφαρμογών όπως οι ομαδοποιημένες εφαρμογές (batch applications) και η ροές δεδομένων (data streaming). Σε αντίθεση με το MapReduce που στοχεύει στην παροχή απεριόριστης πρόσβασης με εύκολο τρόπο στους προγραμματιστές το Dryad δίνει

πλήρη έλεγχο στο γράφο επικοινωνίας των συστημάτων καθώς και στις υπορουτίνες που βρίσκονται στις κορυφές του. Έτσι, ενώ με τη χρήση του MapReduce θυσιάζονται οι υψηλές επιδόσεις, ως αποτέλεσμα της επιδίωξης της απλότητας χειρισμού, το Dryad, μέσω των πολλών δυνατοτήτων που προσφέρει, επιτρέπει τη δημιουργία δικτύων συστημάτων υψηλών επιδόσεων.

Το Dryad δημιουργήθηκε με σκοπό τη μετάβαση από τα μεμονωμένα πολυπύρηνα συστήματα στα κέντρα δεδομένων που περιλαμβάνουν αρκετά υπολογιστικά συστήματα, τα οποία συνδέονται μεταξύ τους μέσω ενός δικτύου. Κατά την περίοδο λειτουργίας του, όλες οι δυσκολίες που προέκυπταν κατά το σχεδιασμό και την κατασκευή ενός κέντρου δεδομένων μπορούσαν να επιλυθούν από το Dryad. Τέτοιες δυσκολίες περιλαμβάνουν τη χρονοδρομολόγηση της χρήσης των υπολογιστών και των CPU του δικτύου, την επαναφορά του δικτύου μετά από σφάλματα επικοινωνίας καθώς και τη μεταφορά των δεδομένων μεταξύ των συστημάτων του δικτύου.

Η αξιοποίηση των δυνατοτήτων του Dryad γίνεται μέσω της χρήσης του DryadLINQ. Το DryadLINQ είναι μια ομάδα επεκτάσεων γλωσσών προγραμματισμού που επιτρέπει τον προγραμματισμό μεγάλης κλίμακας πάνω σε διανεμημένα δεδομένα. Έτσι, τα μέρη ενός προγράμματος που μπορούν να υποστούν παράλληλη επεξεργασία μπορούν να μεταφραστούν αυτόματα και με πλήρη διαφάνεια σε μια μορφή αναγνωρίσιμη από το Dryad και στη συνέχεια να εκτελεστούν παράλληλα από τα συστήματα του δικτύου. [10][48]

2.4.2.5 Apache Spark

Το Spark είναι η πιο δημοφιλής πλατφόρμα για την επεξεργασία ‘’μεγάλων’’ δεδομένων και αποτελεί υπόδειγμα για τη μελλοντική σχεδίαση αντίστοιχων συστημάτων. Αποτελεί μια πλατφόρμα που σχεδιάστηκε από τον Matei Zaharia στο UC Berkeley το 2009, ενώ το 2013 δωρίστηκε στο Apache Software Foundation. Αποτελεί μια διαφορετική πλατφόρμα από το Hadoop και γι’ αυτό το λόγο δεν πρέπει να συγχέονται αυτές οι δύο πλατφόρμες καθώς πλέον συναντάμε και εφαρμογές που γίνεται χρήση του Spark μέσω της πλατφόρμας του Hadoop.

Το Spark είναι ένα ελεύθερης χρήσης πλαίσιο που το βασικότερο χαρακτηριστικό του είναι η δυνατότητα για εντός της μνήμης RAM αποθήκευσης δεδομένων και η παράλληλη επεξεργασία αυτών από ένα δίκτυο συστημάτων (in-memory cluster computing). Προφέρει μια απλή διεπαφή για τη δημιουργία προγραμμάτων όπου οι προγραμματιστές μπορούν να χρησιμοποιήσουν εύκολα τη CPU, τη μνήμη RAM και τα μέσα αποθήκευσης των διασυνδεμένων συστημάτων για την επεξεργασία μεγάλων ομάδων δεδομένων τα οποία βρίσκονται διαμοιρασμένα στα συστήματα του δικτύου.

Ένα από τα βασικά στοιχεία του Spark είναι η χρήση του συστήματος RDD (Resilient Distributed Datasheets) το οποίο λειτουργεί σε συνδυασμό με το σύστημα HDFS και μπορεί να αποθηκεύσει δεδομένα και να διαχειριστεί τα σφάλματα που μπορεί να προκύψουν σε ελάχιστο χρόνο δίχως επαναλήψεις. Επίσης, μία εφαρμογή που εκτελείται στο Spark μπορεί να αποθηκεύσει τα δεδομένα που χρειάζεται στη μνήμη της πλατφόρμας και έτσι να αποφευχθεί η πρόσβαση σε συμβατικά μέσα αποθήκευσης (HDD, SSD), επιταχύνοντας έτσι την εκτέλεσή της σε μεγάλο βαθμό. Επιπλέον, το Spark μπορεί να χρησιμοποιείται ταυτόχρονα για πολλές, διαφορετικής φύσης, διεργασίες όπως υπολογισμούς με χρήση γράφων, επεξεργασία ροών δεδομένων, διαδραστική ανάλυση, ανάλυση ομαδοποιημένων δεδομένων και μηχανική μάθηση.

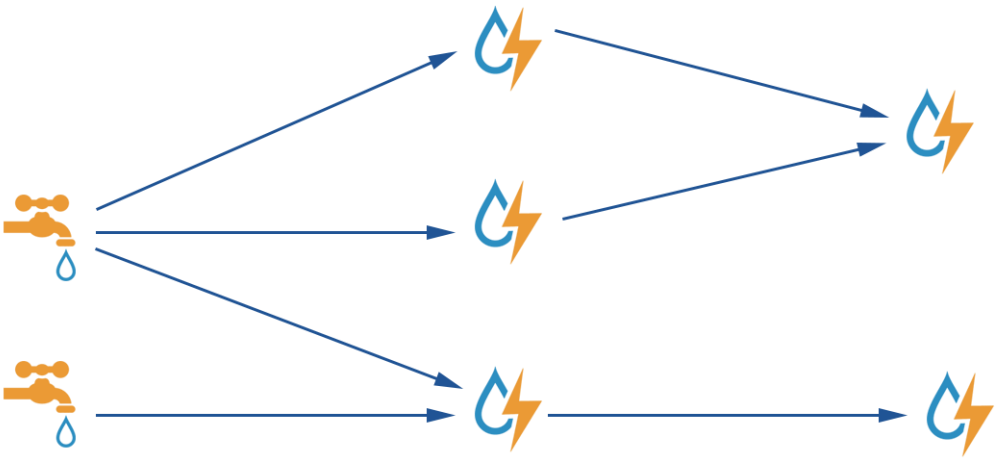
Όλα τα παραπάνω καθιστούν το Spark μία φιλική για το χρήστη, γρήγορη, ανθεκτική και κλιμακώσιμη πλατφόρμα που μπορεί να χρησιμοποιηθεί για πάρα πολλούς σκοπούς και έτσι δικαιολογείται ο υψηλός αριθμός χρηστών που διαθέτει. [10][40][49]

2.4.2.6 Apache Storm

Το Storm είναι ένα σύστημα επεξεργασίας ροών δεδομένων σε πραγματικό χρόνο και πλέον ανήκει στο ίδρυμα Apache. Η αρχική σχεδίασή του έγινε από την εταιρία BackType, μια εταιρία που ασχολούνταν με την ανάλυση δεδομένων ηλεκτρονικών κοινωνικών δικτύων, και αποτελούσε ένα προς ιδιωτική χρήση πρόγραμμα. Το 2011 η BackType εξαγοράστηκε από την Twitter η οποία βελτίωσε το Storm και το κατέστησε ελεύθερο προς χρήση. Η Apache απέκτησε το Storm 2014 και από τότε αποτελεί ένα από τα κορυφαίου επιπέδου προγράμματα του Ιδρύματος.

Το Storm είναι σχεδιασμένο έτσι ώστε να απλοποιεί την επεξεργασία των απεριόριστων ροών δεδομένων πραγματικού χρόνου και χρησιμοποιεί σε πραγματικό χρόνο τα εργαλεία που χρησιμοποιεί το Hadoop για την επεξεργασία ομαδοποιημένων δεδομένων. Είναι ιδιαίτερα προσιτό στους διάφορους χρήστες καθώς είναι απλό στη χρήση και μπορεί να λειτουργήσει χρησιμοποιώντας οποιαδήποτε προγραμματιστική γλώσσα. Το Storm έχει διαφορετική αρχιτεκτονική σε σχέση με άλλα συστήματα που βασίζονται στο πρότυπο MapReduce και χρησιμοποιεί διαφορετικά εργαλεία για την αντιμετώπιση σφαλμάτων, για την επικοινωνία των διασυνδεμένων συστημάτων και για τη διανομή των πόρων του δικτύου στις διάφορες διεργασίες. Όλα αυτά τα εργαλεία είναι ειδικά σχεδιασμένα για τη μεταφορά και επεξεργασία ροών δεδομένων πραγματικού χρόνου με τον πιο αποδοτικό τρόπο.

Στην αρχιτεκτονική του Storm οι εισερχόμενες ροές δεδομένων ονομάζονται spouts ενώ οι δομές που αναλαμβάνουν την επεξεργασία των δεδομένων ονομάζονται bolts. Τα bolts επεξεργάζονται τα δεδομένα τα οποία έχουν μετατραπεί σε ακολουθίες, που περιέχουν δεδομένα διαφόρων τύπων (tuples), οι οποίες προέρχονται είτε από τα spouts από είτε από άλλα bolts. Με αυτή τη λογική γίνεται η επεξεργασία των δεδομένων ασχέτως τύπου και μορφής. Κάθε διεργασία του Storm χαρακτηρίζεται ως ένας Κατευθυνόμενος Ακυκλικός Γράφος (DAG - Directed Acyclic Graph), που αποτελείται από spouts και bolts, και ονομάζεται τοπολογία. Τέλος, για την αποδοτική κατανομή των διεργασιών, το Storm χρησιμοποιεί έναν προεπιλεγμένο προγραμματιστή ο οποίος αναλαμβάνει την ταξινόμηση των τοπολογιών στις διαθέσιμες μονάδες για επεξεργασία με κυκλικό τρόπο, έτσι ώστε να μην υπερφορτώνεται κάποια μονάδα, ή μένει αχρησιμοποίητη, και διατηρώντας την επίδοση στο μέγιστο.



Σχήμα 2.6: Η λογική επεξεργασίας δεδομένων του Storm. Τα δεδομένα υπό τη μορφή των tuples μεταφέρονται από τα spouts (αριστερά) στα bolts (κέντρο και δεξιά) για επεξεργασία. [50]

Το Storm αρχικά σχεδιάστηκε ως μία ανεξάρτητη πλατφόρμα από το Hadoop αλλά πρόσφατα γίνονται προσπάθειες για την ενοποίηση αυτών των δύο συστημάτων με στόχο την εκμετάλλευση της συνδυαστική τους απόδοσης. [10][51]

2.4.2.7 Apache Kafka

Το Apache Kafka είναι μία ελεύθερη προς χρήση τεχνολογία για την επεξεργασία γεγονότων που κατασκευάστηκε από το Ίδρυμα Λογισμικού Apache (Apache Software Foundation). Η επεξεργασία γεγονότων είναι μια μέθοδος εντοπισμού και ανάλυσης ροών δεδομένων που προκύπτουν από ένα συμβάν με σκοπό την αναζήτηση των αιτιών που το προκάλεσαν και την εξαγωγή συμπερασμάτων. Το Kafka διαθέτει ένα επίπεδο αποθήκευσης δεδομένων, που είναι ιδιαίτερα κλιμακώσιμο, ενώ ταυτόχρονα μπορεί να διαχειρίζεται έναν μεγάλο αριθμό ροών δεδομένων πραγματικού χρόνου, κάτι που το κάνει ιδανικό για τέτοιες λειτουργίες.

Κάποια άλλα βασικά χαρακτηριστικά που διαθέτει το Kafka είναι πως μπορεί να διανεμηθεί, να διαιρεθεί ή ακόμη και να αναπαραχθεί με στόχο την αύξηση της ακρίβειάς του και γι' αυτό μπορεί να χρησιμοποιηθεί και ως σύστημα εγγραφών-δημοσιοποιήσεων για επικοινωνία μέσω μηνυμάτων. Επίσης, μπορεί να κλιμακωθεί με ευκολία προσθέτοντας νέες μονάδες σε μια ήδη υπάρχουσα ομάδα συστημάτων. Η ανθεκτικότητα των δεδομένων διασφαλίζεται μέσω της αποθήκευσής τους στα διαθέσιμα αποθηκευτικά μέσα και έτσι τα δεδομένα είναι πάντα έτοιμα για χρήση από τις διασυνδεμένες μονάδες της πλατφόρμας.

Το Kafka λειτουργεί ως μία ομάδα διασυνδεμένων μονάδων που επικοινωνούν μεταξύ τους μέσω μηνυμάτων όπου κάθε μονάδα αναφέρεται ως χρηματιστής (broker). Τα μηνύματα που μεταδίδονται μέσω του Kafka ομαδοποιούνται σε θέματα, και οποιαδήποτε εφαρμογή δημοσιεύει μηνύματα στο Kafka αναφέρεται ως παραγωγός (producers). Οι εφαρμογές ή οι διεργασίες που εγγράφονται σε θέματα του Kafka ονομάζονται καταναλωτές (consumers).

Η επεξεργασία γεγονότων και η ιδιαίτερη δομή λειτουργίας που διαθέτει το Kafka το καθιστούν ένα ιδιαίτερα χρήσιμο εργαλείο με πολλές προοπτικές για την μελέτη ξεσπασμάτων ασθενειών και γενικότερα για την ψηφιακή επιδημιολογία. [10][18]

2.4.2.8 Apache Mahout

Το Mahout είναι ένα ελεύθερο για χρήση πλαίσιο από την Apache. Κατασκευάστηκε με στόχο τη δημιουργία μιας βασικής βιβλιοθήκης αλγορίθμων που έχουν σχέση με τη μηχανική μάθηση και την εξόρυξη δεδομένων (data mining) και λειτουργούσε κατά βάση, στο παρελθόν, πάνω στην πλατφόρμα του Hadoop. Όπως και το Hadoop χρησιμοποιεί τις τεχνολογίες HDFS MapReduce με αποτέλεσμα να έχει ιδιαίτερα καλή απόδοση και σταθερότητα σε λειτουργίες όπως η διαχείριση μεγάλου όγκου δεδομένων, σε αντίθεση με πολλές άλλες πλατφόρμες που αποτυγχάνουν σε αυτόν τον τομέα.

Σήμερα το Mahout συναντάται πιο συχνά σε χρήσεις σε συνδυασμό με το Spark. Το πιο βασικό χαρακτηριστικά του είναι η δυνατότητα υψηλής κλιμακωσιμότητας των αλγορίθμων που αποθηκεύει. Ένα άλλο χαρακτηριστικό που διαθέτει είναι το γεγονός πως λειτουργεί εξίσου καλά σε πλατφόρμες διανεμημένων συστημάτων όσο και σε μεμονωμένα συστήματα.

Οι αλγόριθμοι που αποθηκεύονται στο Mahout αφορούν κυρίως τη λύση μαθηματικών και στατιστικών προβλημάτων, ιδίως στον τομέα της γραμμικής άλγεβρας, και, εξαιτίας αυτού, χρησιμοποιείται κυρίως από επιστήμονες που ασχολούνται με αυτά τα πεδία. [10][52]

2.4.2.9 Apache Flink

Το Flink είναι ένα πλαίσιο διανεμημένης επεξεργασίας ροών δεδομένων (streaming) ελεύθερης χρήσης. Κατασκευάστηκε από την Apache για να είναι πλήρως συμβατό με το Hadoop και είναι σχεδιασμένο για χρήση από απαιτητικές, διανεμημένες, υψηλής επίδοσης και αδιάκοπης λειτουργίας εφαρμογές επεξεργασίας ροών δεδομένων.

Ένα πολύ βασικό χαρακτηριστικό του Flink είναι η υψηλή ανθεκτικότητα και η ταχύτατη επαναφορά του συστήματος μετά από σφάλματα. Επίσης, οι διάφορες διεργασίες του Flink μπορούν να εκτελούνται ταυτόχρονα σε πολλά διασυνδεμένα συστήματα διατηρώντας τις καθυστερήσεις σε χαμηλό επίπεδο, ενώ η ταχύτητα διακίνησης των δεδομένων παραμένει υψηλή. Όπως και το Spark είναι ιδανικό για την εκτέλεση αλγορίθμων μηχανικής μάθησης και εξόρυξης δεδομένων, που βασίζονται σε επαναληπτικές διαδικασίες, ενώ, επίσης, διαχειρίζεται άριστα εντατικές εφαρμογές επεξεργασίας δεδομένων τα οποία είναι διανεμημένα στα συστήματα του δικτύου. Τέλος, υποστηρίζει και εντός της μνήμης RAM αποθήκευσης και παράλληλης επεξεργασίας δεδομένων (in-memory cluster computing), η οποία λειτουργία μειώνει κατά πολύ τον χρόνο εκτέλεσης των διεργασιών αφού αποφεύγονται αναγνώσεις και εγγραφές στο μέσο αποθήκευσης, το οποίο αποτελεί μια πολύ χρονοβόρα διαδικασία.

Για την ταχύτερη λειτουργία του, το Flink, εφαρμόζει ειδικούς μετασχηματισμούς στα σύνολα των παράλληλων δεδομένων που βρίσκονται αποθηκευμένα στα συστήματα του δικτύου, ενώ ταυτόχρονα βελτιστοποιεί τον γράφο του δικτύου των συστημάτων λαμβάνοντας υπ’ όψιν την τοπικότητα των δεδομένων, δηλαδή σε ποια συστήματα υπάρχουν αποθηκευμένα δεδομένα σχετικά με την παρούσα διεργασία. Οι μετασχηματισμοί αυτοί διευκολύνουν την εκτέλεση διεργασιών στα συγκεκριμένα δεδομένα και έτσι αποφεύγονται τυχόν καθυστερήσεις. Κλείνοντας αναφέρουμε πως το Flink είναι πλήρως συμβατό με το MapReduce, εργαλεία του οποίου μπορούν να χρησιμοποιηθούν όποτε αυτό είναι επιθυμητό. [10][53]

2.4.2.10 Yahoo! Simple Scalable Streaming System (S4)

Το Yahoo! S4 είναι μία γενικού σκοπού, διανεμημένη και κλιμακώσιμη πλατφόρμα για την επεξεργασία συνεχών απεριόριστων ροών δεδομένων. Η πλατφόρμα αυτή είναι εμπνευσμένη από τη λειτουργία του MapReduce και χρησιμοποιεί κάποια από τα χαρακτηριστικά του, στοχεύοντας όμως πάντα στην υψηλή απόδοση κατά την επεξεργασία ροών δεδομένων και όχι στην ομαδοποιημένη επεξεργασία που είναι το κύριο χαρακτηριστικό του MapReduce. Από αυτά τα κοινά χαρακτηριστικά προέρχεται και η μεγάλη αντοχή του S4 στα σφάλματα και η μεγάλη ταχύτητα επεξεργασίας.

Η ιδέα της σχεδίασης του S4 προέκυψε από της μηχανές αναζήτησης. Οι μηχανές αυτές χρησιμοποιούν αλγορίθμους μηχανικής μάθησης και εξόρυξης δεδομένων όχι μόνο για το κομμάτι της αναζήτησης αλλά και για την ιδανική απεικόνιση των διαφημίσεων, από τις οποίες προέρχονται και τα βασικά τους έσοδα. Το σε ποιο μέρος της οθόνης και πως και πότε θα απεικονιστεί μία διαφήμιση επηρεάζει σε μεγάλο βαθμό την πιθανότητα να επιλεχθεί για προβολή από το χρήστη. Τα στοιχεία που επεξεργάζονται οι μηχανές αναζητήσεων για την πρόβλεψη της ιδανικής απεικόνισης μιας διαφήμισης περιλαμβάνουν τις προηγούμενες αναζητήσεις του συγκεκριμένου χρήστη, την τοποθεσία στην οποία βρίσκεται και τις τυχόν ρυθμίσεις που έχει κάνει στο παρελθόν. Κατά τη λειτουργία της, μια μηχανή αναζήτησης, πραγματοποιεί χιλιάδες επεξεργασίες δεδομένων σε κάθε δευτερόλεπτο. Αυτή η διαδικασία είναι ιδιαίτερα απαιτητική και πάντα υπάρχουν κάποιοι περιορισμοί καθώς δεν πρέπει να επιβαρύνεται το σύστημα του χρήστη. Το S4 σχεδιάστηκε για να επιταχύνει και να απλοποιεί αυτή τη διαδικασία κάνοντας ακριβέστερες προβλέψεις, με έμμεσο στόχο την αύξηση του κέρδους. Φυσικά το S4 παραμένει μια υψηλών επιδόσεων πλατφόρμα για την επεξεργασία ροών ‘’μεγάλων’’ δεδομένων και έτσι χρησιμοποιείται σε μεγάλο βαθμό γι’ αυτό το σκοπό.

Το S4 γενικά χρησιμοποιείται σε συνδυασμό με ένα σύστημα συγχρονισμού το οποίο ονομάζεται ZooKeeper και παρέχεται από την Apache. Το ZooKeeper είναι μία διανεμημένη ελεύθερη για χρήση υπηρεσία συγχρονισμού για διανεμημένες εφαρμογές και χρησιμοποιείται αρκετά συχνά σε συνδυασμό με διάφορες σύγχρονες πλατφόρμες για τη διαχείριση και τον συγχρονισμό των δεδομένων εξόδου. Στη συνέχεια θα περιγράψουμε τα βασικά χαρακτηριστικά του S4:

- *Αποκεντρωτική κατανομή*: Στο S4 όλοι οι κόμβοι του δικτύου των διασυνδεμένων συστημάτων είναι συμμετρικοί και πανομοιότυποι ενώ δεν υπάρχει κάποιος κόμβος ελέγχου. Έτσι, η δομή του δικτύου είναι ιδιαίτερα απλή και η πλατφόρμα μπορεί να χρησιμοποιηθεί από διαφόρων τύπων συστήματα με ευκολία.

- *Κλιμακωσιμότητα*: Η πρόσθεση νέων συστημάτων στο δίκτυο μπορεί να γίνει με ιδιαίτερη ευκολία, ενώ δεν υπάρχει κάποιος περιορισμός ως προς τον αριθμό των συνολικών μονάδων του δικτύου. Επίσης, η παραγωγικότητα αυξάνεται γραμμικά με την πρόσθεση νέων κόμβων στο δίκτυο.

- *Επεκτάσιμο*: Τα συστατικά και οι εικονικοί πόροι του δικτύου μπορούν να αντικατασταθούν εύκολα για τη δημιουργία μιας πιο προηγμένης και συγκεκριμένων απαιτήσεων εφαρμογής. Επίσης, οι διάφορες εφαρμογές στο S4 δημιουργούνται με μεγάλη ευκολία και μπορούν να εφαρμοστούν σε πληθώρα περιπτώσεων μέσω μιας απλής διεπαφής.

- *Εύκολη διαχείριση του Cluster*: Η διαχείριση του δικτύου Cluster στο S4 είναι ιδιαίτερα απλοποιημένη και όλα τα απαραίτητα εργαλεία είναι ‘’κρυμμένα’’ με τη χρήση ενός επιπέδου επικοινωνίας το οποίο βρίσκεται πάνω από το σύστημα ZooKeeper.

- *Αντοχή σε σφάλματα*: Στο S4 κάποια από τα διασυνδεμένα συστήματα παραμένουν αδρανή κατά την επεξεργασία των δεδομένων ενώ τα υπόλοιπα συστήματα είναι ενεργά. Έτσι, εάν υπάρξει κάποιο σφάλμα σε κάποιο από τα ενεργά συστήματα κατά την επεξεργασία τότε αυτό το σύστημα απενεργοποιείται και τη λειτουργία που εκτελούσε αναλαμβάνει ένα από τα αδρανή συστήματα το οποίο ενεργοποιείται. Η απώλεια δεδομένων εξαιτίας ενός σφάλματος ελαχιστοποιείται με τη χρήση ενός συστήματος ελέγχου και επανάκτησης που περιλαμβάνεται στο S4. [10][54]

2.4.2.11 Google Pregel, Graph Lab, Apache Giraph

Το Pregel είναι ένα δημοφιλές προγραμματιστικό μοντέλο επεξεργασίας με χρήση γράφων δικτύου σχεδιασμένο από τη Google. Είναι ικανό να διαχειριστεί πολύ μεγάλους γράφους που αποτελούνται από δισεκατομμύρια κορυφές και τρισεκατομμύρια ακμές. Ταυτόχρονα, όμως, το Pregel αποτελεί και ένα μοντέλο χαμηλού επιπέδου οπότε οι προγραμματιστές που το χρησιμοποιούν πρέπει να βελτιστοποιούν των κώδικά τους στο μεγαλύτερο δυνατό βαθμό καθώς και, κατά περιόδους, να κάνουν τις απαραίτητες αλλαγές, για να μπορέσουν να εκμεταλλευτούν πλήρως τις δυνατότητές του.

Οι δομές που χρησιμοποιεί το Pregel για την επεξεργασία με χρήση γράφων είναι ειδικά διαμορφωμένες ώστε να μπορούν να αναλύουν μεγάλους διανεμημένους γράφους. Η εκκίνηση ενός προγράμματος του Pregel γίνεται με την μεταφορά ενός γράφου από το σύστημα αποθήκευσης του δικτύου HDFS, από το οποίο παράγεται ο γράφος δικτύου, στη μνήμη του συστήματος. Στη συνέχεια ο προγραμματιστής που χρησιμοποιεί το μοντέλο πρέπει να ορίσει μια καθορισμένη συνάρτηση για την επεξεργασία των δεδομένων καθώς, τον τρόπο εκτέλεσης του προγράμματος και την επικοινωνία μεταξύ των κορυφών του γράφου. Ένας τυπικός υπολογισμός στο Pregel περιλαμβάνει δύο μέρη. Το πρώτο μέρος είναι η παρούσα είσοδος του συστήματος η οποία ακολουθείται από το δεύτερο μέρος, μία αλληλουχία μεγάλων βημάτων που ονομάζονται super-steps. Ένα super-step είναι μία επαναληπτική διαδικασία παράλληλης εκτέλεσης υπολογισμών από τις κορυφές του γράφου οι οποίες επεξεργάζονται τα δεδομένα με τη χρήση μιας καθορισμένης από το χρήστη συνάρτησης. Κατά την εκκίνηση ενός super-step όλες οι κορυφές του γράφου είναι ενεργοποιημένες, που σημαίνει πως εκτελούν υπολογισμούς σε κάθε επανάληψη, ενώ κατά τη διάρκεια των επαναλήψεων αρχίζουν να απενεργοποιούνται όταν δεν υπάρχουν πλέον δεδομένα για να επεξεργαστούν. Τη στιγμή που απενεργοποιούνται όλες οι κορυφές σταματά η επαναληπτική διαδικασία του συγκεκριμένου super-step και ξεκινά το επόμενο με νέα δεδομένα εισόδου προς επεξεργασία. Αυτή η διαδικασία επαναλαμβάνεται μέχρι όλα τα δεδομένα εισόδου να υποστούν επεξεργασία και στη συνέχεια παράγεται η τελική έξοδος του συστήματος. Με τη μέθοδο αυτή επιτυγχάνονται πολύ υψηλές ταχύτητες επεξεργασίας και μειωμένος αριθμός σφαλμάτων κατά την εκτέλεση του προγράμματος.

Το GraphLab και το Apache Giraph είναι κάποια παραδείγματα προγραμμάτων, παρόμοια με το Pregel, που χρησιμοποιούνται επίσης για την επεξεργασία μεγάλης κλίμακας γράφων. Το GraphLab είναι ιδιαίτερα γνωστό για χαρακτηριστικά όπως η χρονοδρομολόγηση των υπολογισμών του

συστήματος, η διαμέριση προγραμμάτων, για παράλληλη εκτέλεση των τμημάτων τους, και ευέλικτος έλεγχος για τη διατήρηση της εγκυρότητας κατά την εκτέλεση ενός προγράμματος. Τα χαρακτηριστικά αυτά το καθιστούν ιδανικό για χρήση σε προγράμματα μηχανικής μάθησης διότι προφέρει υψηλή ταχύτητα και μειωμένο αριθμό σφαλμάτων κατά την εκτέλεση. Το Apache Giraph έχει ιδιαίτερα υψηλές επιδόσεις όταν χρησιμοποιείται σε εφαρμογές ανάλυσης κοινωνικών δικτύων , και γι' αυτό το λόγο χρησιμοποιείται από την Facebook σε αυτόν τον τομέα. [10][55][56]

2.4.2.12 MOA (Massive Online Analysis)

Το MOA είναι ένα πλαίσιο που κάνει χρήση προηγμένων αλγορίθμων και χρησιμοποιείται για την επιμέλεια ροών δεδομένων. Η επιμέλεια δεδομένων είναι μία διαδικασία οργάνωσης και ενσωμάτωσης δεδομένων που προέρχονται από διάφορες πηγές και περιλαμβάνει το σχολιασμό, τη δημοσίευση και την παρουσίαση των δεδομένων, έτσι ώστε η αξία των δεδομένων να διατηρείται με την πάροδο του χρόνου. Με αυτή τη διαδικασία τα δεδομένα μπορούν να χρησιμοποιηθούν πολλές φορές καθώς και να διατηρηθούν για μελλοντική χρήση.

Το MOA επιτρέπει τη δημιουργία και τον πειραματισμό με αλγορίθμους μηχανικής μάθησης έτσι ώστε να μπορεί να παρέχει ταχύτερες απαντήσεις σε ερωτήσεις πάνω στη φύση των επιμεληημένων δεδομένων. Όμως, αν και είναι ένα ιδανικό πλαίσιο για χρήση τεχνικών μηχανικής μάθησης σε μεμονωμένα συστήματα, αυτοί οι αλγόριθμοι δεν υποστηρίζονται σε μεγάλη κλίμακα. Έτσι το σύστημα στο οποίο εκτελείται μια διεργασία του MOA δεν μπορεί να κλιμακωθεί, ακόμα και όταν η διεργασία είναι ιδιαίτερα απαιτητική και οι επεξεργαστικοί πόροι του συστήματος δεν είναι αρκετοί.

Το MOA, όπως και άλλα προγράμματα που χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την ανάλυση ροών δεδομένων, αντιμετωπίζει τα δεδομένα εισόδου οποιαδήποτε μορφής ως ροές δεδομένων. Εφόσον οι ροές δεδομένων είναι απεριόριστες για μια αποτελεσματική ανάλυση πρέπει να πληρούνται κάποιες προϋποθέσεις από αυτά τα πλαίσια, τις οποίες οι χρήστες τους πρέπει να γνωρίζουν. Τα τέσσερα πιο σημαντικά από αυτά τα κριτήρια είναι:

- *Κριτήριο 1*: Εφόσον όλα τα υπολογιστικά συστήματα έχουν περιορισμένη μνήμη, μόνο ένα τμήμα της ροής δεδομένων μπορεί να αποθηκευτεί στη μνήμη και να επεξεργαστεί κάθε φορά.

- *Κριτήριο 2*: Η επεξεργασία και εκτίμηση των δεδομένων μπορεί να γίνει μόνο μία φορά σε μία εκτέλεση της συνολικής διαδικασίας.

- *Κριτήριο 3*: Η διαδικασία επεξεργασίας των δεδομένων μπορεί να διακοπεί οποιαδήποτε στιγμή και παρόλα αυτά να παραχθούν αποτελέσματα, η εγκυρότητα των οποίων είναι αμφιλεγόμενη.

- *Κριτήριο 4*: Η χρονική διάρκεια της επεξεργασίας είναι περιορισμένη, καθορισμένη και σχετίζεται με τις προδιαγραφές του συστήματος.

Κλείνοντας αξίζει να αναφέρουμε πως, με σωστή χρήση, το MOA προσφέρει υψηλή επίδοση και αξιόπιστη λειτουργία ενώ παρέχει δυνατότητες όπως σύγκριση μεταξύ των αποτελεσμάτων, ως προς τον αλγόριθμο και τη μέθοδο που χρησιμοποιήθηκε για την ανάλυση, και τη δημιουργία και

χρήση μεθόδων αξιολόγησης της επίδοσης αλγορίθμων εξόρυξης δεδομένων από τις ροές δεδομένων. [10][57]

2.4.2.13 Apache SAMOA (Scalable Advanced Online Analysis)

Το SAMOA είναι ένα πλαίσιο ελεύθερης χρήσης για την εξόρυξη ‘’μεγάλων’’ δεδομένων σχεδιασμένο από την Apache. Προσφέρει μια συλλογή από διανεμημένους αλγορίθμους για την επεξεργασία ροών δεδομένων. Το SAMOA συνδυάζει τις δύο βασικές προσεγγίσεις για την ανάλυση των ‘’μεγάλων’’ δεδομένων και έτσι αποτελεί ένα σύστημα επεξεργασίας και ανάλυσης ροών δεδομένων το οποίο είναι κλιμακώσιμο και ταυτόχρονα υποστηρίζει τη διανομή του φόρτου εργασίας στα διάφορα συστήματα ενός δικτύου.

Το SAMOA εκτός από πλαίσιο εξόρυξης δεδομένων είναι και βιβλιοθήκη που περιέχει πληθώρα εξελιγμένων αλγορίθμων μηχανικής μάθησης για την ανάλυση ροών δεδομένων. Αποτελεί ένα σύστημα προσιτό στους χρήστες οι οποίοι μπορούν να αποστασιοποιηθούν από τη μηχανή που χρησιμοποιείται για την επεξεργασία των δεδομένων και να εφαρμόσουν τον ίδιο κώδικα σε διαφορετικές περιστάσεις και με διαφορετικές μηχανές επεξεργασίας.

Ένας αλγόριθμος στο SAMOA αντιπροσωπεύεται από έναν κατευθυνόμενο γράφο όπου οι γειτονικοί κόμβοι, δηλαδή τα γειτονικά διασυνδεμένα συστήματα του δικτύου, επικοινωνούν μέσω μηνυμάτων. Σε αυτό το κομμάτι της λειτουργίας του δεν θα εμβαθύνουμε περισσότερο καθώς η λογική είναι ίδια με αυτή του Storm. Επίσης, οι αλγόριθμοι που χρησιμοποιεί είναι παρόμοιοι με αυτούς που περιλαμβάνονται στο MOA και γι’ αυτό το λόγο μοντέλα κατασκευασμένα στο MOA είναι συμβατά με το SAMOA.

Το SAMOA, το οποίο είναι σχεδιασμένο σε Java, ουσιαστικά μπορούμε να πούμε πως αποτελεί έναν πολύ επιτυχημένο συνδυασμό του MOA με το Storm ενώ ταυτόχρονα προσφέρει και επιπλέον χαρακτηριστικά. Τέλος, η ευέλικτη αρχιτεκτονική του, του δίνει τη δυνατότητα να μπορεί να χρησιμοποιηθεί σε πολλές διανεμημένες μηχανές επεξεργασίας ροών δεδομένων, όπως το Storm, το S4 και το Samza, εύκολα και με τη μέγιστη απόδοση. [10][38]

2.4.2.14 H2O

Το H2O είναι ένα ελεύθερο προς χρήση πλαίσιο διανεμημένης μηχανικής μάθησης που εφαρμόζει τεχνικές μηχανικής μάθησης εντός της μνήμης RAM. Το H2O προσφέρει γραμμική κλιμακωσιμότητα και μπορεί να χρησιμοποιηθεί ανεξάρτητα ή και σε συνδυασμό με κάποια άλλη πλατφόρμα όπως το Hadoop και το Spark.

Το H2O περιλαμβάνει αρκετούς τύπους αλγορίθμων μηχανικής μάθησης και ανάλυσης όπως δέντρα αποφάσεων, ‘’βαθιά’’ μάθηση, Naïve Bayes, Random Forest, Gradient Boosting, k-means clustering, ανάλυση κύριων συνιστωσών (PCA) και γενικά μοντέλα γραμμικής ανάλυσης (γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, κτλ.). Προσφέρει διάφορους τρόπους χρήσης και διαθέτει μια διεπαφή που ονομάζεται Flow και είναι ιδιαίτερα φιλική προς το χρήστη. Η πρόσβαση στο Flow γίνεται μέσω του διαδικτύου έτσι ώστε να μπορεί να χρησιμοποιηθεί και από άτομα με λίγη προγραμματιστική εμπειρία. Ταυτόχρονα, όμως, είναι συμβατό και με τις προγραμματιστικές γλώσσες Java, R, Python και Scala έτσι ώστε πιο έμπειροι χρήστες να μπορούν να επηρεάσουν τα χαρακτηριστικά του ανάλογα με τις προτιμήσεις τους.

Ένα βασικό χαρακτηριστικό του H2O είναι ότι μπορεί να χρησιμοποιηθεί είτε σε ένα μεμονωμένο σύστημα, όπως ένας προσωπικός υπολογιστής, είτε σε ομάδες διασυνδεμένων συστημάτων σε συνεργασία με κάποια άλλη πλατφόρμα. Μια άλλη λειτουργία που προσφέρει, που είναι πολύ σημαντική για την επίτευξη υψηλών επιδόσεων, είναι η χρήση τεχνικών συμπίεσης των δεδομένων εντός της μνήμης. Με τη χρήση αυτής της τεχνικής μπορεί να γίνει η επεξεργασία πολύ μεγάλου όγκου δεδομένων σε ένα ιδιαίτερα μικρό χρονικό διάστημα.

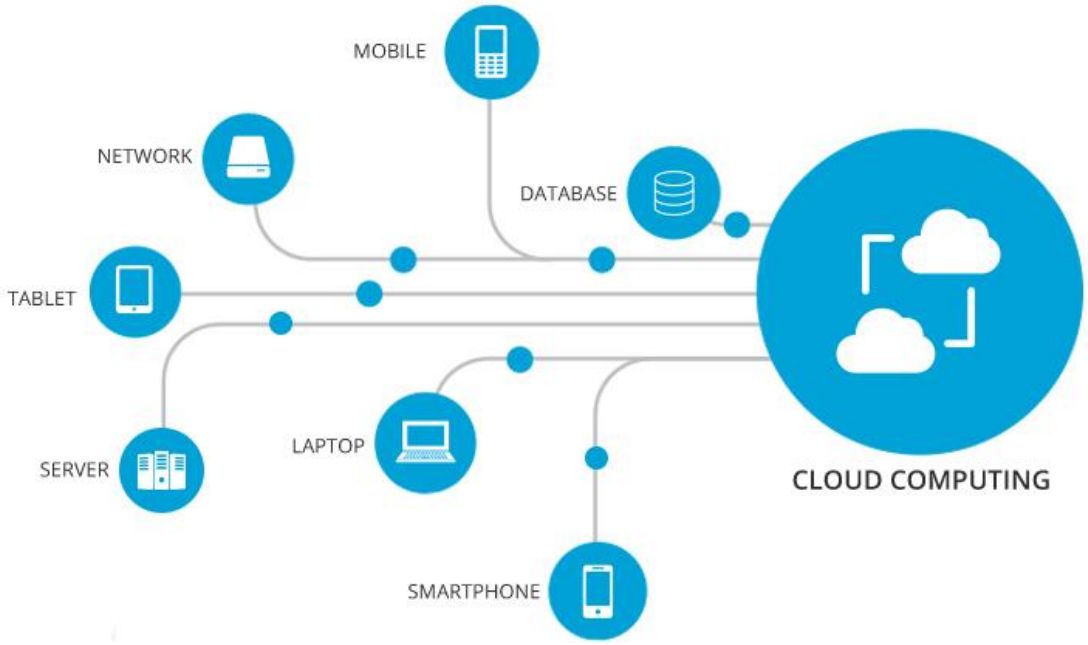
Κλείνοντας, επισημαίνουμε πως όταν το H2O λειτουργεί σε συνδυασμό με κάποια συμβατή πλατφόρμα οριζόντιας κλιμάκωσης, για την επίτευξη μέγιστης επίδοσης, πρέπει να λειτουργεί σε διαφορετικό μηχάνημα από το μηχάνημα στο οποίο βασίζεται το σύστημα HDFS της πλατφόρμας. [10][58]

2.4.3 Cloud computing

Τα τελευταία χρόνια έχει προκύψει μία νέα μέθοδος διασύνδεσης συστημάτων που ονομάζεται Cloud και η διαδικασία της επεξεργασίας δεδομένων χρησιμοποιώντας συστήματα που είναι συνδεδεμένα στο Cloud ονομάζεται Cloud Computing. Το Cloud είναι και αυτό μια μέθοδος οριζόντιας κλιμάκωσης αλλά επειδή έχει αρκετές διαφορές με τις άλλες μεθόδους σε αυτό τον τομέα θεωρήσαμε σωστό να περιγραφεί σε διαφορετική ενότητα.

Το Cloud Computing εμφανίστηκε για πρώτη φορά στα τέλη του 2007 και δίνει τη δυνατότητα διασύνδεσης συστημάτων που μπορεί να βρίσκονται σε οποιοδήποτε μέρος στον κόσμο μέσω του διαδικτύου. Δεν υπάρχει περιορισμός ως προς των τύπο των συστημάτων που διασυνδέονται και τα συστήματα αυτά μπορεί να είναι από μικρής υπολογιστικής ισχύος καταναλωτικά συστήματα μέχρι και εταιρικά μηχανήματα υψηλών επιδόσεων. Η πρόσβαση σε ένα δίκτυο συστημάτων σε Cloud γίνεται εύκολα μέσω του διαδικτύου και έτσι οποιοσδήποτε χρήστης μπορεί να έχει πρόσβαση στα εργαλεία και τις τεχνικές επεξεργασίας που προφέρει το Cloud Computing.

Η βασική ιδέα του Cloud Computing είναι να μεταφέρει την επεξεργασία δεδομένων από ένα μεμονωμένο σύστημα σε ένα δίκτυο συστημάτων, όταν αυτό χρειάζεται, και έτσι ο κάθε χρήστης να μπορεί να χρησιμοποιήσει τους υπολογιστικούς πόρους άλλων συστημάτων για να επιταχύνει την επεξεργασία των δεδομένων. Οι πόροι των συστημάτων που είναι διαθέσιμοι μέσω του Cloud περιλαμβάνουν τον χώρο αποθήκευσης και την επεξεργαστική ισχύ των συστημάτων αυτών. Με αυτή τη μέθοδο, πλέον, ένας χρήστης δεν χρειάζεται να σπαταλήσει μεγάλα κεφάλαια για την αναβάθμιση του συστήματός του, μιας και η χρήση πόρων μέσω του Cloud είναι πολύ πιο οικονομική. Τέλος, συνήθως, ένας χρήστης του Cloud δεν χρειάζεται να κάνει κάποιο αίτημα, ούτε απαιτείται κάποια διαδικασία, για την πρόσβαση στους πόρους των διασυνδεμένων συστημάτων, σε μια χρονική στιγμή, εκτός από μια αρχική συνδρομή που του παρέχει απεριόριστη πρόσβαση σε ένα συγκεκριμένο δίκτυο στο Cloud. Έτσι η χρήση αυτής της τεχνολογίας γίνεται ιδιαίτερα εύκολη και γρήγορη.



Εικόνα 2.1: Η μεγάλη ποικιλία διαφόρων συσκευών που μπορούν να χρησιμοποιηθούν σε ένα δίκτυο Cloud. [59]

Η διασύνδεση συστημάτων μέσω του Cloud έχει κάποιες πολύ σημαντικές διαφορές από τα δίκτυα διασύνδεσης Cluster και γι' αυτό το λόγο πλατφόρμες όπως το Hadoop δεν είναι ιδιαίτερα αποδοτικές για την επεξεργασία δεδομένων στα δίκτυα Cloud, χωρίς τον συνδυασμό τους με κάποια άλλη πλατφόρμα. Στη συνέχεια θα αναφέρουμε τα βασικά χαρακτηριστικά και τις βασικές διαφορές του Cloud Computing και του Cluster Computing.

- ***Cloud Computing:***

  - *Συγκεντρωτική και αποκεντρωτική κατανομή*: Τα συστήματα που συνδέονται σε ένα δίκτυο Cloud μπορεί να βρίσκονται είτε σε κοντινή απόσταση μεταξύ τους είτε να είναι διάσπαρτα σε διάφορα σημεία του κόσμου. Ο έλεγχος του δικτύου μπορεί να γίνεται από ένα συγκεκριμένο σύστημα ή και όχι.

  - *Απαραίτητη η σύνδεση διαδικτύου*: Τα συστήματα που βρίσκονται σε ένα δίκτυο Cloud πρέπει να διαθέτουν μία σταθερή και ικανοποιητικής ταχύτητας σύνδεση στο διαδίκτυο.

  - *Δυναμική διασύνδεση συστημάτων*: Τα συστήματα που είναι συνδεδεμένα σε ένα δίκτυο δεν λειτουργούν πάντα ταυτόχρονα κατά την εκτέλεση μιας διεργασίας. Τα διασυνδεμένα συστήματα βρίσκονται σε μία κατάσταση αδράνειας και κατά την εκτέλεση μιας διεργασίας κάποια από τα συστήματα ενεργοποιούνται και συμμετέχουν στην εκτέλεση. Το ποια και πόσα συστήματα ενεργοποιούνται εξαρτάται από τις απαιτήσεις σε πόρους και τις απαιτήσεις λογισμικού της συγκεκριμένης διεργασίας (μόνο κάποια από τα συστήματα του δικτύου μπορεί να έχουν το απαραίτητο λογισμικό για την εκτέλεση της διεργασίας).

- *Δεν υπάρχουν απαιτήσεις για συγκεκριμένο λογισμικό*: Τα συνδεδεμένα συστήματα μπορεί χρησιμοποιούν διαφορετικά λειτουργικά συστήματα και να διαθέτουν εγκατεστημένα διαφορετικά προγράμματα.

- *Ετερογενές*: Τα συνδεδεμένα συστήματα δεν χρειάζεται να είναι όμοια μεταξύ τους. Έτσι ένα δίκτυο Cloud μπορεί να περιλαμβάνει προσωπικούς υπολογιστές, servers ή ακόμα και φορητές συσκευές.

- *Αυτοελεγχόμενη εκτέλεση*: Πριν την εκτέλεση δεν χρειάζεται να γίνει κάποια ρύθμιση από το χρήστη ως προς τα συστήματα που θα συμμετέχουν στην εκτέλεση και τον τρόπο διαμοιρασμού των δεδομένων. Όλες οι απαραίτητες ρυθμίσεις έχουν οριστεί κατά τη δημιουργία του δικτύου και το δίκτυο λειτουργεί αυτόματα.

- *Δύσκολη υλοποίηση και διαχείριση*: Η μεγάλη ετερογένεια των διασυνδεμένων συστημάτων και η δυναμική λειτουργία καθιστούν ιδιαίτερα δύσκολη τη δημιουργία και τη διατήρηση ενός αξιόπιστου υψηλών επιδόσεων δικτύου Cloud. Ο σχεδιαστής του δικτύου πρέπει να έχει ιδιαίτερα καλές γνώσεις δικτύων έτσι ώστε να μπορεί να προβλέψει τυχόν προβλήματα που μπορεί να προκύψουν και να κάνει τις απαραίτητες βελτιστοποιήσεις όποτε αυτό είναι απαραίτητο.

- *Χαμηλή ασφάλεια και αξιοπιστία*: Λόγω της εύκολης χρήσης τους και της δυναμική τους λειτουργίας τα δίκτυα Cloud έχουν αρκετά χαμηλή ασφάλεια και αξιοπιστία. Έτσι κατά τον σχεδιασμό ενός δικτύου, όταν η ασφάλεια είναι υψηλής σημασίας, χρειάζεται ιδιαίτερη προσοχή ως προς το ποια συστήματα θα συμμετέχουν στο δίκτυο και το πως ένα νέο σύστημα μπορεί να ενταχθεί σε αυτό.

- *Χαμηλό κόστος*: Ένας χρήστης ενός δικτύου Cloud δεν χρειάζεται να επενδύσει σε υλικό και λογισμικό αφού αυτά παρέχονται από το δίκτυο. Έτσι, οι χρήστες δεν χρειάζεται να δαπανήσουν μεγάλα χρηματικά ποσά. Το μόνο που απαιτείται είναι μία, συνήθως μηνιαία συνδρομή για τη χρήση του δικτύου και των πόρων που παρέχει.

- *Μέτριες επιδόσεις*: Λόγω της μεγάλης ετερογένειάς τους και της δυναμικής τους λειτουργίας, τα δίκτυα Cloud δεν παρέχουν σε κάθε χρονική στιγμή τις ίδιες υπολογιστικές επιδόσεις. Έτσι, συνήθως, δεν είναι κατάλληλα για την επεξεργασία πολύ μεγάλων όγκων δεδομένων όταν ο χρόνος και η ταχύτητα εκτέλεσης είναι ύψιστης σημασίας.

- *Σχεδόν απεριόριστος χώρος αποθήκευσης*: Τα δίκτυα Cloud, λόγω του πολύ μεγάλου αριθμού συστημάτων που περιλαμβάνουν και της ευκολίας ένταξης ενός νέου συστήματος στο δίκτυο, μπορούν να παρέχουν σχεδόν απεριόριστο αποθηκευτικό χώρο για τα δεδομένα προς επεξεργασία. Επίσης, η αύξηση του αποθηκευτικού χώρου, όταν απαιτείται, μπορεί να γίνει εύκολα και γρήγορα κατά τη λειτουργία του δικτύου με την πρόσθεση ενός νέου συστήματος στο δίκτυο.

■ ***Cluster Computing:***

- *Συγκεντρωτική κατανομή*: Τα συστήματα που συνδέονται σε ένα δίκτυο Cluster πρέπει να βρίσκονται σε κοντινή φυσική απόσταση μεταξύ τους. Επίσης, ο έλεγχος του δικτύου γίνεται από ένα συγκεκριμένο σύστημα που αποτελεί τη βάση του δικτύου.

- *Δεν απαιτείται σύνδεση διαδικτύου*: Τα Cluster Computing εφαρμόζεται σε μία ομάδα συστημάτων που είναι διασυνδεμένα φυσικά. Στις περισσότερες περιπτώσεις τον σκοπό αυτό αναλαμβάνουν τοπικά δίκτυα (LAN) ειδικά σχεδιασμένα για να επιτυγχάνονται υψηλές ταχύτητες μεταφοράς των δεδομένων μεταξύ των συστημάτων. Έτσι, η χρήση του διαδικτύου δεν είναι απαραίτητη για τη λειτουργία τέτοιων δικτύων.

- *Στατική διασύνδεση συστημάτων*: Σε κάθε εκτέλεση μιας διεργασίας σε ένα δίκτυο Cluster όλα τα συστήματα που συμμετέχουν στο συγκεκριμένο δίκτυο λειτουργούν ταυτόχρονα, ως ένα ενιαίο σύστημα, το καθένα αναλαμβάνοντας ένα μέρος της συγκεκριμένης διαδικασίας. Κατά την εκτέλεση, ο αριθμός των συστημάτων που συμμετέχουν στο δίκτυο δεν μπορεί να αυξηθεί ή να μειωθεί ενώ αλλαγές στο δίκτυο μπορούν να γίνουν μόνο μετά την ολοκλήρωση της εκτέλεσης της διεργασίας και πριν την εκτέλεση της επόμενης. Έτσι, ο αριθμός των συστημάτων που συμμετέχουν στην εκτέλεση μιας διεργασίας παραμένει σταθερός.

- *Όλα τα συστήματα πρέπει να χρησιμοποιούν ίδιο λογισμικό*: Όλα τα διασυνδεμένα συστήματα σε ένα δίκτυο Cluster πρέπει να χρησιμοποιούν το ίδιο λειτουργικό σύστημα. Επίσης, οποιοδήποτε λογισμικό πρόκειται να χρησιμοποιηθεί στην επεξεργασία δεδομένων από το δίκτυο πρέπει να βρίσκεται εγκατεστημένο σε όλα τα συστήματα του δικτύου.

- *Ομογενές*: Τα συστήματα που συμμετέχουν σε ένα δίκτυο Cluster, όπως αναφέραμε, πρέπει να χρησιμοποιούν το ίδιο λειτουργικό σύστημα. Επίσης, στις περισσότερες περιπτώσεις, και το υλικό που χρησιμοποιούν τα συστήματα αυτά πρέπει αν είναι το ίδιο. Έτσι, τα συστήματα πρέπει να χρησιμοποιούν τα ίδια μοντέλα επεξεργαστών και ίδιου τύπου και μεγέθους μνήμη και μέσα αποθήκευσης. Υπάρχουν όμως και εξαιρέσεις, οπότε σε σπάνιες περιπτώσεις συναντώνται δίκτυα Cluster οπού τα διασυνδεμένα συστήματα έχουν διαφορετικό υλικό και λειτουργικό σύστημα.

- *Ο τρόπος εκτέλεσης εξαρτάται από τις ρυθμίσεις του διαχειριστή του δικτύου*: Ο χρήστης του δικτύου καθορίζει τη μέθοδο επεξεργασίας των δεδομένων. Οι ρυθμίσεις που γίνονται από το χρήστη μπορεί να σχετίζονται με τη μέθοδο διανομής των δεδομένων στα συστήματα του δικτύου και την παράλληλη επεξεργασία τους, τη χρονοδρομολόγηση των διεργασιών καθώς και τις εφαρμογές που θα χρησιμοποιηθούν για την επεξεργασία των δεδομένων. Επίσης, μόνο μετά το τέλος της επεξεργασίας των δεδομένων ο χρήστης έχει τη δυνατότητα να αλλάξει τις ρυθμίσεις που αφορούν την επόμενη εκτέλεση.

- *Εύκολη υλοποίηση και διαχείριση*: Η ομογένεια, ο πεπερασμένος αριθμός συστημάτων και η στατική λειτουργία καθιστούν την υλοποίηση και τη διαχείριση των δικτύων Cluster μια εύκολη διαδικασία. Επίσης, ο μεγάλος αριθμός από ελεύθερες για χρήση πλατφόρμες που κυκλοφορούν κάνει ακόμα πιο εύκολη τη διασύνδεση τέτοιων δικτύων και την επεξεργασία δεδομένων σε αυτά.

- *Σχετικά υψηλό κόστος*: Για τη δημιουργία ενός δικτύου Cluster, στις περισσότερες περιπτώσεις, απαιτείται μια σημαντική αρχική επένδυση για τη απόκτηση των συστημάτων που θα διασυνδεθούν. Φυσικά, τα συστήματα αυτά δεν χρειάζεται να είναι πολύ υψηλών προδιαγραφών. Έτσι, το κόστος δεν είναι τόσο μεγάλο όσο για τη δημιουργία ενός υπερυπολογιστή (HPC) όμως σίγουρα είναι μεγαλύτερο από τη συνδρομή σε ένα δίκτυο Cloud.

- *Υψηλές επιδόσεις*: Τα συστήματα Cluster σχεδιάζονται, συνήθως, για κάποιον συγκεκριμένο σκοπό. Έτσι, είναι ιδιαίτερα αποτελεσματικά στην εκτέλεση συγκεκριμένων διεργασιών και σε κάποιες περιπτώσεις πλησιάζουν και ταχύτητες που παλιότερα επιτυγχάνονταν μόνο από υπερυπολογιστές (HPC). Στις υψηλές επιδόσεις τους, επίσης, συμβάλλουν η ομογένεια και στατική διασύνδεση που τα χαρακτηρίζουν.

- *Περιορισμένος χώρος αποθήκευσης*: Τα δίκτυα Cluster, λόγω του πεπερασμένου αριθμού συστημάτων που διασυνδέουν, έχουν καθορισμένο αποθηκευτικό χώρο. Ο χώρος αυτός μπορεί να είναι πολύ μεγάλος αλλά όχι απεριόριστος και ο μόνος τρόπος για να αυξηθεί είναι η ένταξη ενός νέου συστήματος στο δίκτυο ή η προσθήκη επιπλέον αποθηκευτικών μέσων σε ένα σύστημα του δικτύου. Και οι δύο αυτές μέθοδοι, όμως, κοστίζουν και δεν μπορούν να εφαρμοστούν κατά την επεξεργασία δεδομένων από το δίκτυο.

Ο λόγος που πλατφόρμες Cluster Computing, όπως το Hadoop, δεν λειτουργούν βέλτιστα σε δίκτυα Cloud είναι η δυναμική διασύνδεση των συστημάτων. Οι πλατφόρμες είναι σχεδιασμένες ώστε να λειτουργούν βέλτιστα σε στατικά δίκτυα όπου ο αριθμός των διασυνδεμένων συσκευών δεν μεταβάλλεται κατά την επεξεργασία των δεδομένων. Στα δίκτυα Cloud, όπως αναφέραμε ο αριθμός των διασυνδεμένων συσκευών μπορεί να αλλάξει κατά τη διάρκεια της εκτέλεσης των διαφόρων διεργασιών και έτσι απαιτείται η χρήση άλλων τεχνικών.

Κλείνοντας, το Cloud Computing κατά βάση χρησιμοποιείται σε τομείς όπως ο τραπεζικός, ο ασφαλιστικός, οι επιχειρήσεις καθώς και η πρόβλεψη του καιρού. Επίσης, δύο επιπλέον πεδία στα οποία χρησιμοποιείται ιδιαίτερα τα τελευταία χρόνια είναι τα ηλεκτρονικά παιχνίδια και η εξερεύνηση του διαστήματος, όπου χρησιμοποιείται για την επεξεργασία δεδομένων που προέρχονται από τον Διεθνή Διαστημικό Σταθμό (ISS). Αντιθέτως, το Cluster Computing εφαρμόζεται κυρίως στον βιομηχανικό και ερευνητικό τομέα, τον τομέα υγείας καθώς και στον κυβερνητικό τομέα σε υπηρεσίες που παρέχουν παγκόσμια υποστήριξη, οι οποίες απαιτούν τη χρήση αξιόπιστων συστημάτων. Τέλος, τα συστήματα Cluster Computing, λόγω της υψηλής ασφάλειας που παρέχουν, χρησιμοποιούνται και στα συστήματα ασφαλείας πυρηνικών σταθμών παραγωγής ηλεκτρικής ενέργειας. [60][61][62][63]

2.4.3.1 Nephele/PACT

Το Nephele/Pact είναι ένα σύστημα παράλληλης επεξεργασίας δεδομένων που προκύπτει από το συνδυασμό του Nephele με το PACT. Το Nephele είναι ένα κλιμακώσιμο σύστημα παράλληλης επεξεργασίας δεδομένων και χρησιμοποιείται συγκεκριμένα σε δίκτυα συστημάτων στο Cloud, ενώ το PACT είναι ένα προγραμματιστικό μοντέλο.

Το Nephele είναι το πρώτο πλαίσιο παράλληλης επεξεργασίας δεδομένων που είναι ειδικά σχεδιασμένο για χρήση σε δίκτυα Cloud. Αξιοποιεί στο μέγιστο το δυναμικό διαμοιρασμό υπολογιστικών πόρων που είναι μία από τις βασικές λειτουργίες του Cloud Computing και χρησιμοποιείται τόσο για την χρονοδρομολόγηση των διεργασιών όσο και για την επεξεργασία δεδομένων. Επίσης, κατά την επεξεργασία των δεδομένων, το Nephele επιτρέπει την ανάθεση συγκεκριμένων διεργασιών στα εικονικά συστήματα του δικτύου ενώ ταυτόχρονα ελέγχει την εκκίνηση και τη λήξη της επεξεργασίας των δεδομένων.

Το PACT ή αλλιώς Parallelization Contract (‘’σύμβαση’’ παραλληλισμού) είναι ένα προγραμματιστικό μοντέλο που αποτελεί μια γενίκευση του MapReduce. Επεκτείνει τις δυνατότητες του MapReduce κάνοντας χρήση συναρτήσεων δευτέρου βαθμού καθώς και ‘’συμβάσεων’’ εξόδου (Output Contracts) που διασφαλίζουν την συμπεριφορά των συναρτήσεων στο πρόγραμμα.

Στον συναρτησιακό προγραμματισμό οι συναρτήσεις δευτέρου βαθμού και γενικότερα οι συναρτήσεις υψηλότερου βαθμού είναι συναρτήσεις που έχουν ως ορίσματα συναντήσεις ή τα αποτελέσματα μετά την εκτέλεσή τους είναι συναρτήσεις. Έτσι, μια συνάρτηση πρώτου βαθμού έχει ως όρισμα ή έξοδο μία απλή συνάρτηση ενώ μια συνάρτηση δευτέρου βαθμού έχει ως όρισμα μία συνάρτηση πρώτου βαθμού. Με τη χρήση συναρτήσεων δευτέρου βαθμού το PACT επιτρέπει τη φυσική έκφραση των διεργασιών, κατά την επεξεργασία πολύπλοκων όγκων δεδομένων, ενώ τις παραλληλίζει ανεξάρτητα χωρίς την ανάγκη χρήσης κάποιας συγκεκριμένης πλατφόρμας.

Γενικά, το PACT προσφέρει περισσότερες δυνατότητες από το MapReduce. Στη δημιουργία συστημάτων επεξεργασίας δεδομένων, αυτές οι επιπλέον αυτές δυνατότητες περιλαμβάνουν τη δυνατότητα βελτιστοποίησης των προγραμμάτων και τη μεγαλύτερη ευελιξία που διαθέτει ο προγραμματιστής ως προς τον τρόπο εκτέλεσης της επεξεργασίας και τον διαμοιρασμό των δεδομένων. Αντιθέτως, το MapReduce, ενώ μπορεί να χρησιμοποιηθεί για τη δημιουργία στιβαρών υλοποιήσεων, οι υλοποιήσεις αυτές δεν είναι βέλτιστες.

Ο συνδυασμός του Nephele με PACT επιτρέπει τη χρήση των προγραμματιστικών πλεονεκτημάτων του PACT σε συστήματα Cloud Computing. Έτσι, προκύπτουν νέες προοπτικές για την παράλληλη επεξεργασία πολύ μεγάλων όγκων δεδομένων σε δίκτυα Cloud. Έτσι, ενώ έως πρόσφατα για την επεξεργασία και την ανάλυση ‘’μεγάλων’’ δεδομένων χρησιμοποιούνταν κυρίως συστήματα Cluster Computing, πλέον γίνεται δυνατή και επεξεργασία ‘’μεγάλων’’ δεδομένων και στο Cloud. [10][60][64][65]

2.4.4 Κάθετη κλιμάκωση

Όπως έχουμε ήδη αναφέρει η κάθετη κλιμάκωση ενός συστήματος είναι μία διαδικασία αύξησης της υπολογιστικής δύναμης του συγκεκριμένου συστήματος. Η αύξηση της υπολογιστικής δύναμης του συστήματος επιτυγχάνεται με την προσάρτηση επιπλέον υπολογιστικών πόρων στο

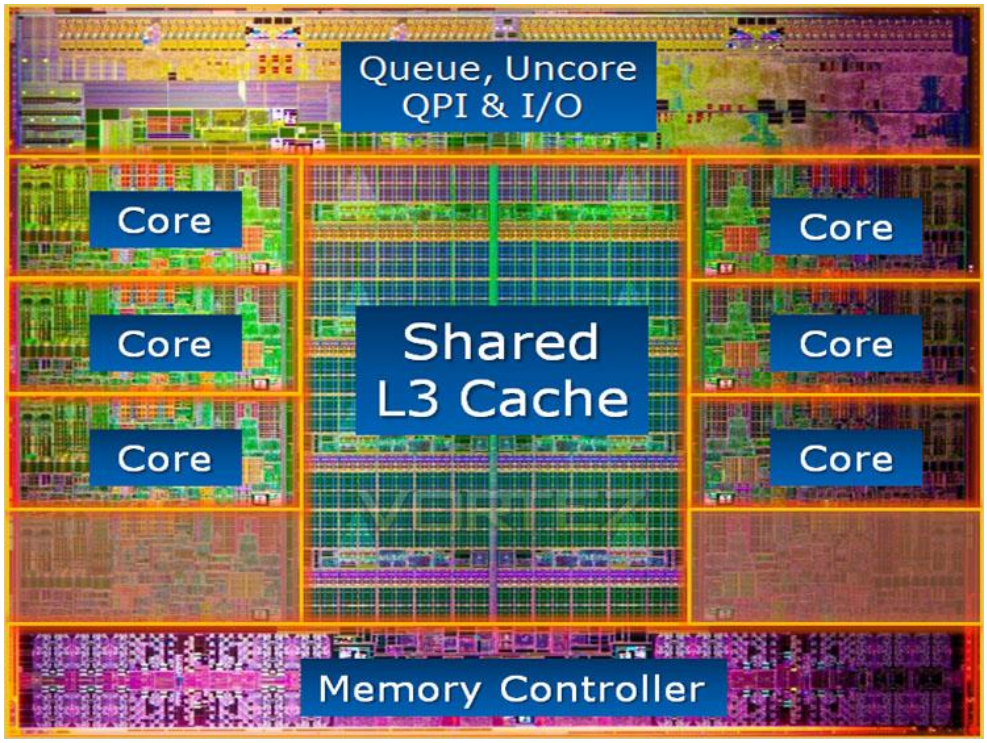
σύστημα όπως μνήμη, επεξεργαστικές μονάδες και χώρο αποθήκευσης. Ο τρόπος που γίνεται αυτού του είδους η κλιμάκωση δηλαδή οι βασικές πλατφόρμες και τα τσιπ που χρησιμοποιούνται στην κάθετη κλιμάκωση θα περιγραφούν στη συνέχεια. [10]

2.4.4.1 Multicore CPU

Αυτή η πλατφόρμα αφορά την εγκατάσταση πολυπύρηνων επεξεργαστών στο υπάρχον σύστημα. Οι πολυπύρηνοι επεξεργαστές αποτελούν μία τεχνολογία που υπάρχει τελευταία περίπου 20 χρόνια και περιλαμβάνει την κατασκευή επεξεργαστών που περιέχουν παραπάνω από έναν υπολογιστικούς πυρήνες. Έτσι, ενώ το τσιπ του επεξεργαστή διατηρεί περίπου το ίδιο μέγεθος, μπορεί να προσφέρει αυξημένες ταχύτητες επεξεργασίας.

Ένας πολυπύρηνος επεξεργαστής μπορεί να έχει 2, 4, 6 ή ακόμα και n υπολογιστικούς πυρήνες, όπου το n είναι πάντα άρτιος αριθμός. Το βασικό πλεονέκτημα που προσφέρουν οι πολυπύρηνοι επεξεργαστές, που είναι και πολύ χρήσιμο στην επεξεργασία ‘’μεγάλων’’ δεδομένων, είναι η παράλληλη επεξεργασία των δεδομένων καθώς υπάρχει η δυνατότητα διαίρεσης μιας διεργασίας σε τμήματα που ονομάζονται νήματα. Τα νήματα ουσιαστικά είναι ακολουθίες εντολών που εκτελούνται ταυτόχρονα, το κάθε ένα από έναν διαφορετικό υπολογιστικό πυρήνα. Έτσι, η συνολική διεργασία εκτελείται τελικά στο 1/n του συνολικού χρόνου που θα χρειαζόταν για την εκτέλεση της από έναν μονοπύρηνο επεξεργαστή, όπου n ο αριθμός των υπολογιστικών πυρήνων. Επιπλέον, πολλοί πολυπύρηνοι επεξεργαστές υποστηρίζουν την ταυτόχρονη επεξεργασία 2,4 ή ακόμα και n νημάτων από έναν υπολογιστικό πυρήνα (όπου n άρτιος αριθμός), μια λειτουργία που ονομάζεται Υπερνηματισμός (Hyperthreading), με αποτέλεσμα να αυξάνεται ακόμα περισσότερο η ταχύτητα επεξεργασίας. Τέλος, αξίζει τα σημειώσουμε πως οι πυρήνες ενός πολυπύρηνου επεξεργαστή μπορεί να έχουν κοινή μνήμη ή διαμοιρασμένη μνήμη, όπου ο κάθε πυρήνας έχει ξεχωριστή μνήμη.

Οι πολυπύρηνοι επεξεργαστές έχουν, όμως, και κάποια μειονεκτήματα. Το βασικότερο από αυτά είναι ο αριθμός των πυρήνων τους που θεωρείται αρκετά μικρός. Σήμερα ο μέγιστος αριθμός πυρήνων σε ένα επεξεργαστή που είναι διαθέσιμος στο κοινό είναι 48, ενώ υπάρχουν ταυτόχρονα άλλου είδους τσιπ που φέρουν χιλιάδες υπολογιστικούς πυρήνες. Ένα άλλο μειονέκτημα των πολυπύρηνων επεξεργαστών είναι η μικρή εσωτερική μνήμη που έχουν, με αποτέλεσμα να απαιτούν πρόσβαση σε εξωτερική μνήμη. Η εξωτερική μνήμη, στην οποία βασίζεται η λειτουργία τους, είναι η μνήμη RAM, η οποία αποτελεί την κύρια μνήμη των υπολογιστικών συστημάτων, και σε συστήματα επεξεργασίας ‘’μεγάλων’’ δεδομένων μπορεί να έχει μέγεθος μέχρι μερικές εκατοντάδες GB. Όμως αυτό το μέγεθος της μνήμης δεν είναι αρκετό καθώς η ταχύτητα ανανέωσης της μνήμης RAM δεν είναι ιδιαίτερα μεγάλη και πολλοί ισχυροί πολυπύρηνοι επεξεργαστές μπορούν να επεξεργαστούν ακόμα περισσότερα δεδομένα από αυτά που διατίθενται κάθε στιγμή από τη RAM. Έτσι δημιουργούνται καθυστερήσεις όποτε τα δεδομένα της κύριας μνήμης δεν είναι αρκετά και οι επεξεργαστές αναγκάζονται να αναστέλλουν τη λειτουργία τους όσο νέα δεδομένα μεταφέρονται από το κύριο μέσο αποθήκευσης στη RAM. [10][66]



Εικόνα 2.2: Η εσωτερική δομή ενός πολυπύρηνου επεξεργαστή. Παρατηρούμε τους πυρήνες (αριστερά και δεξιά), την εσωτερική μνήμη (cache) του επεξεργαστή (κέντρο), τον ελεγκτή της εσωτερικής μνήμης (κάτω) και την είσοδο-έξοδο του επεξεργαστή (πάνω). [67]

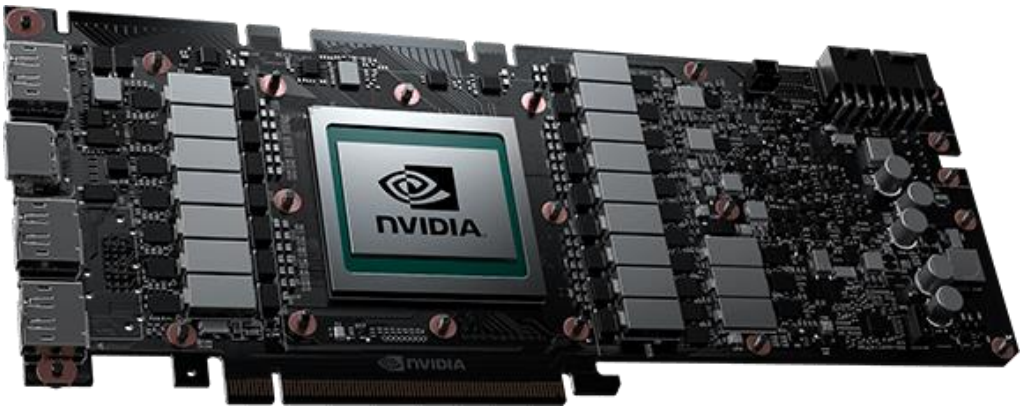
2.4.4.2 Graphics Processing Unit (GPU)

Οι GPU ή αλλιώς Μονάδες Επεξεργασίας Γραφικών είναι ετερογενή πολυπύρηνα τσιπ που εξειδικεύονται στην επεξεργασία γραφικών. Οι πιο συχνές τους χρήσεις μέχρι προσφάτως ήταν η επεξεργασία βίντεο, είτε για επαγγελματικό σκοπό είτε για ηλεκτρονικά παιχνίδια, ή η χρήση τους ως επιταχυντές σε εφαρμογές που απαιτούσαν επεξεργασία γραφικών. Παρόλα αυτά, τα τελευταία χρόνια, η χρήση τεχνικών παράλληλης επεξεργασίας σε διάφορες εφαρμογές, η εξέλιξη των GPU και το γεγονός ότι οι ίδιες οι GPU χρησιμοποιούν τεχνικές παράλληλης επεξεργασίας έχουν οδηγήσει στην εμφάνιση μιας νέας τάσης στην χρήση των GPU. Αυτή η νέα τάση είναι η εφαρμογή υπολογισμών γενικού σκοπού σε GPU (General-Purpose Computing on Graphic Processing Units - GPGPU).

Αυτή η δυνατότητα προέκυψε μετά τη διανομή του πλαισίου CUDA από την Nvidia. Με τη χρήση του CUDA οι προγραμματιστές έχουν, πλέον, εύκολη πρόσβαση στις δυνατότητες των GPU, κάτι που παλιότερα απαιτούσε πολύ καλή γνώση της δομής τους, και τις χρησιμοποιούν για τη δημιουργία ταχύτατων αλγορίθμων μηχανικής μάθησης. Η δημιουργία τέτοιων αλγορίθμων διευκολύνεται ακόμα περισσότερο με τη χρήση νέων βιβλιοθηκών όπως το GPUMiner, οι οποίες χρησιμοποιούν το πλαίσιο CUDA. Για τον έλεγχο της απόδοσης και της ταχύτητας αλγορίθμων μηχανικής μάθησης, που υλοποιούνται σε GPU, έγιναν πειραματικές μελέτες οι οποίες έδειξαν πως υπάρχει όντως αύξηση στην επίδοσή τους σε σχέση με αντίστοιχες υλοποιήσεις σε πολυπύρηνα CPU.

Το πιο βασικό πρόβλημα που έχουν οι GPU είναι η περιορισμένη μνήμη τους. Στα συστήματα πολυπύρηνων CPU η κύρια μνήμη RAM είναι ένα κομμάτι του συστήματος που μπορεί να αντικατασταθεί με μία μεγαλύτερου μεγέθους και υψηλότερης ταχύτητας. Αυτό, όμως, δεν είναι

δυνατό στις GPU καθώς η μνήμη είναι ενσωματωμένη στο ίδιο τσιπ με τους επεξεργαστικούς πυρήνες. Σήμερα η μεγαλύτερη μνήμη που μπορεί να έχει μια GPU είναι τα 12GB, η οποία όμως δεν είναι αρκετή για την αποδοτική επεξεργασία όγκων δεδομένων της τάξης των TB. Γενικά, όταν ο όγκος των δεδομένων προς επεξεργασία ξεπερνά το μέγεθος της μνήμης της GPU, γίνονται αναγκαστικές προσβάσεις στο μέσο αποθήκευσης, που περιέχει τα δεδομένα, γεγονός που δημιουργεί μεγάλες καθυστερήσεις. Ένα άλλο πρόβλημα που παρουσιάζεται στη χρήση των GPU για υπολογισμούς γενικού σκοπού είναι το περιορισμένο λογισμικό και οι αλγόριθμοι που διατίθενται για τέτοιες χρήσεις. Τέλος, το βασικό πλεονέκτημά των GPU, όπως αναφέρθηκε προηγουμένως, είναι η μεγάλη ταχύτητα παράλληλης επεξεργασίας των δεδομένων. Όμως οι περισσότεροι αναλυτικοί αλγόριθμοι που υπάρχουν, τουλάχιστον σήμερα, δεν μπορούν να διαχωριστούν εύκολα σε τμήματα που να μπορούν επεξεργαστούν παράλληλα. Έτσι, η υλοποίησή τους σε GPU είναι ιδιαίτερα δύσκολή και δεν υπάρχει κάποιο κέρδος στην ταχύτητα εκτέλεσης τους από τέτοιου είδους συστήματα. [10][68]



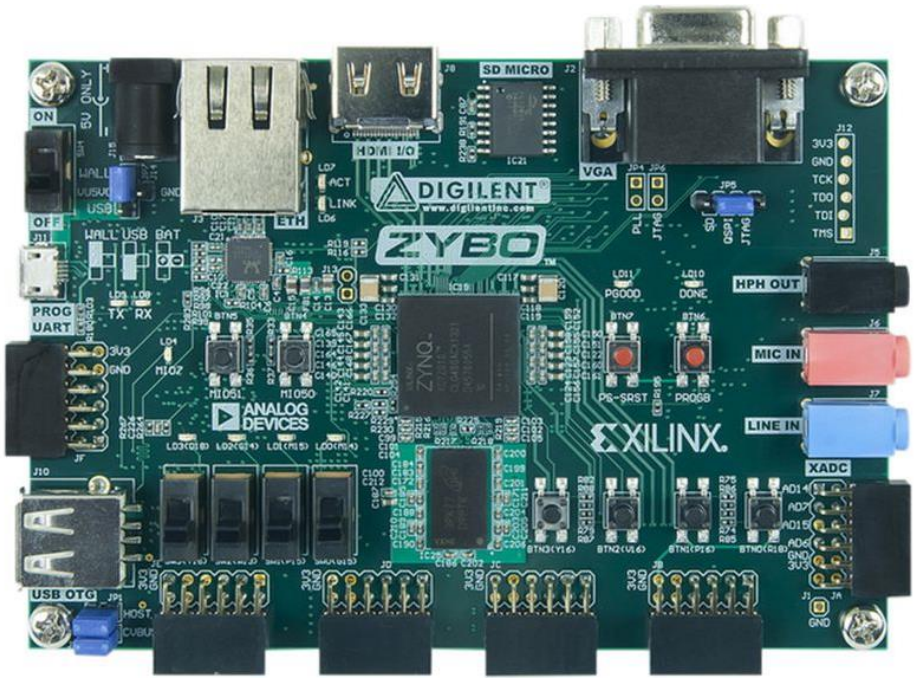
Εικόνα 2.3: Η εσωτερική δομή μιας GPU από την Nvidia. Παρατηρούμε τον τσιπ επεξεργασίας (κέντρο) το οποίο πλαισιώνεται από την μνήμη της GPU (αριστερά και δεξιά). [69]

2.4.4.3 Field Programmable Gate Array (FPGA)

Τα FPGA είναι κι αυτά ένα είδος τσιπ που χρησιμοποιούνται για την επεξεργασία δεδομένων. Υπάρχουν για παραπάνω από είκοσι χρόνια αλλά προσφάτως έγιναν αρκετά πιο δημοφιλή λόγω της εξέλιξης της τεχνολογίας που οδήγησε στην βελτίωση της δομής τους, της αποδοτικότητάς τους και της λογικής με βάση την οποία λειτουργούν.

Τα FPGA είναι στοιχεία στα μπορεί να προγραμματιστεί η συμπεριφορά τους και πολλά διαθέσιμα χαρακτηριστικά τους όπως η χρονοδρομολόγηση των δεδομένων που εισάγουμε σε αυτά και οι λογικοί πόροι που διαθέτουν και θέλουμε να χρησιμοποιήσουμε. Ο προγραμματισμός των FPGA έχει ιδιαίτερα μεγάλη σημασία και επηρεάζει σε μεγάλο βαθμό την τελική τους απόδοση, όπως την ταχύτητά τους και την κατανάλωση ενέργειας κατά τη λειτουργία τους, και χρησιμοποιείται για την προσαρμογή τους έτσι ώστε να λειτουργούν όσο το δυνατόν καλύτερα σε μία συγκεκριμένη εφαρμογή. Είναι σημαντικό να επισημάνουμε πως ένα μετά τον προγραμματισμό του, ένα FPGA, είναι αποδοτικότερο να χρησιμοποιηθεί στην εφαρμογή για την οποία προγραμματίστηκε, καθώς σε άλλες εφαρμογές μπορεί να έχει πολύ χαμηλή απόδοση. Αυτό συμβαίνει διότι κατά τον προγραμματισμό ενός FPGA κάποια χαρακτηριστικά του βελτιώνονται ενώ άλλα ταυτόχρονα γίνονται χειρότερα.

Ο προγραμματισμός των FPGA γίνεται με τη χρήση της γλώσσας περιγραφής υλικού HDL (Hardware Descriptive Language) η οποία απαιτεί κάποιες βασικές γνώσεις υλικού. Η ευελιξία που προσφέρουν δικαιολογεί το αυξημένο κόστος τους σε σχέση με άλλα τσιπ επεξεργασίας, ενώ ο αρχικός προγραμματισμός τους, που πρέπει να γίνει με ακρίβεια, αυξάνει περισσότερο το κόστος υλοποίησης αλγορίθμων σε FPGA. Φυσικά τα FPGA δεν χρησιμεύουν σε όλες τις περιπτώσεις οπότε πριν την χρήση τους για την υλοποίηση μιας εφαρμογής πρέπει να εξεταστεί αν υπάρχει κάποιο κέρδος από πλευρά απόδοσης και ταχύτητας. Οι εφαρμογές στις οποίες συναντώνται συνήθως FPGA είναι σε τείχη προστασίας υλικού (hardware firewalls), ενώ είναι γενικά πιο γρήγορα στον έλεγχο δεδομένων δικτύου από τα τείχη προστασίας λογισμικού (software firewalls). [10][70]



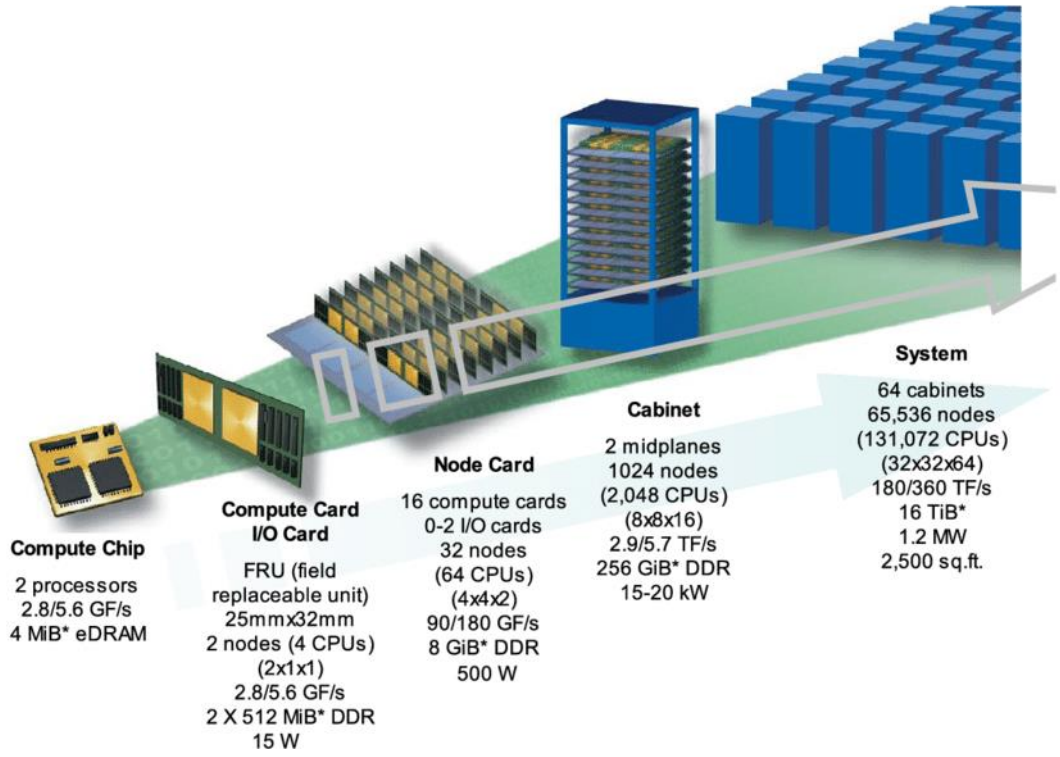
Εικόνα 2.4: Η δομή ενός FPGA. Παρατηρούμε το τσιπ επεξεργασίας (κέντρο). [71]

2.4.4.4 High-performance computing (HPC) clusters

Τα HPC είναι υψηλής επίδοσης υπολογιστικά συστήματα που βασίζονται στη λειτουργία της παράλληλης επεξεργασίας. Αποτελούνται από πολλές μονάδες επεξεργασίας που περιέχουν αρκετά στοιχεία επεξεργασίας (CPU,GPU) και συνδέονται μεταξύ τους μέσω ενός υψηλής ταχύτητας δικτύου. Τα δεδομένα διαχωρίζονται μεταξύ των μονάδων επεξεργασίας και επεξεργάζονται παράλληλα επιτυγχάνοντας έτσι υψηλές επιδόσεις.

Η κατασκευή ενός συστήματος με πολλά HPC και η διασύνδεσή τους αποτελεί την πιο συνηθισμένη πλατφόρμα σχεδιασμού συστημάτων υψηλών επιδόσεων. Σε αυτήν την κατηγορία συστημάτων ανήκουν οι υπερυπολογιστές στους οποίους οι μονάδες επεξεργασίας συνήθως ονομάζονται και λεπίδες (blades) και περιέχουν εκατοντάδες υπολογιστικούς πυρήνες. Οι επεξεργαστές που χρησιμοποιούνται σε τέτοια συστήματα διαχειρίζονται διαφορετικά την εσωτερική μνήμη τους, ενώ χρησιμοποιούνται εξειδικευμένες τεχνικές διασύνδεσης των μονάδων επεξεργασίας και διαφορετική οργάνωση των μέσων αποθήκευσης των δεδομένων του συστήματος. Όλα αυτά σε συνδυασμό με τα πανίσχυρα στοιχεία επεξεργασίας που χρησιμοποιούνται, τα οποία έχουν βελτιστοποιηθεί για μέγιστη ταχύτητα και παραγωγικότητα, έχουν ως αποτέλεσμα τις πολύ υψηλές επιδόσεις του τελικού συστήματος. Τέλος, οι ομάδες

διασυνδεμένων HPC αποτελούν την πιο αποδοτική, ευέλικτη και οικονομική πλατφόρμα για τη δημιουργία συστημάτων προσομοιώσεων. [10]



Σχήμα 2.7: Η δομή ενός HPC και συγκεκριμένα του υπερυπολογιστή BlueGene. Τα τσιπ επεξεργασίας (Compute Chip) τοποθετούνται σε μία κάρτα εισόδου-εξόδου (I/O Card). Οι κάρτες εισόδου-εξόδου τοποθετούνται σε μία κάρτα κόμβου (Node Card) η οποία προσαρμόζεται σε ένα ερμάριο (Cabinet). Το σύνολο των ερμαρίων συνθέτει το συνολικό σύστημα (System). [72]

2.5 Πλατφόρμες επεξεργασίας ‘’μεγάλων’’ δεδομένων (Big data)

Όπως αναφέραμε στην προηγούμενη ενότητα υπάρχουν αρκετές πλατφόρμες κλιμάκωσης των συστημάτων που χρησιμοποιούνται για την επεξεργασία και την ανάλυση ‘’μεγάλων’’ δεδομένων. Οι περισσότερες από αυτές τις πλατφόρμες δεν χρησιμοποιούνται μόνο για την επέκταση και τη διασύνδεση των συστημάτων αλλά και για την ίδια την επεξεργασία των ‘’μεγάλων’’ δεδομένων. Εκτός από αυτές τις πλατφόρμες που εξυπηρετούν διπλούς σκοπούς υπάρχουν και αρκετά άλλα λογισμικά που χρησιμοποιούνται για την ανάλυση ‘’μεγάλων’’ δεδομένων που λειτουργούν σε συνδυασμό με κάποια πλατφόρμα κλιμάκωσης.

Επειδή οι πλατφόρμες αυτής της κατηγορίας είναι πολλές και η περεταίρω ανάλυση της καθεμίας ξεχωριστά ξεφεύγει από το στόχο αυτής της διπλωματικής εργασίας, στην ενότητα αυτή θα τις αναφέρουμε απλώς και θα συμπεριλάβουμε συνοπτικά τα βασικά χαρακτηριστικά τους. Η αναφορά αυτή ξεκινά από την πλατφόρμα με τη μεγαλύτερη βαθμολογία χρηστών και τελειώνει με την πλατφόρμα που έχει τη μικρότερη βαθμολογία. Επίσης, τα χαρακτηριστικά που θα συμπεριληφθούν περιλαμβάνουν τον τύπο κλιμάκωσης του συστήματος με το οποίο λειτουργούν

σε συνδυασμό, τον βασικό σκοπό της χρήσης τους καθώς και κάποιο άλλο χαρακτηριστικό αν είναι απαραίτητο για την σωστή περιγραφή τους. Έτσι λοιπόν αυτές οι πλατφόρμες είναι οι εξής: [73]

- ***Sisense*:** Χρησιμοποιείται σε δίκτυα Cluster και μεμονωμένα συστήματα για την ανάλυση και την ενοποίηση πολύπλοκων δεδομένων από διάφορες πηγές. Χρησιμοποιεί μηχανική μάθηση και αλγορίθμους πρόβλεψης. Είναι συμβατό για χρήση με τις γλώσσες SQL, R και Python.

- ***Sisense for Cloud Data Teams*:** Χρησιμοποιείται σε δίκτυα Cloud για την ανάλυση πολύπλοκων δεδομένων. Χρησιμοποιεί μηχανική μάθηση και αλγορίθμους πρόβλεψης. Είναι συμβατό για χρήση με τις γλώσσες SQL, R και Python.

- ***VMware*:** Χρησιμοποιείται σε δίκτυα Cloud για την υποστήριξη και την επιτάχυνση εφαρμογών ανάλυσης ''μεγάλων'' δεδομένων καθώς και για τη δημιουργία εικονικών συστημάτων. Είναι υλοποιημένο κατά βάση σε C και C^{++} ενώ είναι συμβατό με προγράμματα γραμμένα σε Java και $C^{\#}$.

- ***Microsoft Azure*:** Χρησιμοποιείται σε δίκτυα Cloud για τη δημιουργία, την επιτάχυνση, την αξιολόγηση και τη διαχείριση έξυπνων εφαρμογών που χρησιμοποιούνται στην ανάλυση δεδομένων. Μπορεί να χρησιμοποιηθεί με τις γλώσσες $C^{\#}$, $F^{\#}$, Java, Python, JavaScript, PowerShell καιTypeScript.

- ***Amazon Web Service - AWS*:** Χρησιμοποιείται σε δίκτυα Cloud και διαθέτει εργαλεία για την επεξεργασία, την αποθήκευση και την ανάλυση δεδομένων. Είναι συμβατό με τις γλώσσες Java, JavaScript, $C^{\#}$, Python, Ruby, PHP, Go και C^{++}.

- ***Google BigQuery*:** Χρησιμοποιείται σε δίκτυα Cloud για την αποθήκευση και την ανάλυση δεδομένων πολύ μεγάλου όγκου. Είναι ιδιαίτερα κλιμακώσιμο και διαθέτει δυνατότητες μηχανικής μάθησης. Μπορεί να χρησιμοποιηθεί μόνο με την προγραμματιστική γλώσσα SQL.

- ***MongoDB*:** Χρησιμοποιείται σε δίκτυα Cloud για την αποθήκευση και τη διαχείριση δεδομένων. Είναι συμβατό με το Azure και το Google Cloud και υποστηρίζει γλωσσών προγραμματισμού C, C^{++}, $C^{\#}$, .NET, Erlang ,Haskell, Java, JavaScript, Pearl, PHP, Python, Ruby και Scala.

- ***BlueTalon*:** Χρησιμοποιείται σε όλους τους τύπους δικτύων για την αποθήκευση και την προστασία δεδομένων. Λειτουργεί σε συνδυασμό με κάποια πλατφόρμα κλιμάκωσης όπως το Hadoop και το Spark και είναι ιδιαίτερα ευέλικτο και κλιμακώσιμο.

- ***Google Bigdata*:** Χρησιμοποιείται σε δίκτυα Cloud και διαθέτει διάφορα εργαλεία για την αποθήκευση, την επεξεργασία και την ανάλυση δεδομένων. Χρησιμοποιεί την ίδια υποδομή με τα περισσότερα εργαλεία της Google (όπως το Google Search και το Gmail) και επίσης προσφέρει διάφορες δυνατότητες μηχανικής μάθησης. Υποστηρίζει τις γλώσσες Go, Java, PHP, .NET, Node.js, Ruby και Python.

- ***IBM Big Data*:** Είναι μια πλατφόρμα που χρησιμοποιείται σε δίκτυα Cloud και διαθέτει λειτουργίες μηχανικής μάθησης και τεχνητής νοημοσύνης. Προσφέρει διάφορα εργαλεία για την αποθήκευση, την διαχείριση την επεξεργασία και την ανάλυση δεδομένων, ενώ ταυτόχρονα παρέχει τη δυνατότητα ασφάλισης των δεδομένων αυτών.

- ***Syncsort*:** Μπορεί να χρησιμοποιηθεί σε όλους τους τύπους δικτύων για τη διασφάλιση της ακεραιότητας των δεδομένων. Περιλαμβάνει εργαλεία για τη γρήγορη ταξινόμηση, την ενσωμάτωση, τον εμπλουτισμό και τη διασφάλιση της ποιότητας των δεδομένων. Χρησιμοποιείται σε συνδυασμό με κάποια πλατφόρμα επεξεργασίας και ανάλυσης δεδομένων όπως το Hadoop και είναι συμβατό με τα λειτουργικά συστήματα Microsoft Windows, Unix, Linux. Επίσης λειτουργεί και στα IBM Power Systems.

- ***Wavefront*:** Μπορεί να χρησιμοποιηθεί σε όλους τους τύπους δικτύων σε συνδυασμό με κάποια πλατφόρμα όπως το Hadoop για την ανάλυση ροών δεδομένων. Έχει υψηλή κλιμακωσιμότητα και μπορεί να συλλέξει δεδομένα από διάφορες πηγές. Είναι συμβατό με τις περισσότερες προγραμματιστικές γλώσσες.

- ***Cloudera Enterprise Bigdata*:** Χρησιμοποιείται σε δίκτυα Cloud συνήθως υπό την υποστήριξη του Hadoop ή κάποιας άλλης πλατφόρμας της Apache. Παρέχει ένα συγκεντρωτικά κλιμακώσιμο, ευέλικτο και ασφαλές περιβάλλον για την επεξεργασία και την ανάλυση δεδομένων. Είναι συμβατό με τις περισσότερες γλώσσες προγραμματισμού.

- ***Palantir Bigdata*:** Χρησιμοποιείται σε δίκτυα Cloud και βασικά υπό την υποστήριξη του Amazon Cloud. Αρχικά δεν ήταν ελεύθερα διαθέσιμο και χρησιμοποιούνταν μόνο από τομείς της κυβέρνησης των ΗΠΑ. Παρέχει εργαλεία για την επεξεργασία και την ανάλυση δεδομένων. Είναι συμβατό μόνο με την δική του προγραμματιστική γλώσσα που ονομάζεται Hedgehog.

- ***Oracle Bigdata Analytics*:** Χρησιμοποιείται σε δίκτυα Cloud αλλά και σε μεμονωμένα συστήματα. Παρέχει δυνατότητες διαχείρισης, επεξεργασίας και ανάλυσης δεδομένων ενώ ταυτόχρονα έχει μεγάλη ευελιξία και κλιμακωσιμότητα. Τα εργαλεία που παρέχει είναι συμβατά με τις περισσότερες προγραμματιστικές γλώσσες και κάποια από αυτά μπορούν να εφαρμοστούν και σε κάποιες άλλες πλατφόρμες όπως το Hadoop ενώ διαθέτει και δυνατότητες μηχανικής μάθησης.

- ***DataTorrent*:** Είναι μια εφαρμογή του Hadoop και χρησιμοποιείται για την επεξεργασία ροών δεδομένων. Μπορεί να επεξεργαστεί εκατομμύρια γεγονότα το δευτερόλεπτο και διαθέτη υψηλή αντοχή σε σφάλματα. Επίσης, μπορεί να κλιμακωθεί αυτόματα ανάλογα με τις απατήσεις της εκάστοτε εργασίας. Τέλος, είναι συμβατό με όλες τις προγραμματιστικές γλώσσες όπως και το Hadoop.

- ***Qubole*:** Χρησιμοποιείται σε δίκτυα Cloud για την επεξεργασία και την ανάλυση ροών δεδομένων. Διαθέτει δυνατότητες μηχανικής μάθησης ενώ επίσης μπορεί να χρησιμοποιηθεί για την επιτάχυνση και τη βελτιστοποίηση των εργασιών που εκτελούνται

σε ένα δίκτυο Cloud. Μπορεί να χρησιμοποιηθεί και σε συνδυασμό με το Hadoop ή το Spark.

- ***GoodData*:** Χρησιμοποιείται σε δίκτυα Cloud και παρέχει εργαλεία για την ανάλυση δεδομένων. Έχει τη δυνατότητα εύκολης και γρήγορης ενσωμάτωσης σε διάφορα συστήματα ενώ μπορεί να διαχειριστεί και να διανέμει οποιουδήποτε τύπου και μεγέθους δεδομένα. Επίσης έχει μεγάλη ευελιξία και διαθέτει δυνατότητες μηχανικής μάθησης. Παρέχει υποστήριξη για τις προγραμματιστικές γλώσσες Java, JavaScript, Python και API Reference.

- ***MapR Converged Data Platform*:** Χρησιμοποιείται σε δίκτυα Cloud για την αποθήκευση, τη διαχείριση και την ανάλυση δεδομένων με μεγάλη ταχύτητα, ευελιξία και αξιοπιστία. Διαθέτει ένα διανεμημένο σύστημα αποθήκευσης αρχείων και μπορεί να χρησιμοποιηθεί σε συνδυασμό με το Kafka για την απόκτηση της δυνατότητας επεξεργασίας και ανάλυσης ροών δεδομένων. Επίσης, για την επεξεργασία δεδομένων, μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλες πλατφόρμες όπως το Hadoop ή το Spark

- ***Hortonworks Data Platform*:** Είναι μία ελεύθερη για χρήση διανομή του Hadoop. Παρέχει δυνατότητες αποθήκευσης επεξεργασίας και ανάλυσης δεδομένων σε οποιουδήποτε τύπου δίκτυο. Επίσης παρέχει υψηλή ασφάλεια καθώς και δυνατότητες μηχανικής και ‘’βαθιάς’’ μάθησης. Είναι συμβατό με τις περισσότερες προγραμματιστικές γλώσσες.

- ***Amdocs Insight*:** Είναι ένας κόμβος για την επεξεργασία μεγάλων όγκων δεδομένων στο Cloud ενώ επίσης διαθέτει μια στιβαρή υποδομή διαχείρισης δεδομένων η οποία επιτρέπει τη γρήγορη και αξιόπιστη ανάλυσή τους. Παρέχει δυνατότητες ανάλυσης των δεδομένων με χρήση τεχνητής νοημοσύνης και λειτουργεί σε συνδυασμό με το Google Cloud. Υποστηρίζει τις περισσότερες προγραμματιστικές γλώσσες.

- ***Splunk Bigdata Analytics*:** Είναι μία πλατφόρμα για την επεξεργασία και ανάλυση δεδομένων ‘’μηχανής’’ (Machine Data), τα οποία αποτελούν μια υποκατηγορία των ‘’μεγάλων’’ δεδομένων, σε πραγματικό χρόνο. Τα δεδομένα ‘’μηχανής’’ παράγονται από τη φυσιολογική λειτουργία μιας επιχείρησης όπως τις συναλλαγές, τη χρήση διαφόρων εφαρμογών και τη χρήση του δικτύου που διασυνδέει τα συστήματα της επιχείρησης. Με την ανάλυση αυτού του τύπου δεδομένων μια επιχείρηση μπορεί να αξιολογήσει πόσο καλά λειτουργεί και να εντοπίσει τυχόν προβλήματα στους διάφορους τομείς της. Η πλατφόρμα αυτή μπορεί να εφαρμοστεί είτε εντός του δικτύου της εταιρίας είτε σε ένα εικονικό δίκτυο Cloud και είναι συμβατή με Java, JavaScript, PHP, Python.

- ***Celebrus Technologies*:** Είναι μια πλατφόρμα ανάλυσης δεδομένων καταναλωτών πραγματικού χρόνου. Τα δεδομένα αυτά συγκεντρώνονται από διάφορες πηγές όπως το διαδίκτυο, τα smartphones και τα ηλεκτρονικά μηνύματα και βοηθούν μια εταιρεία να δημιουργήσει ένα προφίλ για τους πελάτες της, να επιλέξει κατάλληλες διαφημίσεις και να διαφυλάξει τα δεδομένα τους. Αυτό το λογισμικό μπορεί να εφαρμοστεί είτε εντός μιας εταιρείας είτε να χρησιμοποιηθεί μέσω ενός δικτύου Cloud.

- ***HPCC Systems Big data*:** Είναι μια πλατφόρμα ελεύθερης χρήσης που μπορεί να εφαρμοστεί σε δίκτυα Cloud που αποτελούνται από διαφόρων τύπων καταναλωτικά συστήματα. Χρησιμοποιείται για την παράλληλη και αυξημένης ταχύτητας επεξεργασία δεδομένων από τέτοιου είδους δίκτυα. Υποστηρίζει τις προγραμματιστικές γλώσσες Java, JavaScript, Python, R, C^{++} καθώς και τη δική της γλώσσα ECL (Enterprise Control Language).

- ***Pachyderm*:** Χρησιμοποιείται για την επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων με στόχο την επίλυση επιστημονικών προβλημάτων. Επίσης, μπορεί να χρησιμοποιηθεί για τη δημιουργία εξειδικευμένων μοντέλων μηχανικής μάθησης βελτιστοποιημένων για χρήση σε συγκεκριμένα προβλήματα. Μπορεί να λειτουργήσει είτε σε δίκτυα Cluster εντός μιας επιχείρησης είτε σε δίκτυα Cloud και προς το παρόν υποστηρίζει τις γλώσσες Java, Python, $C^{\#}$, C^{++},Go και Dart.

- ***Arcadia Data*:** Είναι ένα λογισμικό για την ανάλυση δεδομένων και ροών δεδομένων επιχειρήσεων. Μπορεί να χρησιμοποιηθεί είτε μόνο του, με χρήση κάποιου δικτύου Cloud, είτε σε συνεργασία με το Hadoop, για δίκτυα Cluster. Εκτός από τις δυνατότητες τεχνητής νοημοσύνης υποστηρίζει και τη λειτουργία επεξεργασίας φυσικής γλώσσας.

- ***BigObject*:** Είναι ένα λογισμικό για την επεξεργασία και ανάλυση δεδομένων σε δίκτυα Cloud. Κάποια από τα βασικά του χαρακτηριστικά περιλαμβάνουν τη δυνατότητα επεξεργασίας δεδομένων εντός της μνήμης και το ιδιαίτερα μικρό μέγεθος που καταλαμβάνει στο αποθηκευτικό μέσο του συστήματος που το χρησιμοποιεί. Προς το παρόν είναι συμβατό μόνο με την προγραμματιστική γλώσσα Lua.

- ***Datameer*:** Είναι μια πλατφόρμα διαχείρισης δεδομένων και αρχείων σχεδιασμένη για χρήση από επιχειρήσεις. Μπορεί να λειτουργήσει στο τοπικό δίκτυο μιας επιχείρησης, στο Cloud ή και σε υβριδικά δίκτυα σε συνεργασία με το Azure. Επίσης, διαθέτει τη δυνατότητα εγκατάστασης και σε φορητές συσκευές ενώ υποστηρίζει μόνο τη γλώσσα Java.

- ***SAP Bigdata Analytics*:** Είναι μια πλατφόρμα που παρέχει διάφορα εργαλεία για την επεξεργασία, την ανάλυση και τη διαχείριση δεδομένων σχεδιασμένη για χρήση από επιχειρήσεις. Η πλατφόρμα μπορεί να χρησιμοποιηθεί είτε εντός της επιχείρησης είτε με τη χρήση δικτύου Cloud.

- ***Next Pathway*:** Είναι μια πλατφόρμα διαχείρισης δεδομένων που διευκολύνει την μεταφορά μεγάλων όγκων δεδομένων από υλικά μέσα αποθήκευσης δεδομένων επιχειρήσεων ή άλλων οργανισμών στο Cloud. Υποστηρίζει τις προγραμματιστικές γλώσσες Java, JavaScript, Scala, C^{++} και Python.

- ***CSC Big Data Platform*:** Είναι μία ελεύθερη για χρήση πλατφόρμα που παρέχει εργαλεία για την ανάπτυξη και την εφαρμογή αλγορίθμων επεξεργασίας και ανάλυσης δεδομένων και ροών δεδομένων. Μπορεί να χρησιμοποιηθεί είτε στο τοπικό δίκτυο μιας επιχείρησης σε συνεργασία με το Hadoop είτε σε δίκτυα Cloud.

- ***1010data*:** Είναι μια πλατφόρμα που παρέχει εργαλεία ανάλυσης και διαχείρισης δεδομένων σχεδιασμένη για χρήση από επιχειρήσεις. Τα διάφορα εργαλεία μπορούν να χρησιμοποιηθούν είτε έμμεσα στο τοπικό δίκτυο μιας επιχείρησης ή σε δίκτυα Cloud. Τα διάφορα εργαλεία της πλατφόρμας υποστηρίζουν πολλές προγραμματιστικές γλώσσες και κυρίως τις Java ,C, C^{++}, Python, .NET και VBA.

- ***GE Industrial Internet*:** Προσφέρει μια πλατφόρμα, που ονομάζεται Predix, σχεδιασμένη για χρήση από τον βιομηχανικό τομέα. Προσφέρει διάφορα εργαλεία για την επεξεργασία και ανάλυση ''μεγάλων'' δεδομένων καθώς και διαθέτει και δυνατότητες μηχανικής μάθησης. Η επεξεργασία και ανάλυση δεδομένων γίνεται σε εικονικά δίκτυα στο Cloud ενώ υπάρχει υποστήριξη για τις προγραμματιστικές γλώσσες Java, Node.js, Python, Go, .NET, PHP και Ruby.

- ***DataStax Bigdata*:** Η Datastax προσφέρει ένα πλήθος διαφορετικών λογισμικών κάποια από τα οποία απευθύνονται στο κοινό και κάποια άλλα είναι σχεδιασμένα για εταιρική χρήση. Κάποια από τα λογισμικά αυτά λειτουργούν σε δίκτυα Cloud ενώ κάποια άλλα λειτουργούν υβριδικά, αλλά η κοινή τους χρήση είναι για την επεξεργασία και ανάλυση ''μεγάλων'' δεδομένων και τη δημιουργία ευέλικτων και κλιμακώσιμων βάσεων δεδομένων.

- ***Rubikloud*:** Είναι μια πλατφόρμα για χρήση μέσω του Cloud σχεδιασμένη για χρήση από επιχειρήσεις. Προφέρει εργαλεία για τη μεταφορά, την αποθήκευση, την επεξεργασία και την ανάλυση των επιχειρησιακών δεδομένων. Διαθέτει δυνατότητες μηχανικής μάθησης και τεχνητής νοημοσύνης. Χρησιμοποιείται σε συνδυασμό με τις πλατφόρμες Amazon Web Services, Google Cloud ή Microsoft Azure.

- ***SGI Bigdata*:** Είναι μια πλατφόρμα αποθήκευσης, επεξεργασίας και ανάλυσης δεδομένων σχεδιασμένη για εταιρική χρήση. Προσφέρει όλα τα απαραίτητα εργαλεία ανάλυσης με στόχο την υποστήριξη μιας επιχείρησης και την υγιή της εξέλιξη. Λειτουργεί μέσω του Cloud, όπου αποθηκεύονται και τα δεδομένα της επιχείρησης.

- ***Teradata Bigdata Analytics*:** Προσφέρει μια πλατφόρμα που ονομάζεται Vantage η οποία μπορεί να χρησιμοποιηθεί τόσο από επιχειρήσεις όσο και από καταναλωτές. Η Πλατφόρμα διαθέτει εργαλεία για την αποθήκευση, την επεξεργασία και την ανάλυση δεδομένων στο Cloud. Επίσης διαθέτει συμβατότητα με τις περισσότερες προγραμματιστικές γλώσσες συμπεριλαμβανομένου των SQL, R και Python.

- ***Intel Bigdata*:** Είναι ένα λογισμικό που διαθέτει διάφορα εργαλεία που αφορούν την επεξεργασία και ανάλυση ''μεγάλων'' δεδομένων για διάφορες χρήσεις. Η επεξεργασία και ανάλυση δεδομένων μπορεί να γίνει είτε εντός ενός φυσικού δικτύου συστημάτων είτε στο Cloud.

- ***Guavus*:** Παρέχει μία πλατφόρμα που ονομάζεται Reflex η οποία διαθέτει διάφορα εργαλεία γα την επεξεργασία και ανάλυση δεδομένων και ροών δεδομένων σε πραγματικό χρόνο. Μπορεί να χρησιμοποιηθεί για διάφορες εταιρικές χρήσεις και διαθέτει

δυνατότητες τεχνητής νοημοσύνης. Λειτουργεί εντός του δικτύου συστημάτων Cluster της εταιρείας.

- ***HP Bigdata*:** Παρέχει μια πληθώρα λογισμικών για την αποθήκευση, την επεξεργασία και την ανάλυση δεδομένων τόσο για καταναλωτικές χρήσεις όσο και για εταιρικές. Τα διάφορα εργαλεία και λογισμικά λειτουργούν στο Cloud και υπάρχει υποστήριξη για τις βασικότερες προγραμματιστικές γλώσσες και πλέον και για την R.

- ***Dell Bigdata Analytics*:** Παρέχει μια πλατφόρμα, η οποία ονομάζεται EMC, και διαθέτει διάφορα εργαλεία για την αποθήκευση την επεξεργασία και την ανάλυση δεδομένων σε φυσικά δίκτυα Cluster. Επίσης η πλατφόρμα διαθέτει εργαλεία για την προστασία και ασφάλιση των δεδομένων και παρέχει υποστήριξη για τις περισσότερες προγραμματιστικές γλώσσες.

- ***Pivotal Bigdata*:** Είναι μια πλατφόρμα που παρέχει διάφορα εργαλεία για τη δημιουργία εφαρμογών επεξεργασίας, ανάλυσης και διαχείρισης δεδομένων. Λειτουργεί μέσω του Cloud και διαθέτει συμβατότητα με τις περισσότερες προγραμματιστικές γλώσσες παράλληλης επεξεργασίας όπως PL/R, PL/Python και PL/C.

- ***Cisco Bogdata*:** Είναι ένα λογισμικό για την αποθήκευση, την επεξεργασία και την ανάλυση εταιρικών δεδομένων. Λειτουργεί σε συνδυασμό με το Cloudera και μπορεί να λειτουργήσει είτε εντός του δικτύου συστημάτων Cluster της εταιρείας είτε μέσω του Cloud.

- ***Mu Sigma Bigdata*:** Είναι ένα λογισμικό για την και την ανάλυση, την αξιολόγηση και την αξιοποίηση δεδομένων. Παρέχει ένα σύνολο αναλυτικών αλγορίθμων μηχανικής μάθησης και απευθύνεται όχι μόνο σε εταιρείες αλλά και καταναλωτές. Μπορεί να χρησιμοποιηθεί είτε εντός του δικτύου συστημάτων Cluster μιας εταιρίας είτε μέσω του Cloud και υποστηρίζει τις περισσότερες προγραμματιστικές γλώσσες.

- ***MicroStrategy Bigdata*:** Είναι μια πλατφόρμα για την ομαδοποίηση και την ανάλυση διαφόρων τύπων εταιρικών δεδομένων. Μπορεί να χρησιμοποιηθεί είτε εντός του δικτύου συστημάτων μιας εταιρίας σε συνδυασμό μα το Hadoop, ή κάποιας άλλης παρόμοιας πλατφόρμας, είτε μέσω του Cloud σε συνδυασμό με το Azure ή το AWS. Υποστηρίζει τις περισσότερες προγραμματιστικές γλώσσες.

- ***Opera Solutions Bigdata*:** Παρέχει μία κλιμακώσιμη πλατφόρμα, που ονομάζεται Signal Hub, η οποία παρέχει τα απαραίτητα εργαλεία για την προηγμένη ανάλυση δεδομένων. Επίσης διαθέτει ένα μεγάλο αριθμό αναλυτικών αλγορίθμων μηχανικής μάθησης που μπορούν να εφαρμοστούν στην ανάλυση των δεδομένων.

- ***Informatica PowerCenter*:** Είναι ένα εργαλείο σχεδιασμένο για χρήση από επιχειρήσεις για τη διαχείριση και αποθήκευση των δεδομένων τους. Διαθέτει διάφορα εργαλεία για τη μεταφορά, την αντιγραφή, το συγχρονισμό και την ομαδοποίηση των εταιρικών

δεδομένων. Η διαχείριση και η αποθήκευση των δεδομένων μπορεί να γίνει είτε εντός των συστημάτων της επιχείρησης είτε στο Cloud.

- ***FICO Big Data Analyzer*:** Είναι ένα περιβάλλον το οποίο παρέχει διάφορα εργαλεία ανάλυσης δεδομένων. Διαθέτει δυνατότητες μηχανικής μάθησης και τεχνητής νοημοσύνης. Μπορεί να χρησιμοποιηθεί είτε σε φυσικά δίκτυα συστημάτων είτε μέσω του Cloud.

- ***Attivio*:** Είναι μια πλατφόρμα προηγμένων αναζητήσεων. Διαθέτει δυνατότητες όπως η μηχανική μάθηση και η επεξεργασία φυσικής γλώσσας, έτσι ώστε η ανάλυση των δεδομένων να γίνεται εις βάθος και να παρέχει ακριβή και προσωποποιημένα αποτελέσματα στον χρήστη. Μπορεί να χρησιμοποιηθεί είτε μέσω του Cloud είτε σε φυσικά συστήματα και δίκτυα Cluster.

- ***Kognitio Analytical Platform*:** Είναι μια πλατφόρμα ανάλυσης δεδομένων που μπορεί να χρησιμοποιηθεί είτε μόνη της είτε σε συνεργασία με κάποια άλλη πλατφόρμα. Χρησιμοποιείται σε συνδυασμό με το AWS για παροχή υπηρεσιών μέσω του Cloud ή σε συνδυασμό με το Hadoop σε φυσικά δίκτυα συστημάτων Cluster. Είναι συμβατό με οποιοδήποτε κώδικα, γραμμένο σε οποιαδήποτε προγραμματιστική γλώσσα, που μπορεί να εκτελεστεί σε περιβάλλον Linux.

- ***Zoomdata*:** Είναι μία πλατφόρμα για την ανάλυση δεδομένων και ροών δεδομένων που μπορεί να χρησιμοποιηθεί και από καταναλωτές και από επιχειρήσεις. Η ανάλυση των δεδομένων μπορεί να γίνει είτε μέσω Cloud είτε σε φυσικά συστήματα.

- ***Pentaho Big Data Analytics*:** Είναι μια πλατφόρμα ανάλυσης δεδομένων. Μπορεί να διαχειριστεί ομαδοποιημένα δεδομένα, ροές δεδομένων και δεδομένα πραγματικού χρόνου. Επίσης μπορεί να λειτουργήσει σε συνδυασμό με όλες τις τελευταίες εκδόσεις του Hadoop και να εισάγει σε αυτό τα δεδομένα που προέρχονται από διάφορες πηγές για επεξεργασία. Μπορεί να χρησιμοποιηθεί είτε σε φυσικά συστήματα είτε σε εικονικά δίκτυα στο Cloud και παρέχει υποστήριξη για τις προγραμματιστικές γλώσσες Scala, Java και Python.

3 ΥΛΟΠΟΙΗΜΕΝΑ ΣΥΣΤΗΜΑΤΑ ΠΑΝΩ ΣΤΟ ΠΕΔΙΟ ΤΗΣ ΨΗΦΙΑΚΗΣ ΕΠΙΔΗΜΙΟΛΟΓΙΑΣ

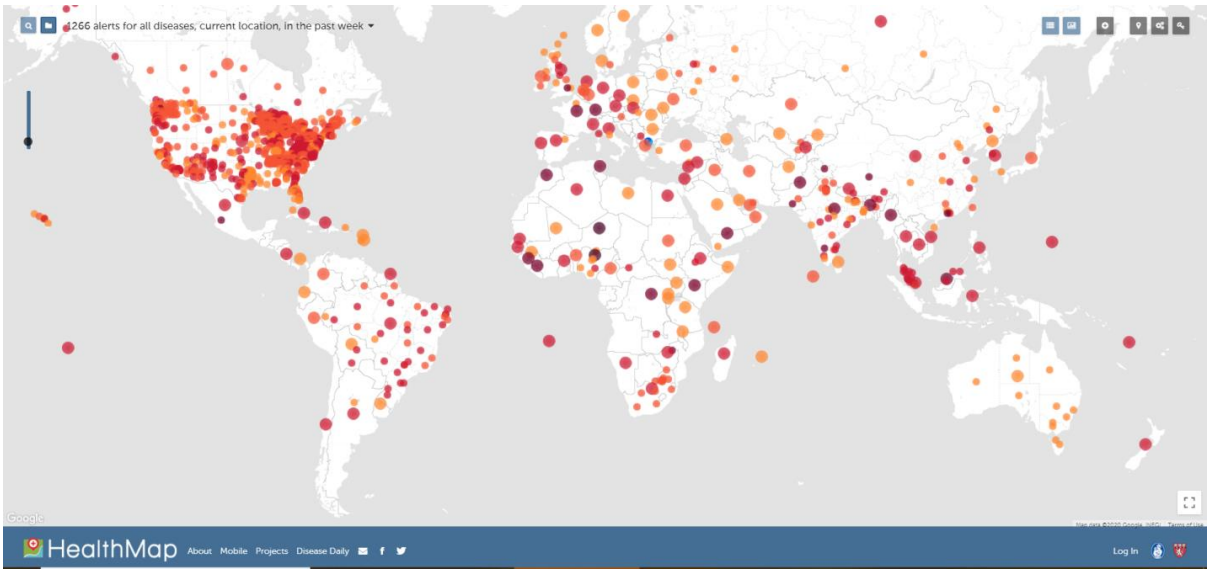
Τα τελευταία 15 χρόνια η ραγδαία αύξηση της διαθεσιμότητας ψηφιακών δεδομένων, της χρήσης κινητών τηλεφώνων, και άλλων ηλεκτρονικών συσκευών, από το μεγαλύτερο μέρος των ανθρώπων στον πλανήτη καθώς και η μαζική αξιοποίηση του διαδικτύου οδήγησε, όπως αναφέραμε προηγουμένως, σε μια σημαντικότατη εξέλιξη του πεδίου της ψηφιακής επιδημιολογίας. Συνεπώς, η εξέλιξη αυτή συνοδεύτηκε από αρκετές υλοποιήσεις σε αυτό το πεδίο, οι οποίες περιλαμβάνουν εφαρμογές όπως ανεξάρτητα συστήματα, που έχουν σχεδιαστεί από εταιρίες για ιδιωτικές χρήσεις, και ηλεκτρονικές σελίδες με διαδραστικούς χάρτες, για την ενημέρωση των πολιτών. Οι υλοποιήσεις αυτές μπορεί να διαφέρουν ως προς τον σκοπό για τον οποίο δημιουργήθηκαν, ως προς τις πηγές δεδομένων που χρησιμοποιούν και ως προς τα συστήματα που διαθέτουν για την συλλογή, επεξεργασία, ανάλυση και αποθήκευση των δεδομένων και των αποτελεσμάτων τους. Στη συνέχεια θα γίνει μια περιγραφική ανάλυση μερικών από τις πιο γνωστές από αυτές τις εφαρμογές, καθώς και του τρόπου λειτουργίας τους και της αποδοτικότητάς τους. Επίσης, θα αναφερθούν οι πηγές δεδομένων που χρησιμοποιούν, θα αιτιολογηθεί η χρήση τους και θα εξηγηθεί η αποχή από τη χρήση κάποιων συγκεκριμένων πηγών δεδομένων.

3.1 HealthMap

Το HealthMap είναι μια αφιλοκερδής υλοποίηση που έχει ως σκοπό την συλλογή και την απεικόνιση σημαντικών πληροφοριών που σχετίζονται με κινδύνους για τη δημόσια υγεία και, βασικά, πληροφοριών για επικείμενες ή εξελισσόμενες επιδημίες, με στόχο την ενημέρωση του ανθρώπινου πληθυσμού. Η πρόσβαση στην πλατφόρμα του HealthMap γίνεται εύκολα είτε μέσω μιας ηλεκτρονικής σελίδας είτε μέσω μιας εφαρμογής για φορητές ηλεκτρονικές συσκευές και είναι δωρεάν, έτσι ώστε όλοι οι διαφορετικής προέλευσης και κοινωνικού επιπέδου χρήστες να έχουν τη δυνατότητα να ενημερωθούν για κινδύνους που αφορούν τη δημόσια υγεία στην περιοχή τους. Το HealthMap δεν αποσκοπεί μόνο στην ενημέρωση των πολιτών, αλλά και των τοπικών συστημάτων υγείας και των κυβερνήσεων, έτσι ώστε να μπορούν έγκαιρα να καταστρωθούν στρατηγικές και να ληφθούν μέτρα αντιμετώπισης μιας επιδημίας. Αποτελεί μια από τις παλιότερες και μακροβιότερες εφαρμογές στο πεδίο της ηλεκτρονικής επιδημιολογίας με πολύ μεγάλο ποσοστό επιτυχίας ως προς το περιεχόμενο που προσφέρει. Δημιουργήθηκε το 2006 στο Boston Children’s Hospital από μια ομάδα ερευνητών, επιδημιολόγων και προγραμματιστών, που μέχρι σήμερα είναι οι υπεύθυνοι για τη λειτουργία του. Στις μέρες μας αποτελεί παγκόσμιο πρωτοπόρο στη χρήση ανεπίσημων ηλεκτρονικών πηγών δεδομένων για τον εντοπισμό και την παρακολούθηση ξεσπασμάτων ασθενειών, και για τον εντοπισμό σε πραγματικό χρόνο κινδύνων για τη δημόσια υγεία που μπορεί να μην προέρχονται αποκλειστικά από ασθένειες.

Η διεπαφή του HealthMap αποτελείται από έναν διαδραστικό παγκόσμιο χάρτη, ο οποίος παρέχει πληροφορίες για τυχόν κινδύνους για την δημόσια υγεία σε κάθε περιοχή, αλλά και

συνολικά στον κόσμο. Το σύστημα εντοπίζει επίσης την τοποθεσία του κάθε χρήστη και, αρχικά, παρέχει πληροφορίες για την κοντινή στο χρήστη περιοχή, επιτυγχάνοντας έτσι τη σωστή ενημέρωση του χρήστη στα επείγοντα ζητήματα που τον αφορούν άμεσα, αποτρέποντας την παράβλεψη ενός κινδύνου. Επίσης ο χρήστης έχει τη δυνατότητα να περιορίσει τις απεικονιζόμενες πληροφορίες κάνοντας πέντε διαφορετικές εξειδικευμένες αναζητήσεις. Οι αναζητήσεις αυτές μπορούν να σχετίζονται με πληροφορίες για μία συγκεκριμένη τοποθεσία, για μια συγκεκριμένη χρονολογία, για μία από τις πηγές μιας ασθένειας ή για μία συγκεκριμένη ασθένεια, καθώς επίσης μπορούν να γίνουν και αναζητήσεις για ασθένειες που δεν προσβάλλουν τους ανθρώπους αλλά κάποιον άλλο έμβιο οργανισμό. Μία άλλη δυνατότητα που παρέχει το HealthMap στον χρήστη είναι η άμεση σύνδεση του με τις πηγές δεδομένων που οδηγούν σε μία ειδοποίηση από στο σύστημα. Έτσι, επιλέγοντας μια συγκεκριμένη ειδοποίηση, ο χρήστης μπορεί να αναζητήσει τα ηλεκτρονικά άρθρα και τις πηγές των πληροφοριών που οδήγησαν στην ειδοποίηση αυτή. Ακόμη, μπορεί και ο ίδιος ο χρήστης, πέρα από την ενημέρωσή του, να αναφέρει έναν κίνδυνο για τη δημόσια υγεία ο οποίος δεν υπάρχει προς το παρόν στο σύστημα, παρέχοντας τις απαραίτητες πληροφορίες. Οι πληροφορίες αυτές στη συνέχεια οποίες ελέγχονται από την εφαρμογή και, αν είναι όντως αληθείς και βάσιμες, απεικονίζονται και αυτές στη διεπαφή.



Εικόνα 3.1: Η διεπαφή του HealthMap με τα κρούσματα ασθενειών σε όλο τον κόσμο. [74]

Το HealthMap αντλεί τις πληροφορίες που απαιτούνται για τη λειτουργία του από ελεύθερες ηλεκτρονικές πηγές δεδομένων που δεν θέτουν περιορισμούς στη χρήση των πληροφοριών που παρέχουν. Η επεξεργασία των πληροφοριών γίνεται μέσω ενός συστήματος τεχνητής νοημοσύνης το οποίο τις επαληθεύει, τις συσχετίζει και τις ομαδοποιεί έτσι ώστε να είναι εύκολα προσβάσιμες μέσω των εξειδικευμένων αναζητήσεων των χρηστών. Μια βασική διαφορά του HealthMap σε σχέση με άλλες αντίστοιχες υλοποιήσεις είναι ότι δεν χρησιμοποιεί δεδομένα από ηλεκτρονικά μέσα κοινωνικής δικτύωσης και, ειδικότερα, χρησιμοποιεί ως επί το πλείστον δεδομένα που προέρχονται από επίσημες πηγές. Αυτή είναι μια επιλογή που κάνει ένας μικρός αριθμός εφαρμογών στο πεδίο της ψηφιακής επιδημιολογίας, και δικαιολογείται από τη μεγάλη δυσκολία επεξεργασίας των δεδομένων αυτής της κατηγορίας. Τα δεδομένα αυτά, όπως έχει προαναφερθεί, έχουν μεγάλη ανομοιομορφία και, λόγω της αναξιοπιστίας που επίσης τα χαρακτηρίζει, η επαλήθευση των αποτελεσμάτων που αντλούμε από αυτά αποτελεί πρόκληση.

Έτσι το HealthMap περιορίζεται στην άντληση δεδομένων από τις παρακάτω πηγές:

- *ProMED Mail*: Είναι ένα πρόγραμμα παρακολούθησης πρωτοεμφανιζόμενων ασθενειών που παρέχεται από την ISID (International Society for Infectious Diseases).

- *World Health Organization (WHO)*: Ο Παγκόσμιος Οργανισμός Υγείας των Ηνωμένων Εθνών.

- *GeoSentinel*: Ηλεκτρονικός φρουρός παρακολούθησης ταξιδιωτών για κλινική μελέτη. Παρέχεται από την ISTM (International Society of Travel Medicine) και το CDC (Center of Disease Control).

- *World Organization of Animal Health (OIE)*: Ο διακυβερνητικός οργανισμός υπεύθυνος για τη βελτίωση της υγείας των ζώων παγκοσμίων.

- *Food and Agriculture Organization of the United Nations (FAO)*: Ο διακυβερνητικός οργανισμός για την εξασφάλιση παγκοσμίως καλή ποιότητα τροφών και αγροτική παραγωγικότητα.

- *EuroSurveilance*: Ένα πρόγραμμα που αφορά τη συλλογή επιστημονικών πληροφοριών για την παρακολούθηση και των έλεγχο ασθενειών στην Ευρώπη που παρέχεται από το ECDC (European Centre for Disease Prevention and Control)

- *Google News*: Μία υπηρεσία που παρέχεται από τη Google και συλλέγει μαζικά ειδήσεις από όλο τον κόσμο και τις αναμεταδίδει.

- *Moreover*: Μία υπηρεσία που παρέχεται από τη VeriSign και συλλέγει μαζικά ειδήσεις από όλο τον κόσμο και τις αναμεταδίδει.

- *Wildlife Data Integration Network (WDIN)*: Ένα δίκτυο παροχής ειδήσεων που ανήκει στο WDIN. Το WDIN είναι ένας ηλεκτρονικός χάρτης που παρέχει πληροφορίες για τις ασθένειες των ζώων κάθε περιοχής και δημιουργήθηκε στη Σχολή Κτηνιατρικής του Μάντισον, που βρίσκεται στο Ουισκόνσιν των ΗΠΑ.

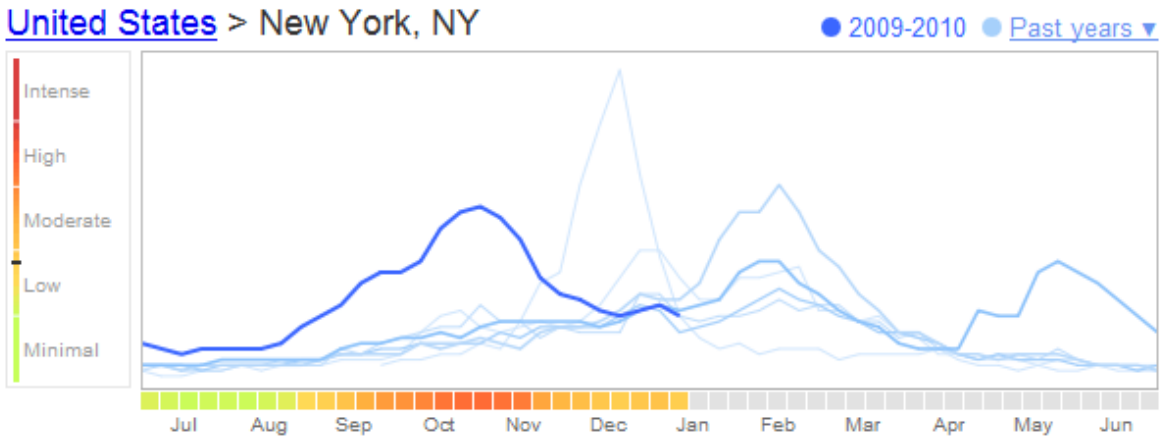
- *Baidu News*: Ένας κόμβος αναμετάδοσης ειδήσεων που παρέχεται από την Κινεζική ηλεκτρονική μηχανή αναζήτησης Baidu, η οποία αποτελεί την πιο δημοφιλή στην Κίνα.

- *SOSO Info*: Ένας κόμβος αναμετάδοσης ειδήσεων που παρέχεται από την Κινεζική ηλεκτρονική μηχανή αναζήτησης Soso.

Για την κατασκευή του HealthMap χρησιμοποιήθηκαν λειτουργικό σύστημα Linux, ο εξυπηρετητής Apache, η γλώσσα προγραμματισμού PHP (που ενδείκνυται για την κατασκευή διαδικτυακών εφαρμογών), καθώς και το εργαλείο MySQL για την κατασκευή της βάσης δεδομένων. Επίσης, η εφαρμογή HealthMap στηρίζεται για τη λειτουργία της στα ελεύθερης χρήσης ηλεκτρονικά προϊόντα Google Maps, GoogleMapAPI for PHP, Google Translate API και xajax PHP AJAX library. Επίσης, για τον διαχωρισμό των επιθυμητών πληροφοριών από της άχρηστες, η εφαρμογή χρησιμοποιεί φίλτρα Μπέις τα οποία εφαρμόζονται με την τεχνική των Φίσερ – Ρόμπινσον (όπως περιγράφεται στο βιβλίο ‘’A Statistical Approach to the Spam Problem’’) και αφαιρούν τις άχρηστες πληροφορίες σε ποσοστό μεγαλύτερο του 99%. [75][76]

3.2 Google Flu Trends

Το Google Flu Trends αποτελούσε μια υπηρεσία της Google η οποία τέθηκε πρώτη φορά σε λειτουργία το 2008. Η υπηρεσία αυτή είχε ως στόχο τη συλλογή πληροφοριών σχετικά με τα κρούσματα γρίπης, που εμφανίζονταν κατά περιόδους σε διάφορα σημεία του κόσμου, με σκοπό την ενημέρωση του ανθρώπινου πληθυσμού. Το Google Flu Trends παρείχε πληροφορίες μέσω χρονικών διαγραμμάτων και γραφικών παραστάσεων για περισσότερες από 25 χώρες και συνέλλεγε πληροφορίες μέσω των αναζητήσεων που έκαναν οι χρήστες της ηλεκτρονικής μηχανής αναζήτησης της Google.

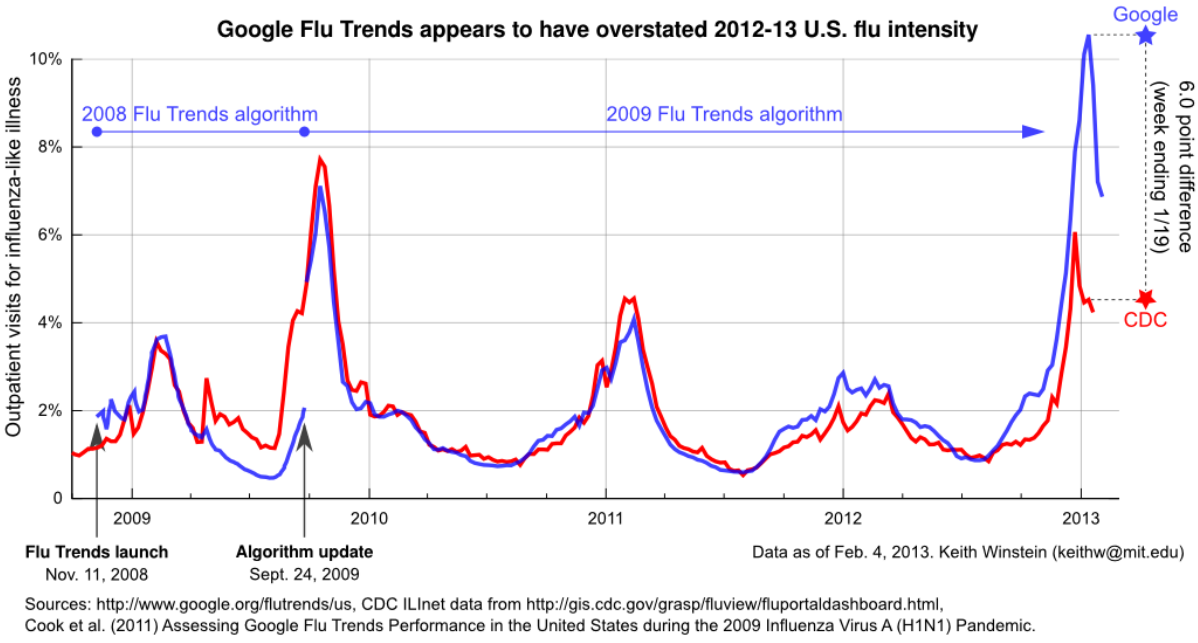


Εικόνα 3.2: Η διεπαφή του Google Flu Trends και μια πρόβλεψη για τα ποσοστά της γρίπης μέσα στην περίοδο Ιούλιος 2009 – Ιούνιος 2010. [77]

Ουσιαστικά το Google Flu Trends στηριζόταν στην ιδέα ότι μπορούσε να επιβεβαιώσει την ύπαρξη γρίπης σε μία περιοχή, και να κάνει προβλέψεις για την εξάπλωση της, όταν οι αναζητήσεις των χρηστών της περιοχής σχετιζόταν, κατά ένα μεγάλο ποσοστό, με τη γρίπη. Σίγουρα δεν μπορεί

κανείς να αρνηθεί πως οι αναζητήσεις των χρηστών του διαδικτύου μπορούν να παρέχουν μία καλή εικόνα για την υγεία, τα ενδιαφέροντα και τον τρόπο ζωής των χρηστών, όμως για τη σωστή λειτουργία μίας ηλεκτρονικής υπηρεσίας που ενημερώνει τον ανθρώπινο πληθυσμό για ένα τόσο σημαντικό θέμα όσο η υγεία πρέπει να διατίθενται και άλλα μέσα επιβεβαίωσης. Αυτή η αβλεψία της Google οδήγησε μετά από επτά χρόνια, στις 9 Αυγούστου του 2015, στην λήξη παροχής δεδομένων από την υπηρεσία Google Flu Trends, καθώς μερικές φορές οι προβλέψεις της απείχαν κατά πολύ από την πραγματικότητα. Επίσης, παρά το γεγονός πως η εφαρμογή χρησιμοποιούσε δεδομένα χρηστών χωρίς όμως να αποκαλύπτει προσωπικά στοιχεία τους και την ταυτότητά τους, επιδιώκοντας την αποφυγή κατηγοριών για μη προστασία ιδιωτικών δεδομένων, αρκετές συστάσεις έγιναν στη Google κατά τη διάρκεια λειτουργίας της εφαρμογής, χωρίς όμως να κατηγορηθεί επίσημα για παραβίαση προσωπικών δεδομένων.

Παρόλα αυτά το Google Flu Trends υπήρξε για κάποια χρόνια ιδιαίτερα επιτυχημένο και σε πολλές περιπτώσεις οι προβλέψεις που έκανε, ακόμα και για δύο εβδομάδες στο μέλλον, ήταν ακριβείς και προηγούνταν κατά πολύ από αυτές του Κέντρου Ελέγχου Ασθενειών(CDC) των ΗΠΑ. Τέλος, για αρκετά μεγάλες περιόδους διατηρούσε ένα μεγάλο ποσοστό ακριβείας της τάξης του 97% σε σχέση με τα πραγματικά δεδομένα που διέθετε το Κέντρο Ελέγχου Ασθενειών. [20][21][78]



Εικόνα 3.3: προβλέψεις του Google Flu Trends σε σχέση με τα πραγματικά δεδομένα που διατίθενται από το Κέντρο Ελέγχου Ασθενειών (CDC) και η μεγάλη απόκλησή τους το 2013. [79]

3.3 BlueDot

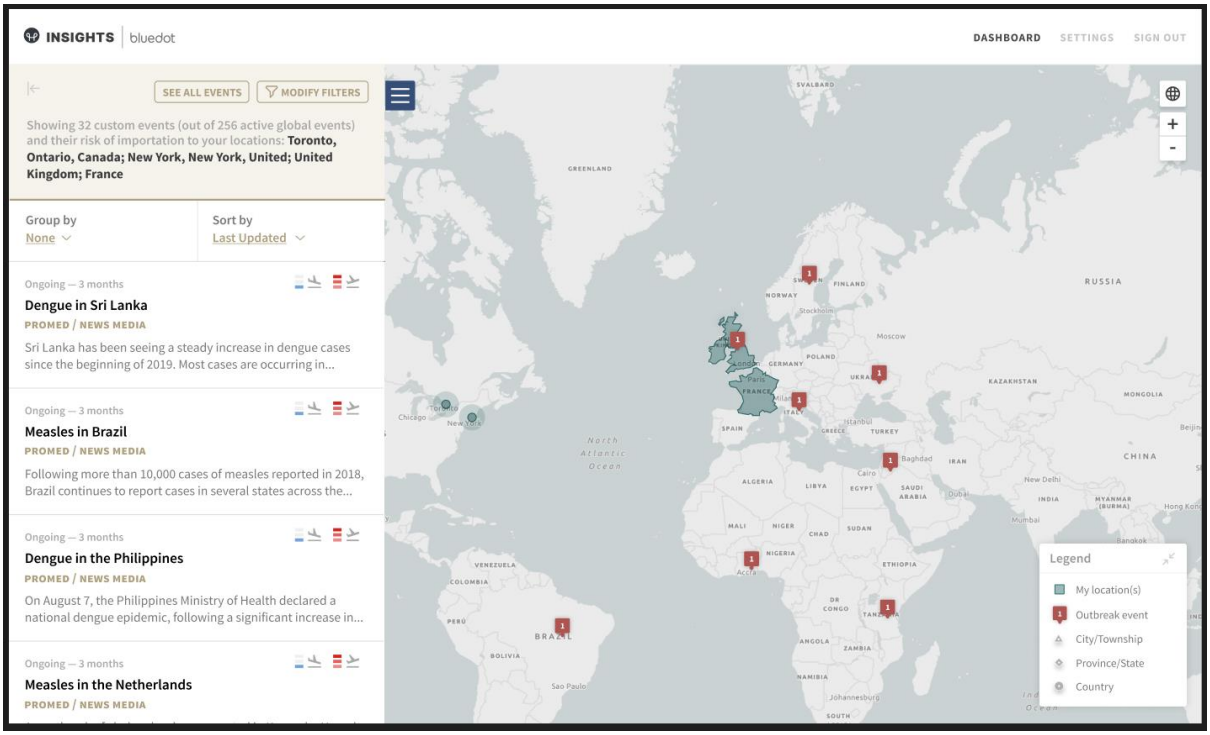
Η BlueDot είναι μία σχετικά νεοϊδρυθείσα εταιρεία και ξεκίνησε να λειτουργεί το 2013 στο Τορόντο του Καναδά. Ο ιδρυτής της Kamran Khan, όντας επιδημιολόγος, συμμετείχε στην πρώτη γραμμή αντιμετώπισης του ξεσπάσματος του Σοβαρού Οξέος Αναπνευστικού Συνδρόμου (SARS) το 2003 στο Τορόντο του Καναδά, και αυτή η εμπειρία τον οδήγησε στην αναζήτηση νέων μεθόδων αντιμετώπισης των επιδημιών. Έτσι, μετά από δέκα χρόνια αποφάσισε να δημιουργήσει την BlueDot, μια εταιρεία που χρησιμοποιεί ένα σύστημα τεχνητής νοημοσύνης για την πρόβλεψη,

παρακολούθηση και αντιμετώπιση μελλοντικών ξεσπασμάτων ασθενειών, των οποίων οι επιπτώσεις είναι τεράστιες τόσο για την οικονομία όσο και για την ανθρώπινη ζωή.

Το σύστημα τεχνητής νοημοσύνης της BlueDot χρησιμοποιεί ένα βαθύ νευρωνικό δίκτυο το οποίο επεξεργάζεται όλες τις εισερχόμενες πληροφορίες από τις διάφορες πηγές δεδομένων, τις ιεραρχεί και βγάζει συμπεράσματα, ενώ ταυτόχρονα εξελίσσεται και γίνεται όλο και πιο έξυπνο. Για την αναγνώριση λεπτομερέστερων μοτίβων και για τη δημιουργία ακόμα πιο δυσδιάκριτων συσχετίσεων, το βαθύ νευρωνικό δίκτυο που χρησιμοποιείται στο συγκεκριμένο σύστημα δεν είναι μονοκατευθυντικό αλλά λειτουργεί με ανάδραση, ανακυκλώνοντας τα δεδομένα μέσα από τα επίπεδα του νευρωνικού δικτύου, οδηγώντας σε εν τω βάθη ανάλυση των δεδομένων. Τέλος, αξίζει να αναφερθεί πως τα αποτελέσματα και οι προβλέψεις του συστήματος τεχνητής νοημοσύνης αντιπαρατίθενται με πραγματικά δεδομένα που προέρχονται από επιδημιολόγους, έτσι ώστε να επιβεβαιωθεί η εγκυρότητα των αποτελεσμάτων, να υπολογιστεί η ακρίβεια του συστήματος και να διασφαλισθεί η σωστή λειτουργία του.

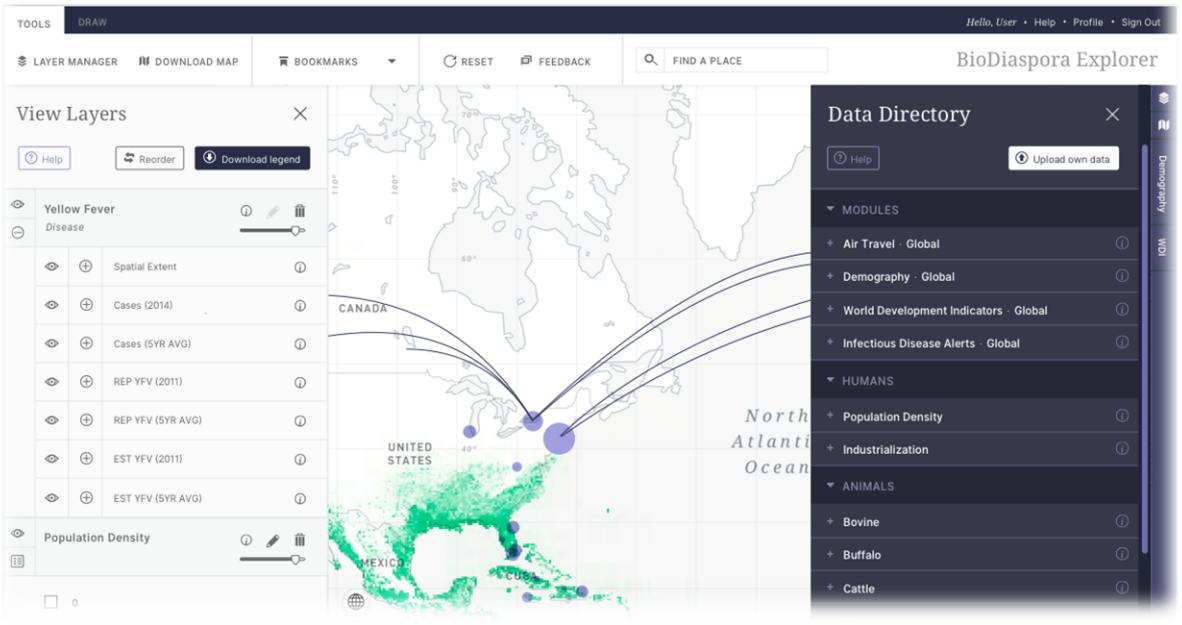
Η εταιρία παρέχει διάφορες υπηρεσίες οι οποίες αφορούν κυρίως φορείς κυβερνήσεων, οργανισμούς και εταιρείες που ασχολούνται με την τη δημόσια υγεία και στοχεύουν στην πρόβλεψη και αντιμετώπιση επιδημιών και στην έγκαιρη ενημέρωση του παγκόσμιου πληθυσμού. Οι υπηρεσίες της BlueDot παρέχονται μέσω δύο εφαρμογών οι οποίες είναι το BlueDot Insights και το BlueDot BioDiaspora Explorer.

Το BlueDot Insights είναι μια εφαρμογή που χρησιμοποιεί το σύστημα τεχνητής νοημοσύνης της BlueDot για να παρέχει σε πραγματικό χρόνο ειδοποιήσεις για την εμφάνιση μεταδοτικών ασθενειών στα διάφορα σημεία του κόσμου. Επίσης, παρέχει πληροφορίες για όλες τις μεταδοτικές ασθένειες που πλήττουν κάθε στιγμή τα διάφορα σημεία του πλανήτη. Οι πληροφορίες και οι ειδοποιήσεις που παρέχονται από το BlueDot Insights μπορούν να προσωποποιηθούν ανάλογα με τις ρυθμίσεις του χρήστη της εφαρμογής και, έτσι, να απεικονίζονται μόνο δεδομένα σχετικά με την ευρύτερη περιοχή στην οποία βρίσκεται ο χρήστης. Ο βασικός στόχος αυτής της εφαρμογής είναι η ταχύτατη ενημέρωση του πληθυσμού, καθώς και των διαφόρων φορέων στον τομέα της υγείας, έτσι ώστε να μπορούν να ληφθούν τα απαραίτητα μέτρα για την αντιμετώπιση και τον περιορισμό της εξάπλωσης μιας ασθένειας όσο το δυνατόν πιο γρήγορα, με στόχο τη μείωση των απωλειών ανθρωπίνων ζωών. Το BlueDot Insights είναι διαθέσιμο ως εφαρμογή για υπολογιστές αλλά και για φορητές συσκευές.



Εικόνα 3.4: *Η διεπαφή του BlueDot Insights.* [80]

Το BlueDot BioDiaspora Explorer είναι μια πλατφόρμα GIS (Geographic Information System Mapping), δηλαδή μια πλατφόρμα που παρέχει πληροφορίες μέσω ενός παγκόσμιου χάρτη, η οποία λειτουργεί μέσω του Cloud. Η πλατφόρμα παρέχει πληροφορίες σχετικά με την πορεία εξέλιξης διαφόρων ασθενειών που πλήττουν τα διάφορα σημεία του κόσμου σχεδόν σε πραγματικό χρόνο. Επίσης το BlueDot BioDiaspora Explorer περιλαμβάνει και περισσότερα από 100 δελτία δεδομένων τα οποία παρέχουν πληροφορίες σχετικές με την εξάπλωση μιας ασθένειας (π.χ. πληροφορίες σχετικές με τις αερομεταφορές). Η πλατφόρμα διαθέτει, επιπλέον, και κάποια εργαλεία και φίλτρα που μπορούν να εφαρμοστούν, τα οποία παρέχουν πληροφορίες σχετικά με τη διασπορά και κάποια άλλα στατιστικά μεγέθη που αφορούν την εξάπλωση μιας συγκεκριμένης ασθένειας. Βασικός στόχος του BlueDot BioDiaspora Explorer είναι, εκτός από την ενημέρωση των διαφόρων χρηστών, η διευκόλυνση στην εκτέλεση ερευνών που στοχεύουν στον υπολογισμό των επιπτώσεων και των κινδύνων που παρουσιάζει κάθε ασθένεια για τον ανθρώπινο πληθυσμό. Αυτές οι πληροφορίες είναι ιδιαίτερα σημαντικές και απαραίτητες για την απόφαση αντιμετώπισης ή όχι μιας ασθένειας, αναλόγως του ρίσκου που παρουσιάζει. Όλες οι κινήσεις και τα μέτρα που λαμβάνονται για την αντιμετώπιση μιας ασθένειας, όπως έχουμε αναφέρει, επηρεάζουν σε πολύ μεγάλο βαθμό την οικονομία μιας χώρας. Έτσι, τέτοιες κινήσεις πρέπει να γίνονται μόνο όταν είναι απολύτως απαραίτητο.



Εικόνα 3.5: Η διεπαφή του BioDiaspora Explorer της BlueDot. [81]

Οι πηγές δεδομένων του συστήματος της BlueDot περιλαμβάνουν ιστοσελίδες μέσων μαζικής ενημέρωσης δεκάδων διαφορετικών γλωσσών, δελτία αναφορών από ηλεκτρονικά δίκτυα παρακολούθησης ασθενειών στη χλωρίδα και την πανίδα του πλανήτη καθώς και δεδομένα αερομεταφορών που έχουν ιδιαίτερα μεγάλη αξία στην πρόβλεψη και την παρακολούθηση της μετάδοσης μιας επιδημίας. Όπως και στην περίπτωση του HealthMap δεδομένα από ηλεκτρονικά μέσα κοινωνικής δικτύωσης δεν χρησιμοποιούνται λόγω της ανομοιομορφίας, της δυσκολίας επεξεργασίας και της αβέβαιης αξίας που τα χαρακτηρίζουν. [21][82][81]

3.4 Global Public Health Intelligence Network (GPHIN)

Το GHPIN είναι ένα δίκτυο εντοπισμού και παρακολούθησης κινδύνων για την παγκόσμια υγεία, το οποίο δημιουργήθηκε το 1997 ως προϊόν της συνεργασίας της κυβέρνησης του Καναδά με τον Παγκόσμιο Οργανισμό Υγείας (WHO). Το 2016, μετά από συμφωνία της Υπηρεσίας Δημόσιας Υγείας (PHAC) και του Εθνικού Συμβουλίου Ερευνών του Καναδά (NRC), κρίθηκε απαραίτητος ο επανασχεδιασμός του, έτσι ώστε να μπορεί να χρησιμοποιεί όλα τα νέα εργαλεία στον τομέα του λογισμικού (επεξεργασία φυσικής γλώσσας, μηχανική μάθηση κτλ.), με στόχο την ταχύτερη επεξεργασία και ανάλυση δεδομένων. Έκτοτε, η πλατφόρμα του GPHIN αναβαθμίζεται ανά τακτά χρονικά διαστήματα και πλέον αποτελεί ένα από τα βασικότερα και δημοφιλέστερα συστήματα και στον τομέα της ανίχνευσης επιδημιών.

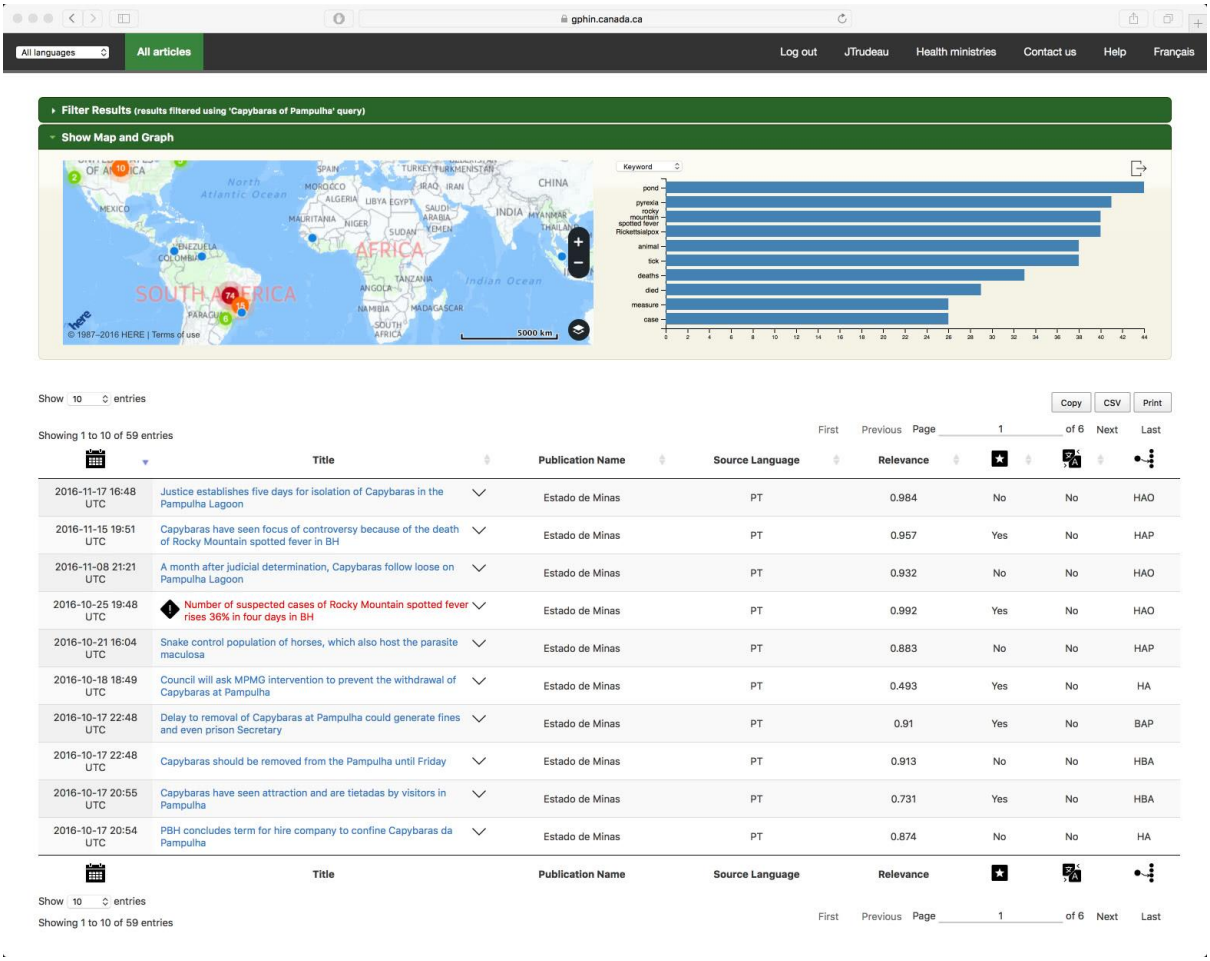
Το GPHIN συλλέγει δεδομένα από σελίδες ηλεκτρονικών μέσων μαζικής ενημέρωσης, από μέσα κοινωνικής δικτύωσης καθώς και από επίσημες αναφορές γεγονότων τα οποία σχετίζονται με κινδύνους ως προς την παγκόσμια υγεία. Τα αρχικά δεδομένα (κυρίως αναφορές και άρθρα), που συλλέγονται από την πλατφόρμα, μπορούν να προέρχονται από 10 διαφορετικές γλώσσες και, πριν από την επεξεργασία τους, μεταφράζονται, μέσω ενός συστήματος επεξεργασίας φυσικής γλώσσας, στα Αγγλικά. Επειδή τα δεδομένα προς μετάφραση περιλαμβάνουν επιστημονική ορολογία σε

διάφορες γλώσσες, για την σωστή μετάφρασή τους, το σύστημα επεξεργασίας φυσικής γλώσσας πρέπει να αναβαθμίζεται συνεχώς. Η αναβάθμιση, όμως, του συστήματος είναι μια δύσκολη και χρονοβόρα διαδικασία και απαιτεί την απενεργοποίηση του GPHIN για μεγάλα χρονικά διαστήματα, κάτι ιδιαίτερα ανεπιθύμητο. Το πρόβλημα αυτό λύνεται με τη χρήση εξωτερικών επιστημονικών λεξικών τα οποία μπορούν να αντικαθίστανται χωρίς να είναι απαραίτητη η απενεργοποίηση της πλατφόρμας. Τέλος, μετά τη μετάφρασή τους, τα δεδομένα, υφίστανται επεξεργασία από την πλατφόρμα και στη συνέχεια ταξινομούνται, ομαδοποιούνται και ιεραρχούνται.

Η ταξινόμηση των δεδομένων γίνεται ως προς την ημερομηνία δημιουργίας τους και τον τόπο προέλευσής τους και, έτσι, δημιουργούνται ομάδες δεδομένων στις οποίες τα δεδομένα έχουν άμεση σχέση μεταξύ τους. Έτσι, με τη μέθοδο αυτή, οι χρήστες του GPHIN έχουν τη δυνατότητα να κάνουν εξειδικευμένες αναζητήσεις, που αφορούν συγκεκριμένες περιοχές του κόσμου, και να λαμβάνουν μόνο τα αντίστοιχα δεδομένα από το σύστημα.

Η ιεράρχηση των δεδομένων έχει να κάνει με τη σοβαρότητα του κινδύνου που παρουσιάζεται για την παγκόσμια υγεία, βάσει των πληροφοριών που λαμβάνονται από μια συγκεκριμένη ομάδα δεδομένων. Η ιεράρχηση γίνεται μέσω ενός δείκτη σχετικότητας, ο οποίος, ουσιαστικά, είναι ένας αριθμός. Ο τρόπος της ανάθεσης ενός δείκτη σχετικότητας σε μία ομάδα δεδομένων από το GPHIN καθορίζεται και μεταβάλλεται από τους αναλυτές που έχουν αναλάβει την υποστήριξη της πλατφόρμας μέσω ενός συστήματος ανάδρασης, έτσι ώστε να γίνεται πάντα σωστή ιεράρχηση των κινδύνων. Επομένως, με τη σωστή χρήση αυτής της μεθόδου, σε μια αναζήτηση ενός χρήστη, προβάλλονται πάντα πρώτα τα δεδομένα υψηλής σημασίας και έτσι διασφαλίζεται η σωστή ενημέρωση των χρηστών και εκμηδενίζεται η πιθανότητα κάποια σημαντική πληροφορία να παραβλεφθεί.

Το GPHIN, σήμερα, χρησιμοποιείται από χρήστες που προέρχονται από παραπάνω από 30 χώρες, οι οποίοι μπορεί να προέρχονται από κυβερνητικούς φορείς υγείας, από μη κυβερνητικούς οργανισμούς διασφάλισης της υγείας και ακόμα και από ιδιωτικές εταιρείες. Οι πληροφορίες που λαμβάνονται από την πλατφόρμα μπορεί να αφορούν μεταδοτικές, ή και όχι, ασθένειες, φυσικές καταστροφές, ανακλήσεις επικίνδυνων φαρμάκων, βιολογικούς κινδύνους, ραδιολογικούς κίνδυνους από ακτινοβολίες κτλ.. Οι πληροφορίες αυτές μπορεί να έχουν τη μορφή στατιστικών πινάκων, διαγραμμάτων ή και διαδραστικών χαρτών, όπου τα δεδομένα απεικονίζονται με χρήση διαφόρων χρωμάτων στις πληγείσες περιοχές του κόσμου.



Εικόνα 3.6: Η διεπαφή του GPHIN. [83]

Το GPHIN αποτελεί ένα από τα δημοφιλέστερα και σημαντικότερα συστήματα στον τομέα της διασφάλισης της υγείας παγκοσμίως και έχει αποδείξει πάρα πολλές φορές την αξία του στα χρόνια λειτουργίας του. Το 2003 εντόπισε το ξέσπασμα του SARS (Severe Acute Respiratory Syndrome) στην Κίνα μέσω άρθρων Κινεζικών εφημερίδων που αναφέρονταν σε αυξημένες πωλήσεις αντιικών φαρμάκων, ενώ επίσης παρείχε πληροφορίες για τη σωστή χορήγηση των φαρμάκων και για το ποια φάρμακα είχαν μεγαλύτερη αποτελεσματικότητα. Στη συνέχεια, το 2009, εντόπισε τα πρώτα κρούσματα της γρίπης του H1N1 μέσω μιας αναφοράς, σε μια εφημερίδα γραμμένη στην Ισπανική γλώσσα, για δύο θανάτους στην πόλη Βερακρούζ του Μεξικό, ενώ το 2012 ήταν το πρώτο σύστημα που εντόπισε οκτώ θανάτους στην Ιορδανία που αργότερα αποδείχθηκε πως οφείλονταν στον ιό MERS-CoV (Middle East Respiratory Syndrome Coronavirus). Τέλος, από το 2014 μέχρι το 2016 που διήρκησε η επιδημία του ιού Ebola στη δυτική Αφρική, το GPHIN παρακολουθούσε την εξέλιξη της επιδημίας και παρείχε πληροφορίες για την εξέλιξή της, για αλλαγές και ακυρώσεις πτήσεων στην περιοχή και για ελέγχους που γίνονταν στα σύνορα των πληγέντων χωρών. [84][85][86]

3.5 Program for Monitoring Emerging Disesase (ProMED-mail)

Το ProMED-mail είναι ένα παγκόσμιο σύστημα παρακολούθησης ξεσπασμάτων ασθενειών και παροχής ειδοποιήσεων. Αποτελεί μια συγκεντρωτική πλατφόρμα βασισμένη στην ανταλλαγή ηλεκτρονικών μηνυμάτων και ξεκίνησε να λειτουργεί το 1994 υπό την αιγίδα της Παγκόσμιας Κοινότητας Μεταδοτικών Ασθενειών (International Society for Infectious Diseases).

Το ProMED-mail λειτουργεί συνδρομητικά και, σήμερα, διαθέτει πάνω από 80000 χρήστες από περισσότερες από 200 χώρες. Η πλατφόρμα είναι ανοικτή για χρήση από οποιονδήποτε ενδιαφερόμενο και, έτσι, διαθέτει χρήστες οι οποίοι μπορεί να προέρχονται από τον ιατρικό κλάδο, από τον τομέα της ενημέρωσης ή ακόμα και από πολιτικούς κύκλους. Το σύστημα παρέχει στους χρήστες του διάφορες εμπεριστατωμένες πληροφορίες και άρθρα που μπορεί να αφορούν ξεσπάσματα νέων ασθενειών, επανεμφανίσεις ασθενειών και αναφορές για επικίνδυνες τοξίνες σε συγκεκριμένες περιοχές.

Τα δεδομένα εισάγονται στο σύστημα του ProMED-mail από τους συνδρομητές του, αλλά και από άλλες πηγές, με τη μορφή άρθρων ή άλλων επίσημων αναφορών. Αυτές οι άλλες πηγές περιλαμβάνουν άρθρα εφημερίδων, αναφορές αγροτών και κτηνοτρόφων, απόψεις ειδικών, κυβερνητικές ή και όχι αναφορές και ιατρικές-ερευνητικές αναφορές. Η επαλήθευση αυτών των δεδομένων γίνεται από μια ομάδα εθελοντών, κυρίως ειδικών, οι οποίοι σήμερα προέρχονται από 32 χώρες ενώ ο αριθμός τους συνεχώς αυξάνεται. Η χρήση του ανθρώπινου στοιχείου στην επαλήθευση των δεδομένων μπορεί να θεωρηθεί ταυτόχρονα πλεονέκτημα και μειονέκτημα, καθώς πάντα είναι επιθυμητή η γνώμη ειδικών σε έναν τόσο ευαίσθητο τομέα όπως η υγεία, αλλά μερικές φορές η άποψη ενός ανθρώπου μπορεί να είναι περισσότερο υποκειμενική παρά αντικειμενική.

Οι πληροφορίες που προβάλλει το σύστημα ομαδοποιούνται σε λίστες. Υπάρχει μία βασική λίστα που προβάλλει όλες τις αναφορές που έχουν ληφθεί πρόσφατα με χρονολογική σειρά, ενώ, επίσης, διατίθεται και ένας μεγάλος αριθμός εξειδικευμένων λιστών οι οποίες διαφοροποιούνται μεταξύ τους ως προς τη γλώσσα, το θέμα ή τη γεωγραφική περιοχή που αφορούν οι αναφορές που περιλαμβάνουν.

Οι χρήστες του ProMED-mail έχουν τη δυνατότητα να κάνουν εξειδικευμένες αναζητήσεις ως προς την περιοχή, την ημερομηνία και τη γλώσσα των πληροφοριών που τους αφορούν, ενώ, επίσης, μπορούν να αναζητήσουν άρθρα που αναφέρονται σε συγκεκριμένα παθογόνα. Η αναζήτηση των πληροφοριών γίνεται μέσω λέξεων-κλειδιών και εκτελείται από ένα ειδικό σύστημα με αρχιτεκτονική αγωγού, σχεδιασμένο σε Python. Το σύστημα αυτό κάνει χρήση συμπτωτικών δικτύων για τη συσχέτιση των δεδομένων, ενώ για την επίτευξη μεγαλύτερης ακρίβειας γίνεται και εφαρμογή γράφων γνώσης στα ήδη συσχετισμένα δεδομένα.

Σήμερα, το ProMED-mail θεωρείται ένα από τα πιο αποδοτικά συστήματα στην παρακολούθηση και αντιμετώπιση επιδημιών. Αυτό γίνεται εύκολα αντιληπτό αν λάβουμε υπόψη πως αρκετά δημοφιλή συστήματα σε αυτόν τον τομέα, όπως HealthMap, το χρησιμοποιούν ως μία από τις πηγές δεδομένων τους. Μέσα στα χρόνια λειτουργίας του η υψηλή αξία του ProMED-mail στη διασφάλιση της παγκόσμιας υγείας έχει επαληθευτεί πολλές φορές όπως το 2003 στον γρήγορο εντοπισμό των πρώτων κρουσμάτων του SARS (Severe Acute Respiratory Syndrome) και στις 14

Μαρτίου του 2014 όπου επισημάνθηκαν από την πλατφόρμα κάποιες πρώτες αναφορές για κρούσματα του Ebola στη Γουινέα της Δυτικής Αφρικής. [87][88][89]

3.6 BioCaster

Το BioCaster είναι ένα συνδρομητικό, μη-κυβερνητικό, ελεύθερο για χρήση, ερευνητικό πρόγραμμα παρακολούθησης ροών δεδομένων από διάφορες πηγές με στόχο τη διαφύλαξη της δημόσιας υγείας. Λειτούργησε από το 2006 μέχρι και το 2012 και αποτελείται από μια διαδικτυακή πλατφόρμα εξόρυξης δεδομένων από κείμενα, με σκοπό τον εντοπισμό νέων ξεσπασμάτων ασθενειών και την παρακολούθηση της πορείας εξάπλωσής τους. Το σύστημα λαμβάνει δεδομένα από 1700 διαδικτυακές ροές δεδομένων RSS (Really Simple Syndication), τα οποία στη συνέχεια τα αναλύει, τα συσχετίζει και τα ιεραρχεί, έτσι ώστε να γίνει δυνατός ο εντοπισμός υγειονομικών κινδύνων που οφείλονται σε μεταδοτικές ασθένειες.

Το BioCaster λαμβάνει ανεπίσημα δεδομένα τα οποία μπορεί να έχουν αρκετές διαφορετικές μορφές και να έχουν συνταχθεί σε πολλές διαφορετικές γλώσσες, ανάλογα με την πηγή από την οποία έχουν προέλθει. Ουσιαστικά, τα δεδομένα αυτά είναι κείμενα τα οποία το σύστημα στη συνέχεια επεξεργάζεται λέξη προς λέξη. Η πλατφόρμα είναι συμβατή με 8 διαφορετικές γλώσσες και για την επεξεργασία των δεδομένων εισόδου χρησιμοποιεί ένα μεγάλης κλίμακας σύστημα ανάλυσης γεγονότων. Το σύστημα ανάλυσης γεγονότων αποτελείται 4 στάδια και σε κάθε στάδιο γίνεται όλο και πιο λεπτομερής ανάλυση των κειμένων προς επεξεργασία, μέχρι τον εντοπισμό ενός συγκεκριμένου συμβάντος. Ένα συγκεκριμένο συμβάν ορίζεται ως μια ομάδα πληροφοριών που περιγράφουν πλήρως ένα γεγονός (πχ. η εμφάνιση μιας συγκεκριμένης ασθένειας σε συγκεκριμένη ημερομηνία και τοποθεσία).

Τα 4 στάδια ανάλυσης του BioCaster είναι τα εξής:

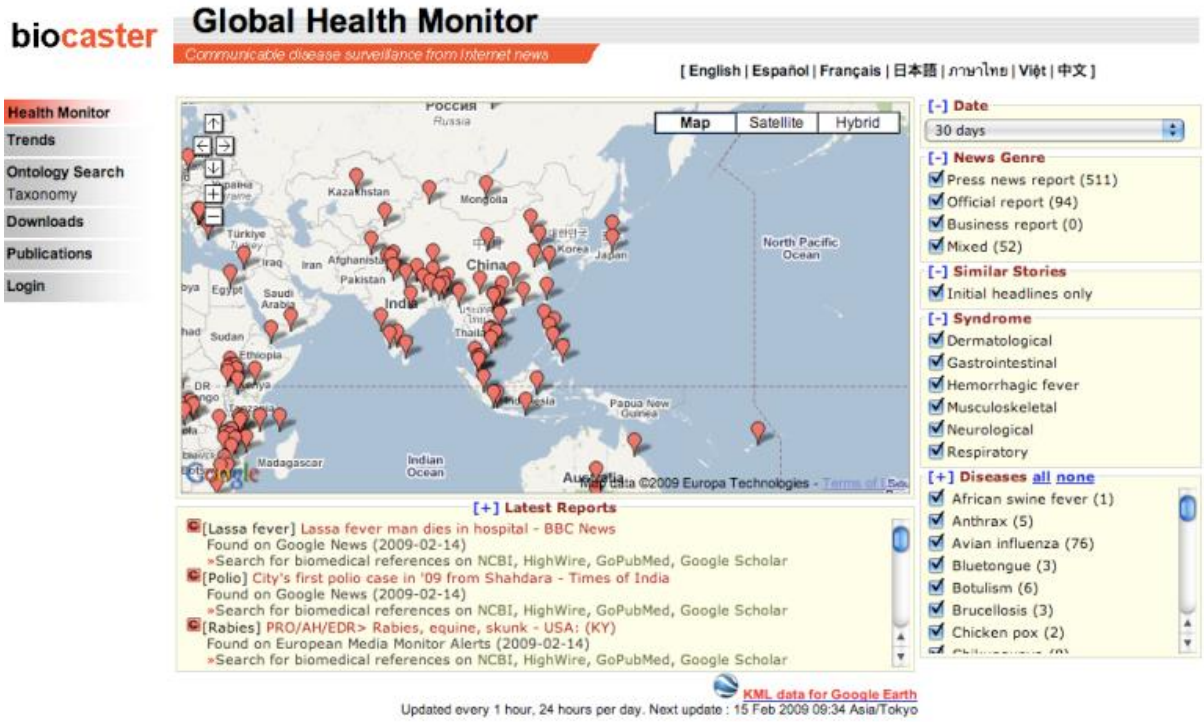
- *Αναγνώριση θέματος*: Στο αρχικό στάδιο αναγνωρίζεται το θέμα το οποίο αφορά το κείμενο υπό επεξεργασία.

- *Αναγνώριση συγκεκριμένων οντοτήτων*: Σε αυτό το στάδιο αναγνωρίζονται συγκεκριμένες χαρακτηριστικές λέξεις ή φράσεις που σχετίζονται με το θέμα του κειμένου υπό επεξεργασία. (πχ. λέξεις σχετικές με ασθένειες ή τοποθεσίες)

- *Εντοπισμός συγκεκριμένης ασθένειας και γεωγραφικής τοποθεσίας*: Σε αυτό το στάδιο εξακριβώνονται οι ασθένειες και οι τοποθεσίας που αναφέρονται μέσα στα κείμενα υπό επεξεργασία.

- *Αναγνώριση γεγονότος*: Στο τελικό στάδιο έχουν εξακριβωθεί όλες οι απαραίτητες πληροφορίες για την περιγραφή ενός συγκεκριμένου συμβάντος. (ασθένεια, ημερομηνία και τόπος)

Για την απεικόνιση των επιβεβαιωμένων συμβάντων το BioCaster χρησιμοποιεί μια διεπαφή βασισμένη στο Google Maps στην οποία απεικονίζει τα επιβεβαιωμένα γεγονότα με βάση τη γεωγραφική τους τοποθεσία, ενώ επίσης είναι δυνατή η ενημέρωση των συνδρομητών του μέσω ηλεκτρονικών μηνυμάτων.



Εικόνα 3.7: Η διεπαφή του BioCaster. [90]

Ένα βασικό προτέρημα του BioCaster έναντι άλλων συστημάτων επεξεργασίας ροών δεδομένων RSS είναι το ότι συνδυάζει 4 πολύ βασικά χαρακτηριστικά, τα οποία είναι:

- Η δυνατότητα της εξόρυξης δεδομένων από κείμενα.
- Η δυνατότητα της αναγνώρισης ενδείξεων επικινδυνότητας μέσα σε κείμενα.
- Η ικανότητα του συστήματος να δημιουργεί λογικές συσχετίσεις μεταξύ των δεδομένων ακόμη και όταν απουσιάζουν βασικές πληροφορίες. (πχ. Η αναφορά σε ένα παθογόνο συσχετίζεται με την ασθένεια την οποία προκαλεί)
- Η ικανότητα του συστήματος να αναγνωρίζει τη σημασία διαφόρων όρων αλλά και τους αντίστοιχους όρους στις 8 γλώσσες με τις οποίες είναι συμβατό.

Το σύστημα του σχεδιάστηκε με τη χρήση ενός λογισμικού που είναι βασισμένο σε Linux και ονομάζεται NPACI Rocks, ενώ συγκεκριμένα ο κώδικας που αναλαμβάνει την συγκέντρωση των δεδομένων και την αναγνώριση νέων ροών δεδομένων σχεδιάστηκε σε Pearl. [91][92][93]

3.7 Europe Media Monitor (EMM) και Medical Information System (MedISys)

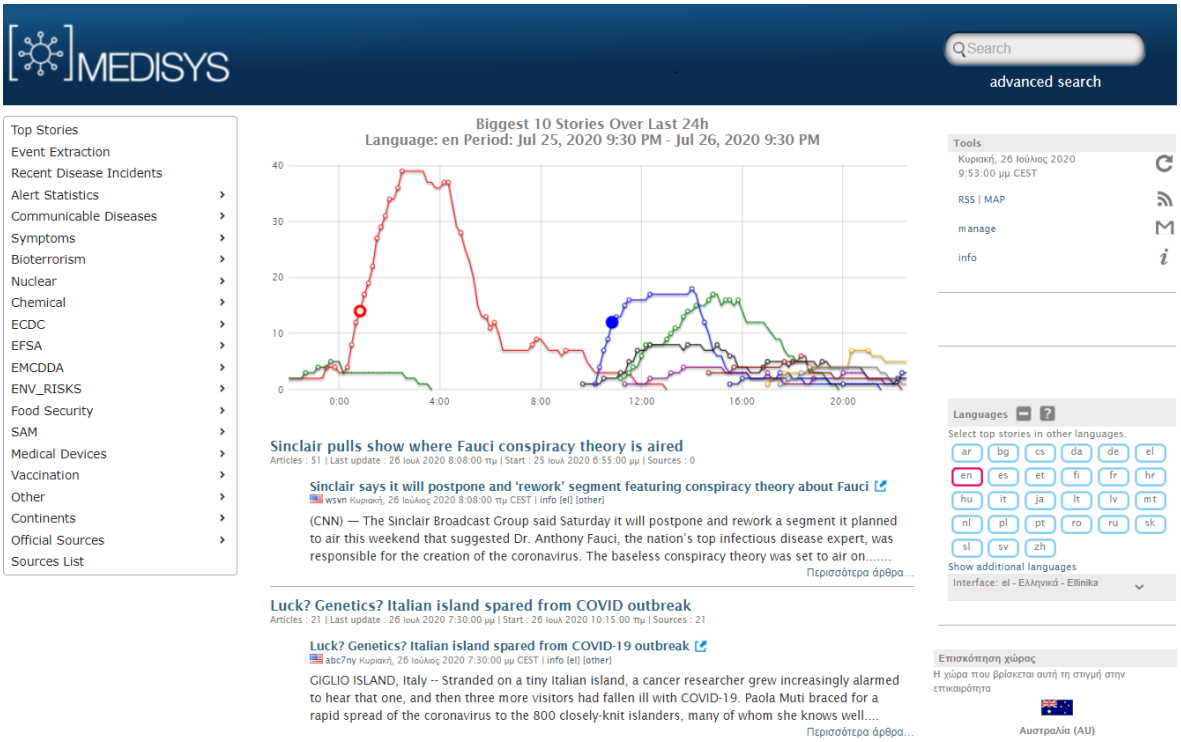
Το EMM είναι ένα Ευρωπαϊκό πρόγραμμα συλλογής και παρακολούθησης ηλεκτρονικών δεδομένων. Το πρόγραμμα συλλέγει δεδομένα υπό τρείς διαφορετικές μορφές οι οποίες είναι οι RSS ροές δεδομένων, η XMM και οι ιστοσελίδες HTML, ενώ δεδομένα άλλων μορφών

μετατρέπονται αυτόματα στη μορφή UTF-8 πριν από την επεξεργασία τους. Στη συνέχεια τα δεδομένα αυτά τροφοδοτούνται στα αρκετά διαφορετικά υποσυστήματα που διαθέτει, το καθένα από τα οποία είναι σχεδιασμένο για διαφορετικό σκοπό. Ένα από αυτά τα υποσυστήματα είναι το MedISys το οποίο αποτελεί ένα πλήρως αυτοματοποιημένο πρόγραμμα εντοπισμού και παρακολούθησης κινδύνων για τη δημόσια υγεία.

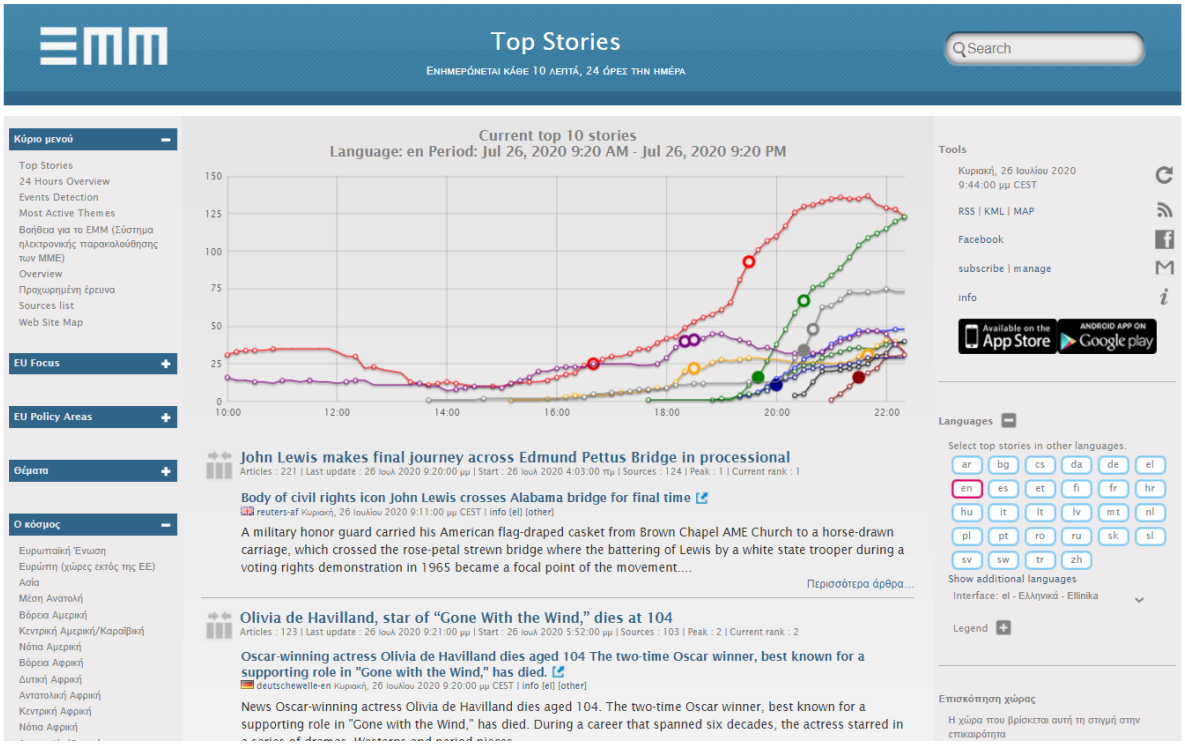
Το MedISys είναι ένα σύστημα ανάλυσης γεγονότων το οποίο δεν περιορίζεται μόνο στον εντοπισμό και την παρακολούθηση ξεσπασμάτων ασθενειών, αλλά επεξεργάζεται και αναλύει ηλεκτρονικά δεδομένα με στόχο τον εντοπισμό κάθε συμβάντος που παρουσιάζει κίνδυνο για την ανθρώπινη ζωή, για την υγεία των φυτών και των ζώων και, γενικότερα, για την ευημερία του πλανήτη. Έτσι, η επισήμανση ενός συμβάντος από το MedISys μπορεί να οφείλεται σε βιολογικούς, χημικούς, ραδιολογικούς ή ακόμα και πυρηνικούς κινδύνους, οι οποίοι σήμερα, εν συντομία, αναφέρονται και ως CBRN (Chemical, Biological, Radiological, Nuclear Warfare). Τέλος, το σύστημα εντοπίζει και μολύνσεις σε τρόφιμα καθώς και διαδικτυακούς ιούς.

Το MedISys συλλέγει δεδομένα από εξειδικευμένες, επίσημες ή και ανεπίσημες, ιατρικές ιστοσελίδες, από ηλεκτρονικά μέσα μαζικής ενημέρωσης καθώς και από κάποια συγκεκριμένα ιστολόγια (blogs). Το σύστημα υποστηρίζει έναν πολύ μεγάλο αριθμό γλωσσών ενώ, επίσης, διαθέτει και τη δυνατότητα μετάφρασης οποιασδήποτε γλώσσας στα Αγγλικά (ακόμα και της Κινεζικής (Μανταρίν) και της Αραβικής). Αξίζει να σημειώσουμε πως από κάθε πηγή δεδομένων συλλέγονται μόνο τα πιο πρόσφατα δεδομένα, έτσι ώστε τα συμβάντα που ανιχνεύει το σύστημα να είναι όσο πιο επίκαιρα γίνεται, αλλά και να μην επιβαρύνεται η πλατφόρμα από την επεξεργασία παρωχημένων δεδομένων. Στη συνέχεια, τα δεδομένα που συλλέγονται από το MedISys, τα οποία είναι κυρίως άρθρα, ταξινομούνται σε προκαθορισμένες κατηγορίες ανάλογα με το θέμα το οποίο αφορούν και τη γλώσσα στην οποία έχουν συνταχθεί. Επιπλέον, για την σωστή κατηγοριοποίηση των άρθρων το σύστημα διαθέτει τη δυνατότητα αναγνώρισης συγκεκριμένων ονομασιών (πχ. ονόματα ανθρώπων, οργανισμών, τοποθεσιών κτλ.). Η συλλογή και η ανάλυση των δεδομένων γίνεται αυτόματα και ασταμάτητα χωρίς την ανάγκη χρήσης αναλυτών και, έτσι, η πλατφόρμα μπορεί να επεξεργάζεται πολύ μεγάλους όγκους δεδομένων σε πολύ μικρά χρονικά διαστήματα. Στο τελικό στάδιο της ανάλυσης, το MedISys, εξάγει συγκεκριμένα συμβάντα τα οποία ιεραρχούνται ανάλογα με το μέγεθος του κινδύνου που παρουσιάζουν.

Οι χρήστες του MedISys αποκτούν πρόσβαση στο πρόγραμμα μέσω μιας εξειδικευμένης διεπαφής, μέσω της οποίας μπορούν να αναζητήσουν και να προβάλουν συγκεκριμένα άρθρα ανάλογα με το θέμα που τους ενδιαφέρει. Επίσης, οι αναζητήσεις αυτές μπορούν να εξειδικευτούν ως προς τη χώρα, τη γλώσσα, την ημερομηνία και την πηγή προέλευσης των άρθρων. Τέλος, η διεπεφή του προγράμματος προφέρει τη δυνατότητα στους χρήστες να παρατηρήσουν την τοπικότητα των δεδομένων που τους ενδιαφέρουν πάνω σε παγκόσμιους χάρτες αλλά και τα αντίστοιχα στατιστικά στοιχεία μέσω γραφικών παραστάσεων.



Εικόνα 3.8: Η διεπαφή του MedISys. [94]



Εικόνα 3.9: Η διεπαφή του Europe Media Monitor. [95]

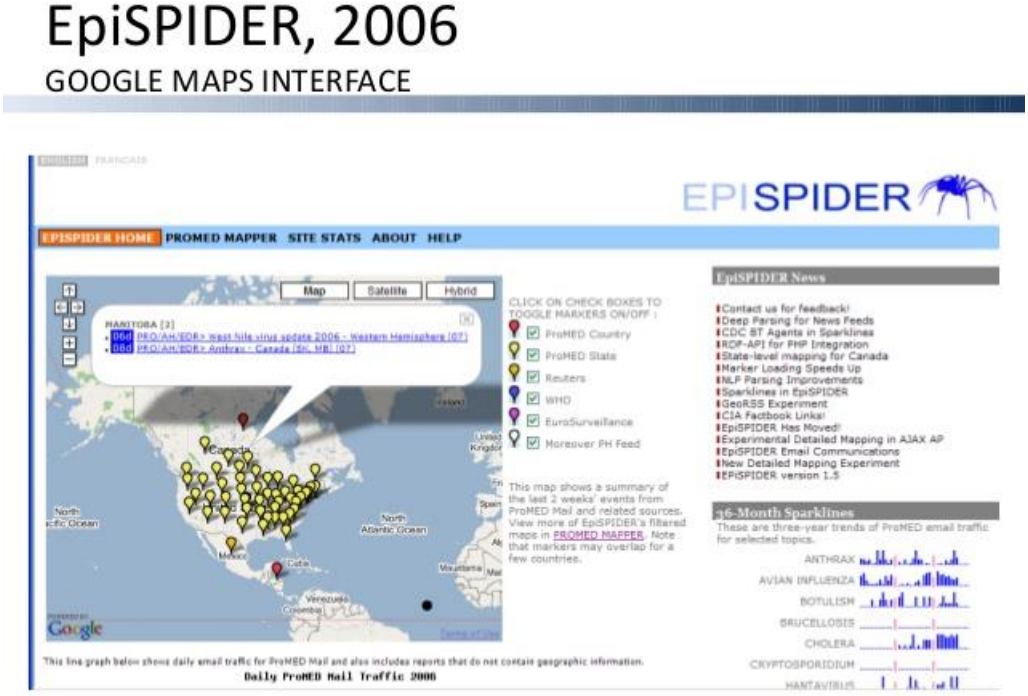
Πέρα από τις δυνατότητες που προσφέρει η διεπαφή του MedISys στους χρήστες του, οι αναλυτές που χρησιμοποιούν το πρόγραμμα και οι εγγεγραμμένοι χρήστες έχουν πρόσβαση και σε ένα εξειδικευμένο εργαλείο που ονομάζεται NewsDesk. Με τη χρήση του NewsDesk, ένας χρήστης μπορεί να κάνει ακόμα πιο εξειδικευμένες αναζητήσεις, να δημιουργήσει δικές του αναφορές και να στείλει ειδοποιήσεις μέσω email και SMS. [96][97][98][99]

3.8 EpiSPIDER

Το EpiSPIDER (Semantic Processing and Integration of Distributed Electronic Resources for Epidemics and Disasters) είναι ένα αυτοματοποιημένο πρόγραμμα συλλογής πληροφοριών σχετικών με ασθένειες σε ανθρώπους και σε ζώα ενώ, επίσης, συλλέγει και πληροφορίες σχετικές με φυσικές καταστροφές. Το πρόγραμμα αρχικά σχεδιάστηκε ως βοήθημα για την ενίσχυση της απόδοσης του ProMED-mail και λειτούργησε για πρώτη φορά τον Ιανουάριο του 2006.

Το EpiSPIDER συλλέγει δεδομένα είτε μέσω του ProMED-mail είτε από ροές δεδομένων RSS, που προέρχονται από ιστοσελίδες μέσων μαζικής ενημέρωσης. Αξίζει να σημειώσουμε πως το μεγαλύτερο μέρος των δεδομένων που συλλέγονται από μέσα μαζικής ενημέρωσης προέρχεται από το κοινωνικό δίκτυο Twitter. Επίσης, ως συμπληρωματικά δεδομένα, συλλέγονται και δημογραφικές πληροφορίες καθώς και πληροφορίες που αφορούν τη δημόσια υγεία. Στη συνέχεια, το σύστημα, επεξεργάζεται, αναλύει και ταξινομεί τα δεδομένα αυτόματα σε διάφορες ομάδες. Η διαδικασία αυτή ενισχύεται με τη χρήση ενός συστήματος επεξεργασίας φυσικής γλώσσας και οι ομάδες, στις οποίες ταξινομούνται τα δεδομένα, σχετίζονται με την ημερομηνία δημιουργίας τους, την γεωγραφική τοποθεσία την οποία αφορούν καθώς και το θέμα στο οποίο αναφέρονται.

Το EpiSPIDER περιλαμβάνει και ένα τελικό στάδιο μετατροπής των ομαδοποιημένων δεδομένων πριν την τροφοδότηση τους στους χρήστες του. Κατά το στάδιο αυτό τα ομαδοποιημένα δεδομένα μετατρέπονται είτε σε ροές δεδομένων RSS είτε σε διαδραστικούς χάρτες. Οι ροές δεδομένων RSS περιλαμβάνουν όλες της απαραίτητες πληροφορίες ως προς την τοποθεσία και το θέμα το οποίο αφορούν, καθώς, επίσης, συνοδεύονται και από γραφικές παραστάσεις σχετικές με τα δεδομένα που περιλαμβάνουν. Οι χρήστες του EpiSPIDER, μπορούν να έχουν πρόσβαση στις ροές δεδομένων RSS και είτε μέσω της πλατφόρμας είτε μέσω του Twitter, ενώ οι διαδραστικοί χάρτες είναι προσβάσιμοι είτε μέσω της πλατφόρμας είτε μέσω ενός blog. Τέλος, για την προβολή των δεδομένων σε διαδραστικούς χάρτες, το EpiSPIDER χρησιμοποιεί τη διεπαφή του Google Maps. [92][100]



Εικόνα 3.10: Η διεπαφή του EpiSPIDER με χρήση του Google Maps. [101]

4 ΚΟΡΟΝΟΪΟΣ

4.1 Εισαγωγή

Μιας και την εποχή που συντάσσεται αυτή η διπλωματική εργασία η ανθρωπότητα μαστίζεται από τη μεγαλύτερη επιδημία όλων των εποχών, θεωρήθηκε σωστό να συμπεριληφθεί ένα κεφάλαιο για την πανδημία του κορονοϊού. Ο κορονοϊός, ή η πανδημία COVID-19 όπως αποκαλείται, αποτελεί τη μεγαλύτερη επιδημία στα χρονικά της ανθρωπότητας, όχι ως προς τη σοβαρότητα των συμπτωμάτων που προκαλεί, αλλά ως προς το εύρος της εξάπλωσής της, που σήμερα καλύπτει κάθε περιοχή του κόσμου.

Η πρώτη επίσημη ανακοίνωση για την ύπαρξη του κορονοϊού προήλθε από Παγκόσμιο Οργανισμό Υγείας (WHO) στις 31 Δεκεμβρίου του 2019, η οποία δήλωνε πως είχε εντοπιστεί μια σειρά κρουσμάτων ιογενούς πνευμονίας αγνώστου προελεύσεως στην πόλη Wuhan, πρωτεύουσα της επαρχίας Hubei, της Κίνας. Όμως, μετά από μελέτες της Κινεζικής κυβέρνησης, αποδεδείχθηκε πως το πρώτο κρούσμα του ιού ήταν ένας πολίτης της επαρχίας Hubei, ηλικίας 55 ετών, στις 17 Νοεμβρίου του 2019. Τελικά, στις 31 Ιανουαρίου του 2020 η Ιταλική κυβέρνηση ανακοίνωσε πως κρούσματα του κορονοϊού είχαν εντοπιστεί και στην Ιταλία και, έκτοτε, ο ιός εξαπλώθηκε ταχύτατα σε όλο τον κόσμο.

Ο κορονοϊός είναι ένας ιός που ανήκει στην υποοικογένεια Orthocoronavirinae της οικογένειας ιών Coronaviridae και επιστημονική του ονομασία είναι SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2). Τα συμπτώματα που προκαλεί επηρεάζουν το αναπνευστικό σύστημα και κυμαίνονται από ήπια έως και πολύ σοβαρά, ακόμα και θανατηφόρα. Οι μελέτες που έχουν γίνει έχουν δείξει πως τα πιο σοβαρά συμπτώματα εκδηλώνονται κυρίως σε μικρά παιδιά, σε ηλικιωμένους και, γενικότερα, σε ανθρώπους που βρίσκονται σε ανοσοκαταστολή, με αδύναμο ανοσοποιητικό σύστημα. [102][103][104][105][106]

4.2 Στατιστικά στοιχεία για τον κορονοϊό

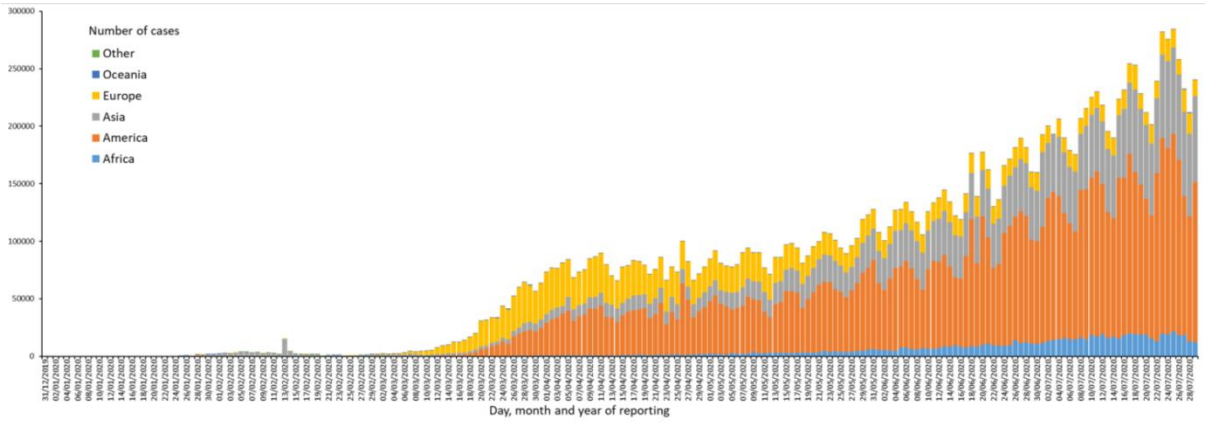
Σε αυτή την ενότητα θα παραθέσουμε κάποια στατιστικά στοιχεία για τον κορονοϊό που αφορούν κυρίως τις ηλικίες των ανθρώπων που προβάλλει, την ένταση των συμπτωμάτων ανά τον αριθμό κρουσμάτων, το ποσοστό θανάτων καθώς και την πορεία εξάπλωσης του ιού.

Όπως αναφέραμε προηγουμένως η ένταση των συμπτωμάτων του κορονοϊού εξαρτάται από διάφορους παράγοντες. Συγκεκριμένα, ως αποτέλεσμα μελετών, έχει διαπιστωθεί πως το 80% των προσβεβλημένων ατόμων εμφανίζουν ήπια ή και καθόλου συμπτώματα, το 15% παρουσιάζει σοβαρά συμπτώματα και υπάρχει ανάγκη χορήγησης οξυγόνου, ενώ το 5% παρουσιάζει πολύ σοβαρά συμπτώματα και οι ασθενείς χρήζουν υποστήριξης της αναπνοής τους με ιατρικά μηχανήματα. Επίσης, το ποσοστό θανάτων κυμαίνεται από 3% έως και 4% των προσβεβλημένων ατόμων, αν και εξαρτάται σε πολύ μεγάλο βαθμό από το βιοτικό επίπεδο του ασθενούς και την ποιότητα του συστήματος υγείας τις χώρας.

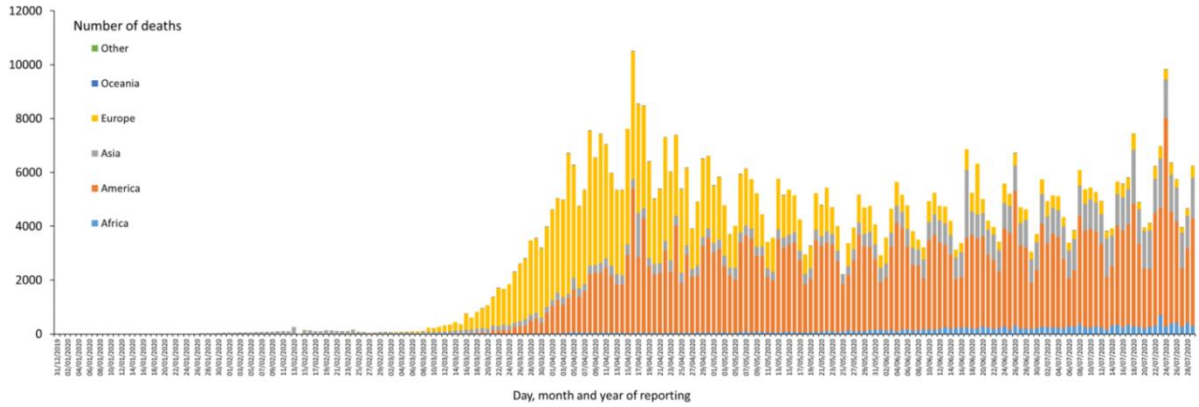
Ο κορονοϊός είναι ένας ιός με πολύ μεγάλη ταχύτητα εξάπλωσης και μεταδίδεται εξ' επαφής, μέσω σταγονιδίων (από το βήχα ή το φτάρνισμα) και μέσω επιφανειών ή αντικειμένων που έχουν έρθει σε επαφή οι φορείς του ιού. Ο δείκτης μετάδοσης του ιού έχει υπολογισθεί μεταξύ 2 και 2.5, όπου ο αριθμός αυτός συμβολίζει τον μέσο αριθμό των νέων κρουσμάτων που προκαλούνται από έναν μεμονωμένο φορέα του ιού. Μετά από έρευνες έχει διαπιστωθεί πως άτομα ηλικίας έως και 19 ετών δεν προσβάλλονται από τον ιό άμεσα, ενώ οι ενήλικες μπορούν να προσβληθούν ως αποτέλεσμα οποιασδήποτε από τις μεθόδους μετάδοσης που αναφέρθηκαν προηγουμένως. Τέλος, έχει παρατηρηθεί το γεγονός πως ο ιός μεταδίδεται από ενήλικες σε νεαρά άτομα, αλλά το αντίστροφο δεν συμβαίνει. Στη συνέχεια περιλαμβάνονται κάποια στοιχεία και διαγράμματα σχετικά με τα κρούσματα, τους θανάτους και την εξάπλωση του κορονοϊού παγκοσμίως. [102]

Χώρα	Συνολικά κρούσματα	Κρούσματα μέσα στις τελευταίες 24 ώρες	Συνολικοί Θάνατοι	Θάνατοι μέσα τις τελευταίες 24 ώρες
Παγκόσμια	16.558.289	215.127	656.093	5.274
ΗΠΑ	4.263.531	54.022	147.449	1.118
Βραζιλία	2.442.375	23.284	87.618	614
Ινδία	1.531.669	48.513	34.193	768
Ρωσία	828.990	5.475	13.673	169
Νότια Αφρική	459.761	7.231	7.257	190
Μεξικό	395.489	4.973	44.022	342
Περού	389.717	4.920	18.418	189
Χιλή	349.800	1.877	9.240	53
Ηνωμένο Βασίλειο	300.696	581	45.878	119
Ιράν	296.173	2.667	16.147	235

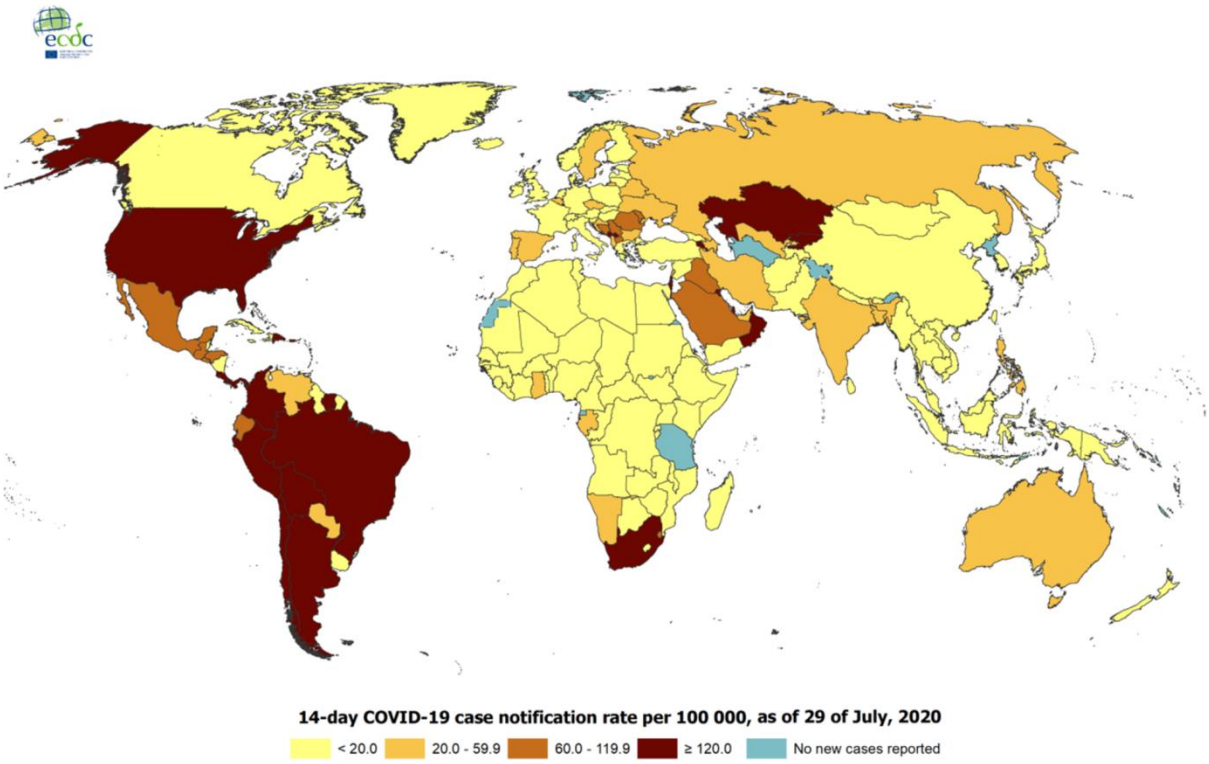
Πίνακας 4.1: Ο αριθμός κρουσμάτων του κορονοϊού και θανάτων, συνολικά και στις 24 τελευταίες ώρες, στις 10 χώρες με τον μεγαλύτερο αριθμό συνολικών κρουσμάτων. (Τελευταία ανανέωση: 29 Ιουλίου 2020, 5.36μμ CEST) [107]



Διάγραμμα 4.1: Η διασπορά των κρουσμάτων του κορονοϊού ανά ήπειρο. (Τελευταία ανανέωση: 29 Ιουλίου 2020, 5.36μμ CEST) [108]

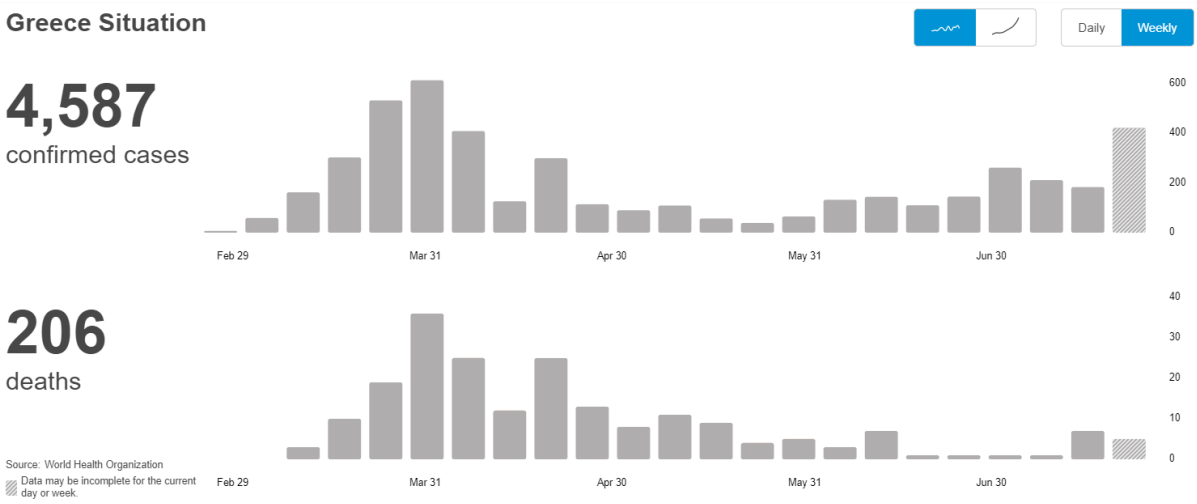


Διάγραμμα 4.2: Η διασπορά των θανάτων εξ' αιτίας του κορονοϊού ανά ήπειρο. (Τελευταία ανανέωση: 29 Ιουλίου 2020, 5.36μμ CEST) [108]



Εικόνα 4.1: Τα κρούσματα του κορονοϊού στις διάφορες περιοχές του κόσμου. (Το χρώμα κάθε περιοχής υποδηλώνει τον αριθμό κρουσμάτων. Όσο πιο σκούρο είναι το χρώμα μιας περιοχής τόσο μεγαλύτερο είναι και το ποσοστό κρουσμάτων σε αυτή την περιοχή) (Τελευταία ανανέωση: 29 Ιουλίου 2020, 5.36μμ CEST) [108]

Αν και πολλές μεγάλες ανεπτυγμένες χώρες έχασαν τον έλεγχο της εξάπλωσης του κορονοϊού, στην Ελλάδα λήφθηκαν τα απαραίτητα μέτρα για την προστασία του πληθυσμού της έγκαιρα και, έτσι, ο συνολικός αριθμός κρουσμάτων και θανάτων περιορίστηκε σε πολύ μεγάλο βαθμό. Στο επόμενο διάγραμμα παρουσιάζονται κάποια στατιστικά στοιχεία για την Ελλάδα συγκεκριμένα.



Διάγραμμα 4.3: Ο αριθμός συνολικών κρουσμάτων και θανάτων εξ' αιτίας του κορονοϊού στην Ελλάδα και τα αντίστοιχα στατιστικά στοιχεία από τη στιγμή του πρώτου κρούσματος (27 Φεβρουαρίου 2020) μέχρι σήμερα. (Τελευταία ανανέωση: 2 Αυγούστου 2020, 2.19μμ CEST) [109]

4.3 Εφαρμογές εντοπισμού και μελέτης του κορονοϊού

Από τις εφαρμογές που περιγράψαμε στο 3ο κεφάλαιο, οι περισσότερες (αυτές που λειτουργούν σήμερα) συμμετείχαν στον εντοπισμό του κορονοϊού και συνεχίζουν να παρακολουθούν την εξάπλωσή του, παρέχοντας πολύ χρήσιμες πληροφορίες στους επιδημιολόγους αλλά και στον παγκόσμιο πληθυσμό. Συγκεκριμένα, λίγο μετά τα μεσάνυχτα στις 30 Δεκεμβρίου, κρούσματα του ιού εντοπίστηκαν από το σύστημα της BlueDot, ενώ την ίδια μέρα ο ιός εντοπίστηκε και από τα συστήματα HealthMap και ProMED-mai. Επίσης, το σύστημα GPHIN και το Ευρωπαϊκό EMM συνεισέφεραν στον εντοπισμό του ιού από τον WHO που οδήγησε στην επίσημη ανακοίνωση στις 31 Δεκεμβρίου του 2019.

Για την παρακολούθηση της πορείας εξέλιξης της πανδημίας COVID-19 αλλά και για την ενημέρωση του παγκόσμιου πληθυσμού, μέσα στους τελευταίους μήνες δημιουργήθηκαν και κάποιες νέες εφαρμογές που αποτυπώνουν τα κρούσματα και τους θανάτους εξ' αιτίας του κορονοϊού στις διάφορες περιοχές του κόσμου, μέσω διαδραστικών χαρτών. Κάποιες από αυτές τις εφαρμογές παρέχουν πληροφορίες σε παγκόσμιο επίπεδο, άλλες έχουν σχεδιαστεί για την τοπική ενημέρωση των κατοίκων μιας χώρας και μερικές από αυτές παρέχουν επίσης στατιστικά στοιχεία και γραφικές παραστάσεις που αφορούν την πορεία εξέλιξης της πανδημίας. Επειδή ο αριθμός αυτών των εφαρμογών είναι ιδιαιτέρα μεγάλος και συνεχίζει να αυξάνεται, στη συνέχεια θα περιγράψουμε μερικές από αυτές τις εφαρμογές ενδεικτικά. [110][111][112][113][86]

4.3.1 WHO Coronavirus Disease (COVID-19) Dashboard

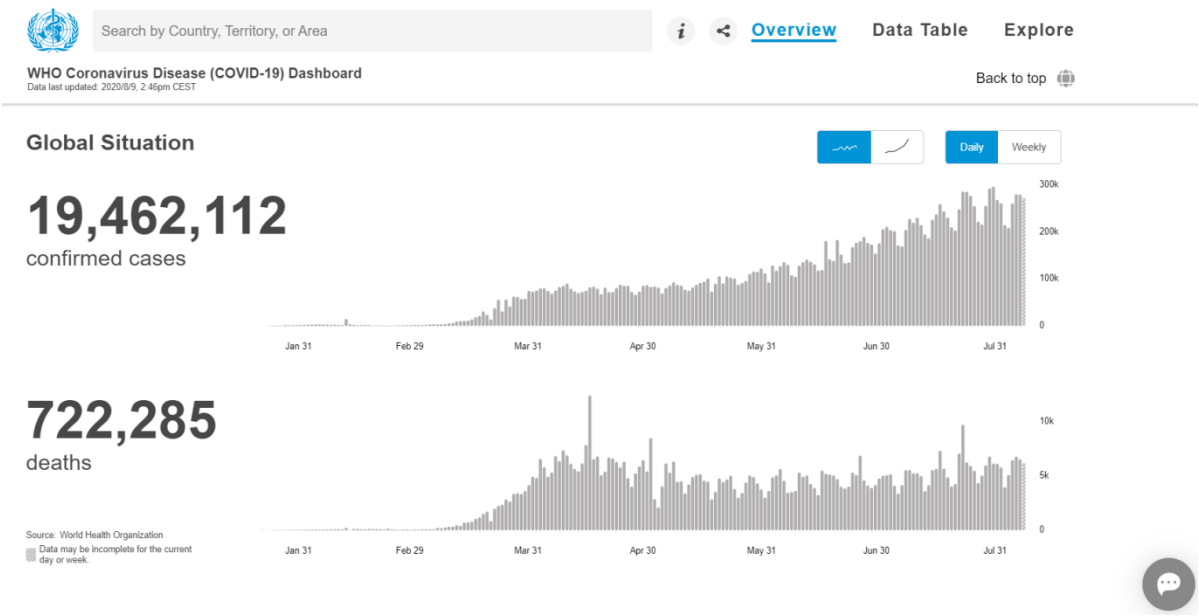
Όπως θα ήταν αναμενόμενο, ο Παγκόσμιος Οργανισμός Υγείας ήταν ο πρώτος φορέας που δημιούργησε μια εφαρμογή για την απεικόνιση δεδομένων σχετικών με την πορεία εξάπλωσης της πανδημίας του κορονοϊού παγκοσμίως. Η εφαρμογή, που ονομάζεται Coronavirus Disease ή COVID-19 Dashboard, ξεκίνησε να λειτουργεί στις 4 Ιανουαρίου του 2020 και έκτοτε παρέχει συνεχή

ενημέρωση για πορεία εξάπλωσης της πανδημίας, ενώ τα δεδομένα που παρέχει να ανανεώνονται καθημερινά.

Το COVID-19 Dashboard περιλαμβάνει έναν διαδραστικό παγκόσμιο χάρτη ο οποίος συνοδεύεται από γραφικές παραστάσεις. Ο διαδραστικός χάρτης απεικονίζει τις διάφορες χώρες του κόσμου με χρήση χρωματικών βαθμίδων. Έτσι, οι χώρες με μικρό ποσοστό κρουσμάτων απεικονίζονται με ανοιχτά χρώματα ενώ οι χώρες με μεγάλο αριθμό κρουσμάτων απεικονίζονται με σκούρα χρώματα. Επίσης, ένας χρήστης της εφαρμογής μπορεί να αναζητήσει περισσότερες πληροφορίες για μια χώρα επιλέγοντάς την στο χάρτη ή μέσω μιας λίστας, κάτι που οδηγεί στην προβολή στοιχείων και γραφικών παραστάσεων για τη συγκεκριμένη χώρα.

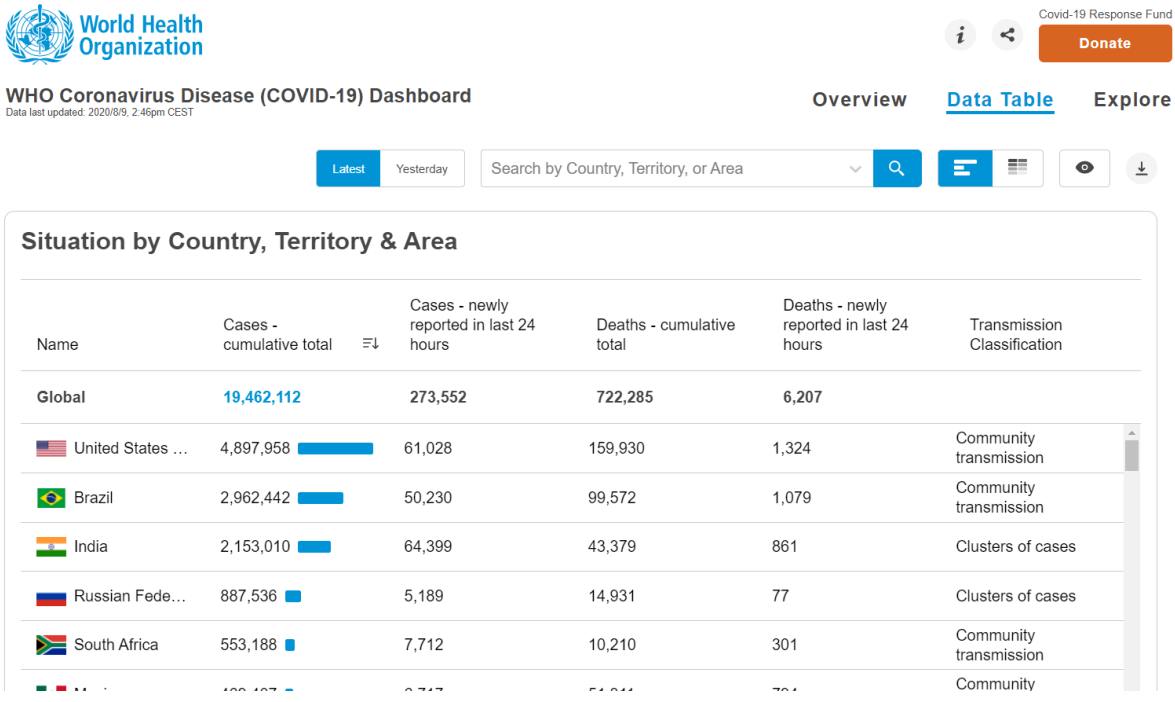
Εικόνα 4.2: Ο διαδραστικός χάρτης του COVID-19 Dashboard του WHO. (Τελευταία ανανέωση: 9 Αυγούστου 2020, 2.46μμ CEST) [109]

Οι γραφικές παραστάσεις που προβάλλονται από το COVID-19 Dashboard περιλαμβάνουν δεδομένα για τα κρούσματα και τους θανάτους εξ' αιτίας του κορονοϊού είτε ημερησίως είτε εβδομαδιαίως. Επίσης, μπορούν να προβληθούν γραφικές παραστάσεις που απεικονίζουν την αθροιστική αύξηση των κρουσμάτων και των θανάτων (Cumulative) ή την ημερήσια-εβδομαδιαία κατανομή τους (Daily-Weekly Change).



Εικόνα 4.3: Η ημερήσια κατανομή των κρουσμάτων και των θανάτων εξ' αιτίας του κορονοϊού, από το COVID-19 Dashboard του WHO. (Τελευταία ανανέωση: 9 Αυγούστου 2020, 2.46μμ CEST) [109]

Το COVID-19 Dashboard περιλαμβάνει, επίσης, και μία άλλη καρτέλα στην οποία προβάλλεται ένας πίνακας με αριθμητικά στοιχεία που αφορούν τα συνολικά κρούσματα και θανάτους παγκοσμίως αλλά και σε κάθε χώρα. Επιπλέον, παρέχονται και πληροφορίες για τα κρούσματα και τους θανάτους μέσα στις τελευταίες 24 ώρες τόσο παγκοσμίως όσο και σε κάθε χώρα. [107][109]



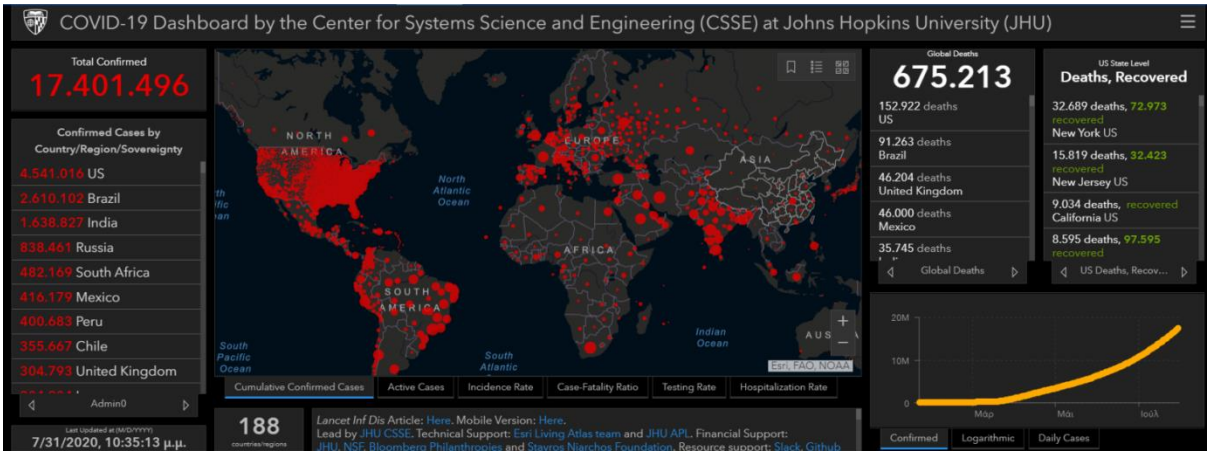
Εικόνα 4.4: Ο πίνακας με τα αριθμητικά στοιχεία των κρουσμάτων και των θανάτων εξ' αιτίας του κορονοϊού, συνολικά και μέσα στις τελευταίες 24 ώρες, παγκοσμίως και σε κάθε χώρα, από το COVID-19 Dashboard WHO. (Τελευταία ανανέωση: 9 Αυγούστου 2020, 2.46μμ CEST) [107]

4.3.2 Johns Hopkins University (JHU) COVID-19 Dashboard

Το COVID-19 Dashboard του JHU είναι μια πλατφόρμα καταγραφής και απεικόνισης επιβεβαιωμένων πληροφοριών σχετικών με την εξάπλωση του κορονοϊού στις διάφορες περιοχές του κόσμου. Η εφαρμογή αποτελεί μέρος του Κέντρου Πληροφοριών για τον κορονοϊό (Coronavirus Resource Center), του Πανεπιστημίου Johns Hopkins, στο οποίο συλλέγονται πληροφορίες για τον ιό και την εξάπλωσή του για διάφορους σκοπούς. Η πλατφόρμα δημιουργήθηκε από την καθηγήτρια Πολιτικών Μηχανικών και Μηχανικών Συστημάτων, του Πανεπιστημίου Johns Hopkins, Lauren Gardner σε συνεργασία με την απόφοιτη μαθήτριά της Ensheng Dong και λειτούργησε για πρώτη φορά στις 22 Ιανουαρίου του 2020. Πλέον, τη διαχείριση και τη διατήρηση της λειτουργίας του συστήματος έχει αναλάβει το Κέντρο Επιστήμης Συστημάτων και Μηχανικής (Center for Science Systems and Engineering – CSSE) του Whiting School of Engineering, το οποίο αποτελεί ένα τμήμα του πανεπιστημίου Johns Hopkins. Τέλος, στην υποστήριξη του συστήματος επίσης συμβάλλουν η εταιρία παροχής γεωγραφικών πληροφοριών ESRI και το Εργαστήριο Εφαρμοσμένης Φυσικής του Πανεπιστημίου Johns Hopkins (Johns Hopkins University Applied Physics Laboratory).

Το COVID-19 Dashboard διαθέτει έναν διαδραστικό παγκόσμιο χάρτη, όπου απεικονίζονται πληροφορίες για τα συνολικά κρούσματα και τους θανάτους εξ’ αιτίας του κορονοϊού στις διάφορες περιοχές του κόσμου. Επίσης, ο χάρτης αυτός συνοδεύεται από γραφικές παραστάσεις, που απεικονίζουν την πορεία αύξησης των κρουσμάτων του ιού, και από αριθμητικά στοιχεία, που αφορούν τα κρούσματα και τους θανάτους παγκοσμίως αλλά και για συγκεκριμένες χώρες. Προς το παρόν το σύστημα παρέχει πληροφορίες για 188 χώρες και οι χρήστες του μπορούν να αναζητήσουν επιπλέον πληροφορίες για την περιοχή που τους ενδιαφέρει επιλέγοντάς την πάνω στο χάρτη. Τέλος, ένα άλλο χαρακτηριστικό της πλατφόρμας είναι ότι εκτός από τον βασικό χάρτη, που απεικονίζει τα συνολικά κρούσματα και τους θανάτους, διατίθενται πέντε επιπλέον χάρτες. Οι τρείς από αυτούς παρέχουν δεδομένα για όλο τον κόσμο και απεικονίζουν τα ενεργά κρούσματα του κορονοϊού, τον αριθμό κρουσμάτων ανά αριθμό κατοίκων και το ποσοστό θανάτων σε κάθε περιοχή. Οι άλλοι δύο χάρτες αφορούν μόνο τις ΗΠΑ και ο ένας παρέχει πληροφορίες για τα ποσοστά των κατοίκων που υπόκεινται σε κλινικό έλεγχο σε κάθε περιοχή, αλλά και για τον αριθμό των συνολικών ελέγχων που έχουν γίνει, ενώ ο άλλος απεικονίζει τα ποσοστά των προσβεβλημένων ατόμων που νοσηλεύονται, καθώς και τον συνολικό αριθμό νοσούντων που έχουν νοσηλευτεί σε κάθε περιοχή.

Το COVID-19 Dashboard του JSU αποτελεί την πιο χρησιμοποιούμενη και προηγμένη εφαρμογή που σχεδιάστηκε με αφορμή την πανδημία του κορονοϊού και παρουσιάζεται στην επόμενη εικόνα. [114]



Εικόνα 4.5: Το COVID-19 Dashboard του Johns Hopkins University. [114]

4.3.3 iMEdD Lab COVID-19 Map

Η iMEdD είναι μια Ελληνική μη κερδοσκοπική δημοσιογραφική εταιρεία που ιδρύθηκε το 2018. Τα κεφάλαια για τη δημιουργία της επιχείρησης προήλθαν από το Ίδρυμα Σταύρος Νιάρχος μέσω μιας ειδικής δωρεάς. Στόχος της iMEdD είναι να προάγει την αξιοπιστία στη δημοσιογραφία, κάτι το οποίο είναι δυσεύρετο σε μια εποχή όπου οι περισσότεροι άνθρωποι ενημερώνονται μέσω κοινωνικών δικτύων. Η εταιρεία συνεργάζεται με το Columbia Journalism School, το CSIS (Center for Strategic and International Studies), το Investigative Europe και το European Data Journalism Network.

Η iMedD περιλαμβάνει τέσσερα τμήματα, που το καθένα έχει διαφορετικό στόχο. Τα τμήματα αυτά είναι το Incubator, το Ideas Zone, το Bridge και το Lab. Το Lab είναι ένα τμήμα το οποίο στοχεύει στην παροχή πρωτογενούς δημοσιογραφικού περιεχομένου προσαρμοσμένου στην ψηφιακή εποχή, χρησιμοποιώντας νέες μεθόδους και εργαλεία. Έτσι, παράγονται και παρουσιάζονται διαδραστικές δημοσιογραφικές έρευνες και αναλύσεις δεδομένων, ενώ ταυτόχρονα ενθαρρύνεται η χρήση αυτών των καινοτόμων μεθόδων από νέους δημοσιογράφους και ερευνητές.

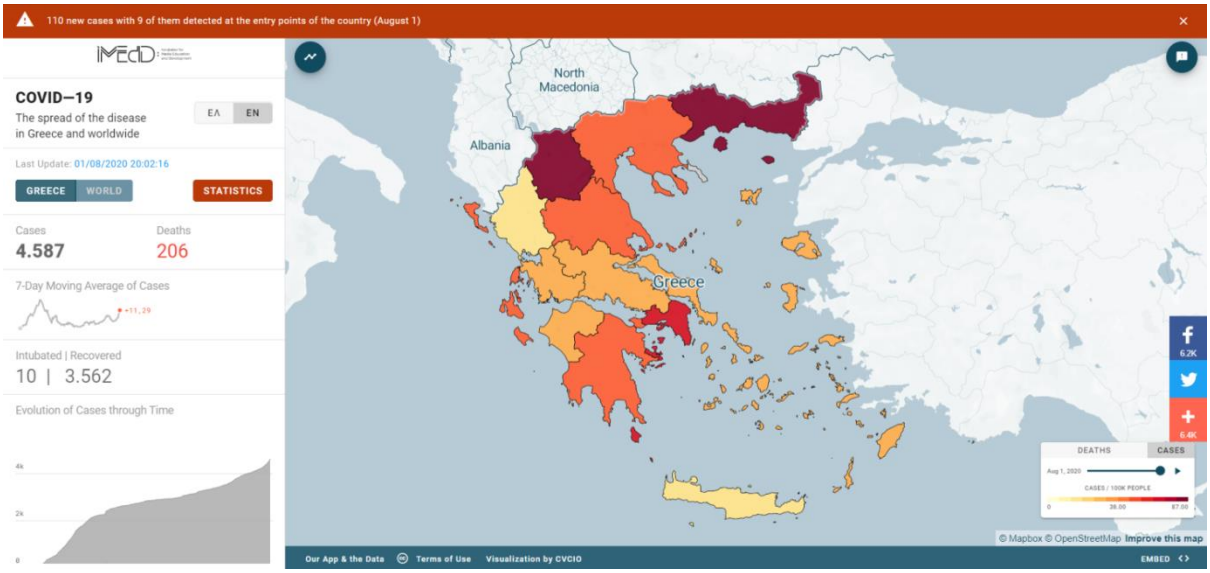
Μία από τις πιο πρόσφατες υλοποιήσεις του iMEdD Lab είναι το COVID-19 Map το οποίο ξεκίνησε να λειτουργεί στις 16 Μαρτίου του 2020 και έκτοτε ανανεώνεται ημερησίως. Το COVID-19 Map είναι μια εφαρμογή απεικόνισης των επιβεβαιωμένων κρουσμάτων και θανάτων εξ' αιτίας του κορονοϊού. Στόχος της εφαρμογής είναι η ενημέρωση της Ελληνικής δημοσιογραφικής κοινότητας και των Ελλήνων πολιτών για την εξέλιξη της πανδημίας όχι μόνο στην Ελλάδα, αλλά και στον κόσμο.

Η εφαρμογή του COVID-19 Map περιλαμβάνει έναν διαδραστικό χάρτη στον οποίο ο αριθμός των κρουσμάτων του κορονοϊού απεικονίζονται σε κάθε περιοχή με χρήση χρωμάτων. Έτσι, τα πιο ανοιχτά χρώματα συμβολίζουν μικρό αριθμό κρουσμάτων ενώ τα πιο σκούρα συμβολίζουν υψηλό αριθμό κρουσμάτων. Επίσης, η εφαρμογή διαθέτει δύο διαφορετικές λειτουργίες όπου στην πρώτη απεικονίζονται δεδομένα μόνο για την Ελλάδα ενώ στην άλλη απεικονίζονται δεδομένα για όλο τον κόσμο. Στη λειτουργία όπου απεικονίζεται η Ελλάδα οι πληροφορίες που παρέχονται είναι πιο λεπτομερείς, καθώς οι διάφορες περιοχές της χώρας απεικονίζονται με διαφορετικά χρώματα. Αντιθέτως, στη λειτουργία που προβάλλεται ο παγκόσμιος χάρτης, κάθε χώρα απεικονίζεται με ένα

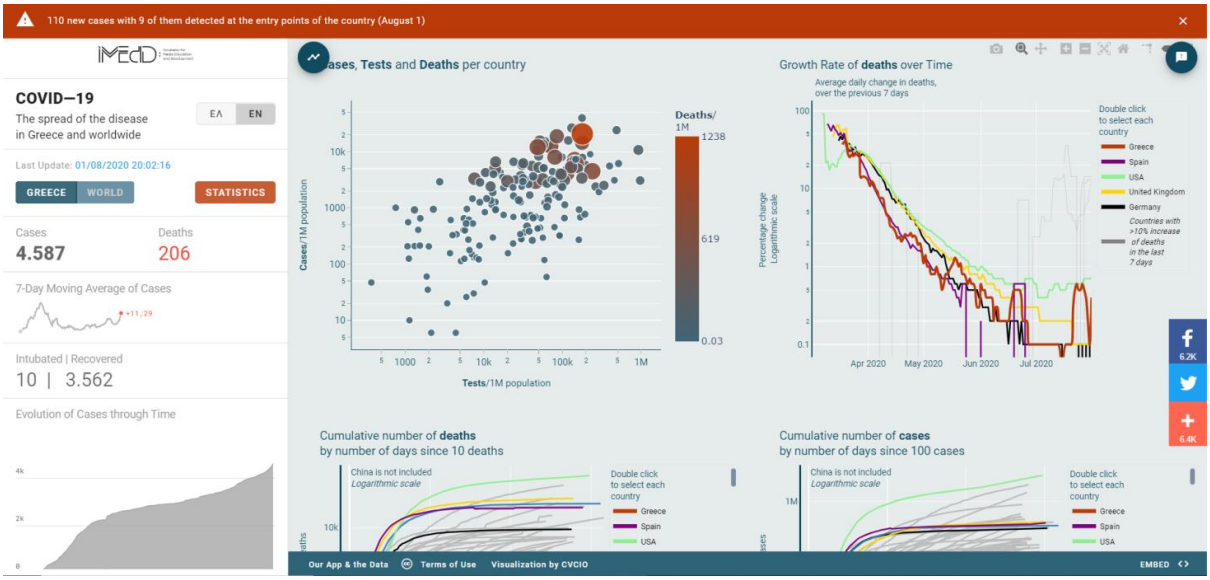
και μόνο χρώμα χωρίς να υπάρχει κάποια αλλαγή χρώματος από περιοχή σε περιοχή μιας χώρας. Επιπλέον, λεπτομερέστερες πληροφορίες για μια περιοχή ή χώρα μπορούν να προβληθούν με την επιλογή ενός συγκεκριμένου σημείου στο χάρτη, το οποίο οδηγεί στην προβολή πληροφοριών για τα συνολικά κρούσματα, για τους θανάτους και για τον αριθμό των ατόμων που έχουν αναρρώσει στη συγκεκριμένη περιοχή. Τέλος, η χρωματική διαβάθμιση των περιοχών μπορεί να εξαρτάται είτε από τα κρούσματα είτε από τους θανάτους εξ' αιτίας του ιού και αλλάζει με το πάτημα ενός κουμπιού στο χάρτη.

Στη διεπαφή του COVID-19 Map, εκτός από τον διαδραστικό χάρτη, προβάλλονται, επίσης, αριθμητικές πληροφορίες για τα κρούσματα και τους θανάτους εξ' αιτίας του κορονοϊου στην Ελλάδα και στον κόσμο, αναλόγως με το ποια από τις δύο λειτουργίες έχει επιλεχθεί. Ακόμη, ένα άλλο χαρακτηριστικό της εφαρμογής είναι η δυνατότητα προβολής διαγραμμάτων και στατιστικών στοιχείων σχετικών με τα ποσοστά θανάτων, κρουσμάτων, αναρρώσεων, διασωληνώσεων και διενεργειών ελέγχων τόσο στην Ελλάδα όσο και στον κόσμο.

Τα δεδομένα που χρησιμοποιεί το COVID-19 Map, που περιλαμβάνουν τα επιβεβαιωμένα κρούσματα και θανάτους αλλά και τη γεωγραφική κατανομή τους, όσον αφορά την Ελλάδα βασίζονται στα επίσημα στοιχεία τα οποία ανακοινώνονται ημερησίως από τον Εθνικό Οργανισμό Δημόσιας Υγείας (ΕΟΔΥ) και από το Υπουργείο Υγείας, ενώ συμπληρώνονται και με διασταυρωμένα στοιχεία που δημοσιεύονται στον ελληνικό Τύπο. Τα παγκόσμια δεδομένα που χρησιμοποιούνται προέρχονται από το αποθετήριο του Johns Hopkins University στο GitHub. [115][116][117][118]



Εικόνα 4.6: Ο χάρτης της Ελλάδος του iMEdD Lab COVID-19 Map όπου οι χρωματικές διαβαθμίσεις αφορούν τον αριθμό κρουσμάτων του κορονοϊού. [118]



Εικόνα 4.7: Προβολή από το iMEdD Lab COVID-19 Map γραφικών παραστάσεων που αφορούν τον αριθμό κρουσμάτων, θανάτων και ελέγχων που έχουν διεξαχθεί, σχετικά με τον κορονοϊό, στις διάφορες χώρες του κόσμου. [118]

4.3.4 Pineza

Το Pineza είναι μια εφαρμογή που δημιουργήθηκε από μια ομάδα τεσσάρων Ελλήνων με εμπειρία στη δημιουργία λογισμικού και την ψηφιακή διαφήμιση. Αρχικά, ο σκοπός δημιουργίας της ήταν η καταγραφή και η απεικόνιση των ποσοστών εγκληματικότητας στην Ελλάδα μέσω ενός διαδραστικού χάρτη. Έτσι, τα εγκλήματα που διαπράττονταν στη χώρα καταγράφονταν και απεικονίζονταν στις διάφορες περιοχές του χάρτη, με στόχο την ενημέρωση και την προστασία του ελληνικού πληθυσμού. Με την έλευση της πανδημίας του κορονοϊού, η ομάδα του Pineza τροποποίησε εθελοντικά την εφαρμογή έτσι ώστε να απεικονίζεται η πορεία εξέλιξης της πανδημίας στη χώρα. Αξίζει να σημειώσουμε πως το Pineza είναι η πρώτη εφαρμογή διαδραστικού χάρτη απεικόνισης δεδομένων, σχετικών με την πανδημία του κορονοϊού, στην Ελλάδα.

Ο διαδραστικός χάρης του κορονοϊού απεικονίζει πληροφορίες για τα κρούσματα του ιού στις διάφορες περιοχές της Ελλάδας, ενώ ταυτόχρονα παρέχονται αριθμητικές πληροφορίες για τα συνολικά κρούσματα, θανάτους και αναρρώσεις ασθενών στη χώρα. Επίσης, παρέχονται και κάποια στατιστικά στοιχεία μέσω γραφικών παραστάσεων, τα οποία αφορούν τα κρούσματα, τις αναρρώσεις και τους θανάτους εξ' αιτίας του κορονοϊού. Ακόμη, μετά από συνεργασία της ομάδας του Pineza με το Bloode, η εφαρμογή παρέχει και πληροφορίες για τα σημεία αιμοδοσίας και τις ανάγκες για αίμα στις διάφορες περιοχές της χώρας. Τέλος, υπάρχει και μελλοντική δυνατότητα απεικόνισης των αποθεμάτων μασκών και αντισηπτικών στις πόλεις της Ελλάδας.

Το Pineza συλλέγει τα δεδομένα που προβάλλει με τη μέθοδο του crowdsourcing, όπου αναφορές που προέρχονται από πολίτες μέσω κοινωνικών δικτύων, συγκεντρώνονται επιβεβαιώνονται και απεικονίζονται σε πραγματικό χρόνο στη διεπαφή της εφαρμογής. [119][120][121]

Εικόνα 4.8: Η διεπαφή του Pineza. [121]

5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΙΝΔΥΝΟΙ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΟΠΤΙΚΕΣ

Η ψηφιακή επιδημιολογία είναι ένας τομέας της επιδημιολογίας με μεγάλη σημασία στη διασφάλιση και στη βελτίωση της ποιότητας ζωής των ανθρώπων. Είναι ένας τομέας που ακόμα και σήμερα βρίσκεται σε ένα πρώιμο στάδιο και θα χρειαστεί να περάσουν κάποια χρόνια μέχρι να αξιοποιηθεί πλήρως. Σήμερα, τα συστήματα που επεξεργάζονται ‘’μεγάλα’’ δεδομένα για τους σκοπούς της ψηφιακής επιδημιολογίας, βρίσκονται σε ένα στάδιο ανάπτυξης και έτσι, προς το παρόν, μπορούν μόνο να εντοπίσουν και να παρακολουθήσουν την πορεία εξάπλωσης μεταδοτικών ασθενειών, ενώ διαθέτουν πολύ περιορισμένες δυνατότητες πρόβλεψης. Τέλος, πολλά από αυτά τα συστήματα, λαμβάνουν επιβεβαιωμένα δεδομένα από διάφορούς οργανισμούς ή άλλες πηγές οπότε, ουσιαστικά, δεν μπορούν να παράγουν κάποια ειδοποίηση πριν να γίνει η επιβεβαίωση αυτών των δεδομένων από άλλους φορείς.

Η συνεχής εξέλιξη της τεχνολογίας, η δημιουργία ταχύτερων υπολογιστικών συστημάτων καθώς και οι βελτιώσεις στον τομέα της τεχνητής νοημοσύνης, μέσα στα επόμενα χρόνια, σίγουρα θα προσφέρουν νέες δυνατότητες στον τομέα της ψηφιακής επιδημιολογίας. Έτσι, θα γίνει δυνατή η δημιουργία νέων συστημάτων που θα μπορούν να επεξεργάζονται και να επιβεβαιώνουν πληροφορίες αυτόνομα και ταχύτερα από άλλους επιδημιολογικούς φορείς. Επίσης, θα γίνει δυνατή η δημιουργία συστημάτων πρόβλεψης τα οποία θα μπορούν να προβλέπουν την έλευση μιας νέας επιδημίας, ακόμα και πριν και από τα πρώτα κρούσματα, ενώ θα υπάρχει και η δυνατότητα δημιουργίας μοντέλων που θα προβλέπουν τη εξάπλωση της. Με τη χρήση τέτοιου είδους συστημάτων, οι φορείς υγείας θα μπορούν, πλέον, να περιορίσουν και να αντιμετωπίσουν τα ξεσπάσματα νέων μεταδοτικών ασθενειών στα πολύ αρχικά τους στάδια, πριν αυτά βγουν εκτός ελέγχου. Ένα παράδειγμα το οποίο επισημαίνει την ανάγκη για τη δημιουργία προηγμένων συστημάτων στον τομέα της ψηφιακής επιδημιολογίας, είναι η πανδημία του κορονοϊού, η οποία θα μπορούσε να αντιμετωπισθεί πολύ ταχύτερα και με καλύτερες μεθόδους αν διαθέταμε ένα μοντέλο πρόβλεψης της εξάπλωσής της.

Ένας άλλος τομέας που μπορεί να ωφεληθεί από τη δημιουργία προηγμένων επιδημιολογικών συστημάτων είναι η φαρμακοβιομηχανία. Οι δοκιμές φαρμάκων και η παρασκευή τους, σήμερα, γίνεται, ως επί το πλείστον, από επιστήμονες. Για την πλήρη αντιμετώπιση μιας μεταδοτικής ασθένειας είναι απαραίτητη η δημιουργία ενός εμβολίου το συντομότερο δυνατό. Η διαδικασία, όμως, αυτή είναι ιδιαίτερα χρονοβόρα, η οποία καθυστερεί ακόμα περισσότερο λόγω του χρονικού διαστήματος ελέγχου του φαρμάκου. Μελλοντικά, ένα επιδημιολογικό σύστημα θα μπορούσε να αναλάβει τη διαδικασία του εντοπισμού των κατάλληλων ουσιών για την αντιμετώπιση ενός παθογόνου μέσα σε ένα πολύ μικρό χρονικό διάστημα και, έτσι, να επιταχυνθεί η διαδικασία παραγωγής ενός εμβολίου. Οι θετικές συνέπειες από την ταχύτερη παραγωγή ενός εμβολίου δεν περιορίζονται μόνο στην μείωση των κρουσμάτων και των θανάτων εξ’ αιτίας μιας ασθένειας, περιλαμβάνουν και την αποσυμφόρηση των νοσοκομείων μιας χώρας, κάτι που όχι μόνο δίνει τη δυνατότητα καλύτερης θεραπείας των ασθενών, αλλά οδηγεί και σε μείωση των κρατικών δαπανών.

Όπως ήδη αναφέραμε, η περεταίρω ανάπτυξη του τομέα της ψηφιακής επιδημιολογίας και, γενικότερα, του τομέα επεξεργασίας και ανάλυσης ‘’μεγάλων’’ δεδομένων, μπορεί να οδηγήσει σε πολλές βελτιώσεις στον τομέα της υγείας και στην ποιότητα ζωής του παγκόσμιου πληθυσμού. Πάντα, όμως, ελλοχεύουν και κάποιοι κίνδυνοι κατά την συλλογή και επεξεργασία δεδομένων από ηλεκτρονικές πηγές. Αρχικά, οι πληροφορίες που συλλέγονται πρέπει να επιβεβαιώνονται και μόνο όταν είναι απόλυτα σίγουρο πως είναι αληθείς να δημοσιεύονται, ειδικά σε πλατφόρμες από τις οποίες ενημερώνονται οι πολίτες των χωρών. Μια λάθος ειδοποίηση από ένα επιδημιολογικό σύστημα, ειδικά όταν χρησιμοποιείται για την ενημέρωση του πληθυσμού, μπορεί να δημιουργήσει πανικό, ενώ η λήψη μέτρων αντιμετώπισης από τους φορείς υγείας οδηγεί σε μη απαραίτητες κρατικές δαπάνες με μεγάλες επιπτώσεις στην οικονομία μιας χώρας. Τέλος, κατά τη συλλογή δεδομένων από ηλεκτρονικά μέσα κοινωνικής δικτύωσης, από smartphones και από άλλες ηλεκτρονικές πηγές μπορεί να προκύψουν ηθικές παραβάσεις. Έτσι, πρέπει πάντα να επιφυλάσσεται το προσωπικό απόρρητο και μόνο μετά τη συγκατάθεση ενός χρήστη να μπορεί ένα σύστημα να λάβει τα δεδομένα του. [7][8][9]

ΣΧΗΜΑΤΑ

ΠΙΝΑΚΕΣ

ΔΙΑΓΡΑΜΜΑΤΑ

ΕΙΚΟΝΕΣ

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] M. Frérot, A. Lefebvre, S. Aho, P. Callier, K. Astruc, and S. A. Glélé, "What is epidemiology? Changing definitions of epidemiology 1978-2017," *PLoS ONE*, vol. 13, no. 12. Public Library of Science, Dec. 01, 2018, doi: 10.1371/journal.pone.0208442.

[2] "Principles of Epidemiology | Boundless Microbiology." https://courses.lumenlearning.com/boundless-microbiology/chapter/principles-of-epidemiology/ (accessed May 08, 2020).

[3] P. M. V. Martin and E. Martin-Granel, "2,500-Year evolution of the term epidemic," *Emerging Infectious Diseases*, vol. 12, no. 6. Centers for Disease Control and Prevention (CDC), pp. 976–980, 2006, doi: 10.3201/eid1206.051263.

[4] "Girolamo Fracastoro | Italian physician | Britannica." https://www.britannica.com/biography/Girolamo-Fracastoro (accessed May 08, 2020).

[5] G. Lippi, C. Mattiuzzi, and G. Cervellin, "Is Digital Epidemiology the Future of Clinical Epidemiology?," *J. Epidemiol. Glob. Health*, vol. 9, no. 2, p. 146, 2019, doi: https://doi.org/10.2991/jegh.k.190314.003.

[6] S. M. S. Islam, T. D. Purnat, N. T. A. Phuong, U. Mwingira, K. Schacht, and G. Fröschl, "Non Communicable Diseases (NCDs) in developing countries: A symposium report," *Global. Health*, vol. 10, no. 1, p. 81, Dec. 2014, doi: 10.1186/s12992-014-0081-9.

[7] "Big Data for Infectious Disease Surveillance and Modeling | The Journal of Infectious Diseases | Oxford Academic." https://academic.oup.com/jid/article/214/suppl_4/S375/2527914 (accessed May 09, 2020).

[8] M. Salathé *et al.*, "Digital Epidemiology," *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002616, Jul. 2012, doi: 10.1371/journal.pcbi.1002616.

[9] T. Eckmanns, H. Füller, and S. L. Roberts, "Digital epidemiology and global health security; An interdisciplinary conversation Tim Eckmanns, Leon Hempel, Kate Polin, Klaus Scheuermann, Edward Velasco," *Life Sci. Soc. Policy*, vol. 15, no. 1, p. 2, Mar. 2019, doi: 10.1186/s40504-019-0091-8.

[10] A. H. Ali and M. Z. Abdullah, "A survey on vertical and horizontal scaling platforms for big data analytics," *Int. J. Integr. Eng.*, vol. 11, no. 6, pp. 138–150, 2019, doi: 10.30880/ijie.2019.11.06.015.

[11] H. A. Park, H. Jung, J. On, S. K. Park, and H. Kang, "Digital epidemiology: Use of digital data collected for non-epidemiological purposes in epidemiological studies," *Healthcare Informatics Research*, vol. 24, no. 4. Korean Society of Medical Informatics, pp. 253–262, Oct. 01, 2018, doi: 10.4258/hir.2018.24.4.253.

[12] "Osiris, Volume 32: Data Histories - Google Books." https://books.google.gr/books?id=AqygDwAAQBAJ&pg=PT52&lpg=PT52&dq=who+defined+the+term+big+data+john+massey&source=bl&ots=wNJAQIuXlT&sig=ACfU3U0z7MuLvbVEmEyoPh_1zZIujk7jSw&hl=en&sa=X&ved=2ahUKEwizmsfJvKzpAhVC8eAKHSgqBugQ6AEwAXoECAsQAQ#v=onepage&q=who defined the term big data john massey&f=false (accessed May 11, 2020).

[13] M. Saecker and V. Markl, "Big data analytics on modern hardware architectures: A technology survey," in *Lecture Notes in Business Information Processing*, 2013, vol. 138 LNBIP, pp. 125–149, doi: 10.1007/978-3-642-36318-4_6.

[14] H. J. Esfahani, K. Tavasoli, and A. Jabbarzadeh, “Big data and social media: A scientometrics analysis,” *Int. J. Data Netw. Sci.*, vol. 3, no. 3, pp. 145–164, 2019, doi: 10.5267/j.ijdns.2019.2.007.

[15] V. Dhawan and N. Zanini, “Big data and social media analytics,” *Res. Matters A Cambridge Assess.*, 2014, Accessed: May 10, 2020. [Online]. Available: www.behaviouralinsights.co.uk.

[16] “Global social media research summary 2020 | Smart Insights.” https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (accessed May 10, 2020).

[17] B. Martens, L. Aguiar, E. Gomez-Herrera, and F. Mueller-Langer, “An economic perspective,” 2018. Accessed: May 13, 2020. [Online]. Available: https://ec.europa.eu/jrc.

[18] S. Y. Yoo, D. H. Kim, S. M. Yang, and O. R. Jeong, “Real-time disease detection and analysis system using social media contents,” *Int. J. Web Grid Serv.*, vol. 16, no. 1, pp. 22–38, 2020, doi: 10.1504/IJWGS.2020.106103.

[19] “Statistics.” https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx (accessed May 13, 2020).

[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009, doi: 10.1038/nature07634.

[21] “How AI, Big Data and Machine Learning can be used against the Corona virus – Ars Electronica Blog.” https://ars.electronica.art/aeblog/en/2020/03/19/ki-corona-part1/ (accessed May 14, 2020).

[22] M. Taddy, “The Technological Elements of Artificial Intelligence,” *Natl. Bur. Econ. Res.*, 2018, doi: 10.3386/w24301.

[23] “Artificial Intelligence & Machine Learning: Policy Paper | Internet Society.” https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/ (accessed May 09, 2020).

[24] “What are the Key Qualities of AI?” https://blog.rossintelligence.com/post/what-are-the-key-qualities-of-ai (accessed May 16, 2020).

[25] “Algorithm Definition.” https://techterms.com/definition/algorithm (accessed May 15, 2020).

[26] “What is a Computer Algorithm? - Design, Examples & Optimization - Video & Lesson Transcript | Study.com.” https://study.com/academy/lesson/what-is-a-computer-algorithm-design-examples-optimization.html (accessed May 15, 2020).

[27] J. Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms,” *Big Data Soc.*, vol. 3, no. 1, p. 205395171562251, Jan. 2016, doi: 10.1177/2053951715622512.

[28] “Definition Machine Learning | Pathmind.” https://pathmind.com/wiki/machine-learning (accessed May 15, 2020).

[29] “A Beginner’s Guide to Neural Networks and Deep Learning | Pathmind.” https://pathmind.com/wiki/neural-network (accessed May 09, 2020).

[30] J. Chan Phooi M’ng and M. Mehralizadeh, “Forecasting East Asian Indices Futures via a Novel Hybrid of Wavelet-PCA Denoising and Artificial Neural Network Models,” *PLoS One*, vol. 11, no. 6, p. e0156338, Jun. 2016, doi: 10.1371/journal.pone.0156338.

[31] A. Tang *et al.*, “Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology,” *Canadian Association of Radiologists Journal*, vol. 69, no. 2. Canadian Medical

Association, pp. 120–135, May 01, 2018, doi: 10.1016/j.carj.2018.02.002.

[32] J. Li, J. H. Cheng, J. Y. Shi, and F. Huang, “Brief introduction of back propagation (BP) neural network algorithm and its improvement,” in *Advances in Intelligent and Soft Computing*, 2012, vol. 169 AISC, no. VOL. 2, pp. 553–558, doi: 10.1007/978-3-642-30223-7_87.

[33] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015, doi: 10.1038/nature14539.

[34] “Deep Learning Definition.” https://www.investopedia.com/terms/d/deep-learning.asp (accessed May 09, 2020).

[35] “A Simple Introduction to Natural Language Processing.” https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32 (accessed May 16, 2020).

[36] “NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part-1).” https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696 (accessed May 16, 2020).

[37] A. D. ’ Ambrosio, “Digital epidemiology. Using the internet for population health. How to listen and what can we discover,” 2016. Accessed: May 16, 2020. [Online]. Available: https://academic.oup.com/eurpub/article-abstract/26/suppl_1/ckw164.081/2448266.

[38] G. De, F. Morales, and A. Bifet, “SAMOA: Scalable Advanced Massive Online Analysis,” 2015. Accessed: Jun. 12, 2020. [Online]. Available: http://mahout.apache.org.

[39] J. Salvador, Z. Ruiz, and J. Garcia-Rodriguez, “Big Data Infrastructure: A Survey,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10338 LNCS, Springer Verlag, 2017, pp. 249–258.

[40] D. Singh and C. K. Reddy, “A survey on platforms for big data analytics,” *J. Big Data*, vol. 2, no. 1, p. 8, Dec. 2015, doi: 10.1186/s40537-014-0008-6.

[41] “Scaling Systems, Fields of study, Abstract, Prinicipal terms.” https://science.jrank.org/computer-science/Scaling_Systems.html (accessed May 17, 2020).

[42] “CUHK Central Research Computing Cluster.” https://www.cuhk.edu.hk/itsc/hpc/overview.html (accessed Jun. 10, 2020).

[43] S. Datta, K. Bhaduri, C. Giannella, H. Kargupta, and R. Wolff, “Distributed data mining in peer-to-peer networks,” *IEEE Internet Comput.*, vol. 10, no. 4, pp. 18–26, Jul. 2006, doi: 10.1109/MIC.2006.74.

[44] “Apache Hadoop.” https://hadoop.apache.org/ (accessed May 20, 2020).

[45] “HDFS Architecture Guide.” https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed May 20, 2020).

[46] K. H. Lee, Y. J. Lee, H. Choi, Y. D. Chung, and B. Moon, “Parallel data processing with MapReduce: A survey,” in *SIGMOD Record*, Dec. 2011, vol. 40, no. 4, pp. 11–20, doi: 10.1145/2094114.2094118.

[47] M. Kang and J. G. Lee, “An experimental analysis of limitations of MapReduce for iterative algorithms on Spark,” *Cluster Comput.*, vol. 20, no. 4, pp. 3593–3604, Dec. 2017, doi: 10.1007/s10586-017-1167-y.

[48] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, *Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks*. 2007.

[49] M. Zaharia, M. Chowdhury, M. J. Franklin, and S. Shenker, “Spark: Cluster Computing with Working Sets.”

[50] “Apache Storm.” https://storm.apache.org/ (accessed May 31, 2020).

[51] “(PDF) Big Data Analysis: Apache Storm Perspective,” Accessed: May 31, 2020. [Online]. Available: https://www.researchgate.net/publication/271196175_Big_Data_Analysis_Apache_Storm_Perspective.

[52] “Apache Mahout.” https://mahout.apache.org/ (accessed May 28, 2020).

[53] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, “Apache Flink™: Stream and Batch Processing in a Single Engine,” *undefined*, 2015.

[54] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, “S4: Distributed Stream Computing Platform,” 2010, doi: 10.1109/ICDMW.2010.172.

[55] “Pregel: A System for Large-Scale Graph Processing – the morning paper.” https://blog.acolyer.org/2015/05/26/pregel-a-system-for-large-scale-graph-processing/ (accessed Jun. 02, 2020).

[56] G. Malewicz *et al.*, *Pregel: A System for Large-Scale Graph Processing*. 2010.

[57] “(PDF) MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering.” https://www.researchgate.net/publication/220320178_MOA_Massive_Online_Analysis_a_Framework_for_Stream_Classification_and_Clustering (accessed Jun. 03, 2020).

[58] Y. Kochura, S. Stirenko, A. Rojbi, O. Alienin, M. Novotarskiy, and Y. Gordienko, “Comparative Analysis of Open Source Frameworks for Machine Learning with Use Case in Single-Threaded and Multi-Threaded Modes.”

[59] “Cloud Computing - SupraITS.” https://www.supraits.com/infrastructure/managed-cloud/hybrid-cloud-3/cloud-computing/ (accessed Jun. 10, 2020).

[60] D. Warneke and O. Kao, “Nephele: Efficient parallel data processing in the cloud,” in *Proceedings of the 2nd ACM Workshop on Many-Task Computing on Grids and Supercomputers 2009, MTAGS ’09*, 2009, doi: 10.1145/1646468.1646476.

[61] N. Sadashiv and S. M. Dilip Kumar, *Cluster, Grid and Cloud Computing: A Detailed Comparison*. .

[62] N. A. Al Etawi, “A Comparison between Cluster, Grid, and Cloud Computing,” 2018.

[63] R. Kumar and S. Charu, “Comparison between Cloud Computing, Grid Computing, Cluster Computing and Virtualization,” 2015, doi: 10.13140/2.1.1759.7765.

[64] M. N. and P. T. . .Chavan, “Implementation of Parallelization Contract Mechanism Extension of Map Reduce Framework for the Efficient Execution Time over Geo-Distributed Dataset,” *Int. J. Eng. Res.*, vol. 3, pp. 745–748, 2014, doi: 10.17950/ijer/v3s12/1208.

[65] A. Alexandrov *et al.*, “Massively parallel data analysis with PACTs on Nephele,” *Proc. VLDB Endow.*, vol. 3, no. 2, pp. 1625–1628, Sep. 2010, doi: 10.14778/1920841.1921056.

[66] “Intel’s 48-core Xeon will go head-to-head with AMD in 2019 | Engadget.” https://www.engadget.com/2018-11-05-intel-48-core-xeon-processor.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAB9wtmXdJYeaP57TLRLXRVSZoCGpJ6sD6Nf2bcY9AMMAi5vasUibnsBuKi6VCQNP0Ub0sheCJvZbC21gsBinKFV24GZa4s3yOM_B6QuuLt9p2KjM3Sb7LqV1XREvSaIodIU_dKeJ8JUyTqvez9B-TIR0gn-hPu6psvMz2wp6CRRQ (accessed May 22, 2020).

[67] “Intel Core i7-3960X Extreme Edition Processor (Sandy Bridge-E) Review - Specifications & Features.” https://www.vortez.net/articles_pages/intel_core_i7_3960x_extreme_edition_processor_sa

ndy_bridge_e,2.html (accessed May 22, 2020).

[68] A. Zibula, H. Kuchen, and P. Ciechanowicz, “General Purpose Computation on Graphics Processing Units (GPGPU) using CUDA within the seminar Parallel Programming and Parallel Algorithms,” 2009.

[69] “NVIDIA Announces “NVIDIA Titan V" Video Card: GV100 for $3000, On Sale Now.” https://www.anandtech.com/show/12135/nvidia-announces-nvidia-titan-v-video-card-gv100-for-3000-dollars (accessed Jun. 28, 2020).

[70] J. Romoth, J. Romoth, M. Porrmann, and R. Ulrich, “Survey of FPGA applications in the period 2000 – Survey of FPGA applications in the period 2000 – 2015,” no. March, 2017, doi: 10.13140/RG.2.2.16364.56960.

[71] “410-279 FPGA Board Ethernet/I^{2}C/SPI/UART/USB Digilent.” https://www.distrelec.de/en/fpga-board-ethernet-spi-uart-usb-digilent-410-279/p/30044350 (accessed Jun. 26, 2020).

[72] “High-Level Schematic of the BlueGene/L Platform | Download Scientific Diagram.” https://www.researchgate.net/figure/High-Level-Schematic-of-the-BlueGene-L-Platform_fig2_228348352 (accessed Jun. 10, 2020).

[73] “Top 53 Bigdata Platforms and Bigdata Analytics Software in 2020 - Reviews, Features, Pricing, Comparison - PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices.” https://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/ (accessed Jun. 15, 2020).

[74] “Flu & Ebola Map | Virus & Contagious Disease Surveillance.” https://healthmap.org/en/ (accessed May 10, 2020).

[75] “About | HealthMap.” http://www.diseasedaily.org/about (accessed Jun. 28, 2020).

[76] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, “HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports,” *J. Am. Med. Informatics Assoc.*, vol. 15, no. 2, pp. 150–157, Mar. 2008, doi: 10.1197/jamia.M2544.

[77] “swine flu graphs and charts - Casada.startupmendoza.co.” http://casada.startupmendoza.co/swine-flu-graphs-and-charts/ (accessed May 11, 2020).

[78] S. Kandula and J. Shaman, “Reappraising the utility of Google Flu Trends,” *PLOS Comput. Biol.*, vol. 15, no. 8, p. e1007258, Aug. 2019, doi: 10.1371/journal.pcbi.1007258.

[79] “Answers to Quora questions.: Q: How accurate is Google Flu Trends?” http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html (accessed May 11, 2020).

[80] “Jen Zhang.” https://www.jenniferzhang.ca/project_insights.html (accessed Jun. 28, 2020).

[81] “Canada commits $192 million to COVID-19 SIF stream, signs deal with BlueDot to help track, make decisions on virus | BetaKit.” https://betakit.com/canada-commits-192-million-to-covid-19-sif-stream-signs-deal-with-bluedot-to-help-track-make-decisions-on-virus/ (accessed May 10, 2020).

[82] “How Canadian AI start-up BlueDot spotted Coronavirus before anyone else had a clue.” https://diginomica.com/how-canadian-ai-start-bluedot-spotted-coronavirus-anyone-else-had-clue (accessed May 09, 2020).

[83] “About - GPHIN - Canada.ca.” https://gphin.canada.ca/cepr/gphinrenewal-renouvellementrmisp.jsp?language=en_CA (accessed Jul. 22, 2020).

[84] M. Dion, P. AbdelMalik, and A. Mawudeku, “Big Data and the Global Public Health Intelligence Network (GPHIN),” *Canada Commun. Dis. Rep.*, vol. 41, no. 9, pp. 209–214, Sep. 2015, doi: 10.14745/ccdr.v41i09a02.

[85] D. Carter, M. Stojanovic, and B. De Bruijn, “Revitalizing the Global Public Health Intelligence Network (GPHIN),” *Online J. Public Health Inform.*, vol. 10, no. 1, May 2018, doi: 10.5210/ojphi.v10i1.8912.

[86] “WHO | Epidemic intelligence - systematic event detection,” *WHO*, 2020, Accessed: Jul. 21, 2020. [Online]. Available: http://www.who.int/csr/alertresponse/epidemicintelligence/en/.

[87] “About ProMED – ProMED-mail.” https://promedmail.org/about-promed/ (accessed Jul. 23, 2020).

[88] J. You, P. Expert, and C. Costelloe, “Using text mining to track outbreak trends in global surveillance of emerging diseases: ProMED-mail,” Cold Spring Harbor Laboratory Press, Jan. 2020. doi: 10.1101/2020.01.10.20017145.

[89] WHO, “Imported from https://www.researchgate.net/publication/279883456_Community-directed_interventions_are_practical_and_effective_in_low-resource_communities_Experience_of_ivermectin_treatment_for_onchocerciasis_control_in_Cameroon_and_Uganda_2004-2010,” *Int. Health*, vol. 11, no. 2, pp. 83–92, 2019, doi: 10.1093/INTHEALTH.

[90] “Detecting Rumors with Web-based Text Mining System | Conflict Early Warning and Early Response.” https://earlywarning.wordpress.com/2009/02/14/detecting-rumors-with-web-based-text-mining-system/ (accessed Jul. 24, 2020).

[91] N. Collier *et al.*, “A multilingual ontology for infectious disease surveillance: Rationale, design and challenges,” *Lang. Resour. Eval.*, vol. 40, no. 3–4, pp. 405–413, Dec. 2006, doi: 10.1007/s10579-007-9019-7.

[92] A. Lyon, M. Nunn, G. Grossel, and M. Burgman, “Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap,” *Transbound. Emerg. Dis.*, vol. 59, no. 3, pp. 223–232, Jun. 2012, doi: 10.1111/j.1865-1682.2011.01258.x.

[93] S. Doan, Q.-H. Ngo, A. Kawazoe, and N. Collier, “Global Health Monitor: A Web-based System for Detecting and Mapping Infectious Diseases,” Nov. 2019, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/1911.09735.

[94] “MediSys.” https://medisys.newsbrief.eu/medisys/clusteredition/el/24hrs_en.html (accessed Jul. 26, 2020).

[95] “NewsBrief.” https://emm.newsbrief.eu/NewsBrief/clusteredition/el/latest_en.html (accessed Jul. 26, 2020).

[96] “MedISys - Medical Information System | EU Science Hub.” https://ec.europa.eu/jrc/en/publication/articles-books/medisys-medical-information-system (accessed Jul. 25, 2020).

[97] J. Mantero, E. Centre, D. Prevention, J. Belyaeva, E. Commission, and J. P. Linge, “How to maximise event-based surveillance web- systems : the example of ECDC / JRC collaboration to improve the performance of MedISys,” Luxembourg, 2011. doi: 10.2788/69804.

[98] J. P. Linge *et al.*, “MedISys: Medical information system,” in *Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks*, IGI Global, 2010, pp. 131–142.

[99] F. Gey *et al.*, “Information Access in a Multilingual World _____ Proceedings of the SIGIR 2009 Workshop Program Committee: Introduction and Overview,”

2009. Accessed: Jul. 25, 2020. [Online]. Available: http://www.infoplosion.nii.ac.jp/info-plosion/ctr.php/m/IndexEng/a/Index/.

[100] “Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance.” https://www.medscape.com/viewarticle/707744_4 (accessed Aug. 08, 2020).

[101] “2009 EpiSPIDER CDC GIS Day.” https://www.slideshare.net/hermantolentino/2009-epispider-cdc-gis-day (accessed Aug. 08, 2020).

[102] “Q&A: Influenza and COVID-19 - similarities and differences.” https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza?gclid=CjwKCAjwmf_4BRABEiwAGhDfSRAl2nDGodM2IlTH1D6Dpn2rklhrNqmU0doG8CqQDZ19Tim0LbqumxoCBHAQAvD_BwE (accessed Jul. 28, 2020).

[103] “Coronaviridae - ScienceDirect.” https://www.sciencedirect.com/science/article/pii/B9780123846846000689?via%3Dihub (accessed Jul. 29, 2020).

[104] “Novel Coronavirus Information Center.” https://www.elsevier.com/connect/coronavirus-information-center?dgcid=_SD_banner#research (accessed Jul. 29, 2020).

[105] “COVID-19 timeline in the Western Pacific.” https://www.who.int/westernpacific/news/detail/18-05-2020-covid-19-timeline-in-the-western-pacific (accessed Jul. 29, 2020).

[106] “Coronavirus: China’s first confirmed Covid-19 case traced back to November 17 | South China Morning Post.” https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back (accessed Jul. 29, 2020).

[107] “WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard.” https://covid19.who.int/table (accessed Jul. 29, 2020).

[108] “COVID-19 situation update worldwide, as of 29 July 2020.” https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases (accessed Jul. 29, 2020).

[109] “WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard.” https://covid19.who.int/?gclid=Cj0KCQjwvb75BRD1ARIsAP6LcqvwPnPTlSoFkLMVfANzuAax39tLneIhVHpFwpoydVnz8GcdBFom6bcaAiGIEALw_wcB (accessed Aug. 09, 2020).

[110] “BlueDot used artificial intelligence to predict coronavirus spread.” https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html (accessed Jul. 30, 2020).

[111] “ProMED, a Free Email Service, Broke the News of the Coronavirus in 2019 | OneZero.” https://onezero.medium.com/a-free-email-service-broke-the-news-of-the-coronavirus-in-2019-ff2b595af606 (accessed Jul. 30, 2020).

[112] “The analyst tracking news on COVID-19 | EU Science Hub.” https://ec.europa.eu/jrc/en/news/analyst-tracking-news-covid-19 (accessed Jul. 30, 2020).

[113] A. Cho, “Artificial intelligence systems aim to sniff out signs of COVID-19 outbreaks,” *Science (80-.).*, May 2020, doi: 10.1126/science.abc7698.

[114] “COVID-19 Map - Johns Hopkins Coronavirus Resource Center.” https://coronavirus.jhu.edu/map.html (accessed Jul. 31, 2020).

[115] “Lab - iMEdD.” https://www.imedd.org/el/imedd-lab/ (accessed Jul. 30, 2020).

[116] “Who we are - iMEdD.” https://www.imedd.org/about/ (accessed Aug. 01, 2020).

[117] “Partners - iMEdD.” https://www.imedd.org/partners/ (accessed Aug. 01, 2020).

[118] “COVID—19.” https://lab.imedd.org/covid19/stats?lang=en (accessed Aug. 01, 2020).

[119] “Pineza - Landing Page | A Crowdsourcing Community to Fight Urban Crime.” https://pineza.eu/ (accessed Aug. 02, 2020).

[120] “Ο Πρώτος Live Χάρτης με τα Κρούσματα του Κορονοϊού στην Ελλάδα.” https://www.vice.com/gr/article/dygkqz/o-prwtos-live-xarths-me-ta-kroysmata-toy-koronoioy-sthn-ellada (accessed Aug. 02, 2020).

[121] “Pineza - Coronavirus.” https://coronavirus.pineza.eu/ (accessed Jul. 30, 2020).