ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

BIG DATA ANALYTICS AND SCIENCE

ΧΡΗΣΤΟΣ ΙΩΑΝΝΗ ΦΟΥΝΑΣ

ΕΠΙΒΛΕΠΟΥΣΑ: ΑΝΑΣΤΑΣΙΑ ΒΕΛΩΝΗ, ΚΑΘΗΓΗΤΡΙΑ, ΛΕΚΤΟΡΑΣ ΕΦΑΡΜΟΓΩΝ

ΑΘΗΝΑ

ΙΑΝΟΥΑΡΙΟΣ 2019

Περιεχόμενα

Περίληψη

Τα μεγάλα δεδομένα τα οποία θεωρούνται στρατηγικός πόρος στους τομείς, όπως είναι η επιστήμη, η υγεία, η βιομηχανία και οι επιχειρήσεις, προσελκύουν όλο και περισσότερο την προσοχή των κρατικών αρχών. Τη σήμερον ημέρα, ορισμένα κράτη επιδιώκουν να επιτρέψουν τη χρήση μεγάλου όγκου δεδομένων, για να αυξήσουν την αποδοτικότητα των διαδικασιών λήψης αποφάσεων και της δραστηριότητας των οργανισμών, για τη δημιουργία νέων υπηρεσιών, για τη δημιουργία νέων ιδεών , ενώ ταυτόχρονα για να αναδεικνύονται και ως ηγέτες στις αγορές. Πλήρης υποστήριξη παρέχεται για την εφαρμογή τεχνολογιών μεγάλων δεδομένων και την επίλυση προβλημάτων αφού οι τεχνολογίες και η επιστήμη των μεγάλων δεδομένων εξελίσσεται ραγδαία με δυνατότητες εξαγωγής πληροφορίας από κείμενα, βίντεο, ήχο, εικόνες και δεδομένων κοινωνικών δικτύων. Τέλος, η παρούσα μελέτη εξετάζει τις μεγάλες πρωτοβουλίες δεδομένων ορισμένων ανεπτυγμένων χωρών στον τομέα αυτό.

ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγή

Τα τελευταία 20 χρόνια, τα δεδομένα έχουν αυξηθεί σε μεγάλη κλίμακα σε διάφορους τομείς. Σύμφωνα με μια έκθεση της International Data Corporation (IDC), το 2011 ο συνολικός όγκος δεδομένων που δημιουργήθηκε και αντιγράφηκε στον κόσμο ήταν 1,8ZB (≈ 1021 B), ο οποίος αυξήθηκε σχεδόν εννέα φορές μέσα σε πέντε χρόνια (Gantz & Reinsel, 2011). Ο αριθμός αυτός θα διπλασιαστεί τουλάχιστον κάθε δύο χρόνια στο εγγύς μέλλον. Κάτω από την εκρηκτική αύξηση των παγκόσμιων δεδομένων, ο όρος των μεγάλων δεδομένων (Big Data) χρησιμοποιείται κυρίως για να περιγράψει τεράστια σύνολα δεδομένων. Σε σύγκριση με τα παραδοσιακά σύνολα δεδομένων, τα μεγάλα δεδομένα συνήθως περιλαμβάνουν μάζες μη δομημένων δεδομένων που χρειάζονται περισσότερη ανάλυση σε πραγματικό χρόνο. Επιπλέον, τα μεγάλα δεδομένα δημιουργούν νέες ευκαιρίες για την ανακάλυψη νέων αξιών, μας βοηθούν να κατανοήσουμε σε βάθος τις κρυφές αξίες και επίσης δημιουργούμε νέες προκλήσεις, π.χ. πώς να οργανώσουμε και να διαχειριστούμε αποτελεσματικά αυτά τα σύνολα δεδομένων.

Πρόσφατα, οι βιομηχανίες ενδιαφέρονται για το υψηλό δυναμικό μεγάλων δεδομένων και πολλές κυβερνητικές υπηρεσίες ανακοίνωσαν μείζονα σχέδια για την επιτάχυνση της έρευνας και εφαρμογών μεγάλων δεδομένων. Επιπλέον, θέματα σχετικά με τα μεγάλα δεδομένα καλύπτονται συχνά από δημόσια μέσα και κορυφαία επιστημονικά περιοδικά, γεγονός που επιβεβαιώνει ότι η εποχή των μεγάλων δεδομένων έχει ξεπεράσει κάθε αμφιβολία (Manyika et al, 2011). Σήμερα, τα μεγάλα δεδομένα που σχετίζονται με την υπηρεσία των εταιρειών του Διαδικτύου αυξάνονται ραγδαία. Για παράδειγμα, η Google επεξεργάζεται δεδομένα εκατοντάδων petabyte (PB), το Facebook παράγει δεδομένα καταγραφής άνω των 10 PB ανά μήνα, η κινεζική εταιρεία Baidu, επεξεργάζεται δεδομένα δεκάδων PB και η Taobao, θυγατρική της Alibaba, παράγει δεδομένα δεκάδων του Terabyte

(TB) για διαδικτυακές συναλλαγές ανά ημέρα. Η μεγάλη έκρηξη των δεδομένων όπως εξελίχθηκε τα τελευταία χρόνια φαίνονται στην παρακάτω εικόνα:

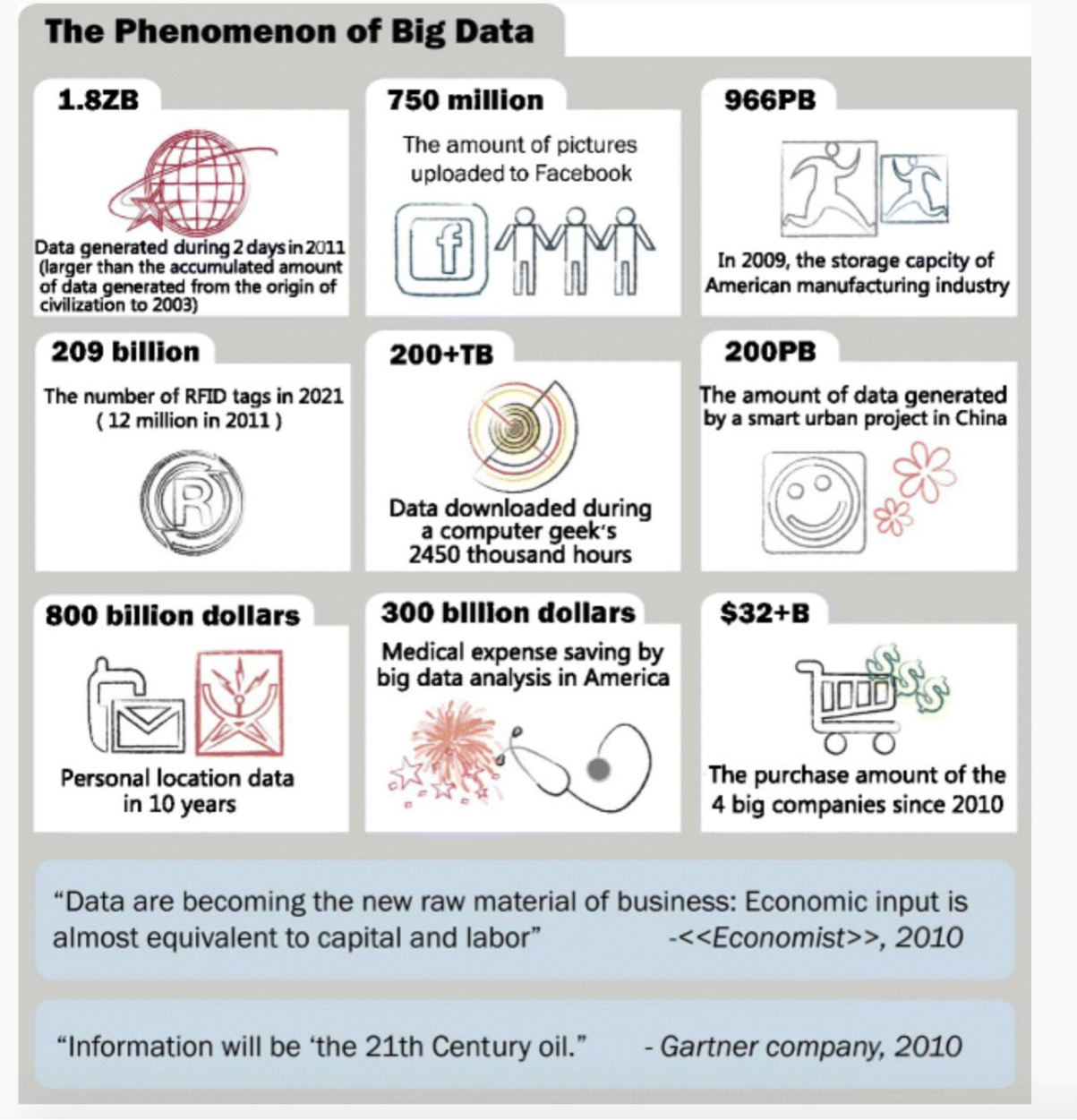


Figure 0 Παρούσα κατάσταση στη βιομηχανία των big data

Είναι επίσης γεγονός ότι οι επαναστάσεις στην επιστήμη έχουν γίνει συχνά μετά από επαναστάσεις στις μετρικές και στα μέτρα απόδοσης. Τα δεδομένα μεγάλου όγκου (Big Data) δημιούργησαν μια ριζική αλλαγή στο πώς σκεφτόμαστε σχετικά με την έρευνα, προσφέροντας μια βαθιά αλλαγή στα επίπεδα της επιστημολογίας και της ηθικής. Τα Big Data επαναπροσδιορίζουν ερωτήσεις-κλειδιά σχετικά με τη σύσταση της γνώσης, τις

διεργασίες της έρευνας, τον τρόπο με τον οποίο πρέπει αναλύονται οι πληροφορίες, τη φύση και τη κατηγοριοποίηση της πραγματικότητας. Όπως συμβαίνει με πολλές γρήγορα αναδυόμενες έννοιες, τα Big Data έχουν οριστεί ποικιλοτρόπως, από τους παλιούς ορισμούς που υποστήριζαν ότι τα Big Data είναι σύνολα δεδομένων πολύ μεγάλα για να «χωρέσουν» σε ένα υπολογιστικό φύλλο Excel ή να αποθηκεύονται σε ένα μόνο μηχάνημα (Strom, 2012), έως τις εκλεπτυσμένες οντολογικές εκτιμήσεις που υπονομεύουν τα δικά τους εγγενή χαρακτηριστικά (Boyd & Crawford, 2012; Mayer-Schonberger & Cukier, 2013). Ένας βασικός ορισμός των Big Data δόθηκε από τον Kitchin (2013) ο οποίος υποστήριξε ότι τα Big data είναι:

- τεράστιου όγκου, αποτελούνται από Terabytes ή Petabytes δεδομένων
- υψηλή ταχύτητα, δημιουργούνται σε ή σχεδόν σε πραγματικό χρόνο
- ποικιλία, τα δεδομένα μπορεί να έχουν δομημένη και αδόμητη φύση
- εξαντλητικό πεδίο, προσπαθώντας να συλλάβει ολόκληρους πληθυσμούς ή συστήματα

Καθώς τα εργαλεία και οι φιλοσοφία των μεγάλων δεδομένων διαδόθηκε, θα αλλάξουν μακροχρόνιες ιδέες σχετικά με την αξία της εμπειρίας, της φύσης της εμπειρογνωμοσύνης και της πρακτικής σε όλες τις επιστήμες.

1.2 **Αντικείμενο της πτυχιακής**

Η παρούσα πτυχιακή εργασία ασχολείται με το θέμα των Big Data και των εφαρμογών της στην επιστήμη και στη βιομηχανία και τη διείσδυσή τους ανά κλάδο και χώρα. Αρχικά, στο Κεφάλαιο 2 θα παρουσιαστεί η εξέλιξη και ορισμός των Big Data, ενώ θα γίνει εκτενής αναφορά με σκοπό την κατανόηση του πρότυπου των 3v. Έπειτα, θα παρουσιαστούν οι κυριότερες τεχνολογίες μεγάλων δεδομένων και οι τεχνολογίες ανάλυσης μεγάλων δεδομένων. Στο Κεφάλαιο 3 θα αναφερθούμε στα πλεονεκτήματα από την εφαρμογή των Big Data analytics, ενώ θα παρουσιαστούν εφαρμογές μεγάλων δεδομένων ανά κλάδο και η διείσδυσή τους ανά αγορές χωρών. Τέλος, θα παρουσιαστούν οι πρωτοβουλίες ανάπτυξης εφαρμογών που βασίζονται στα μεγάλα δεδομένα, στις μεγαλύτερες οικονομικά και τεχνολογικά προηγμένες αγορές/χώρες. Η πτυχιακή εργασία θα καταλήξει παρουσιάζοντας μελλοντικές τάσεις ενώ θα ακολουθήσουν τα συμπεράσματα.

ΚΕΦΑΛΑΙΟ 2 – Εισαγωγή στα μεγάλα δεδομένα

2.1 Εξέλιξη και ορισμός των Big Data

Η αναδυόμενη τεχνολογική εξέλιξη των μεγάλων δεδομένων αναγνωρίζεται ως ένας από τους πιο σημαντικούς τομείς της μελλοντικής τεχνολογίας των πληροφοριών και εξελίσσεται με ραγδαίους ρυθμούς, οδηγούμενο εν μέρει από τα κοινωνικά μέσα και το φαινόμενο του Ίντερνετ των Πραγμάτων (IoT). Οι τεχνολογικές εξελίξεις στη μεγάλη υποδομή δεδομένων, τα αναλυτικά στοιχεία και οι υπηρεσίες επιτρέπουν στις επιχειρήσεις να μετασχηματιστούν σε οργανώσεις με βάση τα δεδομένα. Λόγω της δυνατότητας των μεγάλων δεδομένων να γίνουν ένα παιχνίδι, κάθε επιχείρηση πρέπει να αναπτύξει δυνατότητες και να αξιοποιήσει τα μεγάλα δεδομένα για να παραμείνει ανταγωνιστική. Η IDC (2015) προέβλεψε ότι η μεγάλη τεχνολογία δεδομένων και η αγορά υπηρεσιών θα αυξηθεί με ένα σύνθετο ετήσιο ρυθμό ανάπτυξης 23,1% για την περίοδο 2014-2019, με τις ετήσιες δαπάνες να φτάνουν τα 48,6 δισ. δολάρια το 2019. Ενώ τα δομημένα δεδομένα αποτελούν σημαντικό μέρος των μεγάλων δεδομένων, όλο και περισσότερα δεδομένα δημιουργούνται σε αδόμητες μορφές, όπως βίντεο και εικόνας, τα οποία, οι παραδοσιακές τεχνολογίες διαχείρισης δεν μπορούν να επεξεργαστούν επαρκώς, όπως ένα μεγάλο μέρος των δεδομένων παγκοσμίως που παράγονται από δισεκατομμύρια συσκευές IoT όπως έξυπνες οικιακές συσκευές, φορητές συσκευές και αισθητήρες περιβάλλοντος. Ο Gartner (2015) προέβλεψε ότι 4,9 δισεκατομμύρια συνδεδεμένα αντικείμενα θα χρησιμοποιηθούν το 2015 - αύξηση 30% από το 2014 - και θα φτάσει τα 25 δισεκατομμύρια έως το 2020. Για την κάλυψη του συνεχώς αυξανόμενου χώρου αποθήκευσης και ανάγκης για επεξεργασία μεγάλων δεδομένων, νέες μεγάλες πλατφόρμες δεδομένων αναδύονται, συμπεριλαμβανομένων των βάσεων δεδομένων NoSQL ως εναλλακτική λύση στις παραδοσιακές σχεσιακές βάσεις δεδομένων και του Hadoop ως πλαίσιο ανοιχτού κώδικα για φθηνές, κατανεμημένες συστοιχίες υλικού-λογισμικού.

Ιστορική αναδρομή

Στα τέλη της δεκαετίας του 1970, η έννοια της "μηχανής βάσης δεδομένων" αναδείχθηκε, ως μια τεχνολογία που χρησιμοποιείται ειδικά για την αποθήκευση και την ανάλυση δεδομένων. Με την αύξηση του όγκου δεδομένων, την ικανότητα αποθήκευσης και επεξεργασίας ενός

μόνο κεντρικού υπολογιστή το σύστημα πληροφορικής έγινε ανεπαρκές. Στη δεκαετία του '80, οι άνθρωποι πρότειναν την ιδέα «να μην μοιραστούν τίποτα», και με ένα παράλληλο σύστημα βάσης δεδομένων να ικανοποιήσει τη ζήτηση του αυξανόμενου όγκου δεδομένων. Αυτή η αρχιτεκτονική βασίζεται στη χρήση του cluster και κάθε μηχανή έχει δικό της επεξεργαστή, αποθηκευτική μονάδα και δίσκο. Το σύστημα Teradata ήταν το πρώτο επιτυχημένο εμπορικό σύστημα παράλληλης βάσης δεδομένων. Στις 2 Ιουνίου 1986, υπάρχει ένα γεγονός ορόσημο, όταν η Teradata παρέδωσε την πρώτη παράλληλη βάση δεδομένων με χωρητικότητα αποθήκευσης 1TB στην Kmart ώστε η μεγάλης κλίμακας εταιρεία λιανικής στη Βόρεια Αμερική να επεκτείνει την αποθήκευση των δεδομένων της. Στα τέλη της δεκαετίας του 1990, τα πλεονεκτήματα της παράλληλης βάσης δεδομένων αναγνωρίστηκαν ευρέως στον κλάδο. Ωστόσο, προέκυψαν πολλές προκλήσεις που αφορούσαν δεδομένα μεγάλου όγκου. Με την ανάπτυξη υπηρεσιών διαδικτύου, οι δείκτες και το περιεχόμενο των ερωτημάτων αυξάνονταν ραγδαία. Ως εκ τούτου, οι εταιρείες μηχανών αναζήτησης έπρεπε να αντιμετωπίσουν τις προκλήσεις του χειρισμού αυτών των μεγάλων δεδομένων.

Η Google δημιούργησε πρότυπα προγραμματισμού GFS και MapReduce για να αντιμετωπίσει τις προκλήσεις που προκλήθηκαν από τη διαχείριση δεδομένων και την ανάλυση στην κλίμακα του διαδικτύου. Επιπλέον, το περιεχόμενο που δημιουργείται από χρήστες, αισθητήρες και άλλες πανταχού παρούσες πηγές δεδομένων δημιούργησαν συντριπτικές ροές δεδομένων, γεγονός που απαιτούσε μια θεμελιώδη αλλαγή στην αρχιτεκτονική υπολογιστών και στον ευρείας κλίμακας μηχανισμό επεξεργασίας δεδομένων. Τον Ιανουάριο του 2007, ο Jim Gray, πρωτοπόρος του λογισμικού βάσης δεδομένων, ονόμασε αυτό το μετασχηματισμό ως " The Fourth Paradigm" (Tansley and Tolle, 2009). Αυτός σκέφτηκε επίσης ότι ο μόνος τρόπος αντιμετώπισης αυτού του παραδείγματος ήταν να αναπτύξει μια νέα γενιά εργαλείων πληροφορικής για τη διαχείριση, την εικονογράφηση και την ανάλυση μαζικών δεδομένων. Τον Ιούνιο του 2011,συναντάται άλλο ένα συμβάν ορόσημο: Η EMC / IDC δημοσίευσε μια έρευνα με τίτλο Extracting Values from Chaos, η οποία εισήγαγε την έννοια και το δυναμικό των μεγάλων δεδομένων για πρώτη φορά. Αυτή η ερευνητική έκθεση προκάλεσε το μεγάλο ενδιαφέρον τόσο από τη βιομηχανία όσο και από τον ακαδημαϊκό χώρο για τα μεγάλα δεδομένα.

Τα τελευταία χρόνια, σχεδόν όλες οι μεγάλες εταιρείες, όπως η EMC, η Oracle, η IBM, η Microsoft, η Google, η Amazon και το Facebook κλπ. ξεκίνησαν τα μεγάλα έργα τους. Λαμβάνοντας την IBM ως παράδειγμα, από το 2005, η IBM έχει επενδύσει 16

δισεκατομμύρια δολάρια στις ΗΠΑ σε 30 εξαγορές που σχετίζονται με μεγάλα δεδομένα. Το 2008, το περιοδικό Nature δημοσίευσε ένα ειδικό περιοδικό αφιερωμένο στα Big Data. Το 2012, Ευρωπαϊκή Έρευνα Κοινοπραξία για την Πληροφορική και τα Μαθηματικά (ERCIM) δημοσίευσαν ένα ειδικό τεύχος για τα μεγάλα δεδομένα. Στην αρχή του 2012, μια έκθεση με τίτλο Big Data, Big Impact που παρουσιάστηκε στο Davos Forum στην Ελβετία, ανακοίνωσε ότι τα μεγάλα δεδομένα αποτελούν νέο είδος οικονομικών περιουσιακών στοιχείων, ακριβώς όπως το νόμισμα ή ο χρυσός. Ο Gartner, ο διεθνής οργανισμός έρευνας, εξέδωσε Hype Cycles από το 2012 έως το 2013, στο οποίο ταξινόμησε big data computing, social analysis και stored data analysis σε 48 αναδυόμενες τεχνολογίες που αξίζουν μεγαλύτερη προσοχή. Πολλές εθνικές κυβερνήσεις όπως οι ΗΠΑ έδωσαν επίσης μεγάλη προσοχή στα μεγάλα δεδομένα. Τον Μάρτιο του 2012, η αμερικανική κυβέρνηση ανακοίνωσε μια επένδυση 200 εκατ. Δολαρίων για την έναρξη του προγράμματος " Έρευνας και Ανάπτυξης Μεγάλων Δεδομένων" η οποία ήταν μια δεύτερη σημαντική επιστημονική και τεχνολογική πρωτοβουλία για την ανάπτυξη, μετά την πρωτοβουλία "Information Highway" το 1993. Τον Ιούλιο του 2012, η "Βιώσιμη ΤΠΕ Ιαπωνίας" που εκδόθηκε από το Υπουργείο Εσωτερικών της Ιαπωνίας έδειξε ότι η μεγάλη ανάπτυξη δεδομένων πρέπει να είναι μια εθνική στρατηγική και οι τεχνολογίες εφαρμογής θα πρέπει να είναι το επίκεντρο. Τον Ιούλιο του 2012, τα Ηνωμένα Έθνη εξέδωσαν Έκθεση για τα μεγάλα δεδομένα για την ανάπτυξη, η οποία συνοψίζει τον τρόπο οι κυβερνήσεις χρησιμοποίησαν μεγάλα δεδομένα για καλύτερη εξυπηρέτηση και προστασία τους ανθρώπους τους. Από την παραπάνω ιστορική διαδρομή γίνεται σαφές ότι τα μεγάλα δεδομένα βρίσκονται στο επίκεντρο της βιομηχανίας, της έρευνας αλλά και της κυβερνητικής προσοχής, αναδεικνύοντάς τα ως σημαντικό θέμα στην πληροφορική του 21^{ου} αιώνα.

Ορισμοί

Πολλές εταιρείες-κολοσσοί στο χώρο και ερευνητές έχουν δώσει ορισμούς αναφορικά με τα Big Data. Συγκεκριμένα, η Oracle ισχυρίζεται ότι τα μεγάλα δεδομένα είναι η προέλευση της αξίας από τις παραδοσιακές σχεσιακές βάσεις δεδομένων βασισμένες στις επιχειρηματικές αποφάσεις, που επεκτείνονται με νέες πηγές μη δομημένων δεδομένων. Τέτοιες νέες πηγές περιλαμβάνουν η ιστολογία, κοινωνικά μέσα, δίκτυα αισθητήρων, δεδομένα εικόνας και άλλες μορφές δεδομένων που ποικίλλουν σε μέγεθος, δομή, μορφή και άλλους παράγοντες. Ως εκ τούτου, η Oracle υποστηρίζει έναν ορισμό ο οποίος είναι ένας ορισμός της ένταξης. Υποστηρίζουν ότι τα μεγάλα δεδομένα είναι η συμπερίληψη πρόσθετων πηγών δεδομένων

για την αύξηση των υφιστάμενων λειτουργιών. Αξιοσημείωτο είναι, και ίσως δεν εκπλήσσει, ο ορισμός της Oracle επικεντρώθηκε στην υποδομή. Σε αντίθεση με αυτά που προσφέρονται από άλλους. Η Oracle δίνει έμφαση σε μια σειρά τεχνολογιών που περιλαμβάνουν: NoSQL, Hadoop, HDFS, R και σχεσιακές βάσεις δεδομένων. Με αυτόν τον τρόπο παρουσιάζουν τόσο έναν ορισμό μεγάλων δεδομένων όσο και μια λύση σε μεγάλα δεδομένα. Ενώ ο ορισμός αυτός είναι παράξενος, εφαρμόστηκε πιο εύκολα από ό, τι άλλοι, παρόλο που στερείται ποσοτικοποίησης. Σύμφωνα με τον ορισμό της Oracle δεν είναι σαφές το πότε ακριβώς ο όρος μεγάλα δεδομένα καθίσταται εφαρμοστέος, μάλλον σημαίνει να το γνωρίζεις όταν το βλέπεις.

Η Intel συνδέει μεγάλα δεδομένα με οργανισμούς "που παράγουν ένα μέσο όρο 300 terrabytes (TB) δεδομένων εβδομαδιαία “, περιγράφοντας τα μεγάλα δεδομένα μέσω ποσοτικοποίησης των εμπειριών των επιχειρηματικών εταίρων της. Η Intel υποδηλώνει ότι οι οργανώσεις που συμμετείχαν στην έρευνα ασχολούνται εκτεταμένα με αδόμητα δεδομένα και δίνει έμφαση σχετικά με την εκτέλεση αναλυτικών στοιχείων σχετικά με τα δεδομένα που παράγονται με ρυθμό μέχρι 500 TB την εβδομάδα. Επίσης, επισημαίνεται ότι ο πιο κοινός τύπος δεδομένων που συμμετέχει στην ανάλυση είναι οι επιχειρηματικές συναλλαγές που είναι αποθηκευμένες σε σχεσιακές βάσεις δεδομένων (σύμφωνα με την Oracle's ορισμός), ακολουθούμενη από έγγραφα, ηλεκτρονικό ταχυδρομείο, δεδομένα αισθητήρων, ιστολογικές σελίδες και τα κοινωνικά μέσα.

Η Microsoft παρέχει ένα σαφώς καθορισμένο ορισμό: "Τα μεγάλα δεδομένα είναι ο όρος που χρησιμοποιείται όλο και περισσότερο για να περιγράψει τη διαδικασία εφαρμογής της σοβαρής υπολογιστικής ισχύος και μηχανικής μάθησης σε πολύπλοκα σύνολα πληροφοριών". Αυτός ο ορισμός δηλώνει με αβεβαιότητα ότι μεγάλα δεδομένα απαιτούν την εφαρμογή σημαντικής υπολογιστικής ισχύος. Αυτό γίνεται αντιληπτό σε προηγούμενους ορισμούς, αλλά όχι καθολικά.

Η Google trends αναφέρθηκε στα μεγάλα δεδομένα με διαφορετικό τρόπο, με τους εξής όρους: ανάλυση δεδομένων, Hadoop, No SQL, Google, IBM και Oracle. Από αυτούς τους όρους μια σειρά τάσεις είναι εμφανής. Αρχικά ότι τα μεγάλα δεδομένα είναι άρρηκτα συνδεδεμένα με την ανάλυση δεδομένων και την εξαγωγή γνώσης από τα δεδομένα. Τέλος είναι σαφές ότι υπάρχει ένας αριθμός βιομηχανικών οργανισμών που σχετίζονται με τα μεγάλα δεδομένα.

Τέλος, ο Gartner προτείνει τον ορισμό που περιλαμβάνει τα 3 Vs: Όγκο (volume), Ταχύτητα (Velocity) και Ποικιλία (Variety). Ο ορισμός αυτός έχει καθιερωθεί και περιγράφεται στην επόμενη ενότητα.

2.2 Το πρότυπο των 3v

Όγκος (volume)

Η λέξη «Big» στα μεγάλα δεδομένα ορίζει τον όγκο. Τα υπάρχοντα δεδομένα είναι σε petabytes και αναμένεται να αυξηθούν σε zettabytes στο κοντινό μέλλον. Τα μέσα κοινωνικής δικτύωσης παράγουν τα ίδια δεδομένα κατά σειρά terabytes καθημερινά και αυτή η ποσότητα δεδομένων είναι σίγουρα δύσκολο να αντιμετωπιστούν χρησιμοποιώντας τα υπάρχοντα παραδοσιακά συστήματα. Μια έρευνα που διεξήγαγε η IBM στα μέσα του 2012 αποκάλυψε ότι πάνω από το ήμισυ των 1144 ερωτηθέντων θεωρούσαν δεδομένα δεδομένων πάνω από ένα terabyte ως μεγάλα δεδομένα (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). Ένα terabyte αποθηκεύει τόσες πληροφορίες που θα ταιριάζουν σε 1500 CD ή 220 DVD, αρκετά για να αποθηκεύουν περίπου 16 εκατομμύρια φωτογραφίες στο Facebook. Οι Beaver, Kumar, Li, Sobel και Vajgel (2010) αναφέρουν ότι το Facebook επεξεργάζεται έως και ένα εκατομμύριο φωτογραφίες ανά δευτερόλεπτο. Ένα petabyte ισούται με 1024 terabytes. Σύμφωνα με προηγούμενες εκτιμήσεις, το Facebook αποθηκεύει 260 δισεκατομμύρια φωτογραφίες χρησιμοποιώντας χώρο αποθήκευσης άνω των 20 petabytes.

Ταχύτητα (Velocity)

Η ταχύτητα στα μεγάλα δεδομένα, είναι μια ιδέα που ασχολείται με την ταχύτητα των δεδομένων που προέρχονται από διάφορες πηγές. Αυτό το χαρακτηριστικό δεν περιορίζεται στην ταχύτητα των εισερχόμενων δεδομένων αλλά και στην ταχύτητα κατά την οποία ρέουν τα δεδομένα. Για παράδειγμα, τα δεδομένα από αισθητήρες συσκευών μετακινούνταν συνεχώς στη βάση δεδομένων και αυτό το ποσό δεν είναι μικρό. Έτσι τα παραδοσιακά

συστήματα δεν είναι αρκετά ικανά για την εκτέλεση των αναλύσεων σε δεδομένα που βρίσκονται σε συνεχή κίνηση. Ο πολλαπλασιασμός των ψηφιακών συσκευών, όπως τα smart phones και οι αισθητήρες, έχει οδηγήσει σε έναν πρωτοφανή ρυθμό δημιουργίας δεδομένων και οδηγεί σε αυξανόμενη ανάγκη για αναλύσεις σε πραγματικό χρόνο και σχεδιασμό βάσει στοιχείων. Ακόμη και οι συμβατικοί λιανοπωλητές δημιουργούν δεδομένα υψηλής συχνότητας. Η Wal-Mart, για παράδειγμα, επεξεργάζεται περισσότερα από ένα εκατομμύριο συναλλαγές ανά ώρα (Cukier, 2010). Τα δεδομένα που προέρχονται από κινητές συσκευές και διακινούνται μέσω κινητών εφαρμογών παράγουν χείμαρρους πληροφοριών που μπορούν να χρησιμοποιηθούν για τη δημιουργία εξατομικευμένων προσφορών σε πραγματικό χρόνο για τους καθημερινούς πελάτες. Αυτά τα δεδομένα παρέχουν σωστές πληροφορίες σχετικά με τους πελάτες, όπως η γεωγραφική τοποθεσία, τα δημογραφικά στοιχεία και τα πρότυπα αγορών του παρελθόντος, τα οποία μπορούν να αναλυθούν σε πραγματικό χρόνο για να δημιουργήσουν πραγματική αξία για τους πελάτες.

Ποικιλία (Variety)

Η ποικιλία αναφέρεται στη δομική ετερογένεια σε ένα σύνολο δεδομένων. Οι τεχνολογικές εξελίξεις επιτρέπουν στις επιχειρήσεις να χρησιμοποιούν διάφορους τύπους δομημένων, ημι-δομημένων και αδόμητων δεδομένων. Τα δομημένα δεδομένα, τα οποία αποτελούν μόνο το 5% όλων των υφιστάμενων δεδομένων (Cukier, 2010), αναφέρονται σε πίνακες που υπάρχουν στα υπολογιστικά φύλλα ή στις σχεσιακές βάσεις δεδομένων. Το κείμενο, οι εικόνες, ο ήχος και τα βίντεο είναι παραδείγματα μη δομημένων δεδομένων, τα οποία μερικές φορές στερούνται της δομικής οργάνωσης που απαιτείται από τις μηχανές για ανάλυση. Η μορφή των ημι-δομημένων δεδομένων, που καλύπτει μια συνέχεια μεταξύ πλήρως δομημένων και μη δομημένων δεδομένων, δεν συμμορφώνεται με αυστηρά πρότυπα. Η εκτεταμένη γλώσσα σήμανσης (XML), μια γλώσσα κειμένου για την ανταλλαγή δεδομένων στον Ιστό, είναι ένα τυπικό παράδειγμα ημί-δομημένων δεδομένων. Τα έγγραφα XML περιέχουν ετικέτες δεδομένων καθορισμένες από το χρήστη, οι οποίες τις καθιστούν αναγνώσιμες από το μηχάνημα.

Ένα υψηλό επίπεδο ποικιλίας, ένα καθοριστικό χαρακτηριστικό των μεγάλων δεδομένων, δεν είναι απαραίτητα νέο. Οι οργανισμοί έχουν αποθηκεύσει μη δομημένα δεδομένα από εσωτερικές πηγές (π.χ. δεδομένα αισθητήρων) και εξωτερικές πηγές (π.χ. κοινωνικά μέσα).

Ωστόσο, η εμφάνιση νέων τεχνολογιών διαχείρισης δεδομένων και αναλυτικών στοιχείων, που επιτρέπουν στους οργανισμούς να αξιοποιούν δεδομένα στις επιχειρηματικές τους διαδικασίες, είναι η καινοτόμος πτυχή. Για παράδειγμα, οι τεχνολογίες αναγνώρισης προσώπου εξουσιοδοτούν τους λιανοπωλητές να αποκτήσουν νοημοσύνη σχετικά με την κυκλοφορία των καταστημάτων, την ηλικία ή τη σύνθεση των φύλων των πελατών τους, καθώς και τα πρότυπα κίνησης των καταστημάτων τους. Αυτές οι ανεκτίμητες πληροφορίες αξιοποιούν τις αποφάσεις που σχετίζονται με τις προωθήσεις προϊόντων, την τοποθέτηση και τη στελέχωση. Τα δεδομένα Clickstream παρέχουν πληθώρα πληροφοριών σχετικά με τη συμπεριφορά των πελατών και τα πρότυπα περιήγησης σε διαδικτυακούς εμπόρους λιανικής πώλησης. Το Clickstream παρέχει συμβουλές σχετικά με το χρόνο και τη σειρά των σελίδων που προβάλλει ο πελάτης. Με τη χρήση μεγάλων αναλυτικών στοιχείων, ακόμη και οι μικρομεσαίες επιχειρήσεις (ΜΜΕ) μπορούν να προωθήσουν τεράστιους όγκους ημι-δομημένων δεδομένων για να βελτιώσουν τα σχέδια ιστοσελίδων και να εφαρμόσουν αποτελεσματικά συστήματα cross-selling και εξατομικευμένων συστάσεων για τα προϊόντα.

Εκτός από τα τρία V, έχουν αναφερθεί και άλλες διαστάσεις μεγάλων δεδομένων. Αυτά περιλαμβάνουν:

• Φιλαλήθεια (Veracity). Η IBM δημιούργησε το Veracity ως το τέταρτο V, το οποίο αντιπροσωπεύει την αναξιοπιστία που είναι εγγενής σε ορισμένες πηγές δεδομένων. Για παράδειγμα, τα συναισθήματα των πελατών στα κοινωνικά μέσα ενημέρωσης είναι αβέβαια από τη φύση τους, καθώς συνεπάγονται ανθρώπινη κρίση. Ωστόσο, περιέχουν πολύτιμες πληροφορίες. Έτσι, η ανάγκη αντιμετώπισης ανακριβών και αβέβαιων δεδομένων είναι μια άλλη πτυχή μεγάλων δεδομένων, η οποία αντιμετωπίζεται χρησιμοποιώντας εργαλεία και αναλυτικά στοιχεία που αναπτύσσονται για τη διαχείριση και την εξόρυξη αβέβαιων δεδομένων.

• Μεταβλητότητα (Variability). Το SAS εισήγαγε τη μεταβλητότητα και την πολυπλοκότητα ως δύο επιπλέον διαστάσεις των μεγάλων δεδομένων. Η μεταβλητότητα αναφέρεται στην μεταβολή των ρυθμών ροής δεδομένων. Συχνά, η μεγάλη ταχύτητα δεδομένων δεν είναι συνεπής και έχει περιοδικές κορυφές και γούρνες. Η πολυπλοκότητα αναφέρεται στο γεγονός ότι τα μεγάλα δεδομένα παράγονται μέσω μιας πληθώρας πηγών. Αυτό επιβάλλει μια κρίσιμη πρόκληση: την ανάγκη σύνδεσης, αντιστοίχισης, καθαρισμού και μετατροπής δεδομένων που λαμβάνονται από διαφορετικές πηγές.

• Αξία (Value). Η Oracle παρουσίασε την τιμή ως καθοριστική ιδιότητα των μεγάλων δεδομένων. Με βάση τον ορισμό της Oracle, τα μεγάλα δεδομένα συχνά χαρακτηρίζονται από σχετικά χαμηλή πυκνότητα αξίας. Δηλαδή, τα δεδομένα που λαμβάνονται στην αρχική μορφή συνήθως έχουν χαμηλή τιμή σε σχέση με τον όγκο τους. Ωστόσο, μπορεί να επιτευχθεί υψηλή τιμή με την ανάλυση μεγάλων όγκων τέτοιων δεδομένων.

Η σχετικότητα των μεγάλων όγκων δεδομένων που συζητήσαμε νωρίτερα ισχύει για όλες τις διαστάσεις. Επομένως, τα καθολικά σημεία αναφοράς δεν υπάρχουν για τον όγκο, την ποικιλία και την ταχύτητα που καθορίζουν τα μεγάλα δεδομένα. Τα όρια καθορισμού εξαρτώνται από το μέγεθος, τον τομέα και τη θέση της επιχείρησης και τα όρια αυτά εξελίσσονται με την πάροδο του χρόνου. Επίσης σημαντικό είναι το γεγονός ότι αυτές οι διαστάσεις δεν είναι ανεξάρτητες μεταξύ τους. Καθώς αλλάζει μια διάσταση, αυξάνεται η πιθανότητα ότι μια άλλη διάσταση θα αλλάξει ως αποτέλεσμα. Εντούτοις, υπάρχει ένα «σημείο ανατροπής τριών V» για κάθε επιχείρηση πέραν του οποίου οι παραδοσιακές τεχνολογίες διαχείρισης και ανάλυσης δεδομένων καθίστανται ανεπαρκείς για την απόκτηση έγκαιρης ευφυΐας. Το σημείο ανατροπής 3-V είναι το όριο πέρα από το οποίο οι επιχειρήσεις αρχίζουν να ασχολούνται με μεγάλα δεδομένα. Οι επιχειρήσεις θα πρέπει στη συνέχεια να ανταλλάξουν τη μελλοντική αξία που αναμένεται από τις μεγάλες τεχνολογίες δεδομένων έναντι του κόστους εφαρμογής τους.

2.3 **Ανάλυση στατιστικών μεγάλων δεδομένων**

Η εύρεση δομής στα δεδομένα αλλά και η πρόβλεψη είναι τα σημαντικότερα βήματα στην επιστήμη δεδομένων (Data Science). Εδώ, ειδικότερα, οι στατιστικές μέθοδοι είναι απαραίτητες γιατί είναι σε θέση να χειριστούν πολλά διαφορετικά αναλυτικά καθήκοντα. Σημαντικά παραδείγματα μεθόδων ανάλυσης στατιστικών δεδομένων είναι τα ακόλουθα:

1)Hypothesis testing (Έλεγχος της υπόθεσης). Είναι ένας από τους πυλώνες της στατιστικής ανάλυσης. Στην περίπτωση αυτή πολλές ερωτήσεις που προκύπτουν από τα σχετιζόμενα προβλήματα με τα δεδομένα, μπορούν να μεταφραστούν σε υποθέσεις. Όταν αναφερόμαστε σε υποθέσεις εννοούμε τους φυσικούς δεσμούς που υπάρχουν ανάμεσα στην θεωρία και τα στατιστικά. Δεδομένου ότι οι στατιστικές υποθέσεις σχετίζονται με στατιστικές δοκιμασίες, οι ερωτήσεις και η θεωρία μπορούν να δοκιμαστούν για τα διαθέσιμα δεδομένα. Η πολλαπλή χρήση των ίδιων δεδομένων σε διαφορετικές δοκιμές συχνά συμβαίνει με την ανάγκη

διορθώσεως των επιπέδων σημασίας. Τα μη εφαρμόσιμα στατιστικά στοιχεία, η διόρθωση των πολλαπλών δοκιμών είναι ένα από τα σημαντικότερα προβλήματα, π.χ. στις φαρμακευτικές μελέτες. Η παραίτηση από τέτοιες τεχνικές θα οδηγούσε σε πολλά παραπλανητικά αποτελέσματα από τα δικαιολογημένα.

2)Classification (Ταξινόμηση). Οι μέθοδοι ταξινόμησης είναι βασικές για τον εντοπισμό και την πρόβλεψη υπό πληθυσμών από δεδομένα. Στην αποκαλούμενη περίπτωση μη επίδειξης, τέτοιοι υπό πληθυσμοί πρέπει να βρεθούν από data που δεν έχουν εκ των προτέρων κατανοήσει τις περιπτώσεις τέτοιων υπό πληθυσμών. Αυτό ονομάζεται συχνά ομαδοποίηση. Στην επονομαζόμενη εποπτευόμενη περίπτωση, η ταξινόμηση δεν πρέπει να βρεθεί από ένα σύνολο δεδομένων που να χαρακτηρίζεται για την πρόβλεψη άγνωστων ετικετών, όταν μόνο οι παράγοντες είναι διαθέσιμοι στην περιοχή. Στην εποχή του Μεγάλου Δεδομένου, φαίνεται ότι είναι απαραίτητη μια νέα ματιά στις κλασσικές μεθόδους, δεδομένου ότι το μεγαλύτερο μέρος του χρόνου, η προσπάθεια υπολογισμού των σύνθετων μεθόδων ανάλυσης μεγαλώνει και γίνεται ισχυρότερη από την γραμμική με τον αριθμό των παρατηρήσεων και τον αριθμό των χαρακτηριστικών P. Στην περίπτωση των Big Data, αυτό οδηγεί σε υπερβολικά μεγάλους χρόνους υπολογισμού και σε αριθμητικά προβλήματα. Έτσι καταλήγουμε τόσο στην επιστροφή απλούστερων αλγορίθμων βελτιστοποίησης με χαμηλή χρονική πολυπλοκότητα, όσο και στην επανεξέταση των παραδοσιακών μεθόδων στατιστικής και μηχανικής μάθησης για το Big Data.

3)Regression (Παλινδρόμηση). Οι μέθοδοι παλινδρόμησης είναι το κύριο εργαλείο για την εύρεση παγκόσμιων και τοπικών σχέσεων μεταξύ χαρακτηριστικών όταν μετριέται η μεταβλητή tar-get. Ανάλογα με την κατανομή των υποκείμενων δεδομένων, μπορούν να εφαρμοστούν διαφορετικές προσεγγίσεις. Κάτω από την υπόθεση της κανονικότητας, η γραμμική επαναφορά είναι η πιο κοινή μέθοδος, ενώ γενικευμένη γραμμική παλινδρόμηση χρησιμοποιείται συνήθως για άλλες διανομές από την εκθετική οικογένεια. Οι περισσότερες προχωρημένες μέθοδοι περιλαμβάνουν τη λειτουργική παλινδρόμηση για τα λειτουργικά δεδομένα, την ποσοτική παλινδρόμηση και την παλινδρόμηση που βασίζεται σε λειτουργίες απώλειας. Αυτές οι λειτουργίες στον κόσμο των Big Data, είναι παρόμοιες με εκείνες για τις μεθόδους ταξινομήσεως δεδομένου μεγάλου αριθμού και μεγάλων αριθμών χαρακτηριστικών. Για τη μείωση του αριθμού των πιο ευαίσθητων χαρακτηριστικών, μπορούν να χρησιμοποιηθούν μεταβλητές προσεγγίσεις συρρίκνωσης, διατηρώντας την ποιότητα των χαρακτηριστικών.

4)Time series analysis (Ανάλυση χρόνο-σειρών). Η ανάλυση των χρόνο-σειρών αποσκοπεί στην κατανόηση και στην πρόβλεψη της χρονικής δομής. Οι χρονολογικές σειρές είναι πολύ συνηθισμένες μελέτες των δεδομένων παρατήρησης και η πρόβλεψη είναι πολύ σημαντική πρόκληση για τέτοια δεδομένα. Τυπική εφαρμογή είναι οι επιστήμες συμπεριφοράς και η οικονομία καθώς και οι φυσικές επιστήμες και η μηχανική. Για παράδειγμα, ας δούμε την ανάλυση σήματος, π.χ. ανάλυση ομιλίας ή μουσικής. Εδώ, οι στατιστικές μέθοδοι περιλαμβάνουν την ανάλυση των μοντέλων στις περιοχές χρόνου και συχνότητας. Ο σκοπός της θεματικής είναι η πρόβλεψη μελλοντικών αξιών των ίδιων των χρόνο-σειρών ή των ιδιοτήτων τους. Παραδείγματος χάριν, η δόνηση μιας ακουστικής χρόνο-σειράς μπορεί να διαμορφωθεί έτσι ώστε να προβλεφθεί πραγματικός ο ήχος στο μέλλον και η βασική συχνότητα ενός μουσικού τόνος μπορεί να προκαθοριστεί από κανόνες που ελήφθησαν από παρελθόντα χρονικά διαστήματα. Στις τεχνικές εφαρμογές, ο έλεγχος της διαδικασίας είναι ένας κοινός στόχος της ανάλυσης της χρονικής σειράς.

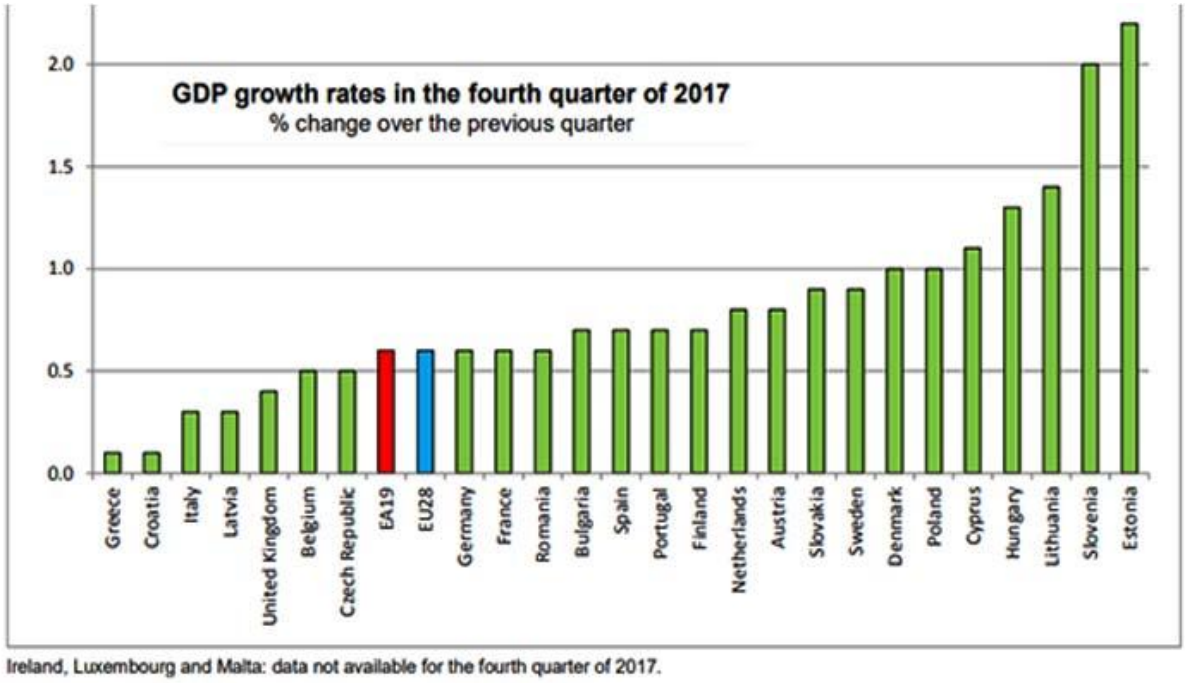


Figure 1 Παράδειγμα στατιστικής Μεγάλων Δεδομένων για την ανάπτυξη των χωρών

2.4 Τεχνολογίες αποθήκευσης μεγάλων δεδομένων

Η τεχνολογία αποθήκευσης δεδομένων τις τελευταίες δεκαετίες έχει μεταβληθεί σημαντικά. Πλέον, τα διάφορα συστήματα βάσεων δεδομένων που έχουν αναπτυχθεί είναι σε θέση να χειρίζονται σύνολα δεδομένων σε διαφορετικές κλίμακες υποστηρίζοντας ταυτόχρονα διάφορες εφαρμογές. Οι παραδοσιακές σχεσιακές βάσεις δεδομένων δεν μπορούν πλέον να ανταποκριθούν στις προκλήσεις σχετικά με τις κατηγορίες και τις κλίμακες που επέφεραν τα Big Data (μεγάλα δεδομένα). Οι βάσεις δεδομένων NoSQL (μη παραδοσιακές σχεσιακές βάσεις δεδομένων) γίνονται όλο και πιο δημοφιλής για τα Big Data. Οι NoSQL διαθέτουν ευέλικτα χαρακτηριστικά και στοιχεία πλήρους υποστήριξης των Big Data μετατρέποντάς τις σε βασική τεχνολογία για τα Big Data. Οι σημαντικότερες βάσεις δεδομένων που χρησιμοποιούνται στην τεχνολογία NoSQL είναι οι εξής:

1. Βάσεις δεδομένων κλειδιού - τιμής (key - value databases). Οι βάσεις δεδομένων κλειδιού - τιμής αποτελούν ένα απλό μοντέλο δεδομένων με τα δεδομένα που αποθηκεύονται να αντιστοιχούν σε βασικές τιμές. Οι συγκεκριμένες βάσεις δεδομένων διαθέτουν απλή δομή και χαρακτηρίζονται από υψηλή επεκτασιμότητα και μικρότερο χρόνο απόκρισης ερωτήματος σε σύγκριση με τις σχεσιακές βάσεις δεδομένων (DeCandia et al., 2007:207-208).

2. Βάση προσανατολισμένη σε στήλες. Η βάση δεδομένων σε στήλες αποθηκεύει και επεξεργάζεται τα δεδομένα σύμφωνα με τις στήλες εκτός από σειρές. Οι συγκεκριμένες βάσεις δεδομένων είναι εμπνευσμένες κυρίως από το Google Big Table. Τα δεδομένα είναι κατανεμημένα και δομημένα σε ένα σύστημα αποθήκευσης δεδομένων, το οποίο έχει σχεδιαστεί για την επεξεργασία δεδομένων μεγάλης κλίμακας (Chang et al., 2008:4-6).

3. Βάση δεδομένων εγγράφων. Η συγκεκριμένη τεχνολογία μπορεί να υποστηρίξει πιο σύνθετες μορφές δεδομένων. Οι βασικότεροι εκπρόσωποι 30 των συστημάτων αποθήκευσης του εγγράφου είναι το MongoDB, το SimpleDB και το CouchDB (Chen et al., 2014:187-189).

2.5 Τεχνολογίες ανάλυσης μεγάλων δεδομένων

Τα μεγάλα δεδομένα είναι άχρηστα αν δεν μπορούν να αναλυθούν και να αξιοποιηθούν. Η δυνητική αξία τους «ξεκλειδώνεται» μόνο όταν χρησιμοποιούνται για τη λήψη αποφάσεων. Για να καταστεί δυνατή η λήψη τέτοιων αποφάσεων βάσει τεκμηριωμένων στοιχείων, οι οργανισμοί χρειάζονται αποτελεσματικές διαδικασίες για να μετατρέψουν μεγάλους όγκους δεδομένων ταχέως κινούμενων και ποικίλων δεδομένων σε χρήσιμες γνώσεις.

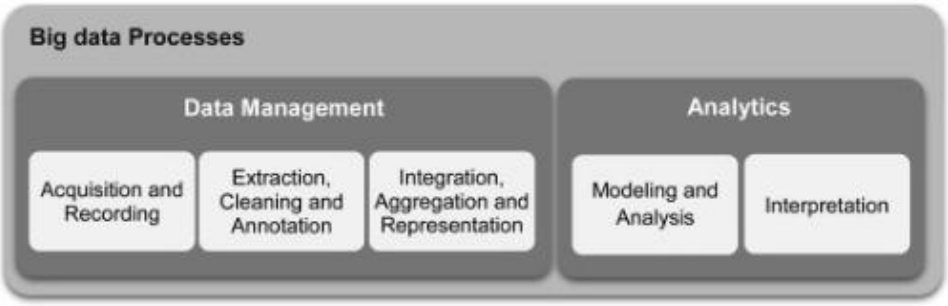


Figure 1 Διαδικασίες Big Data

Η συνολική διαδικασία εξαγωγής πληροφοριών από μεγάλα δεδομένα μπορεί να αναλυθεί σε πέντε στάδια (Labrinidis & Jagadish, 2012), που παρουσιάζονται στην εικόνα 2. Αυτά τα πέντε στάδια αποτελούν τις δύο κύριες υπό-διαδικασίες: διαχείριση δεδομένων και ανάλυση. Η διαχείριση δεδομένων περιλαμβάνει διαδικασίες και υποστηρικτικές τεχνολογίες για την απόκτηση και αποθήκευση δεδομένων και την προετοιμασία και ανάκτηση για ανάλυση. Το Analytics, από την άλλη πλευρά, αναφέρεται σε τεχνικές που χρησιμοποιούνται για την ανάλυση και την απόκτηση πληροφοριών από μεγάλα δεδομένα. Έτσι, οι μεγάλες αναλύσεις δεδομένων μπορούν να θεωρηθούν ως μια υπό-διαδικασία στη συνολική διαδικασία της «εξαγωγής γνώσεων» από μεγάλα δεδομένα. Στις επόμενες υποενότητες, εξετάζουμε εν συντομία τις μεγάλες αναλυτικές τεχνικές δεδομένων για δομημένα και αδόμητα δεδομένα. Δεδομένου του εύρους των τεχνικών, ένας εξαντλητικός κατάλογος τεχνικών είναι πέρα από το πεδίο εφαρμογής ενός ενιαίου εγγράφου. Επομένως, οι ακόλουθες τεχνικές αντιπροσωπεύουν ένα σχετικό υποσύνολο των διαθέσιμων εργαλείων για την ανάλυση μεγάλων δεδομένων.

2.5.1 **Text analytics**

Οι αναλύσεις κειμένου (εξόρυξη κειμένου) αναφέρονται σε τεχνικές που εξάγουν πληροφορίες από δεδομένα κειμένου. Τα feeds των κοινωνικών δικτύων, τα μηνύματα ηλεκτρονικού ταχυδρομείου, οι ιστοσελίδες, τα ηλεκτρονικά φόρουμ, οι απαντήσεις των ερευνών, τα εταιρικά έγγραφα, τα νέα και τα ημερολόγια του τηλεφωνικού κέντρου είναι παραδείγματα κειμένων δεδομένων που τηρούν οι οργανισμοί. Οι αναλύσεις κειμένων περιλαμβάνουν στατιστική ανάλυση, υπολογιστική γλωσσολογία και μηχανική μάθηση. Τα αναλυτικά κείμενα επιτρέπουν στις επιχειρήσεις να μετατρέπουν μεγάλους όγκους κειμένων

που παράγονται από ανθρώπους σε σημαντικές περιλήψεις, οι οποίες υποστηρίζουν τη λήψη αποφάσεων βάσει στοιχείων. Για παράδειγμα, οι αναλύσεις κειμένου μπορούν να χρησιμοποιηθούν για την πρόβλεψη χρηματιστηριακής αγοράς με βάση πληροφορίες που προέρχονται από οικονομικές ειδήσεις (Chung, 2014). Παρουσιάζω μια σύντομη ανασκόπηση των μεθόδων ανάλυσης κειμένου παρακάτω.

Information extraction (IE): Οι τεχνικές εξαγωγής πληροφοριών εξάγουν δομημένα δεδομένα από αδόμητο κείμενο. Για παράδειγμα, οι αλγόριθμοι μπορούν να εξαγάγουν δομημένες πληροφορίες όπως όνομα φαρμάκου, δοσολογία και συχνότητα από ιατρικές συνταγές. Δύο υπο-εργασίες στο IE είναι η Αναγνώριση Οντοτήτων (Entity Recognition) και η Εξόρυξη Συσχετισμού (Relation Extraction) (Jiang, 2012). Το ER βρίσκει ονόματα σε κείμενο και τα ταξινομεί σε προκαθορισμένες κατηγορίες, όπως άτομο, ημερομηνία, τοποθεσία και οργάνωση. Το RE βρίσκει και εξάγει σημασιολογικές σχέσεις μεταξύ των οντοτήτων (π.χ. ατόμων, οργανισμών, φαρμάκων, γονιδίων κλπ.) Στο κείμενο. Για παράδειγμα, δεδομένης της φράσης "Steve Jobs συνιδρυτής της Apple Inc. το 1976", ένα σύστημα RE μπορεί να εξαγάγει σχέσεις όπως ο FounderOf (Steve Jobs, Apple Inc.) ή FoundedIn (Apple Inc., 1976).

Οι τεχνικές συνοπτικής τεκμηρίωσης κειμένων παράγουν αυτόματα μια συνοπτική περίληψη ενός ή πολλαπλών εγγράφων. Η προκύπτουσα περίληψη μεταβιβάζει τις βασικές πληροφορίες στο αρχικό κείμενο. Οι εφαρμογές περιλαμβάνουν επιστημονικά και ειδησεογραφικά άρθρα, διαφημίσεις, μηνύματα ηλεκτρονικού ταχυδρομείου και ιστοσελίδες. Σε γενικές γραμμές, η σύνοψη ακολουθεί δύο προσεγγίσεις: την εξορυκτική προσέγγιση και την αφηρημένη προσέγγιση. Στην εξορυκτική σύνοψη, δημιουργείται περίληψη από τις αρχικές μονάδες κειμένου (συνήθως προτάσεις). Η περίληψη που προκύπτει είναι ένα υποσύνολο του αρχικού εγγράφου. Με βάση την προσέγγιση εξόρυξης, η διατύπωση μιας σύνοψης περιλαμβάνει τον προσδιορισμό των χαρακτηριστικών μονάδων ενός κειμένου και τη σύζευξη τους. Η σημασία των μονάδων κειμένου αξιολογείται με ανάλυση της θέσης και της συχνότητας τους στο κείμενο. Οι τεχνικές εξομάλυνσης της σύνοψης δεν απαιτούν «κατανόηση» του κειμένου. Αντίθετα, οι αφηρημένες τεχνικές συνοπτικής περιγραφής περιλαμβάνουν την εξαγωγή σημασιολογικών πληροφοριών από το κείμενο. Οι περιλήψεις περιέχουν μονάδες κειμένου που δεν υπάρχουν απαραίτητα στο αρχικό κείμενο. Προκειμένου να γίνει ανάλυση του αρχικού κειμένου και να δημιουργηθεί η περίληψη, η αποσπασματική σύνοψη περιλαμβάνει προηγμένες τεχνικές επεξεργασίας φυσικής γλώσσας

(NLP). Ως αποτέλεσμα, τα αποσπασματικά συστήματα τείνουν να παράγουν πιο συνεκτικές περιλήψεις από ό, τι τα εξορυκτικά συστήματα (Hahn & Mani, 2000). Ωστόσο, τα εξορυκτικά συστήματα είναι ευκολότερο να υιοθετηθούν, ειδικά για μεγάλα δεδομένα.

Οι τεχνικές απάντησης ερωτήσεων (Question Answering) παρέχουν απαντήσεις σε ερωτήσεις που τίθενται στη φυσική γλώσσα. Η Siri της Apple και η Watson της IBM αποτελούν παραδείγματα εμπορικών συστημάτων QA. Τα συστήματα αυτά έχουν εφαρμοστεί στην υγειονομική περίθαλψη, τη χρηματοδότηση, το μάρκετινγκ και την εκπαίδευση. Παρόμοια με την αθροιστική σύνοψη, τα συστήματα QA βασίζονται σε σύνθετες τεχνικές NLP. Οι τεχνικές QA ταξινομούνται περαιτέρω σε τρεις κατηγορίες: την προσέγγιση ανάκτησης πληροφοριών (IR), την προσέγγιση που βασίζεται στη γνώση και την υβριδική προσέγγιση. Τα συστήματα QA που βασίζονται σε IR συχνά έχουν τρία υποσυστήματα. Πρώτον, είναι η επεξεργασία ερωτήσεων, που χρησιμοποιείται για τον προσδιορισμό των λεπτομερειών, όπως ο τύπος ερώτησης, η εστίαση ερωτήσεων και ο τύπος απάντησης, οι οποίοι χρησιμοποιούνται για τη δημιουργία ενός ερωτήματος. Δεύτερη, είναι η επεξεργασία εγγράφων που χρησιμοποιείται για την ανάκτηση σχετικών προ-γραπτών αποσπασμάτων από ένα σύνολο υφιστάμενων εγγράφων χρησιμοποιώντας το ερώτημα που διατυπώνεται στην εν λόγω επεξεργασία ερωτήσεων. Η τρίτη, είναι η επεξεργασία των απαντήσεων, η οποία χρησιμοποιείται για να εξάγει τις πιθανές απαντήσεις από την έξοδο του προηγούμενου στοιχείου, να τις ταξινομεί και να επιστρέφει την πιθανή απάντηση με την υψηλότερη κατάταξη ως το αποτέλεσμα του συστήματος QA. Τα συστήματα QA που βασίζονται στη γνώση δημιουργούν μια σημασιολογική περιγραφή της ερώτησης, η οποία στη συνέχεια χρησιμοποιείται για την αναζήτηση δομημένων πόρων. Τα συστήματα QA βασισμένα στη γνώση είναι ιδιαίτερα χρήσιμα σε περιορισμένους τομείς, όπως ο τουρισμός, η ιατρική και οι μεταφορές, όπου δεν υπάρχουν μεγάλοι όγκοι εγγράφων. Τέτοιες περιοχές δεν διαθέτουν δεδομένα πλεονασμού, η οποία απαιτείται για τα συστήματα QA που βασίζονται σε IR. Το Siri της Apple είναι ένα παράδειγμα συστήματος QA που εκμεταλλεύεται την προσέγγιση που βασίζεται στη γνώση. Σε υβριδικά συστήματα QA, όπως ο Watson της IBM, ενώ η ερώτηση αναλύεται σημασιολογικά, οι πιθανές απαντήσεις δημιουργούνται χρησιμοποιώντας τις μεθόδους IR.

Οι τεχνικές ανάλυσης αισθήσεων (εξόρυξης γνώμης) αναλύουν το κείμενο με κατανόηση, το οποίο περιέχει τις απόψεις των ανθρώπων για θέματα όπως προϊόντα, οργανώσεις, άτομα και γεγονότα. Οι επιχειρήσεις καταγράφουν όλο και περισσότερα στοιχεία σχετικά με τα

συναισθήματα των πελατών τους, γεγονός που έχει οδηγήσει στη διάδοση της αντίληψης των συναισθημάτων (Liu, 2012). Το μάρκετινγκ, η χρηματοδότηση και οι πολιτικές και κοινωνικές επιστήμες είναι οι κύριοι τομείς εφαρμογής της ανάλυσης των αισθήσεων.

Το Excel αποτελεί βασικό εργαλείο του Microsoft Office που παρέχει ισχυρή υποστήριξη στην επεξεργασία δεδομένων και στατιστικών με δυνατότητες ανάλυσης.

Το KNIME είναι ένα φιλικό και ανοιχτό προς το χρήστη εργαλείο ενοποίησης, άντλησης, επεξεργασίας και ανάλυσης δεδομένων. Επιτρέπει στους χρήστες να δημιουργήσουν ροές δεδομένων ή σχετικά κανάλια με έναν ορατό τρόπο. Το KNIME είναι γραμμένο σε Java και παρέχει λειτουργίες, διαμέσου των οποίων οι χρήστες μπορούν αν εισάγουν αρχεία δεδομένων.

2.5.2 **Audio analytics**

Οι αναλύσεις ήχου αναλύουν και εξάγουν πληροφορίες από μη δομημένα δεδομένα ήχου. Όταν εφαρμόζεται στην ανθρώπινη ομιλούμενη γλώσσα, οι αναλύσεις ήχου αναφέρονται επίσης ως speech analytics. Δεδομένου ότι αυτές οι τεχνικές έχουν εφαρμοστεί ως επί το πλείστον σε προφορικό ήχο, οι όροι ηχητικές αναλύσεις και αναλυτικά λόγια χρησιμοποιούνται συχνά εναλλακτικά. Επί του παρόντος, τα κέντρα τηλεφωνικής εξυπηρέτησης πελατών και η υγειονομική περίθαλψη αποτελούν τους κύριους τομείς εφαρμογής των αναλυτικών συστημάτων ήχου. Τα τηλεφωνικά κέντρα χρησιμοποιούν αναλυτικά στοιχεία ήχου για αποτελεσματική ανάλυση χιλιάδων ή και εκατομμυρίων ωρών καταγεγραμμένων κλήσεων. Αυτές οι τεχνικές συμβάλλουν στη βελτίωση της πελατειακής εμπειρίας, στην αξιολόγηση της απόδοσης των πρακτόρων, στην αύξηση των ποσοστών κύκλου εργασιών πωλήσεων, στην παρακολούθηση της συμμόρφωσης με διαφορετικές πολιτικές (π.χ. πολιτικές απορρήτου και ασφάλειας), στην απόκτηση γνώσης σχετικά με τη συμπεριφορά των πελατών και στον εντοπισμό ζητημάτων προϊόντων ή υπηρεσιών. Τα συστήματα ανάλυσης ήχου μπορούν να σχεδιαστούν για να αναλύσουν μια ζωντανή κλήση, να διατυπώσουν συστάσεις διασταυρούμενης / επάνω πώλησης με βάση τις παρελθούσες και παρούσες αλληλεπιδράσεις του πελάτη και να παρέχουν ανατροφοδότηση στους πράκτορες σε πραγματικό χρόνο. Επιπλέον, τα αυτοματοποιημένα τηλεφωνικά κέντρα χρησιμοποιούν τις πλατφόρμες Interactive Voice Response (IVR) για τον εντοπισμό και τον χειρισμό των απογοητευμένων καλούντων.

Στην υγειονομική περίθαλψη, οι αναλύσεις ήχου υποστηρίζουν τη διάγνωση και τη θεραπεία ορισμένων ιατρικών καταστάσεων που επηρεάζουν τα πρότυπα επικοινωνίας του ασθενούς (π.χ. κατάθλιψη, σχιζοφρένεια και καρκίνος) (Hirschberg, Hjalmarsson, & Elhadad, 2010). Επίσης, οι αναλύσεις ήχου μπορούν να βοηθήσουν στην ανάλυση των κραυγών ενός βρέφους που περιέχουν πληροφορίες για την υγεία και την συναισθηματική κατάσταση του βρέφους (Patil, 2010). Ο τεράστιος όγκος δεδομένων που καταγράφονται μέσω συστημάτων κλινικής τεκμηρίωσης με γνώμονα την ομιλία είναι ένας άλλος οδηγός για την υιοθέτηση ηχητικών αναλύσεων στην υγειονομική περίθαλψη.

Οι αναλύσεις ομιλίας ακολουθούν δύο κοινές τεχνολογικές προσεγγίσεις: την προσέγγιση που βασίζεται σε μεταγραφή (ευρέως γνωστή ως συνεχής αναγνώριση ομιλίας μεγάλου λεξιλογίου, LVCSR) και η φωνητική προσέγγιση.

• Τα συστήματα LVCSR ακολουθούν μια διαδικασία δύο φάσεων: ευρετηρίαση και αναζήτηση. Στην πρώτη φάση, προσπαθούν να μεταγράψουν το περιεχόμενο ομιλίας του ήχου. Αυτό γίνεται με αλγόριθμους αυτόματης αναγνώρισης ομιλίας (ASR) που ταιριάζουν με τους ήχους σε λέξεις. Οι λέξεις αναγνωρίζονται με βάση ένα προκαθορισμένο λεξικό. Εάν το σύστημα δεν βρει την ακριβή λέξη στο λεξικό, επιστρέφει το πιο παρόμοιο. Η έξοδος του συστήματος είναι ένα αρχείο ευρετηρίου που μπορεί να αναζητηθεί και περιέχει πληροφορίες σχετικά με την ακολουθία των λέξεων που ομιλούνται στην ομιλία. Στη δεύτερη φάση, χρησιμοποιούνται τυπικές μέθοδοι που βασίζονται σε κείμενο για την εύρεση του όρου αναζήτησης στο αρχείο ευρετηρίου.

•Τα φωνητικά συστήματα λειτουργούν με ήχους ή φωνήματα. Τα φωνήματα είναι οι αντιληπτές διαφορετικές μονάδες ήχου σε μια συγκεκριμένη γλώσσα που διακρίνουν μία λέξη από την άλλη. Τα φωνητικά συστήματα αποτελούνται επίσης από δύο φάσεις: φωνητική ευρετηρίαση και αναζήτηση. Στην πρώτη φάση, το σύστημα μεταφράζει την ομιλία εισόδου σε μια ακολουθία φωνημάτων. Αυτό έρχεται σε αντίθεση με τα συστήματα LVCSR όπου η ομιλία μετατρέπεται σε μια ακολουθία λέξεων. Στη δεύτερη φάση, το σύστημα αναζητά την έξοδο της πρώτης φάσης για την φωνητική αναπαράσταση των όρων αναζήτησης.

2.5.3 **Video analytics**

Η ανάλυση βίντεο, γνωστή και ως ανάλυση περιεχομένου βίντεο (VCA), περιλαμβάνει μια ποικιλία τεχνικών για την παρακολούθηση, την ανάλυση και την εξαγωγή σημαντικών πληροφοριών από ροές βίντεο. Παρόλο που οι αναλύσεις βίντεο εξακολουθούν να είναι σε μικρή ηλικία σε σύγκριση με άλλα είδη εξόρυξης δεδομένων (Panigrahi, Abraham, & Das, 2010), έχουν ήδη αναπτυχθεί διάφορες τεχνικές επεξεργασίας βίντεο σε πραγματικό χρόνο καθώς και σε βίντεο που έχουν ήδη εγγραφεί. Η αυξανόμενη επικράτηση των μηχανών κλειστού κυκλώματος τηλεόρασης (CCTV) και η αυξανόμενη δημοτικότητα των ιστότοπων κοινής χρήσης βίντεο αποτελούν τους δύο κορυφαίους συντελεστές στην ανάπτυξη της ηλεκτρονικής ανάλυσης βίντεο. Μια βασική πρόκληση, ωστόσο, είναι το μέγεθος των δεδομένων βίντεο. Για να το θέσουμε σε προοπτική, ένα δευτερόλεπτο ενός βίντεο υψηλής ευκρίνειας, από την άποψη του μεγέθους, είναι ισοδύναμο με πάνω από 2000 σελίδες κειμένου (Manyika et al., 2011). Τώρα θεωρείτε ότι 100 ώρες βίντεο μεταφορτώνονται στο YouTube κάθε λεπτό (Στατιστικά YouTube, n.d.).

Οι μεγάλες τεχνολογίες δεδομένων μετατρέπουν αυτήν την πρόκληση σε ευκαιρία. Αποφεύγοντας την ανάγκη για δαπανηρή και επικίνδυνη χειροκίνητη επεξεργασία, οι μεγάλες τεχνολογίες δεδομένων μπορούν να αξιοποιηθούν για να διεγείρουν αυτόματα και να αντλούν πληροφορίες από χιλιάδες ώρες βίντεο. Ως αποτέλεσμα, η μεγάλη τεχνολογία δεδομένων είναι ο τρίτος παράγοντας που συνέβαλε στην ανάπτυξη της ανάλυσης βίντεο.

Η πρωταρχική εφαρμογή των αναλύσεων βίντεο τα τελευταία χρόνια ήταν στα αυτοματοποιημένα συστήματα ασφάλειας και επιτήρησης. Εκτός από το υψηλό κόστος τους, τα συστήματα επιτήρησης με βάση το εργατικό δυναμικό τείνουν να είναι λιγότερο αποτελεσματικά από τα αυτόματα συστήματα (π.χ. Hakeem κ.ά., έκθεση 2012 ότι το προσωπικό ασφαλείας δεν μπορεί να παραμείνει εστιασμένο στα καθήκοντα επιτήρησης για περισσότερα από 20 λεπτά). Οι αναλύσεις βίντεο μπορούν αποτελεσματικά και αποτελεσματικά να επιτελούν λειτουργίες επιτήρησης όπως ανίχνευση παραβιάσεων απαγορευμένων ζωνών, εντοπισμός αντικειμένων που έχουν αφαιρεθεί ή αφεθεί χωρίς επίβλεψη, ανίχνευση κατακερματισμού σε συγκεκριμένη περιοχή, αναγνώριση ύποπτων δραστηριοτήτων και ανίχνευση παραβίασης κάμερας. Μετά την ανίχνευση μιας απειλής, το σύστημα επιτήρησης μπορεί να ειδοποιεί το προσωπικό ασφαλείας σε πραγματικό χρόνο ή

να ενεργοποιεί μια αυτόματη ενέργεια (π.χ. συναγερμός ήχου, κλειδαριές ή ενεργοποίηση φωτισμού).

Τα δεδομένα που παράγονται από κάμερες CCTV σε καταστήματα λιανικής πώλησης μπορούν να εξαχθούν για επιχειρηματική ευφυΐα. Η διαχείριση του μάρκετινγκ και των επιχειρήσεων είναι οι κύριοι τομείς εφαρμογής. Για παράδειγμα, οι έξυπνοι αλγόριθμοι μπορούν να συλλέξουν δημογραφικές πληροφορίες σχετικά με τους πελάτες, όπως η ηλικία, το φύλο και η εθνικότητα. Ομοίως, οι λιανοπωλητές μπορούν να μετρήσουν τον αριθμό των πελατών, να μετρήσουν την ώρα που παραμένουν στο κατάστημα, να εντοπίσουν τα πρότυπα κίνησης, να μετρήσουν το χρόνο παραμονής τους σε διαφορετικές περιοχές και να παρακολουθήσουν ουρές σε πραγματικό χρόνο. Πολύτιμες ιδέες μπορούν να αποκτηθούν με τη συσχέτιση αυτών των πληροφοριών με τα δημογραφικά στοιχεία των πελατών προκειμένου να ληφθούν αποφάσεις για την τοποθέτηση προϊόντων, την τιμολόγηση, τη βελτιστοποίηση του συνδυασμού, το σχεδιασμό προώθησης, τις πολλαπλές πωλήσεις, τη βελτιστοποίηση της διάταξης και τη στελέχωση.

Μια άλλη πιθανή εφαρμογή της ανάλυσης βίντεο στο λιανικό εμπόριο έγκειται στη μελέτη της αγοραστικής συμπεριφοράς των ομάδων. Μεταξύ των μελών της οικογένειας που ψωνίζουν μαζί, μόνο ένα αλληλεπιδρά με το κατάστημα στην ταμειακή μηχανή, προκαλώντας τα παραδοσιακά συστήματα να χάσουν δεδομένα σχετικά με τα πρότυπα αγορών άλλων μελών. Οι αναλύσεις βίντεο μπορούν να βοηθήσουν τους λιανοπωλητές να αντιμετωπίσουν αυτή τη χαμένη ευκαιρία παρέχοντας πληροφορίες σχετικά με το μέγεθος της ομάδας, τα δημογραφικά στοιχεία του ομίλου και την αγοραστική συμπεριφορά των μεμονωμένων μελών.

Η αυτόματη ευρετηρίαση και ανάκτηση βίντεο αποτελεί έναν άλλο τομέα των εφαρμογών ανάλυσης βίντεο. Η εκτεταμένη εμφάνιση βίντεο σε απευθείας σύνδεση και εκτός σύνδεσης έχει επισημάνει την ανάγκη για ευρετηρίαση περιεχομένου πολυμέσων για εύκολη αναζήτηση και ανάκτηση. Η ευρετηρίαση ενός βίντεο μπορεί να πραγματοποιηθεί με βάση τα διαφορετικά επίπεδα πληροφοριών που είναι διαθέσιμα σε ένα βίντεο, συμπεριλαμβανομένων των δεδομένων, του ηχητικού δίσκου, των μεταγραφών και του οπτικού περιεχομένου του βίντεο. Στην προσέγγιση με βάση τα δεδομένα, τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) χρησιμοποιούνται για την αναζήτηση και ανάκτηση βίντεο. Οι αναλύσεις ήχου και οι τεχνικές ανάλυσης κειμένου μπορούν να

εφαρμοστούν για την ευρετηρίαση ενός βίντεο βάσει των σχετικών μουσικών κομματιών και μεταγραφών, αντίστοιχα. Μια περιεκτική ανασκόπηση των προσεγγίσεων και των τεχνικών για την ευρετηρίαση βίντεο παρουσιάζεται στο Hu, Xie, Li, Zeng και Maybank (2011).

Από την άποψη της αρχιτεκτονικής του συστήματος, υπάρχουν δύο προσεγγίσεις για την ανάλυση βίντεο, δηλαδή βασισμένες σε διακομιστές και βασισμένες στις άκρες:
•Αρχιτεκτονική βασισμένη σε διακομιστές. Σε αυτήν τη διαμόρφωση, το βίντεο που καταγράφεται μέσω κάθε κάμερας οδηγείται πίσω σε έναν κεντρικό και αποκλειστικό εξυπηρετητή που εκτελεί τα αναλυτικά βίντεο. Λόγω των ορίων εύρους ζώνης, το βίντεο που παράγεται από την πηγή συνήθως συμπιέζεται μειώνοντας τους ρυθμούς καρέ και / ή την ανάλυση εικόνας. Η προκύπτουσα απώλεια πληροφοριών μπορεί να επηρεάσει την ακρίβεια της ανάλυσης. Ωστόσο, η προσέγγιση που βασίζεται σε διακομιστές παρέχει οικονομίες κλίμακας και διευκολύνει τη συντήρηση. Ως αποτέλεσμα, όλο το περιεχόμενο της ροής βίντεο είναι διαθέσιμο για την ανάλυση, επιτρέποντας μια πιο αποτελεσματική ανάλυση περιεχομένου. Τα συστήματα με βάση την άκρη, ωστόσο, είναι πιο δαπανηρά για να διατηρήσουν και να έχουν χαμηλότερη ισχύ επεξεργασίας σε σύγκριση με τα συστήματα που βασίζονται σε διακομιστές.

2.5.4 **Social media analytics**

Οι αναλύσεις κοινωνικών μέσων αναφέρονται στην ανάλυση δομημένων και αδόμητων δεδομένων από κανάλια κοινωνικών μέσων. Τα κοινωνικά μέσα είναι ένας ευρύς όρος που περιλαμβάνει μια ποικιλία από ηλεκτρονικές πλατφόρμες που επιτρέπουν στους χρήστες να δημιουργούν και να ανταλλάσσουν περιεχόμενο. Τα κοινωνικά μέσα μπορούν να ταξινομηθούν στους ακόλουθους τύπους: Κοινωνικά δίκτυα (π.χ. Facebook και LinkedIn), blogs (π.χ. Blogger και WordPress), μικροσκόπια (π.χ. Twitter και Tumblr), κοινωνικές ειδήσεις (π.χ. Digg και Reddit) (π.χ., Delicious και StumbleUpon), κοινή χρήση μέσων (π.χ. Instagram και YouTube), wikis (π.χ. Wikipedia και Wikihow), ιστοσελίδες με ερωτήσεις και απαντήσεις (π.χ. Yahoo Answers και Ask.com) Yelp, TripAdvisor) (Barbier and Liu, 2011, Gundecha και Liu, 2012). Επίσης, πολλές εφαρμογές για κινητά, όπως το Find My Friend, παρέχουν μια πλατφόρμα για κοινωνικές αλληλεπιδράσεις και ως εκ τούτου χρησιμεύουν ως κανάλια κοινωνικών μέσων μαζικής ενημέρωσης.

Παρόλο που η έρευνα για τα κοινωνικά δίκτυα χρονολογείται από τις αρχές της δεκαετίας του 1920, ωστόσο, τα social media analytics είναι ένα πεδίο ανάπτυξης που εμφανίστηκε μετά την εμφάνιση του Web 2.0 στις αρχές της δεκαετίας του 2000. Το βασικό χαρακτηριστικό των σύγχρονων κοινωνικών αναλυτικών μέσων είναι η φύση των δεδομένων. Η έρευνα για την ανάλυση κοινωνικών μέσων μαζικής ενημέρωσης εκτείνεται σε διάφορους κλάδους, όπως η ψυχολογία, η κοινωνιολογία, η ανθρωπολογία, η πληροφορική, τα μαθηματικά, η φυσική και η οικονομία. Το μάρκετινγκ αποτελεί την κύρια εφαρμογή των αναλυτικών μέσων κοινωνικής δικτύωσης τα τελευταία χρόνια. Αυτό μπορεί να αποδοθεί στην ευρεία και αυξανόμενη υιοθέτηση των κοινωνικών μέσων από τους καταναλωτές παγκοσμίως (He, Zha, & Li, 2013), στο βαθμό που η Forrester Research, Inc., προβάλλει τα κοινωνικά μέσα ως το δεύτερο ταχύτερα αναπτυσσόμενο κανάλι μάρκετινγκ στις ΗΠΑ μεταξύ 2011 και 2016 (VanBoskirk, Overby, & Takvorian, 2011).

Τα περιεχόμενα που δημιουργούν οι χρήστες (π.χ. αισθήματα, εικόνες, βίντεο και σελιδοδείκτες) και οι σχέσεις και οι αλληλεπιδράσεις μεταξύ των οντοτήτων του δικτύου (π.χ. άτομα, οργανισμοί και προϊόντα) είναι οι δύο πηγές πληροφοριών στα κοινωνικά μέσα. Με βάση αυτή την κατηγοριοποίηση, τα αναλυτικά κοινωνικά μέσα μπορούν να ταξινομηθούν σε δύο ομάδες:

• Content-based analytics: Τα αναλυτικά στοιχεία βάσει περιεχομένου επικεντρώνονται στα δεδομένα που δημοσιεύουν οι χρήστες σε πλατφόρμες κοινωνικών μέσων, όπως ανατροφοδότηση πελατών, κριτικές προϊόντων, εικόνες και βίντεο. Ένα τέτοιο περιεχόμενο στα κοινωνικά μέσα είναι συχνά ογκώδες, αδόμητο, θορυβώδες και δυναμικό. Οι αναλύσεις κειμένου, ήχου και βίντεο, όπως αναφέρθηκε προηγουμένως, μπορούν να εφαρμοστούν για να αντλήσουν γνώση από τέτοια δεδομένα. Επίσης, μπορούν να υιοθετηθούν μεγάλες τεχνολογίες δεδομένων για την αντιμετώπιση των προκλήσεων της επεξεργασίας δεδομένων.

• Structure-based analytics : Αναφέρονται επίσης ως ανάλυση κοινωνικών δικτύων. Αυτός ο τύπος αναλύσεων αφορά τη σύνθεση των δομικών χαρακτηριστικών ενός κοινωνικού δικτύου και την εξαγωγή της νοημοσύνης από τις σχέσεις μεταξύ των συμμετεχόντων οντοτήτων. Η δομή ενός κοινωνικού δικτύου διαμορφώνεται μέσω ενός συνόλου κόμβων και ακμών, που αντιπροσωπεύουν τους συμμετέχοντες και τις σχέσεις, αντίστοιχα. Το μοντέλο μπορεί να απεικονιστεί ως γράφημα που αποτελείται από τους κόμβους και τις άκρες. Ανασκοπούμε δύο τύπους γραφημάτων δικτύου, συγκεκριμένα κοινωνικά γραφήματα και

γραφήματα δραστηριότητας (Heidemann, Klier, & Probst, 2012). Στα κοινωνικά γραφήματα, ένα άκρο μεταξύ ενός ζεύγους κόμβων σημαίνει μόνο την ύπαρξη ενός συνδέσμου (π.χ. φιλίας) μεταξύ των αντίστοιχων οντοτήτων. Τέτοια γραφήματα μπορούν να εξορύσσονται για τον εντοπισμό κοινοτήτων ή τον προσδιορισμό κόμβων (δηλ. Των χρηστών με σχετικά μεγάλο αριθμό άμεσων και έμμεσων κοινωνικών συνδέσεων). Σε δίκτυα δραστηριότητας, ωστόσο, οι άκρες αντιπροσωπεύουν πραγματικές αλληλεπιδράσεις μεταξύ οποιουδήποτε ζεύγους κόμβων. Οι αλληλεπιδράσεις περιλαμβάνουν ανταλλαγές πληροφοριών (π.χ., αρέσει και σχόλια). Τα γραφήματα δραστηριότητας είναι προτιμότερα από τα κοινωνικά γραφήματα, επειδή μια ενεργή σχέση είναι πιο σχετική με την ανάλυση από μια απλή σύνδεση.

Διάφορες τεχνικές έχουν αναδυθεί πρόσφατα για την εξαγωγή πληροφοριών από τη δομή των κοινωνικών δικτύων. Η κοινοτική ανίχνευση (community detection), αποκαλούμενη επίσης κοινοτική ανακάλυψη, εξάγει σιωπηρές κοινότητες μέσα σε ένα δίκτυο. Για τα επιγραμματικά κοινωνικά δίκτυα, μια κοινότητα αναφέρεται σε ένα υπο-δίκτυο χρηστών που αλληλεπιδρούν περισσότερο εκτενώς μεταξύ τους παρά με το υπόλοιπο δίκτυο. Συχνά περιέχουν εκατομμύρια κόμβους και άκρες, τα διαδικτυακά κοινωνικά δίκτυα τείνουν να είναι κολοσσιαία σε μέγεθος. Η κοινοτική ανίχνευση βοηθά να συνοψίσουμε τα τεράστια δίκτυα, τα οποία στη συνέχεια διευκολύνουν την αποκάλυψη των υφιστάμενων συμπεριφορών και προβλέπουν τις αναδυόμενες ιδιότητες του δικτύου. Από την άποψη αυτή, η ανίχνευση της κοινότητας είναι παρόμοια με την ομαδοποίηση (Aggarwal, 2011), μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται για τη διαίρεση ενός συνόλου δεδομένων σε διαφορετικά υποσύνολα με βάση την ομοιότητα των σημείων δεδομένων. Η κοινοτική ανίχνευση έχει βρει διάφορους τομείς εφαρμογών, συμπεριλαμβανομένου του μάρκετινγκ και του World Wide Web (Parthasarathy, Ruan, & Satuluri, 2011). Για παράδειγμα, η κοινοτική ανίχνευση επιτρέπει στις επιχειρήσεις να αναπτύξουν πιο αποτελεσματικά συστήματα συστάσεων για προϊόντα.

2.5.5 **Predictive analytics**

Οι προγνωστικές αναλύσεις περιλαμβάνουν μια ποικιλία τεχνικών που προβλέπουν μελλοντικά αποτελέσματα με βάση τα ιστορικά και τα τρέχοντα δεδομένα. Στην πράξη, τα προβλέψιμα στοιχεία ανάλυσης μπορούν να εφαρμοστούν σε όλους σχεδόν τους κλάδους - από την πρόβλεψη της αποτυχίας των κινητήρων αεριωθουμένων βάσει της ροής δεδομένων

από αρκετές χιλιάδες αισθητήρες έως την πρόβλεψη των επόμενων κινήσεων των πελατών με βάση αυτό που αγοράζουν, όταν αγοράζουν, λένε στα κοινωνικά μέσα ενημέρωσης.

Στον πυρήνα της, οι προβλέψεις αναλύσεων επιδιώκουν να αποκαλύψουν τα πρότυπα και να συλλάβουν τις σχέσεις στα δεδομένα. Οι τεχνικές προληπτικής ανάλυσης υποδιαιρούνται σε δύο ομάδες. Ορισμένες τεχνικές, όπως οι κινητοί μέσοι όροι, προσπαθούν να ανακαλύψουν τα ιστορικά πρότυπα στην μεταβλητή των αποτελεσμάτων και να τα εξαγάγουν στο μέλλον. Άλλοι, όπως η γραμμική παλινδρόμηση, στοχεύουν να καταγράψουν τις αλληλεξαρτήσεις μεταξύ των μεταβλητών και των επεξηγηματικών μεταβλητών και να τις αξιοποιήσουν για να κάνουν προβλέψεις. Με βάση την υποκείμενη μεθοδολογία, οι τεχνικές μπορούν επίσης να ταξινομηθούν σε δύο ομάδες: τεχνικές παλινδρόμησης (π.χ. μοντέλα πολυνομιών λογιών) και τεχνικές μηχανικής μάθησης (π.χ. νευρωνικά δίκτυα). Μια άλλη ταξινόμηση βασίζεται στον τύπο των μεταβλητών έκβασης: τεχνικές όπως οι μεταβλητές συνεχούς έκβασης γραμμικής παλινδρόμησης (π.χ. τιμή πώλησης σπιτιών), ενώ άλλες, όπως τα τυχαία δάση, εφαρμόζονται σε διακριτές μεταβλητές αποτελεσμάτων (π.χ. κατάσταση πιστοληπτικής ικανότητας).

Οι προγνωστικές τεχνικές ανάλυσης βασίζονται κυρίως σε στατιστικές μεθόδους. Διάφοροι παράγοντες απαιτούν την ανάπτυξη νέων στατιστικών μεθόδων για μεγάλα δεδομένα. Πρώτον, οι συμβατικές στατιστικές μέθοδοι στηρίζονται στη στατιστική σημασία: ένα μικρό δείγμα λαμβάνεται από τον πληθυσμό και το αποτέλεσμα συγκρίνεται με την πιθανότητα να εξεταστεί η σημασία μιας συγκεκριμένης σχέσης. Το συμπέρασμα στη συνέχεια γενικεύεται σε ολόκληρο τον πληθυσμό. Αντίθετα, τα μεγάλα δείγματα δεδομένων είναι τεράστια και αντιπροσωπεύουν την πλειοψηφία, αν όχι ολόκληρο, του πληθυσμού. Ως αποτέλεσμα, η έννοια της στατιστικής σημασίας δεν είναι τόσο σημαντική για τα μεγάλα δεδομένα. Δεύτερον, όσον αφορά την υπολογιστική αποτελεσματικότητα, πολλές συμβατικές μέθοδοι για μικρά δείγματα δεν κλιμακώνονται μέχρι τα μεγάλα δεδομένα. Ο τρίτος παράγοντας αντιστοιχεί στα χαρακτηριστικά που είναι εγγενή στα μεγάλα δεδομένα: η ανομοιογένεια, η συσσώρευση θορύβου, οι ψευδείς συσχετισμοί (Fan, Han, & Liu, 2014). Περιγράφω τα παρακάτω.

• Heterogeneity (Ανομοιογένεια): Μεγάλα δεδομένα λαμβάνονται συχνά από διαφορετικές πηγές και αντιπροσωπεύουν πληροφορίες από διαφορετικούς υπο-πληθυσμούς. Ως αποτέλεσμα, τα μεγάλα δεδομένα είναι εξαιρετικά ετερογενή. Τα δεδομένα υπο-πληθυσμού

σε μικρά δείγματα θεωρούνται υπερβολικά μεγάλα λόγω της ανεπαρκούς συχνότητας τους. Ωστόσο, το μέγεθος των μεγάλων συνόλων δεδομένων δημιουργεί τη μοναδική ευκαιρία να μοντελοποιηθεί η ετερογένεια που προκύπτει από δεδομένα υπο-πληθυσμού, τα οποία θα απαιτούσαν εξελιγμένες στατιστικές τεχνικές.

• Noise accumulation (Συσσώρευση θορύβου): Η εκτίμηση προγνωστικών μοντέλων για μεγάλα δεδομένα συχνά περιλαμβάνει την ταυτόχρονη εκτίμηση διαφόρων παραμέτρων. Το συσσωρευμένο σφάλμα εκτίμησης (ή θόρυβος) για διαφορετικές παραμέτρους θα μπορούσε να κυριαρχεί στα μεγέθη μεταβλητών που έχουν πραγματικές επιπτώσεις στο μοντέλο. Με άλλα λόγια, ορισμένες μεταβλητές με σημαντική επεξηγηματική δύναμη θα μπορούσαν να αγνοηθούν ως αποτέλεσμα της συσσώρευσης θορύβου.

• Spurious correlation (Παράλογη συσχέτιση): Για τα μεγάλα δεδομένα, ο ψευδής συσχετισμός αναφέρεται σε μη συσχετισμένες μεταβλητές που έχουν αποδειχθεί ψευδώς συσχετισμένες λόγω του τεράστιου μεγέθους του συνόλου δεδομένων. Οι Fan και Lv (2008) δείχνουν αυτό το φαινόμενο μέσω ενός παραδείγματος προσομοίωσης, όπου ο συντελεστής συσχέτισης μεταξύ ανεξάρτητων τυχαίων μεταβλητών φαίνεται να αυξάνεται με το μέγεθος του συνόλου δεδομένων. Ως αποτέλεσμα, ορισμένες μεταβλητές που είναι επιστημονικά άσχετες (λόγω της ανεξαρτησίας τους) είναι λανθασμένα αποδεδειγμένα συσχετισμένες λόγω της μεγάλης στασιμότητας.

• Incidental endogeneity (Παρεμπόδιση της ενδογένειας): Μια κοινή παραδοχή στην ανάλυση παλινδρόμησης είναι η υπόθεση εξωγένειας: οι επεξηγηματικές μεταβλητές ή οι προβλεπόμενοι δείκτες είναι ανεξάρτητοι από τον υπολειμματικό όρο. Η εγκυρότητα των περισσότερων στατιστικών μεθόδων που χρησιμοποιούνται στην ανάλυση παλινδρόμησης εξαρτάται από αυτή την παραδοχή. Με άλλα λόγια, η ύπαρξη τυχαίας ενδογένειας (δηλαδή η εξάρτηση του υπολειμματικού όρου από ορισμένους από τους προγνωστικούς παράγοντες) υπονομεύει την εγκυρότητα των στατιστικών μεθόδων που χρησιμοποιούνται για την ανάλυση παλινδρόμησης. Παρόλο που η παραδοχή εξωγένειας συναντάται συνήθως σε μικρά δείγματα, η περιστασιακή ενδογένεια είναι συνήθως παρούσα σε μεγάλα δεδομένα. Αξίζει να σημειωθεί ότι, σε αντίθεση με την πλασματική συσχέτιση, η περιστασιακή ενδογένεια αναφέρεται σε μια πραγματική σχέση μεταξύ των μεταβλητών και του όρου σφάλματος. Η έλλειψη σημασίας της στατιστικής σημασίας, οι προκλήσεις της υπολογιστικής αποτελεσματικότητας και τα μοναδικά χαρακτηριστικά των μεγάλων

δεδομένων που συζητήθηκαν παραπάνω υπογραμμίζουν την ανάγκη ανάπτυξης νέων στατιστικών τεχνικών για την απόκτηση γνώσεων από προγνωστικά μοντέλα.

ΚΕΦΑΛΑΙΟ 3 – Εφαρμογές των μεγάλων δεδομένων

3.1 Εφαρμογές μεγάλων δεδομένων ανά κλάδο

Healthcare

Με την ψηφιοποίηση, το συνδυασμό και την αποτελεσματική χρήση των μεγάλων δεδομένων, οργανώσεις υγειονομικής περίθαλψης που μπορεί να κυμαίνονται από γραφεία ιδιωτών γιατρών και ομάδες πολλαπλών παροχών σε μεγάλα νοσοκομειακά δίκτυα και υπεύθυνες οργανώσεις φροντίδας οι οποίες μπορούν να αποκομίσουν σημαντικά οφέλη (Burghard, 2012). Αυτά μπορεί να είναι, η ανίχνευση ασθενειών σε προγενέστερα στάδια όπου αντιμετωπίζονται ευκολότερα και αποτελεσματικότερα, η διαχείριση της ειδικής ατομικής υγείας και η ανίχνευση της απάτης που σχετίζεται με την υγεία πιο γρήγορα και αποτελεσματικά. Επιπλέον, πολλές ερωτήσεις μπορούν να βρουν απάντηση με την εφαρμογή των big data analytics βάσει τεράστιων ποσοτήτων ιστορικών δεδομένων, όπως η διάρκεια διαμονής στο νοσοκομείο (Length Of Stay), ασθενείς που θα επιλέξουν λεπτές χειρουργικές επεμβάσεις, ασθενείς που πιθανόν δεν θα επωφεληθούν από τη χειρουργική επέμβαση, ασθενείς που διατρέχουν κίνδυνο για ιατρικές επιπλοκές, ασθενείς με κίνδυνο για σήψη, ή άλλες νοσοκομειακές λοιμώξεις, προβλέψεις ασθένειας / πρόοδος της νόσου, ασθενείς που κινδυνεύουν να προχωρήσουν σε ασθένειες, αιτιώδης συνάφεια σε παράγοντες ασθένειας και στην εξέλιξη της νόσου και πιθανές συνυπάρχουσες συνθήκες (EMC Consulting).

Η McKinsey εκτιμά ότι τα big data analytics μπορούν να επιτρέψουν περισσότερα από 300.000.000 $ αποταμιεύσεις ετησίως στην υγειονομική περίθαλψη των ΗΠΑ, με μείωση κατά περίπου 8% των εθνικών δαπανών για την υγειονομική περίθαλψη. Οι κλινικές

λειτουργίες και η έρευνα και ανάπτυξη (R&D) είναι δύο από τις μεγαλύτερες περιοχές για πιθανή εξοικονόμηση πόρων με $ 165 δισεκατομμύρια και 108 δισεκατομμύρια δολάρια σπατάλης αντίστοιχα (Manyika et al, 2011). Επίσης, αναφέρει ότι τα μεγάλα δεδομένα θα μπορούσαν να συμβάλουν στη μείωση των εξόδων και της αναποτελεσματικότητας στους ακόλουθους τομείς:

Κλινικές επεμβάσεις (Clinical Operations): Με συγκριτική αποτελεσματικότητα για τον προσδιορισμό περισσότερων κλινικά σημαντικών και οικονομικών αποδοτικών τρόπων διάγνωσης και θεραπείας των ασθενών.

Έρευνα & ανάπτυξη (**Research & Development**):
1) Μοντέλα πρόβλεψης παράγουν πιο απλά, πιο γρήγορα, και πιο στοχευόμενα μέσα Ε&Α για φάρμακα και συσκευές
2) Στατιστικά εργαλεία και αλγόριθμοι συμβάλλουν στη βελτίωση της κλινικής δοκιμής, και της συμμετοχής ασθενών για καλύτερη αντιστοίχιση θεραπείας σε μεμονωμένους ασθενείς, μειώνοντας έτσι τη φάση δοκιμής-αποτυχιών και επιταχύνοντας τις νέες θεραπείες.
3)Ανάλυση των κλινικών δοκιμών και των αρχείων των ασθενών, για τον εντοπισμό των επόμενων ενδείξεων και ανεπιθύμητων ενεργειών πριν τα προϊόντα φθάσουν στην αγορά.

Δημόσια υγεία (**Public health)**:
1) με την ανάλυση των μορφών μιας νόσου, εντοπίζοντας τις εστίες της και τον τρόπο μετάδοσής της ενισχύεται η βελτίωση της δημόσιας υγείας και η ταχύτητα ανταπόκρισης.
2) ταχύτερη ανάπτυξη εμβολίων με μεγαλύτερη ακρίβεια, π.χ., για την επιλογή των ετήσιων στελεχών της γρίπης .
3) μετατροπή μεγάλων ποσοτήτων δεδομένων σε πληροφορίες που μπορούν να ενεργοποιηθούν και να χρησιμοποιηθεί για τον εντοπισμό αναγκών, την παροχή υπηρεσιών για την πρόληψη κρίσεων, ιδίως προς όφελος του πληθυσμού.

Επιπλέον, (IBM, 2012) προτείνει ότι η ανάλυση μεγάλων δεδομένων μπορεί να οδηγήσει σε:
Α)Περίθαλψη βασισμένη σε γεγονότα (**Evidence based medicine)** : Ο συνδυασμός και η ανάλυση δεδομένων με μεγάλη ποικιλία, πχ δομημένων και αδόμητων δεδομένων,

οικονομικών, κλινικών και γονιδιωματικών δεδομένων τα δεδομένα για να ταιριάζουν με τις με τα αποτελέσματα θεραπειών, να προβλέπουν τους ασθενείς που βρίσκονται σε κίνδυνο για ασθένεια με σκοπό την παροχή περισσότερης αποτελεσματικής φροντίδας.

Β) Γονιδιωματική ανάλυση (**genomic analytics**): Εκτέλεση της αλληλουχίας γονιδίων περισσότερο αποτελεσματικά και οικονομικά αποδοτικά ώστε να αποτελέσει μέρος της τακτικής απόφασης ιατρικής περίθαλψης και του αυξανόμενου ιατρικού αρχείου ασθενών (IBM, 2012).

Γ) Παρακολούθηση συσκευής / απομακρυσμένου ελέγχου: Καταγραφή και ανάλυση σε πραγματικό χρόνο μεγάλου όγκους δεδομένων που μετακινούνται γρήγορα μέσα στο νοσοκομείο και στο σπίτι, για την παρακολούθηση της ασφάλειας και πρόβλεψη δυσμενών γεγονότων.

Δ) Αναλυτικά προφίλ ασθενούς: Εφαρμογή προηγμένων αναλυτικών στοιχείων σε προφίλ ασθενών (π.χ., τμηματοποίηση και προγνωστική μοντελοποίηση) για τον προσδιορισμό των ατόμων που θα επωφεληθούν από προληπτική φροντίδα ή αλλαγές στον τρόπο ζωής, για παράδειγμα, της της ασθενείς που διατρέχουν τον κίνδυνο να αναπτύξουν μια συγκεκριμένη ασθένεια (π.χ. διαβήτης) που θα επωφεληθούν από την προληπτική φροντίδα (IBM, 2012).

Σύμφωνα με (IBM, 2013), περιοχές της οποίες τα ενισχυμένα δεδομένα έχουν τα μεγαλύτερα αποτελέσματα είναι: ο εντοπισμός ασθενών που είναι οι μεγαλύτεροι καταναλωτές των υγειονομικών πόρων ή με τον μεγαλύτερο κίνδυνο για δυσμενείς εκβάσεις, η παροχή πληροφοριών που χρειάζονται τα άτομα ώστε να λαμβάνουν τεκμηριωμένες αποφάσεις και να διαχειρίζονται πιο αποτελεσματικά την υγεία της, καθώς και να υιοθετήσουν και να παρακολουθήσουν πιο εύκολα περισσότερο υγιεινές συμπεριφορές, ο προσδιορισμός θεραπειών, προγραμμάτων και διαδικασιών που δεν έχουν μεγάλο κόστος, μείωση των επανεισδοχών με τον εντοπισμό περιβαλλοντικών παραγόντων ή παραγόντων του τρόπου ζωής που αυξάνουν τον κίνδυνο ή προκαλούν ανεπιθύμητα συμβάντα και προσαρμόζουν ανάλογα τα προγράμματα θεραπείας, διαχείριση της υγείας του πληθυσμού από ανίχνευση τρωτών σημείων της πληθυσμούς των ασθενών κατά τη διάρκεια της εξάρσεις ή καταστροφές της νόσου, και η παροχή κλινικών, οικονομικών και επιχειρησιακών δεδομένα για την ανάλυση των πόρων με αξιοποίηση σε πραγματικό χρόνο (IBM, 2013). Το Ιατρικό Κέντρο Mount Sinai στις Η.Π.Α. χρησιμοποιεί τεχνολογίες της Ayasdi, μιας εταιρείας Big

Data, για να μελετά τις γενετικές αλληλουχίες του Escherichia Coli, συμπεριλαμβανομένων πάνω από 1 εκ. παραλλαγών του DNA και για να διερευνά «γιατί τα βακτηρία στελέχη ανθίστανται των αντιβιοτικών». Επίσης, η HealthVault της Microsoft, που ξεκίνησε το 2007, είναι μια εξαιρετική εφαρμογή των ιατρικών Big Data. Στόχος της είναι να διαχειρίζεται ατομικές πληροφορίες αναφορικά με την υγεία των ασθενών. Η εφαρμογή υποδέχεται τις πληροφορίες για την υγεία μέσω κατάλληλων συσκευών είτε μέσω ανοικτής διασύνδεσης με τη χρήση κατάλληλου λογισμικού (Chen et al., 2014:201).

Business Sector

Ένα επιχειρηματικό μοντέλο περιγράφει το σκεπτικό του τρόπου με τον οποίο μια επιχείρηση δημιουργεί, αποδίδει και συλλαμβάνει την αξία των μεγάλων δεδομένων (Osterwalder and Pinner, 2010, σελ. 14), καταγράφοντας την επιχειρηματική λογική του πυρήνα οποιασδήποτε επιχείρησης. Στην πράξη, η ψηφιοποίηση και οι μεγάλες αναλύσεις δεδομένων προκαλούν νέα επιχειρηματικά μοντέλα σε πολλές παραδοσιακές βιομηχανίες. Όχι μόνο οι καθιερωμένες επιχειρήσεις, συχνά αποτυγχάνουν στο να υιοθετήσουν την κατάλληλη ψηφιοποίηση και τις μεγάλες αναλύσεις δεδομένων, αλλά προσπαθούν επίσης να προσαρμόσουν τα επιχειρηματικά τους μοντέλα ώστε να αντικατοπτρίζουν τα συναφή οικονομικά χαρακτηριστικά και τους μηχανισμούς τους (Weill and Woerner, 2015, Westerman κ.ά., 2014). Η βιβλιογραφία για τα στρατηγικά IS (π.χ., Bresnahan et al., 2002, Gable, 2010, Malone et al, 2003, Orlikowski και Barley, 2001, Picot et al., 2008) επικεντρώνεται στο πώς οι ΤΠΕ συνεχίζουν να αναμορφώνουν τις υπάρχουσες δομές εργασίας και οργάνωσης με νέες μορφές καταμερισμού της εργασίας και συνεργασίας στις επιχειρήσεις. Συχνά επικεντρώνεται στην αυξανόμενη αποσύνδεση των διαδικασιών εργασίας από συγκεκριμένες τοποθεσίες (συμπεριλαμβανομένων κτιρίων εργοστασίου ή γραφείων) ή χρόνων (καθορισμένες ώρες εργασίας) και πώς η τεχνολογία επιτρέπει τη χωροταξική και χρονική ευελιξία των ρυθμίσεων. Με την ψηφιοποίηση και τις μεγάλες αναλύσεις δεδομένων, οι διασταυρούμενες ομάδες εργασίας και οι παραδοσιακές ιεραρχικές δομές εργασίας διαλύονται και μετασχηματίζονται σε όλο και πιο ευέλικτες, εσωτερικές και καθαρές δομές σε διάφορες τοποθεσίες (π.χ., Zammuto et al., 2007). Επιπλέον, ο αντίκτυπος στο σχεδιασμό νέων μορφών εργασίας δεν τελειώνει στα εταιρικά όρια, αλλά προσφέρει νέες ευκαιρίες για την ευέλικτη ενσωμάτωση των εξωτερικών ελεύθερων επαγγελματιών ή των εργαζομένων αφενός και για την οργάνωση και ανάπτυξη της συνεργασίας μεταξύ των

επιχειρήσεων. Ένα παράδειγμα εξελίξεων που βασίζονται στην τεχνολογία της πληροφορικής είναι η προσέλκυση ιδεών, διαδικασιών ή χρηματοοικονομικών κεφαλαίων (Majchrzak και Malhotra, 2013), η οποία προσφέρει μια μορφή ενσωμάτωσης των εξωτερικών πόρων της επιχείρησης πέρα από την παραδοσιακή εξωτερική ανάθεση - αλλά πιθανότατα με λιγότερες εταιρικές ή ακόμα και κοινωνικές δεσμεύσεις. Οι αυξημένες βελτιώσεις των καθιερωμένων επιχειρηματικών μοντέλων με βάση την ψηφιοποίηση και τις μεγάλες αναλύσεις δεδομένων στοχεύουν στη βελτιστοποίηση των υφιστάμενων διαδικασιών για τη βελτίωση της συνολικής αποτελεσματικότητας και της ποιότητας των προϊόντων και των υπηρεσιών. Η αυξανόμενη ψηφιοποίηση μειώνει δραματικά το κόστος συναλλαγών για τη συλλογή πληροφοριών, επικοινωνίας και ελέγχου. Μέσω της εύκολης πρόσβασης σε σχεδόν απεριόριστη συλλογή πληροφοριών και εξελιγμένων μεγάλων αναλυτικών στοιχείων, οι επιχειρήσεις μπορούν για παράδειγμα να αναλύσουν την αλληλεπίδραση της διαδικτυακής και αγοραστικής συμπεριφοράς των χρηστών για να προσαρμόσουν τις διαφημίσεις και έτσι να αυξήσουν τη συνολική ζήτηση. Η Wall Mart, αναφορικά με τα αποθέματα των πωλήσεων τους και τα καιρικά δεδομένα, διαπίστωσε ότι οι πελάτες τους αγόρασαν φράουλα Pop Tarts (ζαχαρούχα σνακ) περισσότερο των φακών και των μπαταριών, όταν πλησίασε ένας τυφώνας. Αυτή η διορατικότητα τους επέτρεψε να προσαρμόσουν τα αποθέματά τους εκ των προτέρων ενός τυφώνα και κατά συνέπεια καλύτερη εξυπηρέτηση των αναγκών των πελατών τους (Hays, 2004). Έτσι, οι αυξημένες βελτιώσεις στα καθιερωμένα επιχειρηματικά μοντέλα μέσω της αυξημένης ψηφιοποίησης και των μεγάλων αναλύσεων δεδομένων ενδέχεται να αντικαταστήσουν μακροπρόθεσμα τα λιγότερο αποτελεσματικά επιχειρηματικά μοντέλα (και συνεπώς τις εταιρείες). Ωστόσο, όπως υποστηρίζουν ο Markus και ο Loebbecke (2013), με ένα αυξανόμενο επίπεδο τυποποίησης και ευρείας υιοθέτησης σε ολόκληρη την οικονομία, οι εμπορευματοποιημένες λύσεις μεγάλων δεδομένων ενδέχεται να μην επαρκούν για διαρκή ανταγωνιστικό πλεονέκτημα - όπως και οι προσπάθειες common digitization όπως για τα ERP συστήματα και παρόμοια τους. Η εικονογράφηση και οι μεγάλες αναλύσεις δεδομένων έχουν ήδη οδηγήσει στη διακοπή των καθιερωμένων επιχειρηματικών μοντέλων (Weill and Woerner, 2015). Ορισμένες καθιερωμένες επιχειρήσεις αγωνίζονται να επιβιώσουν. Δημοφιλή παραδείγματα προέρχονται από τα μέσα ενημέρωσης και τις διαφημιστικές βιομηχανίες. Αρκετοί άλλοι κλάδοι συμπεριλαμβανομένης της εκπαίδευσης, φαίνονται να είναι στην επόμενη γραμμή (π.χ., Shirky, 2012 · Westerman et al., 2014). Οι καινοτόμες νεοσύστατες επιχειρήσεις επωφελούνται από τα χαμηλά εμπόδια στην είσοδο, επιτρέποντας τη διακοπή των επιχειρηματικών μοντέλων των εγκατεστημένων επιχειρήσεων (π.χ. Instagram και

Kodak για τη λήψη, αποθήκευση και ανταλλαγή φωτογραφιών, Lucas and Goh, 2009). Οι νέοι επιχειρηματίες προχωρούν σε υπάρχουσες αγορές ή ασκούν ανεξερεύνητες επιχειρηματικές ευκαιρίες με νέα επιχειρηματικά μοντέλα βασιζόμενα στην εκμετάλλευση ψηφιακών καναλιών διανομής, δημιουργώντας και εξυπηρετώντας τη νέα ζήτηση των πελατών, δημιουργώντας νέες μορφές εμπλοκής και σχέσεων με τους πελάτες ή οποιονδήποτε συνδυασμό των τριών (Lucas et al., 2013) . Οι μεγάλες αναλύσεις δεδομένων επιτρέπουν ειδικότερα τη συμπλήρωση ή ακόμα και την υποκατάσταση της εργασίας για μηχανές στο πλαίσιο των διαχειριστικών και προσανατολισμένων προς την απόφαση συμβολών στη δημιουργία αξίας. Η εφαρμογή των Big Data στις επιχειρήσεις πλέον επιτρέπει την ευκολότερη απογραφή, την υλικοτεχνική βελτιστοποίηση και τη συνεργασία με τους προμηθευτές για τον περιορισμό του χάσματος μεταξύ προσφοράς και ζήτησης ενώ και η χρηματοδότηση επιχειρήσεων που στρέφονται στα Big Data έχει αποκτήσει άλλη υπόσταση. Αν υπάρχει ένα στοιχείο όπου οι εταιρείες social media και τα online κοινωνικά δίκτυα εξειδικεύονται είναι τα δεδομένα. Ο μεγάλος όγκος δεδομένων στα social media αντανακλά το πώς οι άνθρωποι αλληλεπιδρούν μεταξύ τους και στο επίκεντρο αυτών των αλληλεπιδράσεων βρίσκονται πολύτιμες πληροφορίες. Πολλές εταιρείες εκτιμούν την ισχυρή φύση των online κοινωνικών δικτύων για αλληλεπίδραση σε προσωπικό επίπεδο με τους πελάτες τους. Το Facebook, το Twitter, το Instagram και το Pinterest προσαρμόζονται στις δυνατότητες που μπορούν να προσφέρουν τα Big Data εξατομικεύοντας κόμη περισσότερο τις δυνατότητές τους προς τους χρήστες. Σε μελέτες πλέον έχει αποδειχθεί ότι το Facebook λόγω του μεγάλου όγκου των δεδομένων που διαχειρίζεται είναι ως ένα βαθμό ισχυρότερο εργαλείο αποτύπωσης της συμπεριφοράς και της προσωπικότητας ενός χρήστη ακόμα και από ανθρώπους του στενού περιβάλλοντος.

Government Sector

Αν και ο επιχειρηματικός τομέας είναι που οδηγεί στην ανάπτυξη μεγάλων δεδομένων εφαρμογών, ο δημόσιος τομέας έχει αρχίσει να αποκομίζει επίσης οφέλη για την υποστήριξη των αποφάσεων σε πραγματικό χρόνο από τα ταχέως αναπτυσσόμενα δεδομένα σε κίνηση από πολλαπλές πηγές, συμπεριλαμβανομένου του ιντερνέτ, βιολογικών και των βιομηχανικών αισθητήρων, βίντεο, ηλεκτρονικό ταχυδρομείο και κοινωνικές επικοινωνίες (Broekema, 2012). Πολλά white papers, άρθρα περιοδικών και επιχειρησιακές εκθέσεις έχουν προτείνει τρόπους με τους οποίους οι κυβερνήσεις μπορούν να χρησιμοποιήσουν

μεγάλα δεδομένα για να εξυπηρετούν τους πολίτες τους και να ξεπερνούν εθνικές προκλήσεις (όπως η άνοδος δαπανών για την υγειονομική περίθαλψη, η δημιουργία θέσεων εργασίας κλπ) (McKinsey, 2011). Τα big data analytics αντιμετωπίζονται ακόμα με σκεπτικισμό ως προς το αν μπορεί στην πραγματικότητα να βελτιώσει τις κυβερνητικές διαδικασίες καθώς οι κυβερνήσεις πρέπει να αναπτύξουν νέες δυνατότητες και να υιοθετήσουν νέες τεχνολογίες (όπως οι Hadoop και NoSQL) για να μετατρέψει σε πληροφορίες τα δεδομένα μέσω της οργάνωσης δεδομένων και των analytics (Chen et al, 2012). Συγκρίνοντας τις εφαρμογές μεγάλων δεδομένων των κορυφαίων χωρών που ηγούνται στην ηλεκτρονική διακυβέρνηση μπορούν να μας κατατοπίσουν που επικεντρώνονται οι τρέχουσες και μελλοντικές εφαρμογές και χρησιμεύουν ως οδηγοί για τις χώρες που ακολουθούν και θέλουν να ξεκινήσουν τις δικές τους εφαρμογές μεγάλων δεδομένων. Τέτοια παραδείγματα χωρών και εφαρμογών είναι:

ΗΠΑ

Για να διαχειριστεί την ανάλυση σε πραγματικό χρόνο των δεδομένων ροής μεγάλου όγκου, η κυβέρνηση των Η.Π.Α. και η IBM συνεργάστηκαν το 2002 για να αναπτύξουν μια μαζικά κλιμακούμενη, συγκεντρωμένη υποδομή (Accenture, 2011). Το αποτέλεσμα είναι το IBM InfoSphere Stream και το IBM Big Data, που είναι ευρέως χρησιμοποιούμενα από κυβερνητικές υπηρεσίες και επιχειρηματικούς οργανισμούς, είναι πλατφόρμες για ανακάλυψη και εικονογράφηση των πληροφοριών από χιλιάδες πηγές σε πραγματικό χρόνο, που περιλαμβάνουν την ανάπτυξη εφαρμογών και τη διαχείριση συστημάτων Hadoop και αποθήκευση δεδομένων. Το 2009, η κυβέρνηση των ΗΠΑ ξεκίνησε το Data.gov ως ένα βήμα προς την κατεύθυνση της διαφάνειας και της λογοδοσίας της κυβέρνησης. Πρόκειται για μια βάση που περιέχει 420.894 σύνολα δεδομένων (από τον Αύγουστο του 2012) που καλύπτουν τις μεταφορές, την οικονομία, την υγειονομική περίθαλψη, την εκπαίδευση και τις ανθρώπινες υπηρεσίες και αποτελούν πηγή δεδομένων για πολλαπλές εφαρμογές: 1,279 από κυβερνήσεις, 236 από πολίτες, και 103 από κινητές συσκευές. Το 2012, η κυβέρνηση Ομπάμα ανακοίνωσε το Big Data Research (OSTP, 2012) μια επένδυση αξίας 200 εκατομμυρίων δολαρίων, με ομοσπονδιακά τμήματα και υπηρεσίες, του οποίου οι κύριοι στόχοι ήταν η υιοθέτηση και ανάπτυξη big data τεχνολογιών, η επιτάχυνση της ανακάλυψης μέσα από την επιστήμη και τη μηχανική, η ενίσχυση της εθνικής ασφάλειας, ο μετασχηματισμός της διδασκαλίας που απαιτείται για την ανάπτυξη και τη χρήση τεχνολογιών Big Data (OSTP, 2012).

Ευρωπαϊκή Ένωση

Το 2010, η Ευρωπαϊκή Επιτροπή ξεκίνησε το "Ψηφιακό θεματολόγιο για την Ευρώπη" για να οριοθετήσει τον τρόπο για την επίτευξη βιώσιμων οικονομικών και με κοινωνικά οφέλη λύσεων για τους πολίτες της ΕΕ μέσω γρήγορων και λειτουργικών εφαρμογών Internet (EC, 2010). Το 2012, στο "Ψηφιακό θεματολόγιο για την Ευρώπη και τις προκλήσεις για το 2012" η Ευρωπαϊκή Επιτροπή δημιούργησε τη στρατηγική για τα big data, δίνοντας έμφαση στο οικονομικό δυναμικό των δημόσιων δεδομένων, διασφαλίζοντας την προστασία των δεδομένων και αυξάνοντας την εμπιστοσύνη των πολιτών, την ανάπτυξη του διαδικτύου των πραγμάτων, την επικοινωνία μεταξύ συσκευών χωρίς άμεση ανθρώπινη παρέμβαση και την εξασφάλιση της ασφάλειας στο Διαδίκτυο και την ασφαλή αντιμετώπιση του δεδομένων και ηλεκτρονικών ανταλλαγών (EC, 2010).

UK

Η κυβέρνηση της Βρετανίας ήταν μία από τις πρώτες χώρες της ΕΕ που υλοποιούν προγράμματα μεγάλων δεδομένων, δημιουργώντας το Κέντρο Σάρωσης του Ηνωμένου Βασιλείου (HSC) το 2004 για να βελτιώσει την ικανότητα της κυβέρνησης να ασχοληθεί με πολύ-επιστημονικές προκλήσεις (Sherry, 2012). Το 2011, οι προσπάθειες της Διεθνούς Διάστασης της Κλιματικής Αλλαγής για την Προοπτική Διερεύνηση της ΚΤΠ αντιμετώπισαν την κλιματική αλλαγή και την επίδρασή της στη διαθεσιμότητα τροφίμων και ύδατος, οι περιφερειακές εντάσεις και η διεθνής σταθερότητα και ασφάλειας με την εκτέλεση σε βάθος ανάλυσης σε πολλαπλά κανάλια δεδομένων. Μια άλλη κυβερνητική πρωτοβουλία της Μεγάλης Βρετανίας ήταν η δημιουργία της δημόσιας ηλεκτρονικής διεύθυνσης http://data.gov.uk το 2009, ανοίγοντας στο κοινό περισσότερα από 1.000 υπάρχοντα σύνολα δεδομένων από επτά κυβερνητικά τμήματα αρχικά, αργότερα αυξήθηκε σε 8.633 σύνολα δεδομένων.

Αποτελεσματικές υπηρεσίες στον κυβερνητικό τομέα

Καθώς οι κυβερνήσεις γίνονται πιο αποτελεσματικές, ο χρόνος και η προσπάθεια που απαιτούνται για την ολοκλήρωση μιας εργασίας έχουν μειωθεί. Εκτός από της κυβέρνησης, οι εταιρείες που βασίζονται σε δεδομένα υπερέβαιναν τον ανταγωνισμό τους και ήταν, κατά μέσο όρο, 5% περισσότερο παραγωγικές και 6% πιο κερδοφόρες. Εάν οι σχετικές εσωτερικές οντότητες μπορούν να συγκεντρώσουν και να αναλύσουν κατάλληλα τα μεγάλα δεδομένα,

θα μειωθεί ο χρόνος που απαιτείται για την παραγωγή αναφορών και την εκτέλεση πρόσθετων και πιο συγκεκριμένων ειδών αναλύσεων. Επιπλέον, η προσπάθεια που απαιτείται για την επεξεργασία θα πρέπει επίσης να μειωθούν με κατανάλωση και ανάλυση μεγάλων δεδομένων. Η χρήση των μεγάλων δεδομένων για την αύξηση της αποτελεσματικότητας της κυβέρνησης με αυτοματοποίηση και επανασχεδιασμό των διαδικασιών ανάλυσης δεδομένων προτείνεται κατά κόρον, καθώς μπορεί να αυξήσει την αποτελεσματικότητα μιας κυβέρνησης μέσω της κατάτμησης των δεδομένων και της διαφάνειας των πληροφοριών (βλέπε σχήμα).

Figure 1. A model for leveraging big data to improve e-government services, ultimately resulting in transformational government. Automating data analysis and redesigning processes can improve efficiency, while segmenting the data and making information more transparent can improve effectiveness.

Αυτοματοποίηση (Automation)

Η αυτοματοποίηση είναι ένας από τους ακρογωνιαίους λίθους της υλοποίησης των ΤΠΕ στις επιχειρήσεις και στο δημόσιο τομέα. Το δημόσιο μπορεί να χρησιμοποιήσει τα μεγάλα δεδομένα και να στηρίξει την ανάλυσή της με τη στόχευση των σημείων συμφόρησης. Για παράδειγμα, ένα συσσωρευμένο χαρτοφυλάκιο αποκαλύπτει μια ευκαιρία για αυτοματοποίηση δεδομένων. Με την υποβολή σε ηλεκτρονική καταχώρηση και όχι σε έντυπη, ορισμένα στοιχεία θα μπορούν να υποστούν αυτόματη επεξεργασία, μειώνοντας έτσι το χρόνο που απαιτείται για την επεξεργασία μιας αξίωσης.

Επανασχεδιασμός (Redesign)

Μέσω της χρήσης μεγάλων αναλυτικών στοιχείων όπως γενετικοί αλγόριθμοι, ανάλυση παλινδρόμησης και τα εργαλεία της συναισθηματικής ανάλυσης, οι διαδικασίες μπορούν να επανασχεδιαστούν. Η ανάλυση παλινδρόμησης εξετάζει τις σχέσεις μεταξύ σαφώς καθορισμένων μεταβλητών και η συναισθηματική ανάλυση επιδιώκει να εξαγάγει την πολικότητα - θετικές ή αρνητικές - από δηλώσεις γνώμης. Αυτοί οι τύποι αναλυτικών εργαλείων μπορούν επίσης να βελτιώσουν την παροχή υπηρεσιών, βοηθώντας τους υπαλλήλους να κατανοήσουν καλύτερα τις ανάγκες των πελατών τους. Για παράδειγμα, η υπηρεσία εσωτερικών εσόδων (IRS) έχει ανασχεδιασμένες διαδικασίες φορολογικής κατάθεσης και χρησιμοποιούν αναλύσεις μεγάλων δεδομένων για τη βελτίωση της ανίχνευσης απάτης.

Κατάτμηση (Segmentation)

Η τμηματοποίηση αποκαλύπτει συγκεκριμένα σύνολα δεδομένων ή ομάδων. Αποτελεί μια κοινή έννοια στο μάρκετινγκ, η οποία χρησιμοποιείται συχνά για τη δημιουργία ομάδων βάση δημογραφικών ή γεωγραφικών περιοχών. Μπορούν να αποκαλυφθούν μέσω της κατάτμησης μεγάλων δεδομένων στον δημόσιο τομέα συστοιχίες που δεν είναι διαισθητικές ή εύκολα ορατές με μια σύντομη εξέταση των δεδομένων. Μεγάλη ανάλυση δεδομένων μπορεί να βοηθήσει τους κυβερνητικούς υπαλλήλους να διαβάσουν το δεδομένα από πολλαπλές προοπτικές για να αποκαλύψουν νέες πληροφορίες. Η αυτοματοποίηση της ανάλυσης δεδομένων και ο επανασχεδιασμός των διαδικασιών μπορούν να βελτιώσουν την αποτελεσματικότητα, ενώ η κατάτμηση των δεδομένων και η δημιουργία διαφανών πληροφοριών μπορούν να βελτιώσουν την αποτελεσματικότητα στο δημόσιο τομέα.

Διαφάνεια (Transparency)

Εργαλεία για ανάλυση μεγάλων δεδομένων μπορούν εύκολα να υποστηρίξουν την υποβολή εκθέσεων σε μεγάλες ποσότητες δεδομένων, κάνοντας τες έτσι διαθέσιμες στο κοινό. Για παράδειγμα, η ανάπτυξη των κοινωνικών μέσων ενημέρωσης στον τομέα της ηλεκτρονικής διακυβέρνησης έχει ήδη αυξήσει τη διαφάνεια και τη μείωση της διαφθοράς σε ορισμένες περιοχές. Δεδομένου ότι η ανάλυση μεγάλων δεδομένων αυξάνεται με την κυβέρνηση, η λήψη αποφάσεων θα οδηγηθεί περισσότερο από τα δεδομένα και λιγότερο από εικασίες, αυξάνοντας την αποτελεσματικότητα της δημόσιας διοίκησης. Η έννοια της ανοικτής διακυβέρνησης απαιτεί την απελευθέρωση περισσότερων πληροφοριών στο δημόσιο.

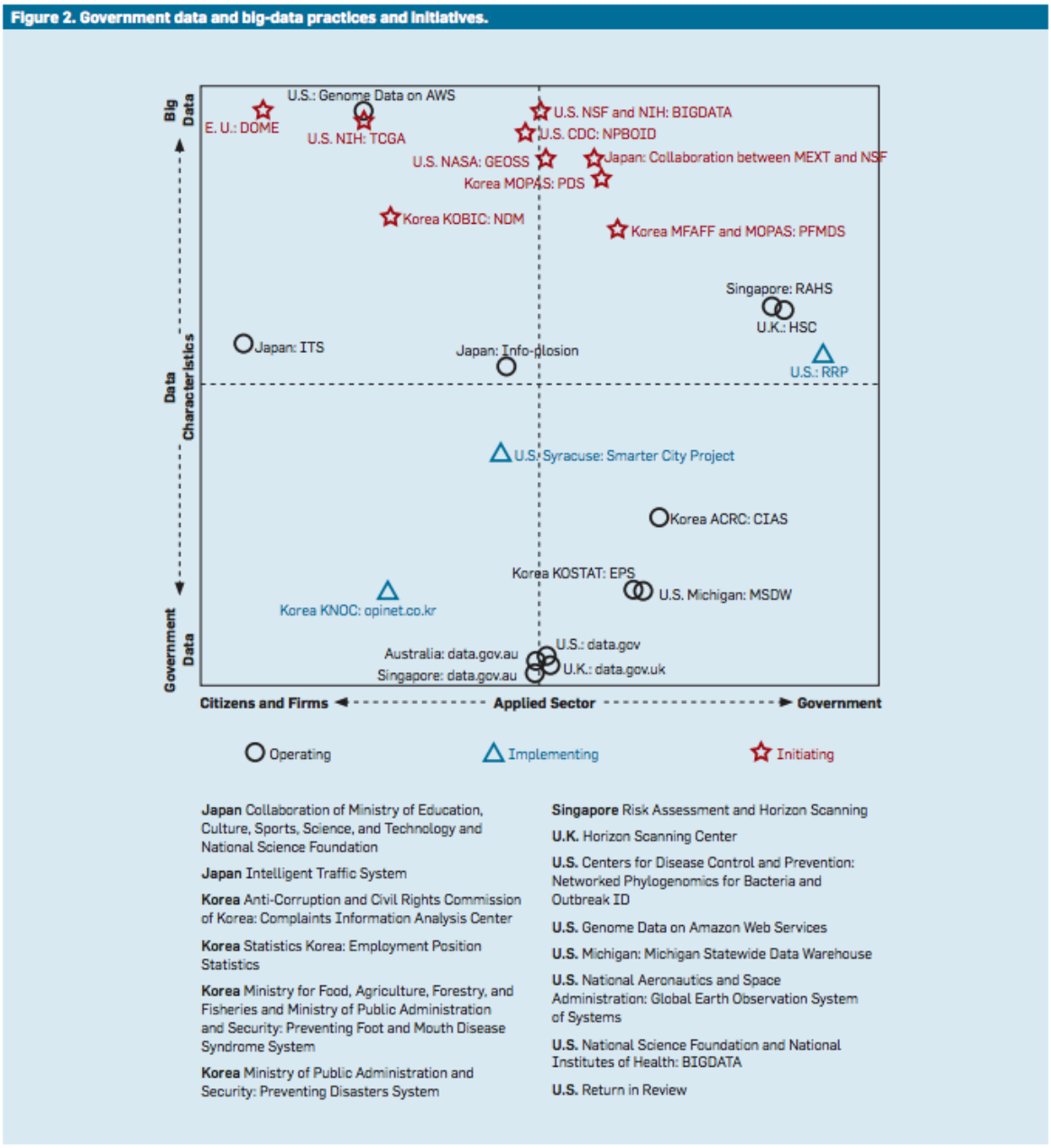


Figure 2 Κυβερνητικά δεδομένα και πρωτοβουλίες μεγάλων δεδομένων ανά χώρα

Με την ανασκόπηση έργων και πρωτοβουλιών μεγάλης κλίμακας δεδομένων σε κορυφαίες χώρες μπορεί κανείς να προσδιορίσει τρεις αξιοσημείωτες μεγάλες τάσεις. Πρώτον, τα περισσότερα έργα που λειτουργούν ή εφαρμόζονται σήμερα μπορεί μόνο οριακά να ταξινομηθούν ως εφαρμογές Big Data, όπως περιγράφονται στο το αριστερό τεταρτημόριο της πάνω εικόνας. Η πλειοψηφία των κυβερνητικών δεδομένων στις χώρες αυτές φαίνεται να μοιράζονται δομημένες βάσεις για την αποθήκευση των δεδομένων και δεν χρησιμοποιούν σε πραγματικό χρόνο. Δεύτερον, τα μεγάλα και σύνθετα σύνολα δεδομένων αναμένεται να είναι ο κανόνας οργανισμούς του δημόσιου τομέα. Οι κυβερνήσεις αναμένουν από τα big data να ενισχύσουν την ικανότητά τους να εξυπηρετούν τους πολίτες και να αντιμετωπίσουν

σημαντικές εθνικές προκλήσεις που αφορούν την οικονομία, την υγειονομική περίθαλψη, την εργασιακή δημιουργία, τις φυσικές καταστροφές και την τρομοκρατία. Ωστόσο, η πλειοψηφία των εφαρμογών Big Data αφορούν στον πολίτη (συμμετοχή σε δημόσιες υποθέσεις) και επιχειρηματικούς τομείς, παρά το πλαίσιο του κυβερνητικού τομέα. Και τρίτον, οι περισσότερες πρωτοβουλίες μεγάλων δεδομένων στον κυβερνητικό τομέα, ιδίως στις Η.Π.Α. (π.χ. το Εθνικό Ίδρυμα Επιστημών και τα Εθνικά Ινστιτούτα Υγείας Πρόγραμμα μεγάλων δεδομένων) είναι σε πρώιμη εξέλιξη ή προγραμματίζονται για μελλοντική εκτέλεση. Αυτό σημαίνει ότι τα προγράμματα εφαρμογής μεγάλων δεδομένων στον κυβερνητικό τομέα βρίσκονται ακόμα σε πρώιμο στάδιο ανάπτυξης, με λίγα λειτουργικά έργα (όπως π.χ. Το RRP των ΗΠΑ, το RAHS της Σιγκαπούρης και το Ηνωμένο Βασίλειο HSC).

Military Intelligence

Η νοημοσύνη βρίσκεται στο επίκεντρο του σχεδιασμού και της εφαρμογής του αμυντικού συστήματος. Δεδομένου ότι τα δεδομένα είναι τώρα διαθέσιμα από κάθε δυνατή γωνία, μια αλλαγή στην προσέγγιση είναι πρωταρχική, ώστε όλες αυτές οι πληροφορίες να μπορούν να αξιοποιηθούν με τον καλύτερο δυνατό τρόπο. Ένα βιβλίο από το Κέντρο για τις Μελέτες Χωροταξίας (CLAWS) παρουσιάζει τις διάφορες δυνατότητες συλλογής πληροφοριών από πηγές τόσο ευρύτερες όσο κοινωνικά μέσα, ανάλυση ιστότοπων αντιπάλων-χωρών, αεροσκάφη και δορυφόρους. Στη σύστασή της για την εφαρμογή της δομής, το έγγραφο προσδιορίζει την πολιτιστική αναθεώρηση Top-Down των Υπηρεσιών Πληροφοριών Άμυνας. Όσον αφορά τα εργαλεία, ο συντάκτης συνιστά την αρχική χρήση των εμπορικών προϊόντων που πρέπει να παρακολουθούνται με εσωτερική Ε & Α. Η ανάπτυξη των δεξιοτήτων πρέπει να επιταχυνθεί και στο εσωτερικό. Το έγγραφο καταλήγει στο συμπέρασμα ότι ένα τέτοιο σύστημα θα έχει την ικανότητα να ανταποκρίνεται και να προσαρμόζεται στην μεταβαλλόμενη φύση των απειλών. Με τα ολοκληρωμένα συστήματα εντολών, ελέγχου, επικοινωνιών, υπολογιστών, πληροφοριών και πληροφοριών (C4I2), δεν υπάρχει έλλειψη δεδομένων. Η ανάγκη της ώρας είναι μια στρατηγική μεγάλων δεδομένων καθώς και αναγνώριση της κρίσιμης σημασίας της από την κορυφή της ιεραρχίας κάτω από τις τάξεις. Ο Αρχηγός του Ινδικού Στρατού έχει επίσης επικεντρωθεί στην ανάγκη για συνεργασίες με τη βιομηχανία. Έχει επίσης αναφερθεί ότι ο Ινδός Στρατός έχει ξεκινήσει το ταξίδι των αναβαθμίσεων του συστήματος προς την κατεύθυνση της ανάπτυξης μιας αρχιτεκτονικής Big Data. Ο Ευρωπαϊκός Οργανισμός Άμυνας έχει διατυπώσει ορισμένες

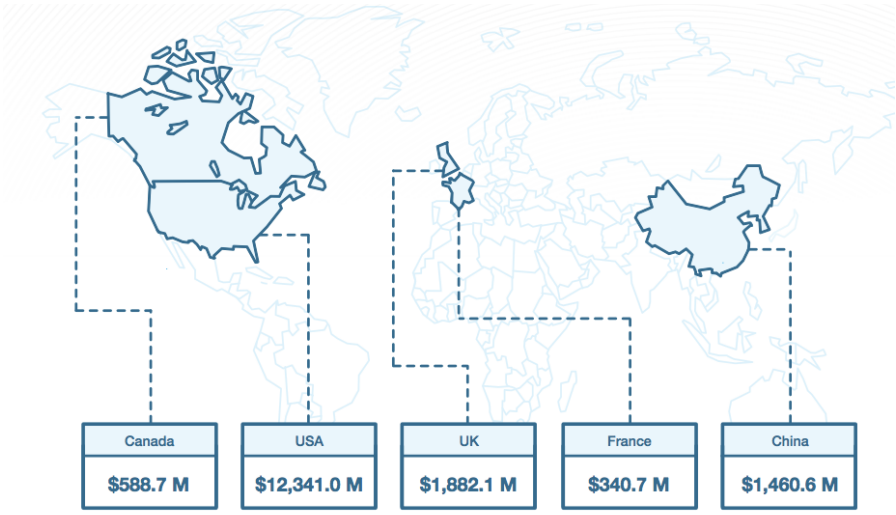
συστάσεις σύμφωνα με μελέτη του 2016. Περιλαμβάνουν την εφαρμογή εφαρμογών Μοντελοποίησης και Προσομοίωσης (M & S) μέσω του Cloud και τη χρήση δεδομένων πρόβλεψης για την ανάλυση των μοντέλων M & S. Τα αναλώσιμα είναι το πιο σημαντικό στοιχείο όταν ένας στρατός λειτουργεί στο πεδίο της μάχης είτε σε στρατηγική θέση είτε σε επιθετικό κίνημα. Σε καταστάσεις όπως αυτές που αντιμετωπίζει η Ινδία στο Doklam, ο προγραμματισμός αποκτά πολλαπλές διαστάσεις. Απαιτεί γρήγορες απαντήσεις για να διατηρούνται ακριβείς αμυντικές θέσεις που παρέχονται επαρκώς. Εάν προκύψει ξαφνική κατάσταση, τα αναλώσιμα πρέπει να είναι άμεσα έτοιμα. Ένα άρθρο της Deloitte επεκτείνεται σε μια διαδικασία που βασίζεται σε δεδομένα και ονομάζεται "Mission Analytics". Ενώ είδαμε ότι η Mission Analytics παρέχει τη δυνατότητα να παρουσιάσει μια ανάλυση SWOT πάνω από πολλές λειτουργικές μεταλλάξεις, οι αποκτήσεις άμυνας στις Η.Π.Α. από το Υπουργείο Άμυνας φαίνεται να υστερούν αρκετά στη χρήση της τεχνολογίας. Αυτό αναφέρθηκε σε μια έκθεση της Υπηρεσίας Έρευνας του Κονγκρέσσο (CRS) με τίτλο: «Χρήση δεδομένων για τη βελτίωση των αμυντικών αποκτήσεων». Η Έκθεση αποκαλύπτει ότι «οι διαδικασίες που βασίζονται στο χαρτί» χρησιμοποιούνται για την παρακολούθηση προγραμμάτων που ξεπερνούν τα δισεκατομμύρια. Δεδομένης της φύσης του Τμήματος, αναμένεται μόνο ότι υπάρχει μεγάλης κλίμακας παραβίαση, όπως αποκαλύφθηκε το 2016. Η Έκθεση επισημαίνει τις ρίζες του θέματος, όπως τα πολλαπλά σημεία αποθήκευσης δεδομένων χωρίς σαφήνεια σε ποιο σύνολο δεδομένων είναι αυθεντικό , ελλιπή και ασυνεπή δεδομένα σχετικά με το κόστος λειτουργίας και υποστήριξης που οδηγούν στην αδυναμία ανάλυσης της αύξησης αυτών των δαπανών. Σύμφωνα με την Έκθεση, το ένστικτο για την προστασία των πληροφοριών οδηγεί σε περιπτώσεις απροθυμίας στην ανταλλαγή δεδομένων, ακόμη και όταν δεν υπάρχει απαίτηση απορρήτου. Ταυτόχρονα, επισημαίνεται επίσης ότι η συλλογή δεδομένων πρέπει να ενημερώνεται από την ακριβή ανάγκη για δεδομένα όσον αφορά τις αποφάσεις που θα επέτρεπε να αποφευχθεί η συγκέντρωση όλων των δεδομένων. Δεδομένου του ποσοστού αστικοποίησης σε όλες τις περιοχές του κόσμου και της επακόλουθης διείσδυσης των smartphone, η ανάλυση των τοποθεσιών των smartphone καθώς και η αξιοποίησή τους στο πεδίο της μάχης εγείρονται σε αυτό το άρθρο από έναν διοικητή πυλών του Αυστραλιανού Στρατού. Όταν εξετάζουμε το αμυντικό περιβάλλον για μια χώρα όπως η Ινδία, είναι πολύ απαραίτητο όχι μόνο να εξετάσουμε όλα τα δομημένα δεδομένα, αλλά αυτό που είναι πιο σημαντικό είναι να αξιοποιήσουμε πλήρως κάθε είδους μη δομημένα δεδομένα εκεί έξω. Σε ένα έθνος ενός δισεκατομμυρίου συν ανθρώπων, τα αναλυτικά εργαλεία Big Data παρέχουν το κρίσιμο πλεονέκτημα για να παραμείνουν στην κορυφή του παιχνιδιού της Άμυνας. Με μια λεπτομερή στρατηγική για

μακροπρόθεσμα, η Ινδία μπορεί να φέρει το βάθος της στην τεχνολογία της πληροφορικής για να αναπτύξει μια λύση για τις προκλήσεις του 21ου αιώνα, που είναι πολύ πιο προηγμένη από οποιαδήποτε άλλη χώρα στη γειτονιά καθώς και σε όλο τον κόσμο.

3.2 **Εφαρμογές μεγάλων δεδομένων ανά αγορές χωρών**

Είναι γεγονός πως δεν υπάρχουν αρκετές διαθέσιμες δωρεάν πληροφορίες για τις εφαρμογές μεγάλων δεδομένων σε αγορές χωρών εξαιτίας των ιδιωτικών φορέων που αυτές αφορούν. Ωστόσο, μια επαρκή εικόνα της γενικής κατάστασης ανά τον κόσμο προσφέρει η έρευνα της OnAudience.com. Η OnAudience.com (https://www.onaudience.com) παρέχει εργαλεία και υπηρεσίες Big Data για ψηφιακό μάρκετινγκ και μετατρέπει με επιτυχία τα Big Data σε έσοδα για διαφημιζόμενους και εκδότες, προσφέροντας υπηρεσίες και προϊόντα που εμπλουτίζουν αποτελεσματικά και δημιουργούν έσοδα από δεδομένα. Η εταιρεία αυτή παρέχει ένα από τα μεγαλύτερα σύνολα δεδομένων που αποτελούνται από πάνω από 27 δις ανώνυμα προφίλ χρηστών από περισσότερες από 200 αγορές σε παγκόσμιο επίπεδο, τα οποία χρησιμοποιούνται κυρίως για τη στόχευση των κατάλληλων κοινών σε προγραμματικές εκστρατείες. Τα δεδομένα που συλλέγονται και επεξεργάζονται από την OnAudience.com δίνουν τη δυνατότητα σε εμπόρους και επιχειρήσεις να εκτελούν εξατομικευμένες διαδικτυακές καμπάνιες και να αναπτύσσουν λύσεις Business Intelligence (Data Enrichment). Είναι μέλος του Cloud Technologies Group, το οποίο ειδικεύεται στο Big Data Marketing και παρέχει λύσεις για τη δημιουργία εσόδων από δεδομένα. Η εταιρεία διαθέτει μοναδικές ικανότητες βελτιστοποίησης των online καμπανιών που βασίζονται στην αυτοματοποιημένη αγορά μέσων (προγραμματική αγορά, προσφορά σε πραγματικό χρόνο). Η Cloud Technologies είναι μία από τις ταχύτερα αναπτυσσόμενες εταιρείες τεχνολογίας στην Ευρώπη, σύμφωνα με την Deloitte Technology Fast 500 EMEA, Deloitte Technology Fast 50 CE και Financial Times 1000. Από τις 18 Μαΐου 2012, η Cloud Technologies S.A. έχει εισαχθεί στην αγορά New Connect που λειτουργεί από το Χρηματιστήριο της Βαρσοβίας. Στην έκθεση Global Size Market της On Audience.com (OnAudience, 2019) αναλύονται τα στοιχεία από τις σημαντικότερες αγορές του κόσμου που δημιουργούν περίπου το 90% των συνολικών διαφημιστικών δαπανών για διαφημίσεις. Πραγματοποιήθηκαν εκτιμήσεις για την αξία των αγορών για τις οποίες υπήρχε διάθεση ανεξάρτητων στοιχείων σχετικά με τον κλάδο της διαδικτυακής διαφήμισης. Χρησιμοποιήθηκαν επίσης και δεδομένα του οργανισμού, με το στατιστικό τους μοντέλο να

περιλαμβάνει τις ψηφιακές διαφημιστικές δαπάνες, το προγραμματικό μερίδιο αγοράς και τη χρήση των δεδομένων. Λαμβάνει επίσης υπόψη το δυναμική των συγκεκριμένων αγορών και περιλαμβάνει προβλέψεις τιμών για τις μελλοντικές περιόδους. Η αγορά δεδομένων αναπτύσσεται ταχύτατα και το 2019 η αξία της αναμένεται να φτάσει τα 26 εκατομμύρια δολάρια παγκοσμίως. Από το 2016 η αγορά δεδομένων διευρύνεται με διψήφιο ρυθμό και η ταχεία ανάπτυξη συνδέεται στενά με τη δυναμική ανάπτυξη της ψηφιακής διαφήμισης και την ψηφιοποίηση των εταιρειών που προχωρούν πολύ γρήγορα. Η αξία της αγοράς διαφημίσεων ψηφιακής προβολής θα φτάσει τα 120,8 εκατομμύρια δολάρια παγκοσμίως αυτό το έτος σύμφωνα με τη μελέτη OnAudience.com. Οι εκτιμήσεις της Zenith είναι παρόμοιες - στις «Προβλέψεις για τις δαπάνες για τη διαφήμιση» μπορούμε να βρούμε ότι η αξία της ψηφιακής προβολής θα φτάσει τα 112,2 εκατομμύρια δολάρια. Η προγραμματική (programmatic) αγορά αναπτύσσεται επίσης ταχέως και ότι το μοντέλο της διαφήμισης χρειάζεται ιδιαίτερα δεδομένα. Σύμφωνα με στην έκθεση της OnAudience.com, οι έμποροι θα δαπανήσουν $ 75 εκατ. για προγραμματικές διαφημίσεις το 2018. Επιπλέον, Η μελέτη eMarketer αποκάλυψε ότι στην αμερικανική αγορά, η οποία είναι η μεγαλύτερη προγραμματική αγορά στον κόσμο με αξία άνω των $ 39 εκατ., οι έμποροι θα δαπανήσουν το 81,5% των προϋπολογισμών ψηφιακής προβολής τους σε προγραμματικές καμπάνιες το 2019. Και οι δύο αγορές προβολής και προγραμματισμού είναι η βενζίνη που οδηγεί την αγορά δεδομένων. Παρά την εισαγωγή του GDPR, το οποίο συζητήθηκε ευρέως στην ψηφιακή βιομηχανία, η αγορά δεδομένων παραμένει ταχύτατα αναπτυσσόμενη. Οι πληροφορίες σχετικά με τους πελάτες αποδεικνύονται καλές για τους εμπόρους, επειδή βοηθούν να προετοιμάσει εξαιρετικά εξατομικευμένα μηνύματα, να βρει τη σωστή ομάδα στόχου και να της στείλει ακριβή μηνύματα τη σωστή στιγμή του ταξιδιού αγοράς τους, αυξάνοντας σημαντικά την αποτελεσματικότητα των διαδικτυακών καμπανιών. Σε όλες σχεδόν τις χώρες που αναλύονται παρακάτω, τα δεδομένα που δαπανώνται το 2019 θα αυξηθούν με διψήφιο ρυθμό, με την τάση τα δεδομένα να είναι σταθερά και έχουν διεθνή εμβέλεια.



TOP 5 world's largest data markets
2017–2019

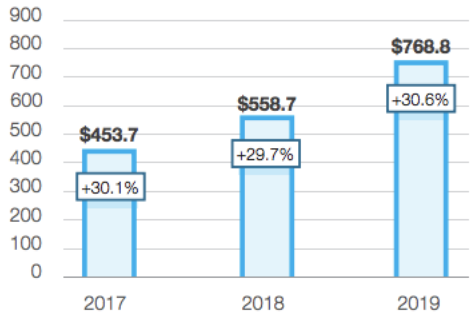
Country	2017	% Change	2018	% Change	2019	% Change
USA	$9,782.3	34.9%	$12,341.0	26.2%	$15,209.0	23.2%
UK	$1,452.4	22.3%	$1,882.1	29.6%	$2,354.9	25.1%
China	$747.2	127.2%	$1,460.6	95.5%	$2,392.6	63.8%
Canada	$453.7	30.1%	$588.7	29.7%	$768.8	30.6%
France	$232.0	56.4%	$340.7	46.8%	$469.5	37.8%

Values in millions

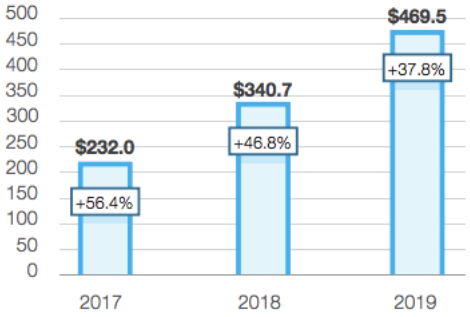
Figure 3 Οι μεγαλύτερες αγορές δεδομένων

Η μεγαλύτερη αγορά δεδομένων παγκοσμίως το 2018 είναι οι ΗΠΑ και αναμένεται ότι θα διατηρήσει τη θέση της και το 2019. Οι εταιρείες στις ΗΠΑ θα δαπανήσουν $ 12,3 εκατ. το 2018 για τα δεδομένα, που είναι σχεδόν το 60% της συνολικής αξίας της αγοράς δεδομένων. Ωστόσο, η αγορά δεδομένων της Κίνας είναι η ταχύτερα αναπτυσσόμενη από τις 5 μεγαλύτερες αγορές δεδομένων στον κόσμο (+ 95,5% σε ετήσια βάση). Το Ηνωμένο Βασίλειο είναι το δεύτερο μεγαλύτερο αγορά το 2018, αλλά θα χάσει τη θέση του το 2019, λόγω της επέκτασης της αγοράς της Κίνας.

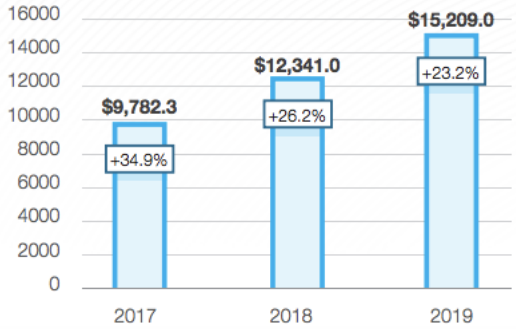
Data market size in **Canada**
2017–2019 ($ millions)



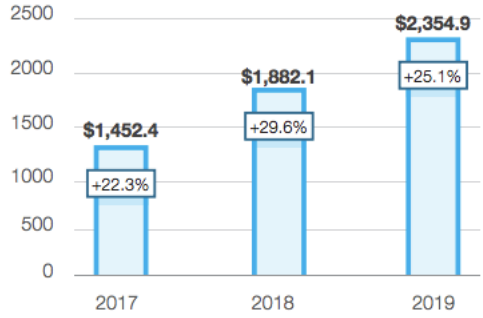
Data market size in **France**
2017–2019 ($ millions)



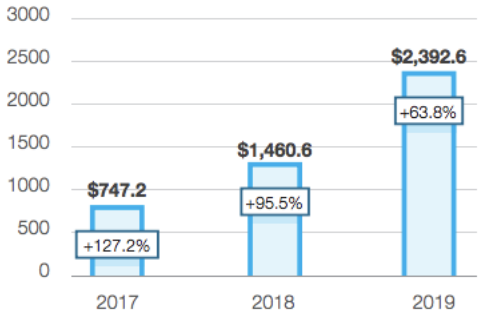
Data market size in the **US**
2017–2019 ($ millions)



Data market size in the **UK**
2017–2019 ($ millions)



Data market size in **China**
2017–2019 ($ millions)



Global data spend
2017–2019

Country	2017		2018		2019	
	Data Value	% Change	Data Value	% Change	Data Value	% Change
Australia	$102.5	104.3%	$220.6	115.3%	$327.6	48.5%
Austria	$10.3	127.6%	$24.5	137.5%	$54.3	121.4%
Canada	$453.7	30.1%	$588.7	29.7%	$768.8	30.6%
China	$747.2	127.2%	$1,460.6	95.5%	$2,392.6	63.8%
Colombia	< $0.1	—	$0.1	78.5%	$0.1	43.6%
Denmark	$80.3	62.3%	$114.2	42.1%	$137.2	20.2%
France	$232.0	56.4%	$340.7	46.8%	$469.5	37.8%
Germany	$127.9	71.3%	$184.0	43.9%	$265.6	44.4%
Hungary	$0.6	351.5%	$1.3	121.7%	$2.1	59.7%
India	$6.3	82.3%	$11.0	75.3%	$24.5	122.0%
Ireland	$11.4	31.6%	$16.5	44.9%	$26.4	59.3%
Italy	$40.7	75.6%	$64.2	57.8%	$88.9	38.4%
Netherlands	$42.1	19.5%	$40.9	-2.7%	$47.7	16.6%
New Zealand	$12.2	—	$17.5	43.4%	$24.6	40.7%
Poland	$11.9	77.5%	$21.0	77.1%	$37.2	76.9%
Romania	$1.2	—	$1.7	38.6%	$3.8	123.6%
Russia	$24.4	84.1%	$34.2	39.8%	$73.3	114.6%
Slovakia	$1.0	113.5%	$1.9	86.5%	$3.4	77.3%
Spain	$14.4	36.7%	$20.0	38.7%	$27.1	35.7%
Sweden	$44.3	143.2%	$63.8	44.1%	$82.0	28.5%
Switzerland	$6.1	301.2%	$21.8	256.3%	$41.6	90.6%
United Kingdom	$1,452.4	22.3%	$1,882.1	29.6%	$2,354.9	25.1%
USA	$9,782.3	34.9%	$12,341.0	26.2%	$15,209.0	23.2%
Global	$15,533.3	36.3%	$20,571.7	32.4%	$26,035.8	26.6%

Figure 4 Global data spend

3.3 **Πρωτοβουλίες μεγάλων δεδομένων σε αναπτυγμένες χώρες**

Πολλές χώρες αναλαμβάνουν πλέον πρωτοβουλίες για να επωφεληθούν από τις ευκαιρίες των μεγάλων δεδομένων σε διάφορους τομείς, όπως η εθνική ασφάλεια, η κοινωνική και

οικονομική ανάπτυξη, η υγεία και ούτω καθεξής. Μπορούμε να πούμε ότι, στο μέλλον, ο οικονομικός και πολιτικός ανταγωνισμός μεταξύ των χωρών θα βασιστεί στη χρήση των πιθανών ευκαιριών των μεγάλων δεδομένων. Με άλλα λόγια, η μελέτη και η εφαρμογή δεδομένων μεγάλης κλίμακας θα είναι απαραίτητα προκειμένου να αυξηθεί η ανταγωνιστικότητα οποιασδήποτε χώρας (Jina et al, 2015). Τα μεγάλα δεδομένα είναι ένα από τα βασικά θέματα που συζητούνται σε διεθνές επίπεδο. Θεωρείται ότι είναι η κινητήρια δύναμη του τομέα των τεχνολογιών των πληροφοριών και των επικοινωνιών (ΤΠΕ) και ονομάζεται "νέο πετρέλαιο". Τα μεγάλα δεδομένα βρίσκονται στο επίκεντρο των διεθνών οργανισμών ως πόρος στρατηγικής σημασίας. Τα τελευταία χρόνια, οι διεθνείς οργανισμοί δίνουν μεγάλη αξία στα μεγάλα δεδομένα και έχουν υιοθετηθεί ορισμένα στρατηγικά σχέδια και σχέδια για τη χρήση των δυνητικών ευκαιριών που προσφέρουν. Το 2012, το Παγκόσμιο Οικονομικό Φόρουμ στο Νταβός αξιολόγησε τα μεγάλα δεδομένα ως νέο οικονομικό πλεονέκτημα και εξέδωσε ένα έγγραφο που αποδεικνύει την ικανότητά του για διεθνή ανάπτυξη. Λαμβάνοντας υπόψη την τεχνολογική εξέλιξη στον κόσμο και τον όγκο και την ποικιλομορφία των πληροφοριών που αυξάνονται ταχέως σε πραγματικό χρόνο, το 2009, τα Ηνωμένα Έθνη ξεκίνησαν την πρωτοβουλία "Global Pulse". Η πρωτοβουλία αυτή αποσκοπεί στην αξιοποίηση των μεγάλων δυνατοτήτων δεδομένων για τη διατήρηση της παγκόσμιας και βιώσιμης ανάπτυξης, την εξάλειψη της φτώχειας και της κρίσης, την αύξηση του βιοτικού επιπέδου και την ανθρωπιστική δραστηριότητα (Alguliyev and Hajirahimova, 2012). Η πρωτοβουλία αποσκοπεί στην υλοποίηση της χρήσης ψηφιακών πηγών πληροφοριών, τεχνολογιών υψηλής ταχύτητας συλλογής και ανάλυσης δεδομένων από τα όργανα λήψης αποφάσεων σε πραγματικό χρόνο, προκειμένου να κατανοηθούν καλύτερα οι παράγοντες που επηρεάζουν τον σχηματισμό ευπαθών τμημάτων του πληθυσμού και να ανακαλύψουν ανωμαλίες, τάσεις και γεγονότα. Τα βασικά ζητήματα του "Global Pulse" περιλαμβάνουν: τη μελέτη καινοτόμων μεθόδων και τεχνικών για την ανάλυση δεδομένων σε πραγματικό χρόνο, την ενσωμάτωση των τεχνολογικών εργαλείων πηγής λογισμικού ελεύθερης και ανοιχτής πηγής για ανάλυση σε πραγματικό χρόνο και ανταλλαγή υποθέσεων, την ανάπτυξη εθνικού παγκόσμιου δικτύου "Pulse Lab" στη χώρα. Μαζί με τους διεθνείς οργανισμούς, τα μεγάλα δεδομένα έχουν προσελκύσει την προσοχή πολλών ανεπτυγμένων χωρών του κόσμου. Η πρώτη πρωτοβουλία στον τομέα αυτό έχει υποβληθεί από τις Ηνωμένες Πολιτείες. Αργότερα, αρκετές δυτικές χώρες, όπως η Αυστραλία, η Κίνα, η Ιαπωνία και άλλοι, έχουν εκτιμήσει τα μεγάλα δεδομένα ως στρατηγικό πόρο ως πετρέλαιο, ενώ η μεγάλη σημασία δίνεται στα προβλήματα στον τομέα αυτό και έχουν εγκριθεί πολλά σχετικά έγγραφα.

ΗΠΑ

Τα μεγάλα δεδομένα έχουν ήδη μεταφερθεί από τη φάση της έρευνας και της ανάπτυξης στη φάση της εφαρμογής στις ΗΠΑ. Η κυβέρνηση του Προέδρου των ΗΠΑ ανακοίνωσε την πρωτοβουλία για την έρευνα και ανάπτυξη μεγάλων δεδομένων τον Μάρτιο του 2012 (White House, 2014). Η πρωτοβουλία αποσκοπούσε στη διεξαγωγή σύνθετων εκδηλώσεων (διασκέψεων, φόρουμ κ.λπ.) για τη χρήση της τεχνολογίας μεγάλων δεδομένων σε βασικούς τομείς της κυβερνητικής πολιτικής των ΗΠΑ και την ανάπτυξη σχεδίων. Επιπλέον, διατέθηκαν 200 εκατομμύρια δολάρια στις αρμόδιες κυβερνητικές υπηρεσίες προκειμένου να οργανωθούν και να αναλυθούν μεγάλοι όγκοι ψηφιακών δεδομένων. Γενικά, το έγγραφο προϋποθέτει την ανάπτυξη 84 σχεδίων. Η πρωτοβουλία στοχεύει στη βελτίωση της νέας υποδομής και ερευνητικής μεθοδολογίας των δεδομένων και στην ενίσχυση των ικανοτήτων για τη χρήση τους για επιστημονικές ανακαλύψεις. Ο Λευκός Οίκος σκοπεύει να χρησιμοποιήσει τα Big Data για τους ακόλουθους σκοπούς:

- ανάπτυξη τεχνολογιών που είναι απαραίτητες για τη συλλογή, την αποθήκευση, την προστασία, τη διαχείριση, την ανάλυση και το μερίδιο μεγάλων δεδομένων.
- επιτάχυνση των επιστημονικών ανακαλύψεων στον τομέα της επιστήμης και της μηχανικής (τεχνολογία), ενίσχυση της εθνικής ασφάλειας και εξάσκηση της τεχνολογίας αυτής για τη ριζική αλλαγή της εκπαίδευσης και της κατάρτισης.
- ενίσχυση του εντοπισμού νέων ταλέντων και της κατάρτισης ειδικών για την ανάπτυξη και χρήση μεγάλων τεχνολογιών δεδομένων.

Το σχέδιο έχει σχεδιαστεί για να εκπαιδεύει επιστήμονες δεδομένων και μηχανικούς, ιδιαίτερα, αναλυτές με υψηλή ικανότητα στον τομέα της εξαγωγής δεδομένων από κείμενα σε οποιαδήποτε γλώσσα. Σύμφωνα με το έγγραφο, τα δεδομένα πρέπει να εφαρμοστούν στους ακόλουθους τομείς: υγειονομική περίθαλψη και κοινωνική προστασία του πληθυσμού, περιβάλλον και αειφόρος ανάπτυξη, αντιμετώπιση έκτακτων περιστατικών, κατασκευής, ρομποτικής και ευφυών συστημάτων, ασφάλειας στον κυβερνοχώρο, ενέργειας, εκπαίδευσης και ανάπτυξης των ανθρώπινων πόρων. Ως μέρος της πρωτοβουλίας, το Εθνικό Ίδρυμα Επιστημών έχει διαθέσει 10 εκατομμύρια δολάρια για το έργο "Expeditions in Computing" που σχεδιάζεται να εφαρμοστεί στα πανεπιστήμια της Καλιφόρνιας εντός 5 ετών. Το έργο έχει σχεδιαστεί για την ενσωμάτωση τριών διαφορετικών προσεγγίσεων, όπως η μηχανική μάθηση, ο υπολογισμός του cloud και η προμήθεια πλήθους για την έρευνα δεδομένων -

μετασχηματισμός δεδομένων σε πληροφορίες (White House, 2014). Στο πλαίσιο της διοργάνωσης μεγάλων δεδομένων, το Υπουργείο Ενέργειας έχει διαθέσει τη χρηματοδότηση ύψους 25 εκατομμυρίων δολαρίων στο Ινστιτούτο Διοίκησης, Ανάλυσης και Οπτικοποίησης (SDAV) για την κλιμάκωση δεδομένων για την υλοποίηση του έργου Scientific Discovery Through Advanced Computing. Υπό την ηγεσία του Εθνικού Εργαστηρίου Lawrence Berkeley του Υπουργείου Ενέργειας, το Ινστιτούτο της SDAV περιλαμβάνει τη χρήση 6 εθνικών εργαστηρίων και την πρακτική 7 πανεπιστημίων για την προετοιμασία ή την ανάπτυξη νέων εργαλείων που θα βοηθήσουν τους ερευνητές στη διαχείριση και απεικόνιση δεδομένων σε υπέρ-υπολογιστές[12]. Η Υπηρεσία Προηγμένων Έργων Έρευνας για την Άμυνα (DARPA) έχει ήδη ξεκινήσει το πρόγραμμα XDATA για την ανάπτυξη διαθέσιμων υπολογιστικών τεχνολογιών και εργαλείων λογισμικού για ανάλυση δεδομένων μεγάλης κλίμακας. Το πρόγραμμα επικεντρώνεται στα βασικά ζητήματα όπως η επέκταση αλγορίθμων κλιμάκωσης για την επεξεργασία δεδομένων και η ανάπτυξη αποτελεσματικών εργαλείων αλληλεπίδρασης ανθρώπου-ηλεκτρονικού υπολογιστή που επιταχύνουν την οπτικοποίηση για διαφορετικούς σκοπούς[3]. Στην υγεία το ερευνητικό κέντρο Institute for Health Metrics and Evaluation (IHME) του University of Washington των ΗΠΑ, ανέπτυξε έναν κατάλογο δεδομένων με το όνομα Global Health Data Exchange (GHDx). Στόχος του GHDx είναι η συλλογή μεγάλου όγκου δεδομένων υγείας από διαφορετικές πηγές από όλο τον κόσμο, προκειμένου οποιοσδήποτε ενδιαφέρεται για θέματα παγκόσμιας υγείας να μπορεί να ανατρέξει, να βρει ή και να μοιραστεί διάφορες πληροφορίες. Στα πλαίσια του έργου, συγκεντρώνει μετρήσεις δεδομένων υγείας από ποικίλες πηγές, από μελέτες, από βάσεις δεδομένων για ασθένειες, από φακέλους νοσοκομείων κ.ο.κ., ώστε να είναι δυνατή η αποτίμηση στρατηγικών, η μέτρση της κατάστασης της υγείας, η αποδοτικότητα των συστημάτων υγείας, κ.ά. Οι πληροφορίες αυτές είναι ανοικτές προς όλους τους ενδιαφερόμενους. Η τρέχουσα έκδοση του GHDx χρησιμοποιεί ένα ανοικτού κώδικα λογισμικό διαχείρισης περιεχομένου και την μηχανή αναζήτησης Apache SOLR. Το National Human Genome Research Institute (NHGRI) χρηματοδοτεί έρευνες μεγάλης κλίμακας στις ΗΠΑ, που σχετίζονται με την εξεύρεση σημαντικών γονιδιωματικών πληροφοριών. Το NHGRI μαζί με το LSAC (Large-Scale Genome Sequencing and Analysis Centers)

[1] https://www.whitehouse.gov/big_data_fact_sheet_final_1.pdf

[2] http://www.science.energy.gov/news/

[3] http://www.darpa.mil/program/xdata

συντονίζουν κάποια ακόμη ερευνητικά κέντρα που εργάζονται στον συγκεκριμένο τομέα. Απώτερος στόχος είναι να παρέχουν δεδομένα μεγάλης κλίμακας, τα οποία θα βοηθήσουν σε πολλούς και διαφορετικούς τομείς την ερευνητική κοινότητα στον τομέα της ιατρικής. Για παράδειγμα, την ανακάλυψη μεταλλάξεων που σχετίζονται με τον καρκίνο ή που χαρακτηρίζουν σύνθετες ασθένειες, ενώ παράλληλα θα βοηθήσουν στην αναζήτηση νέων ερωτημάτων όσον αφορά την διαφοροποίηση των γονιδιωμάτων σε σχέση με την βιολογία και τις ασθένειες, κ.ά. Επίσης, οι συμμετέχοντες έχουν την δυνατότητα, εκτός του όγκου των δεδομένων που παρέχονται, να υιοθετήσουν νέες μεθόδους που προκύπτουν από τις τεχνολογικές εξελίξεις ή που απαιτούνται για την συνέχιση της έρευνας της γονιδιωματικής αλληλουχίας (genome sequencing). Για αυτό το λόγο έχει αναπτύξει το πρόγραμμα Genome Sequencing Program (GSP) που βοηθά, μεταξύ άλλων, στην ανάπτυξη προτύπων και βέλτιστων πρακτικών, κ.ά. Ένα από τα πιο γνωστά οφέλη από την σε μεγάλη κλίμακα έρευνας στα πλαίσια του προγράμματος, είναι η μείωση στα κόστη που σχετίζονται με την αλληλουχία του DNA (DNA sequencing costs). Τα κόστη για την εύρεση της αλληλουχίας ενός γονιδιώματος έφταναν τα 10.000 δολάρια περίπου (για το έτος 2001), παράγοντας περί τα 100GB συμπιεσμένα δεδομένα. Χρησιμοποιώντας όμως το Apache Hadoop, τα συγκεκριμένα κόστη έφτασαν να είναι λιγότερα από 100 δολάρια.

Αυστραλία

Τα τελευταία χρόνια, οι αυστραλιανές κρατικές αρχές έχουν αξιολογήσει τα δεδομένα ως εθνική αξία. Η ανάλυση δεδομένων με την εισαγωγή νέων τεχνολογιών πιστεύεται ότι συμβάλλει τόσο στην αυστραλιανή κυβέρνηση όσο και στον αυστραλιανό λαό. Τα μεγάλα δεδομένα έχουν καταστεί αναπόσπαστο στοιχείο για τη χάραξη πολιτικής, την εισαγωγή νέων υπηρεσιών και τη δημιουργία ευκαιριών καινοτομίας. Το 2013, η κυβέρνηση της Αυστραλίας υιοθέτησε τη στρατηγική της Αυστραλιανής δημόσιας υπηρεσίας για μεγάλα δεδομένα[4]. Η στρατηγική περιλαμβάνει τα σχέδια για την υλοποίηση των μεταρρυθμίσεων στον τομέα των δημόσιων υπηρεσιών και την παροχή αποτελεσματικότερης δημόσιας πολιτικής και ασφάλειας των πολιτών μέσω της χρήσης μεγάλων κλιμάκων δεδομένων. Ο κύριος στόχος εδώ είναι η αύξηση της αποδοτικότητας των υπηρεσιών που παρέχονται στον δημόσιο τομέα και η τοποθέτηση της Αυστραλίας μεταξύ των κορυφαίων χωρών του κόσμου για τη χρήση της ανάλυσης δεδομένων στον τομέα της συνεργασίας και της καινοτομίας. Η

[4] Australian Pubic Service Big Data Strategy,2013, http://www.finance.gov.au

έννοια αποσκοπεί στην επέκταση των υπηρεσιών, ευκαιρίες για νέες υπηρεσίες και επιχειρηματικές εταιρικές σχέσεις, βελτιωμένη πολιτική εξέλιξη και προστασία της ιδιωτικής ζωής των δεδομένων, καθώς και ενίσχυση της κρατικής στήριξης για επενδύσεις στον τομέα των ΤΠΕ. Η ανάπτυξη της στρατηγικής προσδιορίστηκε αρχικά στην Τεχνολογική Στρατηγική για την Επικοινωνιακή Πληροφορική της Αυστραλίας 2012-2015. Το κύριο σημείο αυτού του εννοιολογικού εγγράφου ορίζεται γενικά ως εξής: παροχή καλύτερων υπηρεσιών με τη βελτίωση των υπηρεσιών και την αύξηση της ποιότητάς τους. βελτίωση της αποτελεσματικότητας της κυβερνητικής δραστηριότητας μέσω της βέλτιστης επένδυσης και της επέκτασης της καινοτομίας · άμεση συμμετοχή στη δημιουργία γνώσης και αποτελεσματική συνεργασία. Επιπλέον, στη στρατηγική προτείνονται δύο άλλα έργα:

1) Ανάπτυξη μητρώου περιουσιακών στοιχείων

2) Παρακολούθηση της τεχνικής ανάπτυξης σε μεγάλη ανάλυση δεδομένων.

Η επίλυση αυτών των ζητημάτων εξαρτάται από την αποτελεσματική χρήση και ανάλυση των δεδομένων που ακολουθούν έξι αρχές, οι οποίες είναι:

• βάση δεδομένων (η οποία είναι κοινή για όλους) είναι εθνικός πλούτος.

• ιδιωτικά χαρακτηριστικά του έργου.

• πληρότητα και διαφάνεια των διαδικασιών.

•διανομή εμπειριών, πόρων και δυνατοτήτων που αποκτήθηκαν.

• συνεργασία μεταξύ βιομηχανίας και ερευνητικών ιδρυμάτων.

• αύξηση των δημόσιων δεδομένων.

Επί του παρόντος, η Υπηρεσία Τελωνείων και Προστασίας των Συνόρων της Αυστραλίας εκμεταλλεύεται τις δυνατότητες των μεγάλων δεδομένων για τον εντοπισμό των υπόπτων[5]. Μελλοντικά, η Αυστραλία οραματίζεται την υλοποίηση πολλών δυνατοτήτων σε διάφορους τομείς όπως, να παρέχει καλύτερες πληροφορίες σχετικά με τα αποτελέσματα παροχής υπηρεσιών και να ενημερώνει το μέλλον με μοντέλα για την παροχή αυτών των υπηρεσιών καθώς και τον εντοπισμό των ελλείψεων. Έτσι να επιτρέπουν στις κυβερνητικές υπηρεσίες να στοχεύουν καλύτερα σε εκείνους που τις χρειάζονται, επιτρέποντας την αποτελεσματικότερη παροχή υπηρεσιών και να τους επιτρέψει να βελτιώσουν τις υπηρεσίες τους προσαρμόζοντας την παροχή υπηρεσιών με βάση τις ατομικές ανάγκες των επιχειρήσεων και των κοινοτήτων. Επιπρόσθετα, να δοθούν νέες ευκαιρίες επιχειρηματικών συνεργασιών, με την ανάλυση μεγάλων δεδομένων που αναμένεται να οδηγήσει στην ανάπτυξη νέων υπηρεσιών με βάση τις πληροφορίες που προκύπτουν από τη διαδικασία

[5] http://www.cio.com.au/article/524794/

ανάλυσης. Και τέλος τις εξελίξεις στη βιομηχανία και την ωριμότητα εργαλείων και υπηρεσιών που χρησιμοποιούν μεγάλες βιομηχανίες με βάση τη χρήση ανοικτών κυβερνητικών δεδομένων.

Μεγάλη Βρετανία

Η Μεγάλη Βρετανία είναι μια χώρα που μετατρέπει τη μεγάλη επανάσταση δεδομένων σε ένα από τα κύρια γεγονότα του 21ου αιώνα. Ως εκ τούτου, η χώρα αυτή έχει παγκόσμιας κλάσης επιχειρηματικούς τομείς όπως η αεροδιαστημική, η αυτοκινητοβιομηχανία, η αγροτική τεχνολογία, η υγειονομική περίθαλψη, τα μέσα ενημέρωσης, οι τηλεπικοινωνίες και άλλα. Προσφέρει τεράστιες ευκαιρίες στον ανθρώπινο δυναμικό, την υποδομή και τα δεδομένα, και το μεγαλύτερο μερίδιο στον τομέα των υπολογιστικών συστημάτων υψηλής απόδοσης στον κόσμο. Είναι επίσης παγκόσμιος ηγέτης στην επιστήμη των δεδομένων και στην επιστήμη των υπολογιστών, ενώ ιδιαίτερη προσοχή δίνεται στην ανάπτυξη της ηλεκτρονικής υποδομής. Υψηλής απόδοσης υπολογιστές, cloud computing και άλλες σύγχρονες τεχνολογίες που εξασφαλίζουν την επίλυση πολύ περίπλοκων και χρονοβόρων προβλημάτων αποτελούν τη βάση της ηλεκτρονικής υποδομής στη Μεγάλη Βρετανία. Η κυβέρνηση χορήγησε επένδυση ύψους 158 εκατομμυρίων στερλινών για πόρους HPC το 2011 και 375 εκατομμύρια στερλίνες το 2012[6].

Η Μεγάλη Βρετανία κατέχει ηγετική θέση στον κόσμο για ανοικτά δεδομένα. Από το 2010, το data.gov.uk έχει δημιουργήσει περισσότερα από 10.000 σύνολα δεδομένων. Το πρώτο Ινστιτούτο Ανοιχτών Δεδομένων παγκοσμίως ιδρύθηκε επίσης εδώ[7]. Για να βελτιωθεί η ποιότητα των παρεχόμενων υπηρεσιών στη Μεγάλη Βρετανία, τον Οκτώβριο του 2012, η Data Services ιδρύθηκε με την οικονομική υποστήριξη του Συμβουλίου Οικονομικής και Κοινωνικής Έρευνας (ESRC). Η υπηρεσία δεδομένων είναι μια εθνική υπηρεσία που παρέχει στους κοινωνιολόγους, τους ερευνητές και τους επαγγελματίες την πρόσβαση σε δεδομένα απογραφής, σε κοινωνικά και οικονομικά σύνολα δεδομένων που χρηματοδοτούνται από το κράτος. Η υπηρεσία συνδυάζει τα στοιχεία της Υπηρεσίας Οικονομικών και Κοινωνικών Δεδομένων που δημιουργήθηκε από το ESRC, την υπηρεσία Secure Data, την ιστοσελίδα

[6] Seizing the data opportunity. A strategy for UK data capability, https://www.gov.uk/

[7] Industrial Strategy: government and industry in partnership, https://www.gov.uk/

Census.ac.uk, συμπεριλαμβανομένου του προγράμματος απογραφής[8]. Τον Οκτώβριο του 2013 δόθηκε πρόσθετη οικονομική υποστήριξη στην Υπηρεσία Δεδομένων από τις κυβερνητικές υπηρεσίες του Ηνωμένου Βασιλείου και άλλες αρχές για το συντονισμό του Δικτύου Ερευνών Διοικητικών Δεδομένων (ADRN), το οποίο παρέχει πρόσβαση σε διοικητικά δεδομένα. Το ADRN θεωρείται ένα από τα στάδια του μεγάλου δικτύου δεδομένων του ESRC και η κύρια λειτουργία του είναι να εξασφαλίζει στους ερευνητές, τους δασκάλους, τους φοιτητές, τους τοπικούς πολιτικούς, τους φιλάνθρωπους και τους ιδιοκτήτες επιχειρήσεων «υψηλής ποιότητας κοινωνικά και οικονομικά δεδομένα». Η κυβέρνηση της Μεγάλης Βρετανίας διέθεσε 64 εκατομμύρια λίρες χρηματοδότησης για το δίκτυο δεδομένων ESRC για τη βελτιστοποίηση των δεδομένων ως πόρου. Διάφορα μεγάλης κλίμακας δεδομένα που συλλέγονται από τα κρατικά τμήματα, τις επιχειρήσεις και τους οργανισμούς αποτελούν σημαντικό πόρο που μπορεί να χρησιμοποιηθεί για τα ερευνητικά ιδρύματα, τους οργανισμούς και για ολόκληρη την κοινότητα[9].

Τα Big Data, τα οποία έχουν τη δυνατότητα να αλλάξουν κάθε επιχειρηματικό τομέα και το χώρο της επιστήμης, είναι μία από τις οκτώ μεγάλες τεχνολογίες (μεγάλα δεδομένα και ενεργειακά αποδοτικοί υπολογιστές, δορυφόροι και διαστημικά χωρικά προγράμματα, ρομποτική και αυτόνομα συστήματα, συνθετική βιολογία, αναγεννητική ιατρική, την επιστήμη, τα πιο προηγμένα υλικά και τη νάνο-τεχνολογία, την ενέργεια και τη συντήρησή της) της Μεγάλης Βρετανίας. Το 2012, η κυβέρνηση επένδυσε 189 εκατομμύρια λίρες στα κέντρα δεδομένων και στον ενεργειακά αποδοτικό υπολογιστή για να λύσει μεγάλα ζητήματα δεδομένων. Τον Ιούνιο του 2013, η βρετανική κυβέρνηση ενέκρινε τη βιομηχανική στρατηγική: κυβέρνηση και βιομηχανία στην εταιρική σχέση (Στρατηγική για την Οικονομία της Πληροφορίας) με στόχο την ανάπτυξη της οικονομίας της πληροφορίας. Η στρατηγική τονίζει το τεράστιο δυναμικό των μεγάλων δεδομένων για την αλλαγή όλων των τομέων της οικονομίας και δείχνει τη σημασία της ηλεκτρονικής υποδομής και των επιστημονικών δεδομένων για την αξιοποίηση των δεδομένων μεγάλης κλίμακας. Η στρατηγική εξασφαλίζει επίσης την αποτελεσματικότερη χρήση της τεχνολογίας των πληροφοριών και των δεδομένων και παρέχει στους πολίτες τη δυνατότητα να επωφεληθούν από τον ψηφιακό αιώνα[10]. Τον Οκτώβριο του 2013, ως συνέχεια της προαναφερθείσας στρατηγικής, η

[8] http://www.statslife.org.uk/features/22-introducing-the-uk-data-service.

[9] UK Data Service-http://ukdataservice.ac.uk

[10] https://www.gov.uk/government/publications/information-economy-strategy

κυβέρνηση ενέκρινε τη στρατηγική Επικράτηση της ευκαιρίας των δεδομένων: Μια στρατηγική για την ικανότητα δεδομένων του Ηνωμένου Βασιλείου. Η στρατηγική που αναπτύχθηκε με τη συνεργασία της βιομηχανίας και της επιστημονικής κοινότητας αποσκοπούσε στο να καταστήσει το Ηνωμένο Βασίλειο παγκόσμιο ηγέτη για τη χρήση χρήσιμων πληροφοριών, οι οποίες προέρχονται από τα δεδομένα, από τους πολίτες, τις επιχειρήσεις και τους ακαδημαϊκούς κύκλους, συμπεριλαμβανομένου του δημόσιου και του ιδιωτικού τομέα. Η στρατηγική περιλαμβάνει τις ακόλουθες πτυχές:

- Ανθρώπινο κεφάλαιο: ειδικευμένο εργατικό δυναμικό και πολίτες με αυτοπεποίθηση των δεδομένων
- Διαθέσιμα εργαλεία και υποδομή για την αποθήκευση και ανάλυση δεδομένων
- Δεδομένα ως παράγοντας που επιτρέπει την προσβασιμότητα και την κοινή χρήση των σχετικών δεδομένων από τους καταναλωτές, την επιχειρηματική και την επιστημονική κοινότητα.

Για το σκοπό αυτό, η στρατηγική έχει σχεδιαστεί για να λάβει μέτρα, όπως η δημιουργία ικανοτήτων στην επιχείρηση, στον ακαδημαϊκό και στον δημόσιο τομέα. την ενίσχυση των δεξιοτήτων που απευθύνονται σε σχολεία, πανεπιστήμια και στην περαιτέρω εκπαίδευση υποστηρίζοντας την ικανότητα των δεδομένων έρευνας και ανάπτυξης στη Μεγάλη Βρετανία.

Η αγκαλιά ανοιχτών δεδομένων του Τμήματος Μεταφορών συνέβαλε στη βελτίωση των δημόσιων υπηρεσιών και επέτρεψε στον ιδιωτικό τομέα να αναπτύξει νέα προϊόντα όπως η εφαρμογή μεταφοράς CityMapper. Η TfL έχει κερδίσει επαίνους παρέχοντας δωρεάν σε οποιονδήποτε, εύκολη πρόσβαση σε δρομολόγια, κατάσταση υπηρεσίας και πληροφορίες διακοπής. Οι προγραμματιστές χρησιμοποιούν αυτές τις πληροφορίες για τη δημιουργία νέων προϊόντων και υπηρεσιών μεταφοράς, γεγονός που βοηθά την TfL να επεκτείνει την εμβέλεια των δικών της καναλιών πληροφοριών σε σταθμούς, στάσεις λεωφορείων και σε απευθείας σύνδεση. Το τοπικό κυβερνητικό όργανο υποστηρίζει ότι περισσότερες από 600 εφαρμογές που χρησιμοποιούνται από το 42% των Λονδρέζων τροφοδοτούνται τώρα από την επιλογή πάνω από 80 ανοιχτών τροφοδοσιών δεδομένων, τα οποία είναι διαθέσιμα μέσω ενοποιημένου API.

Η έρευνα που πραγματοποιήθηκε από την TfL και διεξήχθη από την Deloitte εκτιμά ότι η παροχή των δεδομένων αυτών ανέρχεται σε 130 εκατομμύρια λίρες στερλίνες ετησίως στην οικονομία του Λονδίνου.

Στον τομέα της υγειονομικής περίθαλψης και τους χρήστες του, το NHS Digital καθιστά διαθέσιμα στο κοινό δεδομένα υγείας και κοινωνικής φροντίδας που διατίθενται βάσει της άδειας κυκλοφορίας. Αυτό περιλαμβάνει στατιστικές δημοσιεύσεις, δεδομένα που παράγονται ως απόκριση σε αιτήματα FOI, δαπάνες και δομικά δεδομένα και οργανωτικά δεδομένα μέσω της υπηρεσίας δεδομένων NHS Digital, συγκεντρώνει δεδομένα σύμφωνα με τα εθνικά πρότυπα και κανονισμούς μέσω του δικτυακού της τόπου και μέσω του data.gov.uk. Το 2017, ο Υγειονομικός Γραμματέας ξεκίνησε την πρόκληση MyNHS ανοικτών δεδομένων, ένα ταμείο 100.000 λιρών στερλινών για να ανταμείψει τις πιο δημιουργικές εφαρμογές και ψηφιακά εργαλεία για τη βελτίωση των υπηρεσιών. Η Defra έχει επίσης επαινεί για τις προσπάθειές της. Το 2015, το τμήμα απελευθέρωσε 8.000 σύνολα δεδομένων, 1.000 από τα οποία είναι στην γεωργία, τα οποία οι άνθρωποι του γεωργικού τομέα μπορούν να χρησιμοποιήσουν για να κατανοήσουν την υγεία των ζώων και να βρουν την καλύτερη γη για καλλιέργεια. "Θεωρούν τα δεδομένα ως δημόσιο περιουσιακό στοιχείο, το οποίο είναι ακριβώς το σωστό πράγμα", δήλωσε τότε ο επικεφαλής του GDS Mike Bracken.

Γαλλία

Η Γαλλία είναι μια βιομηχανική χώρα, όπου έχουν αναπτυχθεί επιχειρήσεις, επιστημονικές εφευρέσεις και επιχειρηματικότητα. Οι έξυπνες και δικτυακές τεχνολογίες, το λογισμικό, το cloud computing, τα μεγάλα συστήματα ασφάλειας δεδομένων και πληροφοριών κ.ο.κ. θεωρούνται προτεραιότητα στη Γαλλία. Σε αυτή τη χώρα, τα δεδομένα έχουν γίνει το νέο οικονομικό πλεονέκτημα της κυβέρνησης, των επιχειρήσεων, των περιφερειών και των πόλεων. Από το 2011, η κυβέρνηση παρείχε πρόσβαση στα δημόσια δεδομένα[11] δίνοντας ιδιαίτερη προσοχή στην ανάπτυξη της ψηφιακής οικονομίας. Για το σκοπό αυτό, το Φεβρουάριο του 2013, η γαλλική κυβέρνηση ενέκρινε το σχέδιο «Ψηφιακό οδικό χάρτη» που αφορούσε την υλοποίηση επτά έργων, μεταξύ των οποίων τα μεγάλα δεδομένα. Στο πλαίσιο

[11] http://s244543015.onlinehome.fr/2014/11/clerc_big-data-icci-2014.pdf

του «Προγράμματος Επενδύσεων για το Μέλλον», έχει διατεθεί για το σχέδιο μια επένδυση ύψους 11,5 εκατομμυρίων ευρώ[12]. Το σχέδιο περιλαμβάνει τρεις κύριους τομείς: 1) την ανάπτυξη ικανοτήτων της ψηφιακής οικονομίας για τη νεότερη γενιά, 2) την ενίσχυση της ανταγωνιστικότητας των γαλλικών εταιρειών εις βάρος της ψηφιακής οικονομίας και 3) την προώθηση των αξιών της ψηφιακής κοινωνίας και της οικονομίας.

Τα τελευταία χρόνια, η γαλλική κυβέρνηση έδωσε έμφαση στην ανάπτυξη πολιτικής στον τομέα των μεγάλων δεδομένων. Έτσι, το στρατηγικό πρόγραμμα "Νέα Βιομηχανική Γαλλία" που εγκρίθηκε τον Σεπτέμβριο του 2013 περιλάμβανε το Big Data ως ένα από τα 34 μεγαλύτερα αλληλένδετα έργα ανασυγκρότησης της γαλλικής βιομηχανίας (ιατρική βιοτεχνολογία, ψηφιακό νοσοκομείο, cloud computing, online εκπαίδευση, νάνο-ηλεκτρονική, διαδίκτυο των πραγμάτων, υπηρεσίες χωρίς σύνδεση, υπολογιστές, ρομπότ, ασφάλεια στον κυβερνοχώρο, φυτά του μέλλοντος και άλλες περιοχές). Το σχέδιο Big Data εγκρίθηκε τον Ιούλιο του 2014 ως μέρος αυτής της στρατηγικής. Ο στόχος του σχεδίου είναι να καταστεί η Γαλλία παγκόσμιος ηγέτης σε αυτόν τον τομέα. Η πρωτοβουλία καλύπτει τους τομείς, όπως η δημιουργία κέντρων τεχνολογικών πόρων, η ανάπτυξη εκπαιδευτικών προγραμμάτων και νέων επιχειρήσεων επιστημόνων και η υποστήριξη της επιστημονικής έρευνας. Το σχέδιο καλύπτει κυρίως τρία σύνολα δραστηριοτήτων: την ανάπτυξη μεγάλου οικοσυστήματος δεδομένων στη Γαλλία, πρωτοβουλίες στον τομέα των μεγάλων δεδομένων (που περιλαμβάνει έργα στον δημόσιο και στον ιδιωτικό τομέα) αξιολόγηση των κανονισμών (που περιλαμβάνει κανόνες προσωπικών δεδομένων).

Αρζεμπαϊτζάν

Αν και το Αζερμπαϊτζάν δεν είναι μια από τις μεγαλύτερες χώρες, είναι όμως μία από τις χώρες που κατέχουν υψηλή θέση για την ολοκληρωμένη χρήση των ΤΠΕ στη δημόσια διοίκηση. Πρέπει να σημειωθεί ότι το Αζερμπαϊτζάν έχει ήδη ξεκινήσει τις σπουδές στον τομέα των μεγάλων δεδομένων, η οποία είναι μια από τις μεγαλύτερες προκλήσεις του 21ου αιώνα, και τη χρήση αυτής της τεχνολογίας σε ορισμένους τομείς δραστηριότητας. Έτσι, ο όγκος των δεδομένων που συλλέγονται και επεξεργάζονται μέσω της πύλης ηλεκτρονικής διακυβέρνησης αναπτύσσεται ραγδαία. Πρέπει να σημειωθεί ότι η δικτυακή πύλη ηλεκτρονικής διακυβέρνησης ήταν αυτή η οποία ιδρύθηκε το 2012 και σήμερα συνδέει 80

[12] The new face of industry in France, http://www.entreprises.gouv.fr/

οργανισμούς, παρέχοντας πάνω από 650 ηλεκτρονικές υπηρεσίες στους πολίτες σε γενικές γραμμές, και ο αριθμός των χρηστών του έφτασε τα 2,5 εκατομμύρια. Αυτό με τη σειρά του, προκαλεί σοβαρά προβλήματα για την ανάλυση των κρατικών δεδομένων για διαφορετικούς σκοπούς. Προκειμένου να αντιμετωπιστεί το πρόβλημα, το Data Center, το οποίο είναι το πρώτο στην περιοχή, είναι έτοιμο για λειτουργία. Πρέπει να σημειωθεί ότι το Ινστιτούτο Πληροφορικής της ANAS διερευνά το προβλήματα εξάσκησης γνώσης από μεγάλα σύνολα δεδομένων. Ταυτόχρονα, AzScienceNet Data Center με την τεράστια μνήμη και υπολογιστικούς πόρους (200 terabytes μνήμης, 14 Tflops υπολογιστική ισχύς) λειτουργεί στο ινστιτούτο. Πρέπει να σημειωθεί ότι το AzScienceNet συνδέεται περίπου με 4000 υπολογιστές ερευνητικών ιδρυμάτων της ANAS. Το κέντρο, το οποίο θεωρείται βασικό στοιχείο υποδομής για τα μεγάλα δεδομένα, προσφέρει υπηρεσίες cloud για τα περίπλοκα επιστημονικά ζητήματα που απαιτούν μεγάλους υπολογιστικούς πόρους και παρέχουν πόρους αποθήκευσης για τα ινστιτούτα και τα ιδρύματα οργανώσεις της ANAS. Συνολικά, η ευρεία έρευνα και εφαρμογή των μεγάλων τεχνολογιών δεδομένων και η υποστήριξη που παρέχεται από το κράτος για την επίλυση των προβλημάτων σε αυτόν τον τομέα συνιστάται. Η εφαρμογή τους ακολουθώντας τα μέτρα του κράτους προσπαθεί:

1) Να αναλύσει τις διάφορες καταστάσεις που σχετίζονται με την εφαρμογή μεγάλων δεδομένων στη χώρα, και για τον προσδιορισμό των τομέων προτεραιότητας αυτής της τεχνολογίας.
2) Να αναπτύξει την εθνική στρατηγική μεγάλων δεδομένων. Η προετοιμασίας αυτής της στρατηγικής ΤΠΕ θα πρέπει να βασίζεται στην εμπειρία των αναπτυγμένων χωρών για να υπάρχει η κατάλληλη αναπτυξιακή στρατηγική της χώρας.
3) Να αναπτύξει την ηλεκτρονική υποδομή. Η ανάπτυξη του απαραίτητου λογισμικού και υποδομής συλλογής, για αποθήκευση, προστασία, διαχείριση, ανάλυση και κοινή χρήση δεδομένων μεγάλης κλίμακας.
4) Ανάπτυξη και υποστήριξη του ανοιχτού δημόσιου δικτύου που εξασφαλίζει την αποτελεσματική χρήση μεγάλων δεδομένων για την κοινωνία, η οποία συλλέγεται από τις κυβερνητικές υπηρεσίες, τις επιχειρήσεις και τους οργανισμούς.
5) Να επιταχύνει την κατάρτιση νέων ταλέντων στην ανάπτυξη και τη χρήση μεγάλων τεχνολογιών δεδομένων, για να υπάρχουν εμπειρογνώμονες και να βελτιώσουν τις δυνατότητες χρήσης τους για επιστημονικές ανακαλύψεις.

6) Να οργανώσει τη διδασκαλία της ακαδημαϊκής πειθαρχίας στον επιστήμονα δεδομένων στην τριτοβάθμια εκπαίδευση πτυχίων, τη δημιουργία κέντρων τεχνολογικών πόρων και την υποστήριξη της έρευνας και ούτω καθεξής.
7) Να συντονίζει και να οργανώνει τις κοινές δραστηριότητες των κρατικών, βιομηχανικών, ακαδημαϊκών και μη κερδοσκοπικών οργανισμών ώστε να αξιοποιούν στο έπακρο τις μεγάλες ευκαιρίες δεδομένων και τη χρήση των δυνατοτήτων τους.

Οι μεγάλες τεχνολογίες δεδομένων πρέπει να εφαρμοστούν στους ακόλουθους τομείς: ασφάλιση του πληθυσμού, διασφαλίζοντας τη διαφάνεια, το περιβάλλον και την αειφόρο ανάπτυξη, την ανταπόκριση σε καταστάσεις έκτακτης ανάγκης και φυσικές καταστροφές, μεταφορές και ενέργεια, εκπαίδευση και εργατικό δυναμικό τη διαχείριση και την ανάπτυξη, καθώς και τη δημιουργία νέων επιχειρήσεων και υπηρεσιών.

Ιαπωνία

Η Ιαπωνία κατέχει την τρίτη θέση στον κόσμο για το μέγεθος της οικονομίας[13]. Η Ιαπωνία διαθέτει υποδομή ΤΠΕ υψηλού επιπέδου και κορυφαίες πηγές υπηρεσιών στον τομέα των ΤΠΕ παγκοσμίως, όπως οι Fujitsu, Hitachi, NTT Data και Nec. Πάνω από το 86% του πληθυσμού της Ιαπωνίας είναι χρήστες του διαδικτύου και των smart phone και κατέχει τον υψηλότερο ρυθμό ανάπτυξης σε αυτόν τον τομέα σε όλο τον κόσμο[14]. Τα μεγάλα δεδομένα αποτελούν μία από τις κύριες οικονομικές προτεραιότητες για την ιαπωνική κυβέρνηση και έχουν υιοθετηθεί από την κυβέρνηση ορισμένες στρατηγικές στον τομέα αυτό.

Το 2012, η κυβέρνηση υιοθέτησε στρατηγική ανοικτών κυβερνητικών δεδομένων για να συμβάλει στη διαφάνεια στην Ιαπωνία. Το Συμβούλιο για τη Ρυθμιστική Μεταρρύθμιση ξεκίνησε την καθιέρωση των βασικών αρχών για τη χρήση δεδομένων μεγάλης κλίμακας από τις μεγάλες τοπικές εταιρείες χωρίς να παραβιάζουν τους νόμους περί ιδιωτικής ζωής. Η ιαπωνική κυβέρνηση διέθεσε 13,2 δισεκατομμύρια γιεν για την εφαρμογή αυτής της στρατηγικής. Επιπλέον, η κυβέρνηση ενέκρινε το 2012 και το 2013 τις εθνικές μεγάλες

[13] World Bank: India Overtakes Japan as World's Third Largest Economy, http://thediplomat.com/2014/05/

[14] http://www.internetlivestats.com/internet-users/japan/

στρατηγικές δεδομένων, για την ολοκληρωμένη στρατηγική για τις ΤΠΕ 2020 και τη διακήρυξη ως το πλέον προηγμένο κράτος της Πληροφορικής παγκοσμίως. Οι στρατηγικές αυτές αποσκοπούσαν στην ανάπτυξη της τεχνολογίας των πληροφοριών της Ιαπωνίας κατά την περίοδο 2013-2020 μέσω ανοικτών δημόσιων δεδομένων και μεγάλων δεδομένων. Ο κύριος στόχος είναι η απόκτηση του καθεστώτος μιας χώρας με υψηλά πρότυπα για την εκτεταμένη χρήση μεγάλων δεδομένων στον κλάδο της πληροφορικής της Ιαπωνίας[15].
Η στρατηγική αξιολογεί τη χρήση μεγάλων δεδομένων για τη δημιουργία νέων επιχειρήσεων και τη δημιουργία θέσεων εργασίας, καθώς και για νέες υπηρεσίες στον τομέα της οικολογίας, της εκπαίδευσης, των μεταφορών και άλλων τομέων μέσω της συνεργασίας με τον ιδιωτικό και τον δημόσιο τομέα. Το έγγραφο τονίζει τη σημασία της επέκτασης της πρόσβασης των ιδιωτικών τομέων στα δημόσια δεδομένα και τη στήριξη της δημιουργίας νέων επιχειρηματικών τομέων και υπηρεσιών σε βάρος της χρήσης μεγάλων δεδομένων.

Τον Ιούνιο του 2013, η ιαπωνική κυβέρνηση υιοθέτησε τη στρατηγική της Ιαπωνίας για την αναγέννηση. Η στρατηγική περιγράφει τη δημιουργία ισχυρών υποδομών και εγκαταστάσεων για τη σύνδεση με την αγορά υπηρεσιών δεδομένων και σχεδιάζει να καταστήσει την Ιαπωνία παγκόσμιο ηγέτη στον τομέα της πληροφορικής. Ο στόχος της στρατηγικής είναι να ενισχύσει τη θέση της ως παγκοσμίως ηγετικής χώρας πληροφορικής. Η κυβέρνηση προβλέπει ότι ο όγκος της αγοράς υπηρεσιών δεδομένων θα φθάσει τα 51 δισ. Λίρες μέχρι το 2020. Επιπλέον, έχουν διατεθεί 87,5 εκατ. Λίρες για τα έργα για την επέκταση των μελετών και την ανάπτυξη τεχνολογιών virtualization δικτύων και προγραμμάτων ανάλυσης δεδομένων κ.ο.κ. Η νεοσύστατη τεχνολογική υποδομή επικεντρώνεται στη βελτίωση της βιομηχανικής ανταγωνιστικότητας της Ιαπωνίας και στη δημιουργία νέων τομέων παραγωγής και καινοτομίας[16].

Νότια Κορέα

Η Νότια Κορέα έχει μεγάλη εμπειρία στη χρήση της εξελιγμένης τεχνολογίας. Η οικονομική ανάπτυξη της χώρας, η καινοτομία και το ανταγωνιστικό πλεονέκτημα είναι επίτευγμα της χρήσης αυτών των τεχνολογιών. Αυτή η τάση παρατηρείται ιδιαίτερα στην χρήση των μεγάλων δεδομένων. Όπως και άλλες τεχνολογίες, έτσι και τα μεγάλα δεδομένα διαδραματίζουν πολύ σημαντικό ρόλο στην ανάπτυξη της χώρας. Η χρήση των μεγάλων

[15] Declaration to be the world's most advanced IT nationt, 2013, http://japan.kantei.go.jp/policy/it/2013/0614_declaration.pdf

[16] https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/en_saikou_jpn_hon.pdf

δεδομένων έχει προσελκύσει το ενδιαφέρον επιστημονικών, κοινοτικών, ιδιωτικών και δημόσιων φορέων. Το συμβούλιο Προεδρίας της Εθνικής Στρατηγικής ΤΠΕ, ίδρυσε μια ομάδα εργασίας στο πλαίσιο της πρωτοβουλίας για τα μεγάλα δεδομένα το 2011. Η πρωτοβουλία ήταν με στόχο την ανάλυση δημόσιων δικτύων και συστημάτων μεγάλων δεδομένων, τη σύγκλιση των δεδομένων μεταξύ του δημόσιου και του ιδιωτικού τομέα, τα διαγνωστικά συστήματα για ανοικτά μεγάλα δεδομένα και τη δημιουργία διαχείρισης και αναλυτικές τεχνολογίες στον δημόσιο και τον ιδιωτικό τομέα. Εκτός αυτού, η πρωτοβουλία περιλαμβάνει επίσης τη συνεργασία του Big Data Strategy Center του Οργανισμού για την Εθνική Κοινωνία της Πληροφορίας με το μεγάλο ινστιτούτο δεδομένων, το οποίο λειτουργεί στο πλαίσιο του Εθνικού Κράτους της Σεούλ.

Τον Νοέμβριο του 2012, η εθνική Επιτροπή Επιστήμης και Τεχνολογίας της Νότιας Κορέας ανέπτυξε το Big Data Master Plan. Το σχέδιο εξετάζει τη χρήση μεγάλων δεδομένων στη χάραξη της δημόσιας πολιτικής, στον τομέα των δημόσιων υπηρεσιών, στην προστασία της ανταγωνιστικότητας των μικρών και μεσαίων επιχειρήσεων, την ανάπτυξη ερευνητικών δεξιοτήτων και τη δημιουργία νέων υποδομών. Ο στόχος του σχεδίου είναι να συνειδητοποιήσουμε την ιδέα του έξυπνου Έθνους. Ως μέρος αυτού του σχεδίου, πιλοτικό πρόγραμμα για μεγάλα δεδομένα στη χρήση οι επίσημες στατιστικές έχουν θεσπιστεί από την κυβέρνηση. Η Νότια Κορέα αναπτύσσει συστήματα πληροφοριών μεγάλης βάσης δεδομένων για την πρόβλεψη φυσικών καταστροφών με βάση τις πληροφορίες που λαμβάνονται από διάφορες πηγές. Η κυβέρνηση σκοπεύει να χρησιμοποιήσει μεγάλα δεδομένα για την απογραφή. Σε σύγκριση με την παραδοσιακή απογραφή, αυτή είναι μια εξοικονόμηση 140 εκατομμυρίων USD.

Ένα πολύ σημαντικό πλεονέκτημα λόγω της ταχείας ανάλυσης μεγάλων όγκων μη δομημένων δεδομένων από το Facebook, Twitter και άλλα κοινωνικά δίκτυα, ο αυξανόμενος αριθμός αυτοκτονιών στη Νότια Κορέα τα τελευταία χρόνια έχει περιοριστεί λόγο της καλής πρόληψης που είναι αποτέλεσμα συλλογής μεγάλων δεδομένων.

Κίνα

Η διεξαγωγή μιας σειράς εκδηλώσεων που συνδέονται με τη Διάσκεψη Κορυφής για την Τεχνολογία Δεδομένων, τη Σύνοδο Κορυφής για την Καινοτομία Big Data & Analytics, το Νομικό Συνέδριο Μεγάλων Δεδομένων της Κίνας, το Big Data Asia Showcase, το Big Data World Forum και άλλες που λαμβάνουν χώρα στην Κίνα εξηγούν την τεχνολογική πρόοδο σε αυτόν τον τομέα (Hoi-Wai et al, 2014). Η Κίνα, η οποία διαθέτει 1.2 δισεκατομμύρια συνδρομητές κινητής τηλεφωνίας, κατέχει τη μεγαλύτερη αγορά κινητής τηλεφωνίας στον κόσμο και περισσότερους από 700 εκατομμύρια χρήστες του Διαδικτύου. Τα κοινωνικά μέσα ενημέρωσης στην Κίνα παίρνουν πιο ενεργό θέση, ενώ σύμφωνα με επίσημες εκτιμήσεις, περισσότερα από 250 εκατομμύρια άτομα χρησιμοποιούν κοινωνικά μέσα, όπως blogs, κοινωνικά δίκτυα, ιστολογία και άλλα. Οι μη κυβερνητικοί οργανισμοί εκτιμούν ότι ο αριθμός αυτός είναι κοντά στα 590 εκατομμύρια.

Η αύξηση του όγκου των πληροφοριών, συμπεριλαμβανομένου του αυξανόμενου αριθμού των κινητών τηλεφώνων, των κοινωνικών μέσων και των χρηστών του Διαδικτύου στην Κίνα, δημιούργησε ευνοϊκές συνθήκες για τη χρήση μεγάλων δεδομένων για τον εντοπισμό προβλημάτων στη χώρα και την ανάπτυξη της χώρας. Τώρα, ο δημόσιος τομέας έχει αρχίσει να διερευνά τις δυνατότητες των μεγάλων δεδομένων. Τον Ιούνιο του 2014, το Πολιτικό Συμβουλευτικό Συμβούλιο του Κινεζικού Λαού πρότεινε την πρωτοβουλία να αξιοποιηθεί το δυναμικό των μεγάλων τεχνολογιών δεδομένων για την αύξηση της διοικητικής ικανότητας στο φόρουμ που διοργάνωσε το συμβούλιο. Η πρωτοβουλία εξετάζει επίσης τη χρήση μεγάλων δεδομένων στους τομείς όπως η βελτίωση των επιδόσεων του δημόσιου τομέα, ανάπτυξη του σχεδιασμού των αστικών μεταφορών, κατανόηση των κοινωνικοοικονομικών τάσεων, αξιολόγηση της φτώχειας, υποστήριξη της διάθεσης ηλεκτρονικών απορριμμάτων και τον εντοπισμό καυτών σημείων ρύπανσης στις αστικές περιοχές. Για να χρησιμοποιηθούν αποτελεσματικά τα μεγάλα δεδομένα είναι σημαντικό να δημιουργηθούν οι προϋποθέσεις για την πραγματοποίηση κοινών εργασιών του δημόσιου και του ιδιωτικού τομέα για την επίλυση πολυάριθμων προβλημάτων (ανάλυση δεδομένων, προβλήματα συστήματος κλπ.).

3.4 **Προβλήματα μεγάλων δεδομένων**

Πέραν των όποιων ωφελειών προσφέρουν τα Big Data, υπάρχουν μια σειρά προβλημάτων που πρέπει να αντιμετωπιστούν. Το σημαντικότερο πρόβλημα είναι το θέμα της προστασίας της ιδιωτικής ζωής και της ιδιοκτησίας των δεδομένων. Αν και η εποχή των Big Data βρίσκεται ακόμα στην απαρχή της, ωστόσο έχουν σημειωθεί περιπτώσεις παραβίασης της ιδιωτικής ζωής. Το συγκεκριμένο πρόβλημα ανακύπτει από ορισμένες δυσκολίες των υπολογιστικών συστημάτων να ορίσουν τις πληροφορίες που θεωρούνται προσωπικά δεδομένα. Επίσης, υφίσταται δυσκολία στη τήρηση διακριτικότητας σε ορισμένα στοιχεία όπως το φύλο, η ηλικία, οι καταναλωτικές προτιμήσεις κλπ, δημιουργώντας πολλές φορές δυνατότητες επηρεασμού της καταναλωτικής συμπεριφοράς (Hill, 2014). Τα συστήματα που σχετίζονται με την καταναλωτική συμπεριφορά δεν είναι τα μοναδικά που δημιουργούν μεγάλη ανησυχία για το απόρρητο των δεδομένων. Η υγειονομική περίθαλψη, η κοινωνική δικτύωση και τα κυβερνητικά συστήματα περιέχουν επίσης μεγάλες ποσότητες ευαίσθητων πληροφοριών. Κάθε άτομο στον ανεπτυγμένο κόσμο μπορεί να συνδέεται με τουλάχιστον ένα γεγονός σε μια ηλεκτρονική βάση δεδομένων που ο οποιοσδήποτε θα μπορούσε να

χρησιμοποιήσει για εκβιασμό, διακρίσεις, παρενόχληση, οικονομικές βλάβες (Lazer et al., 2014:1203- 1204). Ένα άλλο πρόβλημα με τα Big Data είναι ότι μπορεί να είναι παραπλανητικές οι πληροφορίες που παρουσιάζουν εξαιτίας των αρνητικών συσχετίσεων τους. Ουσιαστικά, παρουσιάζονται πληροφορίες που πολλές φορές δεν είναι αληθείς αλλά παρουσιάζουν μια αληθοφάνεια ή δεν έχουν κανένα απολύτως νόημα που να σχετίζεται με τις αναζητούμενες πληροφορίες. Στα Big Data ορισμένες φορές παρουσιάζονται στρεβλώσεις νοημάτων οι οποίες όχι μόνο δεν συνεισφέρουν στην πληροφόρηση αλλά αποτελούν βασικό εμπόδιο. Τέλος, η λειτουργία των Big Data απαιτεί κατάλληλες υποδομές και υπολογιστικά συστήματα τα οποία καταλαμβάνουν μεγάλο χώρο και απαιτούν αντίστοιχα μεγάλη ισχύ και μεγάλη δικτύωση. Τα συστήματα πληροφορικής αντιπροσωπεύουν σήμερα περίπου το 10% της παγκόσμιας χρήσης ηλεκτρικής ενέργειας. Ο αριθμός αυτός είναι βέβαιο ότι θα αυξηθεί ραγδαία, καθώς όλο και περισσότερο χρησιμοποιούνται αποθηκευτικοί δίσκοι και επεξεργαστές για τα δεδομένα που αποθηκεύονται. Η εκθετική αύξηση των 33 δεδομένων έχει οδηγήσει σε αύξηση των απαιτήσεων ενέργειας καθιστώντας τα Big Data ως ιδιαίτερα σπάταλα σε ενέργεια (Hoy, 2014:5-6).

4.1 **Μελλοντικές εφαρμογές και κατευθύνσεις των Big Data**

Τα συστήματα μεγάλων δεδομένων μπορούν να αποσυντεθούν σε τέσσερις διαδοχικές ενότητες και συγκεκριμένα στην παραγωγή, στην απόκτηση, στην αποθήκευση και στην ανάλυση δεδομένων. Κάθε συνιστώσα αυτής της αλυσίδας αξίας παρουσιάζει διάφορες προκλήσεις που απαιτούν βαθιά έρευνα, κυρίως εξαιτίας του ετερογενούς και πολύπλοκου χαρακτήρα των δεδομένων. Ως διεπιστημονικό ή οριοθετημένο θέμα, τα μεγάλα δεδομένα έχουν τη δυνατότητα να προσελκύουν μια όλο και μεγαλύτερη προσοχή ανά κλάδο, όπως των βιομηχανικών κοινοτήτων ή του διοικητικού και οικονομικού τομέα που σχετίζονται με τη βιομηχανία. Μερικές πιθανές μελλοντικές τάσεις των μεγάλων δεδομένων για σύγχρονη βιομηχανία περιλαμβάνουν, αλλά δεν περιορίζονται σε:

• Νέες τεχνικές και βελτιώσεις για μεγάλη ανάλυση και εξόρυξη δεδομένων
• Η λύση που βασίζεται σε υπηρεσίες υπολογιστικού νέφους για μεγάλη αποθήκευση και μετάδοση δεδομένων
• Λύσεις μεγάλων δεδομένων που επικεντρώνονται στον έλεγχο και στην παρακολούθηση
• Βελτιστοποίηση και πρόγνωση σε μεγάλη κλίμακα σε επίπεδο εργοστασιακής βάσης

• Λύσεις για την παροχή μεγάλων δεδομένων στα συστήματα διαχείρισης κινδύνου
• Θεωρία μεγάλων δεδομένων για μοντέρνες βιομηχανικές εφαρμογές
• Λύσεις μεγάλων δεδομένων για έξυπνες εφαρμογές και καθαρά συστήματα ισχύος

Ορισμένες από τις συγκεκριμένες κατευθύνσεις που έχουν ορισθεί ως μελλοντικό πεδίο ανάπτυξης των εφαρμογών μεγάλων δεδομένων μπορούν να κατηγοριοποιηθούν ως εξής:

Α) Μεγάλα δεδομένα και ανοικτός κώδικας

Οι εφαρμογές ανοιχτού κώδικα όπως το Apache Hadoop, το Spark και άλλες που έρχονται να κυριαρχήσουν στο χώρο των μεγάλων δεδομένων. Σύμφωνα με τη Forrester, η χρήση Hadoop αυξάνεται κατά 32,9% ετησίως, ενώ οι ειδικοί λένε ότι το 2019, πολλές επιχειρήσεις θα επεκτείνουν τη χρήση των τεχνολογιών Hadoop και NoSQL. Συνεπώς, είναι αναγκαίες τεχνολογίες που τους επιτρέπουν την πρόσβαση σε δεδομένα με ανταπόκριση σε πραγματικό χρόνο.

Β) Τεχνολογία μνήμης

Μία από τις τεχνολογίες που οι εταιρείες διερευνούν σε μια προσπάθεια να επιταχύνουν την επεξεργασία μεγάλων δεδομένων είναι η τεχνολογία in-memory. Σε μια παραδοσιακή βάση δεδομένων, τα δεδομένα αποθηκεύονται σε συστήματα αποθήκευσης εξοπλισμένα με σκληρούς δίσκους ή μονάδες SSD. Η τεχνολογία της μνήμης αποθηκεύει τα δεδομένα στη μνήμη RAM, η οποία είναι πολλές φορές ταχύτερη. Σύμφωνα με μια έκθεση της Forrester Research, το υλικό των δεδομένων στη μνήμη θα αυξηθεί κατά 29,2% ετησίως.

Γ) Μηχανική μάθηση

Δεδομένου ότι οι δυνατότητες ανάλυσης μεγάλων δεδομένων έχουν προχωρήσει, ορισμένες επιχειρήσεις έχουν αρχίσει να επενδύουν στη μηχανική μάθηση (ML). Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης που εστιάζει στο να επιτρέψει στους υπολογιστές να μάθουν νέα πράγματα χωρίς να έχουν προγραμματιστεί ρητά. Με άλλα λόγια, αναλύει τα υπάρχοντα δεδομένα για να καταλήξει σε συμπεράσματα που αλλάζουν τον τρόπο με τον οποίο συμπεριφέρεται η εφαρμογή. Σύμφωνα με την Gartner, η μηχανική μάθηση είναι μία από τις 10 κορυφαίες τάσεις στρατηγικής τεχνολογίας για το 2019. Τα πιο προηγμένα σύγχρονα συστήματα μάθησης και τεχνητής νοημοσύνης κινούνται πέρα από

τους παραδοσιακούς αλγόριθμους βασισμένους σε κανόνες για τη δημιουργία συστημάτων που κατανοούν, μαθαίνουν, ενδεχομένως λειτουργεί αυτόνομα.

Δ) Ευφυείς εφαρμογές μεγάλων δεδομένων

Ένας άλλος τρόπος με τον οποίο οι επιχειρήσεις χρησιμοποιούν τη μηχανική μάθηση και τις τεχνολογίες AI είναι η δημιουργία έξυπνων εφαρμογών. Αυτές οι εφαρμογές συχνά ενσωματώνουν αναλύσεις μεγάλων δεδομένων, αναλύοντας τις προηγούμενες συμπεριφορές των χρηστών προκειμένου να παρέχουν εξατομίκευση και καλύτερη εξυπηρέτηση. Ένα παράδειγμα που έχει γίνει πολύ γνωστό είναι οι μηχανισμοί σύστασης που τώρα τροφοδοτούν πολλές εφαρμογές ηλεκτρονικού εμπορίου και ψυχαγωγίας. Στη λίστα των Top 10 στρατηγικών τεχνολογικών τάσεων για το 2019, η Gartner κατέγραψε τις έξυπνες εφαρμογές δεύτερης γενιάς. Κατά τα επόμενα 10 χρόνια, σχεδόν κάθε εφαρμογή, εφαρμογή και υπηρεσία θα ενσωματώσει κάποιο επίπεδο AI, αποτελώντας μια μακροπρόθεσμη τάση που συνεχώς θα εξελίσσεται και θα επεκτείνει την εφαρμογή του AI και της μηχανικής μάθησης για άλλες εφαρμογές και υπηρεσίες.

Ε) Ευφυής ασφάλεια

Πολλές επιχειρήσεις ενσωματώνουν επίσης αναλύσεις μεγάλων δεδομένων στη στρατηγική ασφαλείας τους. Τα στοιχεία του ημερολογίου ασφαλείας των οργανισμών παρέχουν πληροφορίες σχετικά με τις απόπειρες κυβερνητικής πολιτικής που οι οργανισμοί μπορούν να χρησιμοποιήσουν για την πρόβλεψη, πρόληψη και μετριασμό των μελλοντικών προσπαθειών. Ως αποτέλεσμα, ορισμένοι οργανισμοί ενσωματώνουν το λογισμικό πληροφοριών ασφαλείας και διαχείρισης συμβάντων (SIEM) με μεγάλες πλατφόρμες δεδομένων όπως το Hadoop, μια τάση που θα αναπτυχθεί περισσότερο στο μέλλον.

Ζ) Μεταφορά σκοτεινών δεδομένων στο Cloud

Οι πληροφορίες που πρόκειται να μετατραπούν σε ψηφιακή μορφή ονομάζονται σκουρόχρωμα δεδομένα και είναι μια τεράστια δεξαμενή που είναι αναξιοποίητη. Αυτές οι αναλογικές βάσεις δεδομένων αναμένεται να ψηφιοποιηθούν και να μεταφερθούν στο cloud, έτσι ώστε να μπορούν να χρησιμοποιηθούν για αναλυτικές προβλέψεις που ωφελούν τις επιχειρήσεις.

Η) Quantum Computing

Το να αναλύουμε και να ερμηνεύουμε τεράστια ποσά δεδομένων μπορεί να πάρει πολύ χρόνο με την τρέχουσα τεχνολογία που χρησιμοποιούμε. Εάν μόνο μπορούμε να έχουμε δισεκατομμύρια δεδομένα ταυτόχρονα μέσα σε λίγα λεπτά, μπορούμε να μειώσουμε το χρόνο επεξεργασίας πάρα πολύ, δίνοντας στις εταιρείες την ευκαιρία να λάβουν έγκαιρες αποφάσεις για να επιτύχουν τα επιθυμητά αποτελέσματα. Αυτή η τεράστια επιχείρηση μπορεί να γίνει δυνατή μόνο μέσω της κβαντικής πληροφορικής. Παρά το γεγονός ότι είναι στις αρχές, πραγματοποιούνται πειράματα σε κβαντικούς υπολογιστές σε μια προσπάθεια να βοηθήσουν στην πρακτική και θεωρητική έρευνα σε διάφορες βιομηχανίες. Πολύ σύντομα, μεγάλες εταιρείες τεχνολογίας όπως η Google, η IBM και η Microsoft θα ξεκινήσουν τη δοκιμή κβαντικών υπολογιστών για να τις ενσωματώσουν στις επιχειρηματικές διαδικασίες τους.

Θ) Πιο έξυπνα Chatbots

Με την πιο έξυπνη τεχνολογία AI, τα chatbots αναπτύσσονται τώρα από τις εταιρείες για να χειρίζονται τα ερωτήματα των πελατών για να παρέχουν πιο εξατομικευμένες αλληλεπιδράσεις, ενώ εξαλείφουν την ανάγκη για πραγματικό ανθρώπινο δυναμικό. Τα μεγάλα δεδομένα έχουν πολλά να κάνουν με την παροχή πιο ευχάριστης εμπειρίας στον πελάτη, καθώς τα bots επεξεργάζονται μεγάλα ποσά δεδομένων για να παρέχουν σχετικές απαντήσεις με βάση τις καταχωρημένες λέξεις-κλειδιά από τους πελάτες στα ερωτήματά τους. Κατά τη διάρκεια των αλληλεπιδράσεων, είναι επίσης σε θέση να συλλέγουν και να αναλύουν πληροφορίες σχετικά με τους πελάτες από συνομιλίες. Αυτή η διαδικασία μπορεί να βοηθήσει τους εμπόρους να αναπτύξουν μια πιο απλοποιημένη στρατηγική για την επίτευξη καλύτερων μετατροπών.

Ι) Ταχέως αναπτυσσόμενα δίκτυα IoT

Είναι όλο και συχνό το γεγονός ότι τα smartphones μας χρησιμοποιούνται για τον έλεγχο των οικιακών συσκευών μας, χάρη στην τεχνολογία που ονομάζεται Διαδίκτυο των πραγμάτων (IoT). Με τις έξυπνες συσκευές όπως το Google Assistant και το Microsoft Cortana που τείνουν σε σπίτια για να αυτοματοποιήσουν συγκεκριμένα καθήκοντα, η αυξανόμενη τρέλα του IoT σχεδιάζει εταιρείες να επενδύσουν στην ανάπτυξη της τεχνολογίας. Οι περισσότερες οργανώσεις θα αξιοποιήσουν την ευκαιρία παρέχοντας καλύτερες λύσεις IoT. Αυτό θα οδηγήσει σε περισσότερους τρόπους για τη συλλογή τεράστιων ποσοτήτων δεδομένων, μαζί με τα μέσα διαχείρισης και ανάλυσης. Η ανταπόκριση του κλάδου είναι να προωθήσει

περισσότερες νέες συσκευές που είναι πιο ικανές να συλλέγουν, να αναλύουν και να επεξεργάζονται δεδομένα.

4.2 **Η σχέση Internet of Things (IoT) και Big Data**

Προσωπικά θεωρώ ότι ένα από τα μελλοντικά μοντέλα στο χώρο της πληροφορικής το οποίο θα εξελίσσεται συνεχώς είναι το IoT, το οποίο έχει άμεση σύνδεση με τα Big Data. Στην περίπτωση του Internet of Things (Διαδίκτυο των Πραγμάτων), πολλά στοιχεία που συγκεντρώνονται προέρχονται από δεδομένα του πραγματικού κόσμου από κατάλληλες συσκευές και αισθητήρες που τα καταγράφουν. Οι αισθητήρες ενσωματώνονται σε διάφορες συσκευές και μηχανήματα στον πραγματικό κόσμο συλλέγοντας διάφορα είδη δεδομένων, όπως περιβαλλοντικά στοιχεία, γεωγραφικά δεδομένα, αστρονομικά δεδομένα, διοικητικά στοιχεία κλπ. Για την καταγραφή δεδομένων στο IoT μπορεί να αξιοποιηθεί και κινητός εξοπλισμός, εγκαταστάσεις 17 μεταφορών, δημόσιες εγκαταστάσεις και οικιακές συσκευές (Gubbi et al). Τα Big Data που παράγονται από IoT έχουν διαφορετικά χαρακτηριστικά σε σύγκριση με τα γενικά Big Data λόγω του διαφορετικού τύπου των δεδομένων που συλλέγονται με χαρακτηριστικά στοιχεία την ετερογένεια, την ποικιλία, τον αδόμητο χαρακτήρα. Παρά το γεγονός ότι τα τρέχοντα δεδομένα από το IoT δεν αποτελούν το κυρίαρχο μέρος των Big Data, από το 2030 όπου ο αριθμός των αισθητήρων θα φτάσει το 1 τρισεκατομμύριο τα δεδομένα του IoT θα αποτελούν το βασικότερο τμήμα των Big Data. Μια έκθεση από την Intel επισήμανε ότι τα δεδομένα στο IoT έχουν τρία χαρακτηριστικά που συμμορφώνονται με το πρότυπο των Big Data: α) μεγάλος αριθμός τερματικών που δημιουργούν πληθώρα δεδομένων, β) τα δεδομένα που προέρχονται από το IoT είναι συνήθως ημί-δομημένα ή αδόμητα και γ) τα δεδομένα του IoT είναι χρήσιμα μόνο όταν αναλύονται (Chen et al., 2014). Προς το παρόν, η ικανότητα επεξεργασίας δεδομένων του IoT βρίσκεται ακόμα σε χαμηλότερα επίπεδα έναντι των Big Data καθιστώντας αναγκαία την εισαγωγή των νέων τεχνολογιών δεδομένων για την προώθηση της ανάπτυξης του IoT. Πολλοί φορείς εκμετάλλευσης του IoT έχουν κατανοήσει τη σημασία των Big Data για την τελική του επιτυχία μέσω αποτελεσματικής ενσωμάτωσής τους και αξιοποίησης του Cloud. Υπάρχει επιτακτική ανάγκη να υιοθετηθούν νέες τεχνολογίες για το IoT ενώ και η ανάπτυξη των Big Data μέσω IoT βρίσκεται σε πρώιμο στάδιο. Έχει αναγνωριστεί ότι οι δύο τεχνολογίες είναι αλληλένδετες και θα πρέπει να αναπτυχθούν από κοινού: από τη μία πλευρά, η ευρεία διάδοση του IoT οδηγεί σε υψηλή αύξηση των Big Data, τόσο σε ποσότητα και σε είδη παρέχοντας έτσι νέες ευκαιρίες στην εφαρμογή και ανάπτυξή τους ενώ από την άλλη πλευρά η απαίτηση νέων δεδομένων στο IoT επιταχύνει την έρευνα και την πρόοδο στα επιχειρηματικά μοντέλα (Zaslavsky , 2012).

4.3 **Συμπεράσματα**

Η εξέλιξη των τεχνολογιών επεξεργασίας δεδομένων και αναλύσεων έχουν βελτιώσει την αξία των δεδομένων. Αυτό, με τη σειρά του, προσέλκυσε την προσοχή των χωρών για να κάνουν τη μέγιστη χρήση των δεδομένων. Επειδή τα μεγάλα δεδομένα μπορούν να διαδραματίσουν σημαντικό ρόλο στην επίλυση των προβλημάτων όπως η πρόληψη των μολυσματικών ασθενειών, της τρομοκρατίας, των φυσικών καταστροφών και των παγκόσμιων κινδύνων, καθώς και στη λήψη σωστών αποφάσεων σε θέματα υγείας,

κοινωνικής ασφάλισης κ.ο.κ σε κρατικό επίπεδο, οι σημαντικοί διεθνείς οργανισμοί και οι ανεπτυγμένες χώρες έχουν υιοθετήσει ορισμένα έγγραφα στον τομέα αυτό. Οι διεξαγόμενες μελέτες μας επιτρέπουν να δηλώσουμε ότι αυτά τα κράτη έχουν υιοθετήσει μεγάλες στρατηγικές δεδομένων σε διάφορους τομείς (επιστημονικό, κοινωνικό, οικονομικό, υγεία, ασφάλεια κλπ.) που τους δίνει σημαντικό τεχνολογικό και αναπτυξιακό πλεονέκτημα αξιοποιώντας την οικονομική ευκαιρία των μεγάλων δεδομένων.

Αναφορές

Abraham, A., & Das, S. (Eds.). (2010). Computational intelligence in power engineering (Vol. 302). Springer.

Accenture. Build It and They Will Come? Chicago, 2012; http://www.accenture.com/ SiteCollectionDocument

Aggarwal, C. C. (2011). Social network data analytics, Chapter An introduction to social network data analytics. IBM TJ Watson Research Center Hawthorne, NY 10532, 13.

Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In Social network data analytics (pp. 243-275). Springer, Boston, MA.

Alguliyev R.M., Hajirahimova M. Sh. "Big Data" phenomenon: Challenges and Opportunities // Problems of Information Technology, 2014, No2, pp. 3-16.

Barbier, G., & Liu, H. (2011). Data mining in social media. In Social network data analytics (pp. 327-352). Springer, Boston, MA.

Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010, October). Finding a Needle in Haystack: Facebook's Photo Storage. In OSDI (Vol. 10, No. 2010, pp. 1-8).

Boyd D and Crawford K (2012) Critical questions for big data. Information, Communication and Society 15(5): 662–679.

Bresnahan, T., Brynjolfsson, E., Hitt, L., 2002. Information technology, work organization and the demand for skilled labor: firm-level evidence. Quarterly Journal of Economics 117 (1), 339–376

Broekema, C.P. et al. DOME: Towards the ASTRON and IBM Center for Exascale Technology. In Proceedings of the 2012 Workshop on High-Performance Computing for Astronomy Data, 2012, 1–4.
Broekema, C.P. et al. DOME: Towards the ASTRON and IBM Center for Exascale Technology. In Proceedings of the 2012 Workshop on High-Performance Computing for Astronomy Data, 2012, 1–4.
Burghard C: Big Data and Analytics Key to Accountable Care Success. 2012, IDC Health Insights

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188.

Chen, H., Chiang, R.H.L., and Storey, V.C. Business intelligence and analytics: From big data to big impact. MIS Quarterly 36, 4 (Dec. 2012), 1165–1188.

Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. International Journal of Information Management, 34(2), 272-284.

Economist, T. (2010). Data, data everywhere: A special report on managing information. The Economist.

European Commission. A Digital Agenda for Europe. Brussels, Aug. 26, 2010; http://ec.europa.eu/ digital-agenda/

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849-911.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Gable, G., 2010. Strategic information systems research: an archival analysis. Journal of Strategic Information Systems 19 (1), 3–16.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1142(2011), 1-12.

Gundecha, P., & Liu, H. (2012). Mining social media: a brief introduction. In New Directions in Informatics, Optimization, Logistics, and Production (pp. 1-17). Informs.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. Computer, 33(11), 29-36.

Hakeem, A., Gupta, H., Kanaujia, A., Choe, T. E., Gunda, K., Scanlon, A., ... & Haering, N. (2012). Video analytics for business intelligence. In Video analytics for business intelligence (pp. 309-354). Springer, Berlin, Heidelberg.

Hays, C., 2004. What Wal-Mart Knows About Customers' Habits, New York Times (Nov 14). (retrieved 20.06.15)

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.

Heidemann, J., Klier, M., & Probst, F. (2012). Online social networks: A survey of a global phenomenon. Computer networks, 56(18), 3866-3878.

Hirschberg, J., Hjalmarsson, A., & Elhadad, N. (2010). "You're as Sick as You Sound": Using Computational Approaches for Modeling Speaker State to Gauge Illness and Recovery. In Advances in speech recognition (pp. 305-322). Springer, Boston, MA.

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6), 797-819.

IBM big data platform for healthcare." Solutions Brief. 2012,http://public.dhe.ibm.com/common/ssi/ecm/en/ims14398usen/IMS14398USEN.

IBM: Data Driven Healthcare Organizations Use Big Data Analytics for Big Gains. 2013,http://www03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf,Google Scholar

Jackie Hoi-Wai C. Big Data for Development in China : UNDP China Working Paper, 2014, http://www.cn.undp.org/

Jiang, J. (2012). Information extraction from text. In Mining text data (pp. 11-41). Springer, Boston, MA.

Jina X., Benjamin W. W., Chenga X., Wanga Y. Significance and Challenges of Big Data Research // Big Data Research, 2015, vol. 2, no. 2, pp. 59–64

Kitchin R (2014) The real-time city? Big data and smart urbanism. GeoJournal 79: 1–14.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management, 34(3), 387-394.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. journal of the Association for Information Science and Technology.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Lucas, H., Agarwal, R., Clemons, E., El Sawy, O., Weber, B., 2013. Impactful research on transformational information technology: an opportunity to inform new audiences. MIS Quarterly 37 (2), 371–382.

Lucas, H., Goh, J., 2009. Disruptive technology: how Kodak missed the digital photography revolution. Journal of Strategic Information Systems 18 (1), 46– 55.

Majchrzak, A., Malhotra, A., 2013. Towards an information systems perspective and research agenda on crowdsourcing for innovation. Journal of Strategic Information Systems 22 (4), 257–268.

Malone, T., Crowston, K., Herman, G., 2003. Organizing Business Knowledge: The MIT Process Handbook. MIT Press, Cambridge, MA.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Markus, L., Loebbecke, C., 2013. Commoditized digital processes and business community platforms: new opportunities and challenges for digital business strategies. MIS Quarterly 37 (2), 649–653.

Mayer-Schonberger V and Cukier K (2013) Big Data: A Revolution that Will Change How We Live, Work and Think. London: John Murray.

Office of Science and Technology Policy, Executive Office of the President. Fact Sheet: Big Data Across the Federal Government. Washington, D.C., Mar. 29, 2012; http://www.whitehouse.gov/administration/eop/ostp

Office of Science and Technology Policy, Executive Office of the President. Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments. Washington, D.C., Mar. 29, 2012; http://www.whitehouse.gov/administration/eop/ostp

OnAudience, Global Data Market Size, 2019 https://www.onaudience.com/files/OnAudience.com_Global_Data_Market_Size_2017-2019.pdf

Orlikowski, W., Barley, S., 2001. Technology and institutions: what can research on information technology and research on organizations learn from each other? MIS Quarterly 25 (2), 145–165.

Osterwalder, A., Pigneur, Y., 2010. Business Model Generation. A Handbook for Visionaries, Game Changers, and Challengers. John Wiley & Sons, Hoboken, NJ.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

Parthasarathy, S., Ruan, Y., & Satuluri, V. (2011). Community discovery in social networks: Applications, methods and emerging trends. In Social network data analytics (pp. 79-113). Springer, Boston, MA.

Patil, H. A. (2010). "Cry Baby": Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant's Cry. In Advances in speech recognition (pp. 323-348). Springer, Boston, MA.

Picot, A., Reichwald, R., Wigand, R., 2008. Information, Organization and Management. Springer, Berlin-Heidelberg.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report. IBM Institute for Business Value and Said Business School at the University of Oxford.

Sun, J., & Tang, J. (2011). A survey of models and algorithms for social influence analysis. In Social network data analytics(pp. 177-214). Springer, Boston, MA.

Tang, L., & Liu, H. (2010). Community detection and mining in social media. Synthesis lectures on data mining and knowledge discovery, 2(1), 1-137.

Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). A. J. Hey (Ed.). Redmond, WA: Microsoft research.

VanBoskirk, S., Overby, C. S., & Takvorian, S. (2011). US interactive marketing forecast, 2011 to 2016. BCAMA, Marketing Association of BC.

Weill, P., Woerner, S., 2015. Thriving in an increasingly digital ecosystem. MIT Sloan Management Review 56 (4), 27–34

Weill, P., Woerner, S., 2015. Thriving in an increasingly digital ecosystem. MIT Sloan Management Review 56 (4), 27–34.

White House, 2014, Big data is a big deal, 2012,http://www.whitehouse.gov/blog/2012/03/29/

Zammuto, R., Griffith, T., Majchrzak, A., Dougherty, D., Faraj, S., 2007. Information technology and the changing fabric of organization. Organization Science 18 (5), 749–762.
International Journal of Data Science and Analyticshttps://doi.org/10.1007/s41060-018-0102-5, statistics Claus Weihs Katja ,Ickstadt
Future of Defence: Big Data And Military Intelligence, Nikunj Thakkar.