ΤΕΧΝΟΛΟΓΙΚΟ ΙΔΡΥΜΑ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μεγάλα Δεδομένα & Επιχείρηση

Μάλλιος Λ. Ανδρέας

Επιβλέποντες: **Βασίλειος Νικολαΐδης**, **Καθηγητής**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μεγάλα Δεδομένα & Επιχείρηση

Μάλλιος Ανδρέας

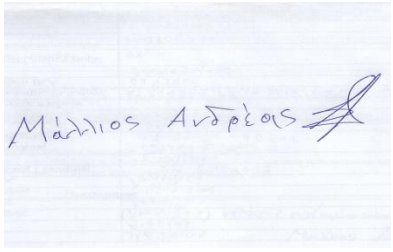
ΑΜ: 2013166

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΉΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας. Επίσης αναλαμβάνω πλήρως και αποκλειστικά την ευθύνη για τις απόψεις, αποτελέσματα, συμπεράσματα και γενικότερα το περιεχόμενο που παρουσιάζεται στην Εργασία αυτή, η οποία και δεν πρέπει να ερμηνευτεί πως αντιπροσωπεύει απόψεις του επιβλέποντα εκπαιδευτικού, του Τμήματος, της Σχολής ή του Ιδρύματος.

Όνομα & Επώνυμο Συγγραφέα (Με Κεφαλαία): ΑΝΔΡΕΑΣ ΜΑΛΛΙΟΣ

Υπογραφή (Ολογράφως, χωρίς μονογραφή):



Ημερομηνία (Ημέρα – Μήνας – Έτος): 20 / 11 / 2018

Περίληψη

Επί των ημερών μας, οι επιχειρήσεις και οι οργανισμοί δεν θέλουν να γνωρίζουν μόνο τι συνέβη και γιατί, αλλά επίσης θέλουν να γνωρίζουν τι συμβαίνει αυτή τη στιγμή και τι είναι πιθανό να συμβεί στο μέλλον (LaValle κ.α., 2011). Από τη στιγμή που οι επιχειρήσεις και οι οργανισμοί διψάνε για αυτού του είδους την πληροφορία και την υιοθέτηση της ιδέας του παγκόσμιου ιστού, η παραγωγή των δεδομένων και η ταχύτητα συλλογής αυτών αυξήθηκε με εκθετικό ρυθμό. Δεδομένου της ακρίβειας του νόμου του Moore, όπου η επεξεργαστική ισχύ θα διπλασιάζεται κάθε δύο χρόνια, δίχως σημαντική μεταβολή στη τιμή, είχε ως αποτέλεσμα την επίτευξη σπουδαίων κατορθωμάτων. Το 1994, το κόστος για την αποθήκευση δεδομένων μεγέθους ενός Gigabyte είχε $1000, ενώ το αντίστοιχο κόστος το 2010 ήταν μόλις $0.10.

Η μεγάλη ζήτηση για αυτού του είδους την πληροφορία οδήγησε τις εταιρίες στην αναζήτηση, την αποθήκευση και την ανάλυση αυτού του τεράστιου όγκου δεδομένων. Ταυτόχρονα, η ραγδαία τεχνολογική ανάπτυξη βοήθησε σημαντικά τις επιχειρήσεις και τους οργανισμούς ώστε να μπορούν να επεξεργαστούν τον τεράστιο όγκο δεδομένων και ως αποτέλεσμα αυτού, έστρεψαν το ενδιαφέρον τους στην εξερεύνηση και στην εκμετάλλευση της διαθέσιμης πληροφορίας. Το φαινόμενο αυτό αποκαλείται: «Μεγάλα Δεδομένα». Τη σήμερον μέρα, οι επιχειρήσεις βλέπουν τα μεγάλα δεδομένα ως ένα περιουσιακό στοιχείο και χαρακτηριστικά ορισμένες συγκρίνουν τα μεγάλα δεδομένα με το πετρέλαιο, όπου και τα δύο θα πρέπει πρώτα να επεξεργαστούν για να αποκτήσουν αξία.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μεγάλα Δεδομένα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Πληροφορία, Επιχείρηση, Apache Spark, Ιδιωτικότητα

Abstract

Nowadays, businesses and organizations do not only want to know what happened and why it happened, but they also want to know what is happening right now and what is likely to happen in the future (LaValle κ.α., 2011). By the time that businesses and organizations hunger for these insights and the adoption of the World Wide Web (www), the generation of data and collection speed has benn increased exponentially (Chen κ.α., 2012). The roughly biannual doubling of computing power and storage for the same price, also known as Moore's law, has also done remarkable things. In 1994, people paid $1000 for a Gigabyte of storage, while in 2010 the costs of a Gigabyte of storage was only $0,10.

The great demand of this type of information led organizations and businesses to capture, store and analyze large amounts of data. At the same time, the rapid technological breakouts helped organizations to being able to process these huge amount of data and as a result, they shifted their focus on exploring and exploiting all this available information. This phenomenon is called "Big Data". Nowadays, organizations and businesses see Big Data as an asset and characteristically some of them make the comparison with oil, as like oil, Big Data should be refined before it gets a value.

SUBJECT AREA: Big Data

KEYWORDS: Information, Business, Apache Spark, Privacy

Αφιερώνω την προσπάθειά μου στους γονείς μου, που προσέφεραν τη δυνατότητα σπουδών καθώς και για την υποστήριξη τους.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω την οικογένεια μου, η οποία μου προσέφερε τη δυνατότητα να σπουδάσω και να φτάσω σήμερα στο σημείο να παραδίδω την πτυχιακή αυτή εργασία. Επιπρόσθετα θα ήθελα να τους ευχαριστήσω για την πολύτιμη ψυχολογική στήριξη καθ' όλη την διάρκεια των σπουδών μου.

θα ήθελα να ευχαριστήσω όλους τους καθηγητές και τις καθηγήτριες που με δίδαξαν στα μαθήματα της σχολής και που μου έδωσαν τα κατάλληλα κίνητρα και τις απαραίτητες γνώσεις ώστε να φθάσω στο στάδιο της εκπόνησης της πτυχιακής μου εργασίας. Τέλος, θα ήθελα να ευχαριστήσω τόσο την οικογένεια μου όσο και τους φίλους μου για την κατανόηση και τη συμπαράσταση που έδειξαν ολόκληρη την περίοδο της εκπόνησης της εργασίας αυτής.

Περιεχόμενο Δεδομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Πρόλογος

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Πελοποννήσου στο τμήμα της Λογιστικής και Χρηματοοικονομικής. Αντικείμενο της εργασίας είναι η μελέτη των μεγάλων δεδομένων και η σημασία τους για τις εταιρίες και τους οργανισμούς.

Τα μεγάλα δεδομένα αποτελούν ένα από τα πιο καυτά θέματα στη σημερινή εποχή και η σημασία τους για τις εταιρίες και τους οργανισμούς είναι σημαντική τόσο όσον αφορά το marketing, αλλά και όσον αφορά την αύξηση της ποιότητας των προϊόντων και των υπηρεσιών τους.

Στα πλαίσια της εργασίας αυτής θα γίνει περιγραφή της έννοιας των μεγάλων δεδομένων και της αξίας τους για τις επιχειρήσεις. Επιπρόσθετα θα γίνει αναφορά σε ορισμένες από τις σημαντικότερες προκλήσεις που δημιουργούνται, όπως είναι η δυνατότητα επεξεργασίας ενός τεράστιου και πολύπλοκου όγκου δεδομένων και η προστασία των προσωπικών δεδομένων των χρηστών.

Εισαγωγή

Επί του παρόντος, μεγάλη ποσότητα δεδομένων δημιουργείται κάθε μέρα. Με τη ραγδαία αύξηση των δεδομένων μετακινούμαστε από την εποχή των Terabyte στην εποχή των Petabyte. Στο Facebook υπάρχουν αρκετές χιλιάδες tables αποθήκευσης δεδομένων με πάνω από 700 terabyte δεδομένων [Ashish Thusoo κ.α., 2009]. Όσο το μέγεθος των δεδομένων συνεχίζει να αυξάνεται, οι εφαρμογές αναγκάζονται να αναζητούν όλο και περισσότερους υπολογιστικούς πόρους και πόρους αποθήκευσης. Ωστόσο, η αύξηση τόσο των δεδομένων πραγματικού χρόνου, όσο και των ιστορικών δεδομένων αυξάνεται με γρηγορότερο ρυθμό από ότι οι υπολογιστικοί πόροι. Ως αποτέλεσμα αυτού, οι ερευνητές συνειδητοποιούν ότι η παράλληλη επεξεργασία είναι ο μόνος τρόπος για την αντιμετώπιση αυτού του γεγονότος.

Η σημασία των μεγάλων δεδομένων έχει αρχίζει να γίνεται εμφανείς σε διάφορες επιχειρήσεις και οργανισμούς, όπως είναι η ανώτερη εκπαίδευση. Η στρατηγική χρήση και οι εφαρμογές των big data στην ανώτερη εκπαίδευση θα οδηγούσαν σε ανώτερη εκπαιδευτική ποιότητα και καλύτερη εμπειρία φοιτητών αλλά και προσωπικού. Σημαντική επίσης είναι η συμβολή των μεγάλων δεδομένων στην υγεία, με σημαντικότερη ίσως πρόκληση των ανάπτυξη αλγορίθμων μηχανικής μάθησης, όπου θα προβλέπουν ασθένειες και θα προτείνουν θεραπείες με μεγαλύτερη ακρίβεια από ότι μπορεί να πετύχει ο τομέας της ιατρικής τη σήμερον μέρα.

Η συνεχόμενη αύξηση των μεγάλων δεδομένων έχει οδηγήσει την επιστημονική κοινότητα στην δημιουργία frameworks για τη διαχείριση των δεδομένων. Ορισμένα από τα σημαντικότερα open source Big Data frameworks σήμερα είναι το Apache Spark [apache-spark], Hadoop MapReduce [map-reduce] και το Apache Flink [flink].

Η πτυχιακή εργασία είναι οργανωμένη ως εξής: Στο κεφάλαιο 1 θα γίνει προσδιορισμός της έννοιας των Big Data και της σημασίας τους για την κοινωνία. Στη συνέχεια, στο κεφάλαιο 2 θα γίνει αναφορά παραδειγμάτων όσον αφορά τον τρόπο αξιοποίησης των μεγάλων δεδομένων από συγκεκριμένες επιχειρήσεις και οργανισμούς. Στο κεφάλαιο 3 θα γίνει αναφορά στην εξόρυξη δεδομένων, ενώ στο κεφάλαιο 4 θα γίνει αναφορά στο πιο διάσημο Big Data framework στις μέρες μας, το Spark. Τέλος στο κεφάλαιο 5 θα γίνει αναφορά σε μία πολύ σημαντική συνιστώσα, την ιδιωτικότητα, και πως αυτή απειλείται από την μη ασφαλή διαχείριση των διαθέσιμων δεδομένων.

1. Big Data

1.1. Η πηγή των μεγάλων δεδομένων: Πληροφορία

Ο πρώτος λόγος πίσω από τη γρήγορη επέκταση των μεγάλων δεδομένων είναι ο εκτεταμένος βαθμός με τον οποίον δημιουργούνται, μοιράζονται και χρησιμοποιούνται τα δεδομένα τα τελευταία χρόνια. Η ψηφιοποίηση, δηλαδή ο μετασχηματισμός του αναλογικού σήματος σε ψηφιακό, έγινε πολύ δημοφιλής στις αρχές της δεκαετίας του '90. Εκείνη την εποχή, ξεκίνησε η δημιουργία πολύ δημοφιλών εργαλείων που χρησιμοποίησαν τα ψηφιακά δεδομένα και αποτέλεσαν την πρώτη μαζική ομάδα ψηφιοποίησης. Ένα πολύ χαρακτηριστικό παράδειγμα είναι το Google Books Library Project (2015) [google-books-library], το οποίο ξεκίνησε το 2004 και ο βασικός του στόχος ήταν η πλήρης ψηφιοποίηση πάνω από 15 εκατομμύρια βιβλίων που υπήρχαν στις βιβλιοθήκες πανεπιστημίων, συμπεριλαμβανομένων του Harvard, του Stanford και του Oxford.

Μόλις τα σήματα μετατραπούν σε ψηφιακή μορφή, μπορούν να οργανωθούν σε πιο δομημένα σύνολα δεδομένων. Αυτό το περαιτέρω βήμα, το οποίο οι Mayer-Schönberger και Cukier (2013) ονόμασαν «datafication» [datafication], είναι σε θέση να προσφέρει μια μοναδική μακροοικονομική προσέγγιση για τη μελέτη σχετικών τάσεων και προτύπων, το οποίο θα ήταν αδύνατο να επιτευχθεί, εάν όλα τα δεδομένα παρέμεναν σε αναλογική μορφή. Στην περίπτωση του προαναφερθέντος προγράμματος μαζικής ψηφιοποίησης της Google, η επεξεργασία δεδομένων ξεκίνησε όταν ένα τεράστιο ποσό από συμβολοσειρές κειμένων μετατράπηκε σε αλληλουχίες συνεχών λέξεων (n-grams), για τις οποίες ήταν εφικτή η παρακολούθηση του επιπέδου εμφάνισης κατά τη διάρκεια πέρασης του χρόνου. Με τον τρόπο αυτό, οι ερευνητές κατάφεραν να βρουν πληροφορίες για διαφορετικά πεδία, όπως είναι η γλωσσολογία, η ετυμολογία, η κοινωνιολογία και η ιστορική επιδημιολογία, χρησιμοποιώντας τα σύνολα δεδομένων των Google's Books [Michel κ.α., 2011].

Η ιεραρχία των data-information-knowledge-wisdom προσφέρει μια εναλλακτική μεθοδολογία, σύμφωνα με την οποία οι πληροφορίες εμφανίζονται ως δεδομένα που είναι δομημένα κατά τρόπο που να είναι χρήσιμα και συναφή προς ένα συγκεκριμένο σκοπό [Rowley, 2007]. Με αυτή την προσέγγιση, η πληροφορία μετατρέπεται σε πλούτο γνώσης που μπορεί να δημιουργήσει αξία για τις επιχειρήσεις [Cricelli and Grimaldi, 2008]. Ως εκ τούτου, μπορούμε να συμπεράνουμε ότι η πληροφορία και όχι τα δεδομένα, αποτελεί το θεμελιώδες καύσιμο του σημερινού φαινομένου των Big Data.

Σύμφωνα με τον Prescott (2013), οι κατάλογοι βιβλιοθηκών μπορεί να θεωρηθούν ότι αντιπροσωπεύουν μια πρώιμη συνάντηση με τα μεγάλα δεδομένα. Στην πραγματικότητα, οι κατάλογοι βιβλιοθηκών χαρακτηρίζονται

επίσης από ένα ορισμένο επίπεδο "ετερογένειας" λόγω ανθρώπινων λαθών και την ανάπτυξη διαφορετικών προτύπων για την καταλογογράφηση με την πάροδο του χρόνου. Οι μέθοδοι μεγάλων δεδομένων μπορούν να χρησιμοποιηθούν για τον εντοπισμό των διαφόρων διαδικασιών καταλογογράφησης των περιουσιακών στοιχείων της βιβλιοθήκης με την πάροδο του χρόνου και για την εξεύρεση νέων ασυμφωνιών στα δεδομένα. Για παράδειγμα, διαφορετικά “στρώματα” δεδομένων μπορούν να αναγνωριστούν στον κατάλογο της βρετανικής βιβλιοθήκης λόγω της προοδευτικής αναδιαμόρφωσης: “Οι τεχνικές των μεγάλων δεδομένων μπορούν να επιτρέψουν κάτι σαν μια «αρχαιολογία» δεδομένων σε καταλόγους βιβλιοθηκών”.

Παρατηρούμε μια ισχυρή σύνδεση μεταξύ των repositories των μεγάλων δεδομένων και των ψηφιακών βιβλιοθηκών (DLs). Σύμφωνα με τον Candela (2007):

> τα DL είναι οργανώσεις που μπορεί να είναι εικονικές, να συλλέγουν, να διαχειρίζονται και να διατηρούν το μακροπρόθεσμο πλούσιο ψηφιακό περιεχόμενο και να προσφέρουν στις κοινότητες χρηστών τους εξειδικευμένες λειτουργίες σχετικά με αυτό το περιεχόμενο, σε κωδικοποιημένες πολιτικές.

Σύμφωνα με τον Jansen κ.α. (2013), η ετερογένεια του περιεχομένου που μπορεί να βρεθεί σε μια ψηφιακή βιβλιοθήκη, και η οποία κυμαίνεται από τις ψηφιοποιημένες εκδόσεις των τυπωμένων βιβλίων σε περιεχόμενα που παράχθηκαν ψηφιακά, απαιτεί προηγμένη τεχνολογία διαχείρισης δεδομένων.

Ένα ιδιαίτερο στοιχείο πολυπλοκότητας προέρχεται από το γεγονός ότι το ψηφιακό περιεχόμενο μπορεί να βρίσκεται σε διαφορετικά επίπεδα συντακτικής και σημασιολογικής αφαίρεσης. Οι τεχνικές των μεγάλων δεδομένων παρέχουν αρκετή ευελιξία για να αντιμετωπίσουν τέτοιες εγγενώς ετερογενείς πληροφορίες. Όταν το μέγεθος ενός DL όσον αφορά τον όγκο, τη ταχύτητα και την ποικιλία του περιεχομένου, της βάσης των χρηστών ή οποιασδήποτε άλλης πτυχής απαιτεί "εξειδικευμένες τεχνολογίες ή προσεγγίσεις", τότε μπαίνουμε στον τομέα των πολύ μεγάλων DLs (VLDLs) [Candela κ.α., 2012] . Είναι ενδιαφέρον να παρατηρήσουμε πως ο ορισμός των VLDLs είναι συνεπής με αυτόν που προτείνουμε για τα Big Data. Αυτό δηλώνει το πόσο σημαντική είναι η τεχνολογία των μεγάλων δεδομένων και οι μέθοδοι αυτών και είναι αναγκαία για τη συνεχή ανάπτυξη των βιβλιοθηκών και της διαχείρισης των πληροφοριών.

Ένας επιπρόσθετος σημαντικός λόγος για την αυξανόμενη διαθεσιμότητα των πληροφοριών είναι ο πολλαπλασιασμός των προσωπικών συσκευών που είναι συνδεδεμένες με το Διαδίκτυο και οι οποίες διαθέτουν ψηφιακούς αισθητήρες (όπως κάμερες, συσκευές εγγραφής ήχου και εντοπιστές GPS). Τέτοιοι αισθητήρες καθιστούν δυνατή την ψηφιοποίηση, ενώ η σύνδεση τους στο διαδίκτυο επιτρέπει τη συλλογή δεδομένων, τη μετατροπή και, τελικά, την οργάνωση τους ως πληροφορία. Εκτιμάται ότι σε κάποια χρονική στιγμή μεταξύ του 2008 και του 2009 η ποσότητα των συνδεδεμένων συσκευών

ξεπέρασε τον αριθμό των ανθρώπων [Evans, 2011] και ο Gartner (2014) προέβλεπε ότι μέχρι το 2020 θα υπάρχουν 26 δισεκατομμύρια συσκευές στη γη, το οποίο θα ισοδυναμεί σε 3 συσκευές ανά ζωντανό άτομο. Το σενάριο στο οποίο τα τεχνητά αντικείμενα, εξοπλισμένα με μοναδικά αναγνωριστικά στοιχεία, αλληλεπιδρούν μεταξύ τους με κοινούς στόχους, χωρίς καμία ανθρώπινη αλληλεπίδραση, εμπίπτουν στο όνομα του Διαδικτύου των πραγμάτων, του IoT [Atzori κ.α., 2010, Estrin κ.α., 2002], και αποτελεί μια πολλά υποσχόμενη πηγή πληροφοριών στην εποχή των Big Data. Ένας αυξανόμενος όγκος δεδομένων θα προκύψει επίσης από τη συνεργασία μεταξύ επιχειρήσεων μέσω εργαλείων που βασίζονται στο Internet [Michelino κ.α., 2008].

Ένα σημαντικό χαρακτηριστικό των δεδομένων που παράγονται και χρησιμοποιούνται σήμερα είναι η επέκταση της ποικιλίας της μορφής αυτών. Παραδοσιακοί αλφαριθμητικοί πίνακες ξεπερνούνται από την αυξανόμενη διαθεσιμότητα λιγότερο δομημένων πηγών δεδομένων, όπως βίντεο, εικόνες και κείμενων που παράγονται από ανθρώπους [Russom, 2011]. Η πολλαπλότητα των τύπων δεδομένων και η συνύπαρξή τους είναι μία από τις σημαντικότερες προκλήσεις που σχετίζονται με το χειρισμό των μεγάλων δεδομένων σήμερα [Manyika κ.α., 2011].

1.2. Απαραίτητη προϋπόθεση για τη χρήση του Big Data: Τεχνολογία

Μια σημαντική πτυχή που αφορά τα μεγάλα δεδομένα και συναντάται έντονα στη βιβλιογραφία σχετίζεται με τα συγκεκριμένα τεχνολογικά ζητήματα που συμβαδίζουν με τη χρήση της μεγάλης ποσότητας δεδομένων. Η επεξεργασία των μεγάλων δεδομένων με τη «σωστή» ταχύτητα συνεπάγεται σημαντικές υπολογιστικές απαιτήσεις και απαιτήσεις αποθήκευσης που ένα μέσο σύστημα πληροφορικής μπορεί να μην είναι σε θέση να προσφέρει.

Το Hadoop [hadoop] είναι ένα framework ανοιχτού κώδικα που σχεδιάστηκε ειδικά για να υποστηρίξει την επεξεργασία των μεγάλων δεδομένων. Τα βασικά συστατικά του Hadoop είναι το HDFS [hdfs] και το MapReduce. Και οι δύο αυτές τεχνολογίες αναπτύχθηκαν αρχικά από την Google (Ghemawat κ.α., 2003), προτού γίνει open source κάτω από την αιγίδα της Apache [apache]. Το HDFS (Hadoop Distributed File System) επιτρέπει σε πολλαπλούς, εξ' αποστάσεως υπολογιστές να συνεργαστούν απρόσκοπτα προς ένα κοινό υπολογιστικό στόχο [Shvachko κ.α., 2010]. Αντίθετα, το MapReduce αποτελεί ένα μοντέλο προγραμματισμού που αποσκοπεί στην αποτελεσματική διάσπαση των λειτουργιών σε ξεχωριστές λογικές μονάδες [Dean and Ghemawat, 2008].

Το Hadoop και το MapReduce έχουν αποδειχθεί πολύ αποτελεσματικά όσον αφορά την εξερεύνηση και τη διαχείριση μεταδεδομένων μεγάλων βιβλιοθηκών. Ο Powell (2012) πρότεινε μια εφαρμογή για την εξαγωγή και την

αντιστοίχιση ονομάτων δημιουργών αξιοποιώντας τη τεχνολογία των μεγάλων δεδομένων. Ένα ακόμη παράδειγμα που δηλώνει την αποτελεσματικότητα των frameworks των μεγάλων δεδομένων για την υλοποίηση τεράστιων DIs είναι το PuntStore [Wang κ.α., 2013]. Το PuntStore υποστηρίζει την ενσωμάτωση πολλών μηχανών αποθήκευσης και μηχανών αναζητήσεων για να μεγιστοποιήσει την αποτελεσματικότητα της χρήσης μεγάλων συλλογών ψηφιακών αρχείων αντικειμένων.

Εκτός από την πολυπλοκότητα που προκύπτει όσον αφορά την επεξεργασία των μεγάλων δεδομένων, ένα ακόμη θεμελιώδες τεχνολογικό ζήτημα, που προκαλείται λόγω της διασκορπισμένης φύσης των μηχανών, είναι η μετάδοσή τους. Τα δίκτυα επικοινωνίας πρέπει να υποστηρίξουν μεγαλύτερες και ταχύτερες μεταφορές δεδομένων και τα συστήματα απαιτούν ειδικές τεχνικές συγκριτικής αξιολόγησης για την αξιολόγηση της συνολικής τους απόδοση [Xiong κ.α., 2013].

Μια πρόσθετη τεχνολογική απαίτηση που συνδέεται με τη χρήση των μεγάλων δεδομένων είναι η ικανότητα αποθήκευσης μεγαλύτερης έκτασης δεδομένων σε μικρότερες συσκευές. Ο νόμος του Moore (2006) αναφέρει ότι ο αριθμός των τρανζίστορ που μπορούν να τοποθετηθούν σε ένα τσιπ πυριτίου τείνει να διπλασιάζεται κάθε 18 έως 24 μήνες και αυτό σημαίνει ότι η χωρητικότητα αποθήκευσης μνήμης αυξάνεται εκθετικά. Ωστόσο, τα δεδομένα αυξάνονται επίσης εκθετικά [Hilbert and López, 2011] και το ζήτημα της αποθήκευσης μεγάλου όγκου δεδομένων εξακολουθεί να αποτελεί μία κρίσιμη τεχνολογική πρόκληση στην εποχή των μεγάλων δεδομένων.

1.3. Τεχνικές επεξεργασίας μεγάλων δεδομένων: Μέθοδοι

Τα τεράστια ποσά δεδομένων πρέπει να υποβάλλονται σε επεξεργασία με πιο σύνθετες μεθόδους από τις συνήθεις στατιστικές διαδικασίες. Δυστυχώς, μια συγκεκριμένη αρμοδιότητα σχετικά με τις δυνατότητες και τους περιορισμούς αυτών των τεχνικών δεν είναι άμεσα προσβάσιμη στην αγορά εργασίας επί του παρόντος.

Οι αναλυτικές μέθοδοι των μεγάλων δεδομένων έχουν επισημανθεί από τους Manyika κ.α. (2011) και Chen κ.α. (2012). Έχουν αποκτήσει μια λίστα με τις πιο συνηθισμένες διαδικασίες που περιλαμβάνουν: ανάλυση συμπλέγματος, γενετικούς αλγορίθμους, φυσική επεξεργασία γλώσσας, μηχανική μάθηση, νευρωνικά δίκτυα, πρόβλεψη μοντέλων, μοντέλα παλινδρόμησης, ανάλυση κοινωνικών δικτύων, ανάλυση συναισθημάτων, επεξεργασία σήματος και οπτικοποίηση δεδομένων.

Σύμφωνα με τους Chen κ.ά. (2012), δεδομένου του σημερινού δομημένου επιχειρηματικού περιβάλλοντος που επικεντρώνεται στις πληροφορίες, οι επιχειρήσεις πρέπει να επενδύσουν σε διεπιστημονική εκπαίδευση επιχειρηματικών πληροφοριών και αναλύσεων, προκειμένου να καλύψουν τις "κρίσιμες αναλυτικές δεξιότητες και δεξιότητες πληροφορικής, γνώση

επιχειρήσεων και τομέα και δεξιότητες επικοινωνίας". Ταυτόχρονα, η διαδικασία αυτή πρέπει να συνοδεύεται από μια πολιτισμική αλλαγή, με τη συμμετοχή ολόκληρου του πληθυσμού της εταιρείας, προτρέποντας τα μέλη της να «διαχειρίζονται αποτελεσματικά τα δεδομένα σωστά και να τα ενσωματώνουν στις διαδικασίες λήψης αποφάσεων» [Buhl κ.α., 2013].

Νέες επαγγελματικές δεξιότητες θα μπορούσαν να προκύψουν από μια τέτοια καινοτόμο εκπαίδευση που θα βοηθούσε στην κατάρτιση εμπειρογνωμόνων για την αφομοίωση των διαφόρων κλάδων [Mayer-Schönberger και Cukier, 2013]. Αυτοί οι επιστήμονες δεδομένων μπορούν να θεωρηθούν ως υβριδικοί ειδικοί που μπορούν να διαχειριστούν τόσο την τεχνολογική γνώση όσο και την ακαδημαϊκή έρευνα [Davenport και Patil, 2012]. Υπάρχει ένα κενό στην εκπαίδευση για αυτό το επαγγελματικό προφίλ [Manyika κ.α., 2011] και απαιτούνται νέα παραγωγικά μαθήματα και μέθοδοι μάθησης για τη διδασκαλία μελλοντικών ειδικών δεδομένων.

Επιπλέον, πρέπει να σημειωθεί ότι η ανάπτυξη των μεγάλων δεδομένων έχει αλλάξει τη μέθοδο της λήψης αποφάσεων από μια στατική διαδικασία σε μια δυναμική. Πράγματι, η ανάλυση των σχέσεων μεταξύ των πολλών γεγονότων που προέρχονται από τα δεδομένα πληροφοριών αντικατέστησε την αναζήτηση παραδοσιακών, λογικών συνδέσεων. Είναι λογικό να υποθέσουμε ότι η συνέπεια της εφαρμογής των μεγάλων δεδομένων σε εταιρείες, ερευνητικά και πανεπιστημιακά ιδρύματα θα μπορούσε να τροποποιήσει τόσο τους κανόνες λήψης αποφάσεων [McAfee κ.α., 2012] όσο και την επιστημονική μέθοδο [Anderson, 2007].

Η μάθηση σχετικά με τα ισχυρά και τα αδύνατα σημεία της εφαρμογής των μεθόδων Big Data αντιπροσωπεύει έναν αναμφισβήτητο πόρο για τα δημόσια και τα ιδιωτικά ιδρύματα κατά τη διεξαγωγή στρατηγικών διαδικασιών λήψης αποφάσεων [Boyd and Crawford, 2012]. Είναι προφανές ότι η διορατικότητα των μελλοντικών δυνατοτήτων που προωθούνται από τις εφαρμογές του μεγάλων δεδομένων θα πρέπει να επαληθευτεί προσεκτικά, με απόλυτη γνώση της πολυπλοκότητάς τους.

1.4. Ο Αντίκτυπος των μεγάλων δεδομένων στη ζωή μας

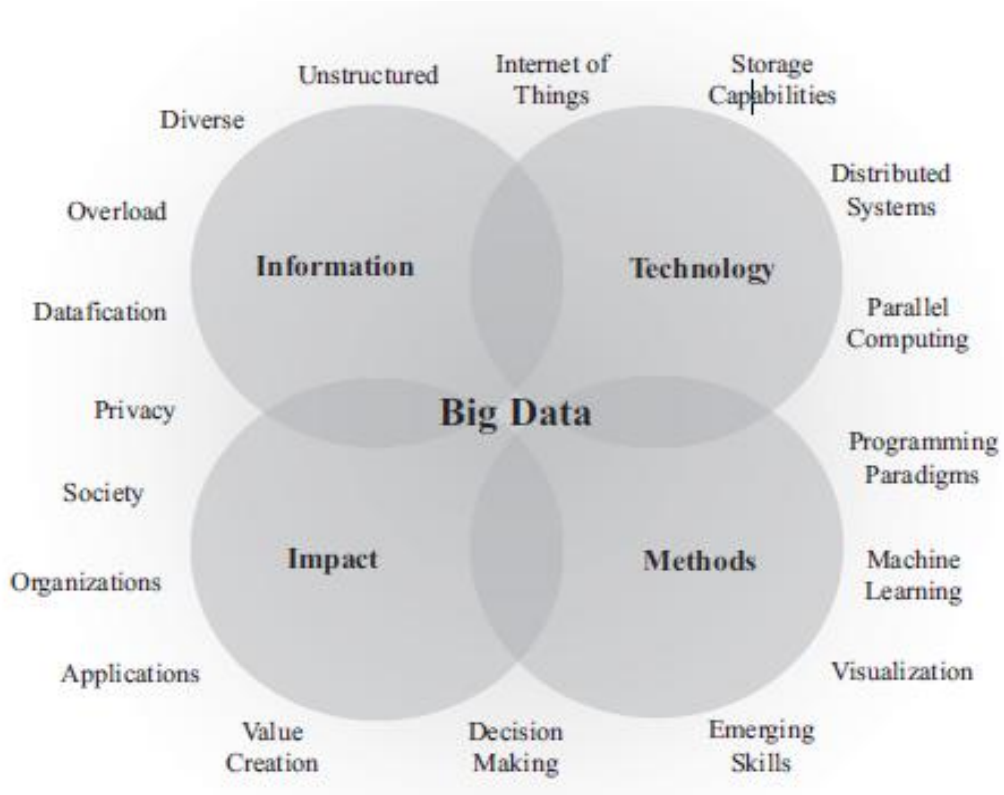
Η αξιοποίηση και η διαχείριση των μεγάλων δεδομένων επηρεάζουν πολλούς τομείς δραστηριοτήτων της κοινωνίας μας. Οι εφαρμογές των μεγάλων δεδομένων έχουν δείξει ένα συνεκτικό επίπεδο προσαρμοστικότητας στις διαφορετικές απαιτήσεις που απορρέουν από διαφορετικούς επιστημονικούς τομείς και βιομηχανικούς οργανισμούς. Τα προβλήματα που προέρχονταν από πολύ απομακρυσμένες περιοχές λύθηκαν μερικές φορές χρησιμοποιώντας τις ίδιες τεχνικές και τύπους δεδομένων. Ένα παράδειγμα αυτού είναι η εφαρμογή της ανάλυσης του συσχετισμού στα αρχεία καταγραφής αναζήτησης της Google που παρήγαγαν γνώμες που εφαρμόζονται σε μια σειρά τομέων, από την επιδημιολογία έως και την

οικονομία [Askitas και Zimmermann, 2009, Ginsberg κ.α., 2009, Guzman, 2011].

Τα μεγάλα δεδομένα ταυτόχρονα αποτελούν πηγή ανησυχίας, καθώς η ταχεία ανάπτυξή τους προηγήθηκε της θέσπισης εξαντλητικών οδηγιών για την προστασία των ιδιωτικών πληροφοριών [Boyd και Crawford, 2012]. Για παράδειγμα, είναι απαραίτητο να αποφευχθεί οποιαδήποτε πιθανή αναγνώριση των προσωπικών δεδομένων μέσω αλγορίθμων ανωνυμοποίησης με στόχο την υπεράσπιση της ιδιωτικής ζωής.

Επιπλέον, η προσβασιμότητα των πληροφοριών θα πρέπει να ρυθμίζεται σωστά και αμερόληπτα, ώστε να αποφεύγονται οι αντί-ανταγωνιστικές επιχειρηματικές πρακτικές [Manovich, 2012] που θα μπορούσαν να ενισχύσουν τις δεσπόζουσες θέσεις στην αγορά. Η δημιουργία ενός νέου ψηφιακού χάσματος μεταξύ των επιχειρήσεων, λόγω του διαφορετικού επιπέδου πρόσβασης στα δεδομένα, αποτελεί πιθανό εμπόδιο στην πρόοδο της καινοτομίας [Boyd και Crawford, 2012].

Τα Big Data επηρεάζουν επίσης τις επιχειρήσεις σε βάθος, καθώς αναγκάζονται να επανεξετάσουν την οργάνωσή τους και όλες τις επιχειρησιακές διαδικασίες τους υπό το πρίσμα της διαθεσιμότητας νέων πληροφοριών που θα μπορούσαν να μετατραπούν σε ανταγωνιστικό πλεονέκτημα σε μια αγορά δεδομένων [McAfee κ.α. 2012, Pearson and Wegener, 2013] (Εικόνα: Θέματα που αφορούν τα μεγάλα δεδομένα).



Εικόνα 1: Θέματα που αφορούν τα μεγάλα δεδομένα

1.5. Τι είναι τα Big Data?

Οι αναδυόμενοι κλάδοι συχνά αντιμετωπίζουν έλλειψη συμφωνίας όσον αφορά τον ορισμό των βασικών εννοιών. Πράγματι, το επίπεδο συναίνεσης που επιδεικνύει μια επιστημονική κοινότητα κατά τον ορισμό μιας έννοιας είναι ένα υποκατάστατο της ανάπτυξης ενός πειθαρχείου [Ronda-Pupo και Guerras-Martin, 2012]. Η γρήγορη και χαοτική εξέλιξη της βιβλιογραφίας των μεγάλων δεδομένων έχει εμποδίσει την ανάπτυξη ενός καθολικά και τυπικά αποδεκτού ορισμού όσον αφορά τα μεγάλα δεδομένα. Στην πραγματικότητα, αν και αρκετοί συγγραφείς έχουν προτείνει τους δικούς τους ορισμούς για τα Big Data [Beyer και Laney, 2012, Dijcks, 2013, Dumbill, 2013, Mayer-Schönberger και Cukier, 2013, Schroeck κ.α., 2012, Shneiderman, Suthaharan, 2014 και Ward and Barker, 2013], καμία από τις προτάσεις αυτές δεν εμπόδισε τα μεταγενέστερα έργα να τροποποιήσουν ή να αγνοήσουν τους προηγούμενους ορισμούς και να προτείνουν μία νέα [Ward and Barker, 2013]. Αυτή η έλλειψη συμφωνίας και ομοιογένειας, αν και δικαιολογείται από τη σχετική νεότητα των μεγάλων δεδομένων ως έννοια, περιορίζει την ορθή ανάπτυξη του πειθαρχικού πλαισίου.

2. Παραδείγματα εφαρμογών των Big Data

2.1. Τα Big Data και οι ερευνητικές επιπτώσεις για την ανώτερη εκπαίδευση στο Ηνωμένο Βασίλειο

Ο Vincent Koon Ong (2016) χρησιμοποιώντας την παραγωγή από τα UK JISC's BI projects, κάνει κριτική και δίνει σε γενικές γραμμές μερικές περιπτώσεις ανάλυσης των Big data στα ανώτερα εκπαιδευτικά ιδρύματα του Ηνωμένου Βασιλείου, καθώς και κάποιες ερευνητικές επιπτώσεις για την μελλοντική έρευνα των big data στην ανώτερη εκπαίδευση.

Με την αυξανόμενη δημιουργία τεράστιων ποσών από δομημένα και αδόμητα data σε ένα όλο και πιο διανεμημένο επιχειρηματικό περιβάλλον, οι οργανισμοί αναγκάζονται να αναζητήσουν νέες τεχνικές λύσεις και διαχειριστικές προσεγγίσεις για να διαχειριστούν τα δεδομένα. Ως αποτέλεσμα αυτού, οι έννοιες και οι εφαρμογές των μεγάλων δεδομένων έχουν κατακτήσει τον επιχειρηματικό κόσμο. Οι οργανισμοί που διαχειρίζονται τα μεγάλα δεδομένα έχουν αρχίσει να αποκομίζουν απτά και μη απτά οφέλη από αυτά. Για παράδειγμα η συλλογή των μεγάλων δεδομένων δίνει στον οργανισμό την ικανότητα να προσφέρει σε κάθε πελάτη μία εξατομικευμένη εμπειρία και με αυτή την εμπειρία ο οργανισμός μπορεί να βελτιώσει τη διαχείριση της σχέσης του με τον πελάτη. Ο Chen κ.α. (2012) αναγνωρίζουν πέντε περιοχές ως περιοχές «μεγάλης επιρροής» στην έρευνα των Big data:

1) Ηλεκτρονικό εμπόριο και market intelligence

2) Ηλεκτρονική κυβέρνηση και πολιτική

3) Επιστήμη και τεχνολογία

4) Smart health και ευεξία

5) Προστασία και δημόσια ασφάλεια.

Η ανώτερη εκπαίδευση είναι μια άλλη περιοχή που θα πρέπει να προστεθεί στην λίστα κατά τον Vincent Koon Ong. Σύμφωνα με την τελευταία μεγάλη έρευνα των Πανεπιστημίων του Ηνωμένου Βασιλείου στην επιρροή του τομέα της ανώτερης εκπαίδευσης πάνω στην οικονομία του Ηνωμένου Βασιλείου το 2011-2012, τα Πανεπιστήμια τώρα παράγουν 73 δις Λίρες σε παραγωγή – πάνω, κατά 24% από τα 59 δις Λίρες που παρήγαγαν όταν δημοσιεύθηκε η τελευταία έρευνα το 2009. Αυτό θέτει την ανώτερη εκπαίδευση μπροστά από πολλούς άλλους τομείς στο Ηνωμένο Βασίλειο, συμπεριλαμβανομένων της διαφήμισης και της έρευνας της αγοράς, τις νόμιμες υπηρεσίες, την κατασκευή υπολογιστών, βασικών φαρμάκων και των αέριων μεταφορών. Τα

Πανεπιστήμια επίσης παράγουν περισσότερα GDP ανά μονάδα δαπάνης από πολλούς άλλους τομείς συμπεριλαμβανομένων της υγείας, της δημόσιας διοίκησης και των κατασκευών.

Ο τομέας της ανώτερης εκπαίδευσης βασίζεται πολύ στα δεδομένα των φοιτητών για να λάβει σημαντικές και στρατηγικές αποφάσεις. Τα κολέγια και τα πανεπιστήμια έχουν συλλέξει και εντοπίσει περισσότερα δεδομένα φοιτητών από ποτέ πριν, από την εισαγωγή των φοιτητών ως την αναχώρηση των φοιτητών, ακόμα και μετά την αναχώρηση, όπως δεδομένα αιτήσεων, δεδομένα εγγραφής σε μαθήματα, δεδομένα παρακολούθησης, δεδομένα online learning, δεδομένα επίδοσης, δεδομένα εκτός προγράμματος, δεδομένα πρακτικής και εργασιακής απασχόλησης. Η στρατηγική χρήση και οι εφαρμογές των μεγάλων δεδομένων στην ανώτερη εκπαίδευση θα μπορούσε να οδηγήσει σε ανώτερη εκπαιδευτική ποιότητα και καλύτερη εμπειρία για φοιτητές και προσωπικό. Ωστόσο, η ανάλυση των Big data γίνεται κυρίως για να ικανοποιήσει πιστοποιήσεις ή απαιτούμενα αντί να απευθύνεται σε στρατηγικά θέματα και τα περισσότερα από τα δεδομένα που συλλέγονται δε χρησιμοποιούνται καθόλου.

Το UK JISC, μία αναγνωρισμένη φιλανθρωπική οργάνωση και πρωταθλητής στην χρήση της ψηφιακής τεχνολογίας στην εκπαίδευση και την έρευνα στο Ηνωμένο Βασίλειο, ξεκίνησε το πρόγραμμα BUSINESS INTELLIGENCE PROGRAMME ανάμεσα στο 2011-2012 ως μέρος της στρατηγικής του JISC να βοηθήσει τα ιδρύματα να αναπτυχθούν και να χρησιμοποιούν τα εταιρικά και επιχειρησιακά συστήματα αποτελεσματικά. Το Business intelligence programme στοχεύει στο να βοηθήσει περισσότερο τα εκπαιδευτικά ιδρύματα να αναγνωρίσουν την ωριμότητά τους στην επιχειρηματική νοημοσύνη «Business intelligence (BI)» και να φτιάξουν τα κατάλληλα BI συστήματα ή λύσεις τις οποίες θα παρέχουν στους ανώτερους διευθυντές καλύτερη και πιο έγκαιρη πρόσβαση σε ακριβή δεδομένα με αποτέλεσμα βελτιωμένες προβλέψεις, συγκριτική αξιολόγηση και άλλες αναφορές έτσι ώστε να επιτευχθούν οι επιχειρησιακοί στόχοι.

2.1.1. Big data analytics

Ο Goes (2014) βλέπει ευκαιρίες για IS έρευνα στον τομέα των μεγάλων δεδομένων σε 3 επίπεδα :

1) ΥΠΟΔΟΜΗ ΤΩΝ BIG DATA : Αυτό εστιάζει σε τεχνικά θέματα που σχετίζονται με τη δημιουργία και συλλογή των δεδομένων, τα 4 V (volume, velocity, variety, veracity) των πηγών των δεδομένων, την εκτέλεση και διαχείριση των μεγάλων δεδομένων.

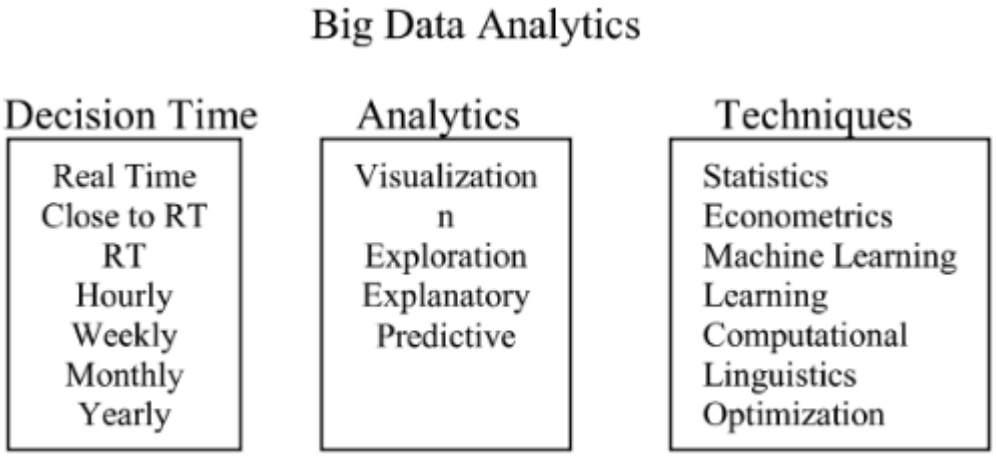
2) BIG DATA ANALYTICS : Αυτό απαιτεί την διαδικασία και τεχνικές data driven για υποστήριξη της λήψης αποφάσεων μέσω της διορατικότητας που κερδίζεται από την ανάλυση των δεδομένων

3) Μεταμόρφωση και επιρροή: Αυτό πληροφορεί τους οργανισμούς ή τα άτομα για θέματα όπως IT Business alignment, τη διαδικασία μεταμόρφωσης επιχείρησης, την ανάπτυξη και χρήση των συστημάτων μεγάλων δεδομένων.

Η ανάλυση των big data βασίζεται στην συλλογή των δεδομένων και στατιστική ανάλυση, και είναι στενά συνδεδεμένη με την επιχειρηματική νοημοσύνη στο πλαίσιο της λήψης αποφάσεων που καθοδηγείται από τα δεδομένα.

Για τον λόγο αυτό η ανάλυση των big data θα ήταν πιο βασική για τα ανώτερα εκπαιδευτικά ιδρύματα. Η μελέτη του Bischel (2012) δείχνει ότι η ανάλυση των δεδομένων είναι ένα ενδιαφέρον θέμα ή μία βασική προτεραιότητα για τα περισσότερα Πανεπιστήμια και μπορεί να βοηθήσει σημαντικά την πρόοδο ενός ιδρύματος σε τέτοιες στρατηγικές περιοχές όπως την κατανομή των πηγών, την επιτυχία των φοιτητών και τα οικονομικά.

Ο Goes (2014) παρουσιάζει μία ταξινομία για να περιγράψει τις εργασίες έρευνας στην ανάλυση των big data όπως φαίνεται στην εικόνα: «Ταξινομία της ανάλυσης των Big Data». Αυτή η ταξινομία θα χρησιμοποιηθεί για να γίνει κριτική στα 11 projects BI του JISC στο Ηνωμένο Βασίλειο.



Εικόνα 2: Ταξινομία της ανάλυσης των Big Data (Goes, 2014)

Στα πλαίσια της ανώτερης εκπαίδευσης, ο χρόνος απόφασης είναι βασικός καθώς τα δεδομένα των φοιτητών παράγονται σε αληθινό χρόνο, όπως αληθινός χρόνος των δραστηριοτήτων μάθησης στο διαδίκτυο εβδομαδιαίως, παρακολούθηση της τάξης, και ετησίως, όπως στατιστικές της προόδου. Οι οργανισμοί τώρα ενδιαφέρονται περισσότερο στο να συμμετέχουν σε πραγματικό χρόνο ή κοντά στον πραγματικό χρόνο στην λήψη αποφάσεων καθώς τους δίνει ένα πιο ανταγωνιστικό πλεονέκτημα. Η ανάλυση είναι η χρήση των δεδομένων, η στατιστική ανάλυση και επεξηγηματικά και προβλεπτικά μοντέλα για να κερδίσουν διορατικότητα και να παρουσιάσουν

δεδομένα μέσα από ποικίλους τύπους οπτικοποίησης. Οι τύποι οπτικοποίησης μπορούν να είναι από απλά ραβδογράμματα και διαγράμματα πίτας ως πιο περίπλοκοι και διαδραστικοί τύποι οπτικοποίησης όπως ο διαδραστικός χάρτης δέντρου και το δυναμικό διάγραμμα φυσαλίδων. Ο στόχος της οπτικοποίησης είναι ότι ο χρήστης μπορεί εύκολα να κατανοήσει τεράστια και περίπλοκα δεδομένα. Οι τεχνικές αναφέρονται στη χρήση της στατιστικής, της οικονομετρίας, της γλωσσολογίας ή προηγμένων υπολογιστικών και machine learning μεθόδων για να αναλύσουν, να κατανοήσουν ή να προβλέψουν τάσεις ή εξαιρέσεις στα δεδομένα. Τα δεδομένα πρέπει να επεξεργάζονται στο κατάλληλο format για την εφαρμογή αυτών των τεχνικών.

2.2. Big data analytics στην ανώτερη εκπαίδευση στο Ηνωμένο Βασίλειο

Αυτή η ενότητα παρουσιάζει την κριτική του Vincent Koon Ong (2016) σε 11 JISC's BI projects από το Φεβρουάριο του 2011 ως τον Αύγουστο του 2012. Πρώτα, ζητήθηκε από κάθε πανεπιστημιακό ίδρυμα να κάνει κριτική στο δικό του BI επίπεδο ωριμότητας. Δεύτερον, χρησιμοποιώντας την ταξινομία ανάλυσης των big data του Goes, γίνεται κριτική στον βαθμό της εκτέλεσης της ανάλυσης των big data σε αυτά τα ιδρύματα. Ο πίνακας: «Περίληψη των JISC's των BI projects» παρέχει την περίληψη αυτών των project. Περισσότερες πληροφορίες για αυτά τα project μπορούν να βρεθούν στην ιστοσελίδα JISC infoNET.

Πίνακας 1: Περίληψη των JISC's B1 projects

Πανεπιστήμιο και θεματικές περιοχές του project	**Περίληψη του project**
University of Central Lancashire Φοιτητική επίδοση, διατήρηση και πρόοδος	Το project εκτίμησε τις απαιτήσεις για business intelligence σε σχέση με την επίδοση της διοίκησης / KPIs και ανέπτυξε ένα μοναδικό BI σύστημα που υποστηρίζει την ανταπόκριση σε αυτές τις απαιτήσεις.
University of Bolton Προϋπολογισμός και σχεδιασμός του φόρτου εργασίας	Το project υπολόγισε τα μέσα με τα οποία τα αποτελεσματικά data συλλέγονται, την συλλογή, την ελευθέρωση και ξαναχρησιμοποίησή τους που μπορούν και να ανταποκριθούν και στις ανάγκες της

	στήριξης των αποφάσεων μέσα στον οργανισμό, αλλά και για εκείνες των εξωτερικών πρακτορείων.
University of East London Φοιτητικό lifecycle και επίδοση της συγκριτικής αξιολόγησης	Το project έχει βελτιώσει τρεις BI εφαρμογές (student lifecycle, benchmarking and corporate performance) για την υποστήριξη των διαδικασιών του στρατηγικού σχεδιασμού των ανώτερων στελεχών.
University of Sheffield Εισαγωγή των φοιτητών και πρόοδος	Το project ανέπτυξε μια μεθοδολογία και λύση που δίνει τη δυνατότητα στους χρήστες να αναζητούν και να συνδέουν τα δεδομένα εισαγωγής των φοιτητών με εξωτερικές πηγές δεδομένων, όπως το HESA, από ένα μοναδικό σημείο πρόσβασης.
University of Durham Επίδοση της συγκριτικής αξιολόγησης	Το project συνέλεξε ένα σύνολο από metadata από ιδρύματα Ανώτερης Εκπαίδευσης που επιτρέπουν στις τρέχουσες δομές μέσα στα set εθνικών data να χαρτογραφηθούν σε δομές τμημάτων μέσα σε κάθε ίδρυμα.
University of Glasgow Ενδιαφέρον για έρευνα και αποτελέσματα	Το project συμπεριέλαβε, οπτικοποίησε και αυτοματοποίησε την παραγωγή πληροφοριών για τις ομάδες έρευνας στο Πανεπιστήμιο της Γλασκώβης βελτιώνοντας την πρόσβαση σε αυτά τα data προς στήριξη της λήψης στρατηγικών αποφάσεων, της δημοσιότητας, βελτιώνοντας τη συνεργασία και τη διεπιστημονική έρευνα, και έρευνα data reporting.
University of Manchester Εγκαταστάσεις και βελτιστοποίηση της χρησιμότητας	Το project ανέπτυξε ένα BI σύστημα που υποστηρίζει τη συλλογή, ανάλυση και διαχείριση των Estates Management Statistics (EMS) και που παρέχει στα ανώτερα στελέχη του Πανεπιστημίου καλύτερη και έγκαιρη πρόσβαση σε ακριβή δεδομένα, αναφορές και προβλεπτικές αναλύσεις σε σχέση με

	δεδομένα περιβάλλοντος και βιωσιμότητας.
Liverpool University Επίδοση της συγκριτικής αξιολόγησης	Το project σχεδίασε και εφάρμοσε μια λύση Διοίκησης Πληροφοριών, συνδυάζοντας την τεχνολογία με την ακεραιότητα των data, τη βελτίωση στις διαδικασίες της επιχείρησης και την αλλαγή της διοίκησης.
Open University Φοιτητική αφοσίωση, διατήρηση και πρόοδος	Το project ανέπτυξε προβλεπτικές modelling τεχνικές για να αναγνωρίζει τους φοιτητές σε κίνδυνο και ανέπτυξε μια διαδικασία για την αυτόματη καθημερινή εξαγωγή των VLE data στην αποθήκευση των τρεχόντων πανεπιστημιακών data για να συνδέονται με αυτές τις πηγές δεδομένων.
University of Bedfordshire Φοιτητική αφοσίωση, διατήρηση και πρόοδος	Το project έδειξε πως η υιοθέτηση ενός συστήματος εντοπισμού της φοιτητικής αφοσίωσης μπορεί να υποστηρίξει και να βελτιώσει τη λήψη αποφάσεων του ιδρύματος με αποδείξεις σε big data analytics.
University of Huddersfield Επίδοση της έρευνας	Το project ένωσε τις εσωτερικές πληροφορίες που εξάγονται από την αποθήκη δημοσιεύσεων ενός ιδρύματος με εξωτερικές πληροφορίες (ορισμούς ακαδημαϊκών μαθημάτων, ποιότητα των αποτελεσμάτων και δημοσιεύσεις) , για καταχώρηση σε ένα εργαλείο οπτικοποίησης.

Α. BI Επίπεδο Ωριμότητας

Το JISC του Ηνωμένου Βασιλείου έχει δώσει ένα πλαίσιο για το BI Επίπεδο Ωριμότητας στα πανεπιστημιακά ιδρύματα ώστε να υπολογίζουν το επίπεδο ωριμότητάς τους σε σχέση με την πρωτοβουλία και την εκτέλεση του BI. Έχουν αναγνωριστεί 6 στάδια στο μοντέλο ωριμότητας BI:

- ✓ 1ο στάδιο – τα δεδομένα κατακερματίζονται και καταστρέφονται-διασκορπισμένα σε παραδοσιακές, συχνά τοπικές πηγές δεδομένων-

χειρόγραφες αναφορές είναι διαθέσιμες στη διοίκηση του τμήματος, της σχολής και του ιδρύματος.

- ✓ 2ο στάδιο – οι πληροφορίες γίνονται όλο και πιο συναφείς, βρίσκονται σε κεντρικά διαχειριζόμενα συστήματα με ξεκάθαρη τοπική ευθύνη για την εισαγωγή και την ποιότητα των δεδομένων. Οι περισσότερες αναφορές είναι ακόμη χειρόγραφες.
- ✓ 3ο στάδιο – ένα Business Intelligence (BI) project ξεκινάει, και επιλέγονται ένας προμηθευτής και σύστημα.
- ✓ 4ο στάδιο – ένα αρχικό BI System μπαίνει σε εφαρμογή το οποίο επιτρέπει στους διευθυντές σε κάθε επίπεδο να έχουν πρόσβαση στα δεδομένα όταν τα χρειάζονται.
- ✓ 5ο στάδιο – το BI System και οι σύνδεσμοί του στις πηγές δεδομένων αυτοματοποιούνται όλο και περισσότερο- οι αναφορές γίνονται πιο εξελιγμένες και εξαπλώνονται σε ευρύτερο πληθυσμό χρηστών.
- ✓ 6ο στάδιο – τα συστήματα χρησιμοποιούνται για λήψη αποφάσεων που βασίζεται σε αποδείξεις και για προβλέψεις, μοντέλα και αξιολόγηση των μελλοντικών επιλογών.

Από τα 11 πανεπιστημιακά ιδρύματα, 3 (27,3%) βρίσκονται στο 1ο στάδιο BI Ωριμότητας, ενώ άλλα 3 (27,3%) βρίσκονται στο 2ο στάδιο. 1 ίδρυμα (9%) βρίσκεται στο 3ο στάδιο, 2 ιδρύματα (18,2%) στο 4ο στάδιο και άλλα 2 (18,2%) στο 5ο στάδιο. Αυτό δείχνει ότι η πλειοψηφία των πανεπιστημιακών ιδρυμάτων στο Ηνωμένο Βασίλειο είναι ακόμα σε πρώιμο στάδιο ως προς τη χρησιμοποίηση των δεδομένων του ιδρύματος για big data analytics. Τα δεδομένα οπωσδήποτε γίνονται πιο συναφή και βρίσκονται σε κεντρικά διαχειριζόμενα συστήματα. Για παράδειγμα, στο Πανεπιστήμιο του Bolton τα δεδομένα βρίσκονται στο κεντρικό δίκτυο του πανεπιστημίου με έναν αριθμό μονολιθικών κεντρικών εφαρμογών, όπως το Tribal SITS System για τα δεδομένα των φοιτητών, το Trend για τα δεδομένα των ανθρώπινων πόρων, και αυτά θεωρούνται ως οι κεντρικές ‘μοναδικές πηγές της αλήθειας’ για τα ποικίλα στοιχεία με τα οποία σχετίζονται. Όμως, η χρησιμοποίηση των δεδομένων στα περισσότερα ιδρύματα είναι ακόμα περιορισμένη στην παρουσίαση των λειτουργικών αποφάσεων. Σε μερικές περιπτώσεις, οι αναφορές, παρουσιάσεις γίνονται ακόμη χειρόγραφα. Για παράδειγμα, στο Πανεπιστήμιο της Glasgow παρόλο που τα δεδομένα για τα project και τα αποτελέσματά τους αποθηκεύονταν σε κεντρικά διαχειριζόμενα συστήματα, οι πραγματικές ερευνητικές ομάδες δεν αποθηκεύονταν. Η ανάλυση των ερευνητικών ομάδων γινόταν συμπληρώνοντας χειρόγραφα μια λίστα του σχετικού προσωπικού και των δεδομένων που εξάχθηκαν. Ακόμη, στο Πανεπιστήμιο του Bedfordshire τα αποτελέσματα των δεδομένων δεν θεωρούνται λειτουργικώς χρήσιμα ούτε προσανατολισμένα προς δράση αυτή τη στιγμή. Ωστόσο, όλα τα πανεπιστημιακά ιδρύματα έχουν μετακινηθεί σε υψηλότερο στάδιο BI ωριμότητας. Για παράδειγμα, στο Πανεπιστήμιο του Huddersfield έχουν εισάγει ένα εργαλείο που μπορεί να ενσωματωθεί με το RIMS System του πανεπιστημίου με σκοπό να καταστεί ένα εσωτερικό εργαλείο για τη λήψη ανώτερων διοικητικών αποφάσεων, ενώ στο Πανεπιστήμιο του Manchester κατά τη διάρκεια του project και με την

προσθήκη του αποθηκευτικού χώρου δεδομένων INGRiD και του πίνακα οργάνων, η αξιολόγηση της BI ωριμότητας μετακινήθηκε από τα επίπεδα 1, 2 και 3 στα πρώτα στάδια του επιπέδου 4.

Πάντως, αυτά τα ανώτερα εκπαιδευτικά ιδρύματα συνάντησαν έναν αριθμό προκλήσεων στην ανάπτυξη και την προσθήκη του BI. Πρώτα απ΄όλα, οι φορείς έχουν σημαντική επιρροή στην ανάπτυξη και την προσθήκη. Υπήρχε αναντιστοιχία προσδοκιών ανάμεσα στην ανώτερη διοικητική ομάδα, τους BI πωλητές και παρόχους, το ακαδημαϊκό προσωπικό και τους διαχειριστές, οι οποίες προκαλούσαν διαμάχες για τα συμφέροντα και τις προτεραιότητες της ανάπτυξης και της προσθήκης του BI. Μια ξεκάθαρη διοικητική δομή πρέπει να τεθεί, ακολουθούμενη από τακτική και συνεχή επικοινωνία ανάμεσα στους εμπλεκόμενους φορείς, ώστε να εφαρμοστεί με επιτυχία αυτή η αλλαγή στη διοίκηση. Αν εξωτερικοί πωλητές μπορούν να εμπλακούν στην ανάπτυξη και εισαγωγή του BI, τα κριτήρια επιλογής και η διαδικασία, όπως επίσης και το συμβόλαιο – συμφωνία πρέπει να περιγραφούν αναλυτικά, για να διατηρηθεί καλή σχέση με τον προμηθευτή και να μεγιστοποιηθεί το επενδυτικό όφελος. Έχει αποδειχτεί ότι η λύση του BI ίσως αλλάξει επίσης και την στρατηγική κατεύθυνση εξαιτίας των αλλαγών στη χρήση των δεδομένων και εξαιτίας των τεχνολογικών περιορισμών. Η προσβασιμότητα των δεδομένων, η ιδιοκτησία των δεδομένων, η ποιότητα των δεδομένων και οι προθεσμίες είναι θέματα-κλειδιά στην ανάπτυξη και την εισαγωγή του BI. Τα ‘datasets’, ειδικά τα δεδομένα σε πραγματικό χρόνο και τα εξωτερικά δεδομένα ίσως να μην είναι προσβάσιμα και ικανά να χρησιμοποιηθούν. Συνεπώς, ο ‘καθαρισμός’ των δεδομένων και η εκ νέου διαμόρφωσή τους θα χρειάζονται για να αυξηθεί η ποιότητα και η ακρίβεια των δεδομένων. Είναι σημαντικό να διεξαχθούν ορισμένοι έλεγχοι της εφαρμοσιμότητας όσο το δυνατό νωρίτερα στο αρχικό BI project. Η τεχνολογία θα μπορούσε επίσης να καταστεί εμπόδιο στην επιτυχία του BI project. Η διαθεσιμότητα και η δυνατότητες των τεχνολογικών λύσεων θα μπορούσαν να καθορίσουν το βαθμό της εξέλιξης των BI systems, όπως ‘έξυπνα’ εργαλεία και τεχνικές για προληπτική και έξυπνη επεξεργασία των πληροφοριών, διαδραστική και δυναμική ψηφιακή οπτικοποίηση. Τέλος, η ανάπτυξη και η εισαγωγή της διαδικασίας πρέπει να έχει ως κέντρο της το χρήστη από την αρχή ως το τέλος. Αυτό θα δώσει κίνητρο και θα ενθαρρύνει τους ανώτερους διευθυντές, το ακαδημαϊκό προσωπικό και τους διαχειριστές να συνειδητοποιήσουν την αξία και την χρησιμότητα του BI System για καλύτερη διοίκηση και στρατηγικό σχεδιασμό.

B. Big data analytics

Αναφορικά με την ταξινομία της Εικόνας 2 (Goes, 2014), τα περισσότερα πανεπιστημιακά ιδρύματα δεν χρησιμοποιούν στατιστικές αναλύσεις των big data σε βάση πραγματικού χρόνου. Τα περισσότερα ιδρύματα χρησιμοποιούν δεδομένα συναλλαγών και δεδομένα που παράγει το σύστημα. Σε μερικές περιπτώσεις, οι αποθήκες δεδομένων υπάρχουν ως ένας τρόπος να ενσωματωθούν, να οργανωθούν και να συνοψισθούν τα μεγάλα datasets. Οι αναφορές συνήθως απαιτούνται σε εβδομαδιαία ή μηνιαία βάση, όπως οι

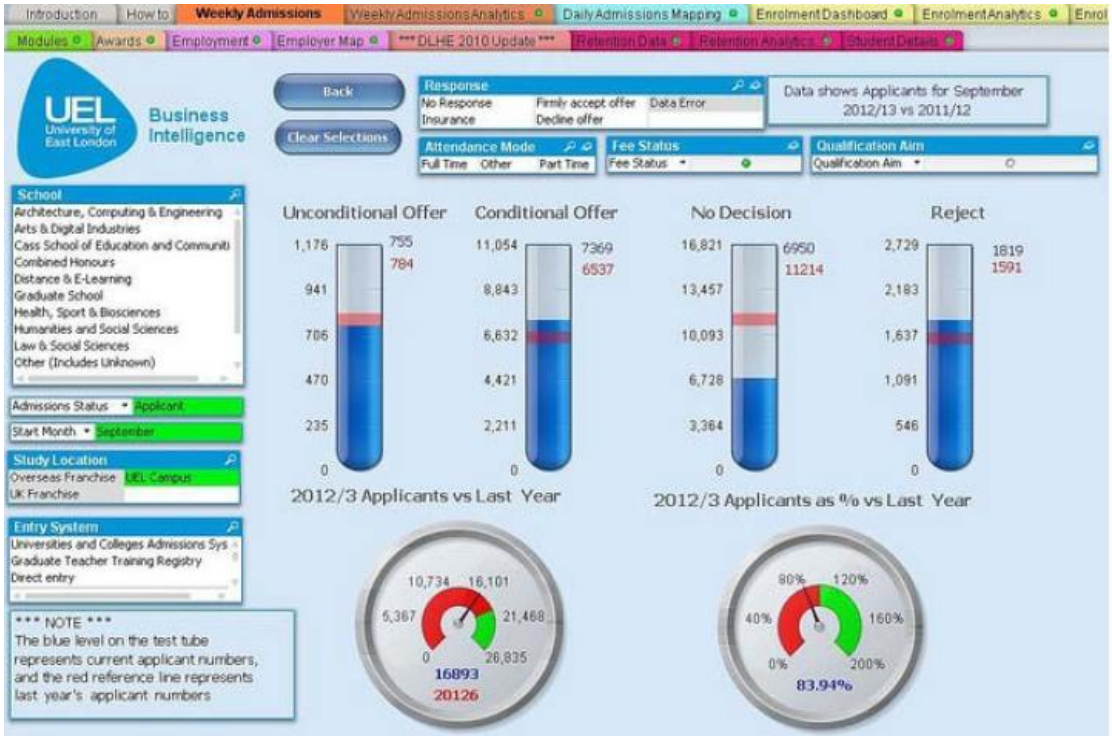
εβδομαδιαίες αναφορές εισαγωγής των φοιτητών, οι μηνιαίες αναφορές της αφοσίωσης των φοιτητών, και δεδομένα εργασιακής απασχόλησης. Στο Πανεπιστήμιο του East London, για παράδειγμα, τα δεδομένα εξάγονται τυπικά εβδομαδιαίως, συχνά καθημερινά κατά τη διάρκεια των φορτωμένων περιόδων και μηνιαίως για δομικά συστατικά όπως το DLHE. Ακόμη, στο πανεπιστήμιο του Bedfordshire οι αναφορές πραγματοποιούνται είτε μετά από αίτημα του χρήστη είτε ανανεώνονται περιοδικά μέσα από ένα αυτοματοποιημένο πρόγραμμα (εβδομαδιαίως, κάθε δεκαπενθήμερο, μηνιαίως), ενώ στο Πανεπιστήμιο του Liverpool γίνεται μια μηνιαία απογραφή των φοιτητικών δεδομένων. Η παραγωγή δεδομένων του ιστορικού έχει καταστεί πρόκληση για τα ιδρύματα ώστε να αναγνωρίζονται οι φοιτητές σε κίνδυνο και να λαμβάνονται αμέσως τα κατάλληλα μέτρα.

Ωστόσο, αυτό αρχίζει να αλλάζει καθώς όλο και περισσότερα πανεπιστημιακά ιδρύματα στο Ηνωμένο Βασίλειο χρειάζονται και αναγνωρίζουν την αξία των έγκαιρων δεδομένων για την αποτελεσματική λήψη αποφάσεων. Για παράδειγμα οι αιτήσεις μέσω διαδικτύου (online) έχουν ευρέως εισαχθεί από τα ιδρύματα ανώτερης εκπαίδευσης έτσι ώστε η ομάδα που ασχολείται με την εισαγωγή των φοιτητών να απαντά στις αιτήσεις έγκαιρα. Ο έλεγχος της αφοσίωσης – συμμετοχής των φοιτητών σε πραγματικό χρόνο, ή έστω κοντά στον πραγματικό χρόνο θα απαιτείται για να είναι ανταποκρίνεται στα απαιτούμενα του UK Border Agency (UKBA), αλλά και σε άλλα νομικά απαιτούμενα. Με την αύξηση της χρήσης της τεχνολογίας των μέσων κοινωνικής δικτύωσης (social media), οι φοιτητές επικοινωνούν και αναρτούν τη δική τους ανατροφοδότηση όλο και περισσότερο μέσω αυτών. Επομένως, μπορεί να δοθεί άμεση ανταπόκριση αλλά και λήψη μέτρων ώστε να ενισχυθεί η εμπειρία των φοιτητών.

Με όρους στατιστικής ανάλυσης, μερικά από αυτά τα ιδρύματα ανώτερης εκπαίδευσης έχουν αναπτύξει ποικίλα στατιστικά εργαλεία για να αναπτύξουν διορατικότητα στη λήψη αποφάσεων. Για παράδειγμα το Ανοιχτό Πανεπιστήμιο έχει αναπτύξει εργαλεία με προβλεπτικά μοντέλα για να προβλέψει την επίδοση των φοιτητών, ιδιαίτερα για φοιτητές που αντιμετωπίζουν τον κίνδυνο να αποτύχουν σε κάποιο μάθημα. Στα αναπτυσσόμενα προβλεπτικά μοντέλα, έχει βρεθεί ότι ένα συγκεκριμένο ποσό δοκιμής και λάθους (trial and error) απαιτείται για να καθοριστεί ποιοι παράμετροι παρέχουν τις περισσότερες πληροφορίες για να φτιαχτεί το μοντέλο. Το Πανεπιστήμιο του Bedfordshire έφτιαξε ένα συντελεστή μέτρησης εξατομικευμένης φοιτητικής αφοσίωσης – συμμετοχής, ο οποίος επιτρέπει στους χρήστες να επιλέξουν και να συνδυάσουν διαφορετικούς τύπους δεδομένων συμμετοχής που εντόπισε το σύστημα. Επιπλέον, οι χρήστες μπορούν να χρησιμοποιήσουν τη Διαδικασία Στατιστικής Ιεραρχίας (Analytic Hierarchy Process – AHP) για να καθορίσουν τις πιο βαρύνουσας σημασίας παραμέτρους που θεωρούν οι ίδιοι ότι αντανακλούν τη σημασία και την προτεραιότητα των διαφορετικών τύπων δεδομένων συμμετοχής. Για παράδειγμα η συμμετοχή στην τάξη μπορεί να θεωρείται πιο σημαντική από την συμμετοχή σε κάποια δραστηριότητα εκτός του κύκλου μαθημάτων.

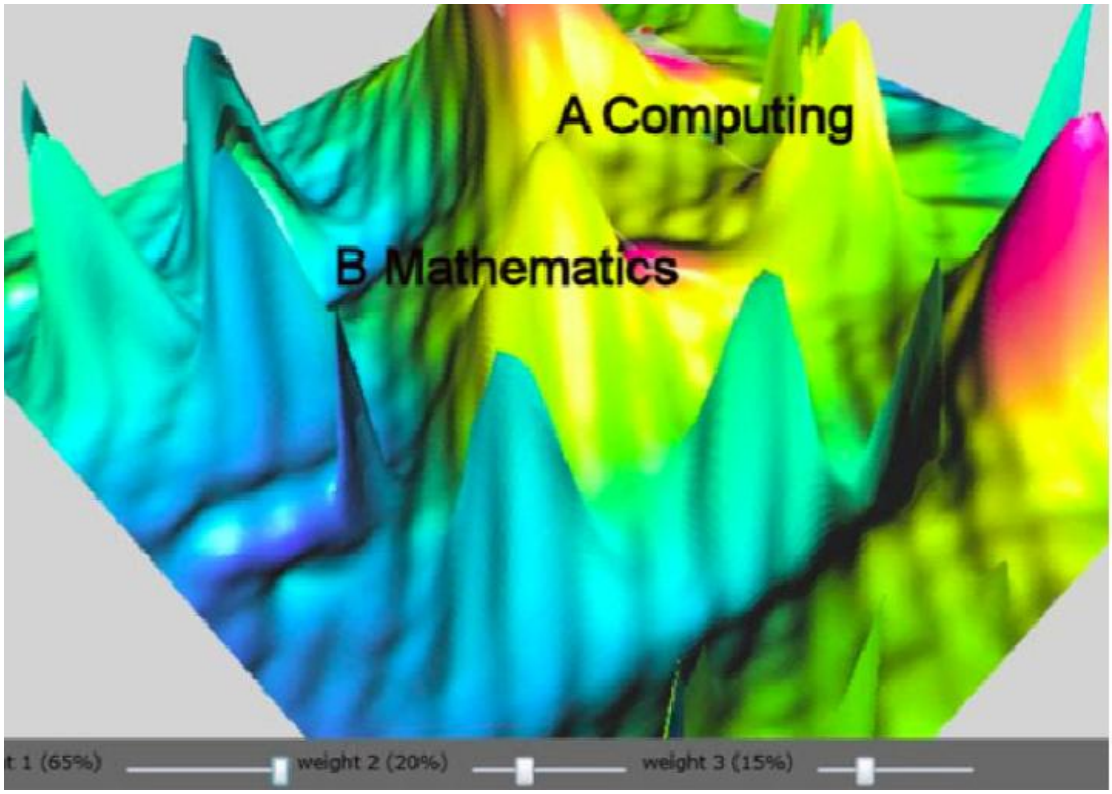
Η στατιστική ανάλυση των big data έχει να κάνει με την κατανόηση τεράστιων και πολύπλοκων datasets. Τα πανεπιστημιακά ιδρύματα πρέπει να αρχίσουν να ασχολούνται με ποικίλους τύπους στατιστικής ανάλυσης, όπως η ανακάλυψη, η επανάληψη, οι ευέλικτες δυνατότητες, η πρόβλεψη και η διοίκηση της λήψης αποφάσεων (an Oracle White Paper, 2013). Εξετάζοντας το παράδειγμα της φοιτητικής αφοσίωσης – συμμετοχής, η διαδικασία της ανακάλυψης περιλαμβάνει την εξερεύνηση και τη σύνδεση διαφορετικών δεδομένων αφοσίωσης, όπως η παρακολούθηση στην τάξη, άλλες δραστηριότητες, η συμμετοχή στην ψηφιακή βιβλιοθήκη, τα email του Πανεπιστημίου κ.ά. για να προκύψουν διάφορα μοτίβα, όπως οι παράγοντες κινδύνου συγκράτησης των φοιτητών, είναι απαραίτητο να περάσουμε στη διαδικασία της επανάληψης, να μάθουμε από κάθε επανάληψη και σταδιακά να περάσουμε στο επόμενο επίπεδο στατιστικής ανάλυσης, όπως η περαιτέρω δυνατότητες και η προβλεπτική στατιστική ανάλυση με σκοπό να αποκτήσουμε ισχυρή διορατικότητα για την συμμετοχή, αφοσίωση των φοιτητών και την συγκράτησή τους στο πανεπιστήμιο. Η στατιστική ανάλυση των big data να διεξαχθεί αυτόματα όταν χρειάζεται να ληφθεί μια συγκεκριμένη απόφαση. Για παράδειγμα, το ακαδημαϊκό προσωπικό μπορεί αυτόματα να λαμβάνει ειδοποιήσεις για τους φοιτητές που δεν συμμετέχουν και αντιμετωπίζουν τον κίνδυνο να εγκαταλείψουν το πανεπιστήμιο.

Οι ανώτεροι διευθυντές δεν μπορούν να δουν ένα μοτίβο χωρίς μια καλή οπτικοποίηση των δεδομένων. Επιπρόσθετα, μια μοναδική οθόνη δεν αρκεί για να παρουσιαστεί ο αριθμός των μεγάλων datasets. Για τον λόγο αυτό, ο τύπος της οπτικοποίησης των δεδομένων πρέπει να είναι κατάλληλος και φιλικός προς τον χρήστη, έτσι ώστε οι χρήστες να μπορέσουν να κατανοήσουν αποτελεσματικά τα συμπεράσματα που προκύπτουν από την στατιστική ανάλυση. Η προηγμένη οπτικοποίηση των δεδομένων περιλαμβάνει δυναμικό περιεχόμενο δεδομένων, οπτικά ερωτήματα, οπτικοποίηση πολλαπλών διαστάσεων, γλαφυρή οπτικοποίηση, εξατομίκευση και ειδοποιήσεις για ανάληψη δράσης της επιχείρησης. Μερικές προηγμένες οπτικοποιήσεις δεδομένων έχουν αναπτυχθεί μέσα στο JISC’s BI Programme. Το Ανοιχτό Πανεπιστήμιο, το Πανεπιστήμιο του Bedfordshire, το Πανεπιστήμιο του Manchester και το Πανεπιστήμιο του Central Lancashire χρησιμοποίησαν μια δυναμική ράβδο και γραμμές γραφικών ως πίνακα ελέγχου ψηφιακής διοίκησης για να παρουσιάσουν ποικίλους τομείς των δεικτών επίδοσης (Key Performance Indicators – KPIs) και άλλους στρατηγικούς ελέγχους και το σχεδιασμό πληροφοριών. Το Πανεπιστήμιο της Glasgow χρησιμοποίησε το διάγραμμα Venn για να εκφράσει το βαθμό στον οποίο ίσως ταιριάζουν τα ενδιαφέροντα των ερευνητών. Το Πανεπιστήμιο του East London χρησιμοποίησε ‘dials’ και ‘test tubes’ για να παρουσιάσει τη στατιστική ανάλυση δεδομένων που έχουν να κάνουν με τον κύκλο ζωής των φοιτητών (Student Lifecycle), τον πίνακα ελέγχου και την εισαγωγή των φοιτητών (βλ. εικόνα: Student Lifecycle – Admissions and Dashboard).



Εικόνα 3: Student Lifecycle – Admissions and Dashboard

Το Πανεπιστήμιο του Huddersfield έχει προσπαθήσει να αναπτύξει μια πιο προηγμένη οπτικοποίηση δεδομένων μέσα από τον έλεγχο από τον χρήστη της εξατομικευμένης οπτικοποίησης σε τρισδιάστατη (3D) μορφή (βλ. εικόνα**:** Τρισδιάστατη οπτικοποίηση).



Εικόνα 4: Τρισδιάστατη οπτικοποίηση

Με τεχνικούς όρους στατιστικής ανάλυσης των big data, όλα τα BI projects βασίζονται πολύ στην προσβασιμότητα, την ποιότητα και τον βαθμό πιστότητας των δεδομένων που χρειάζονται για να φτιάξουν κατάλληλες τεχνικές για τα BI μοντέλα. Για παράδειγμα, το Πανεπιστήμιο του East London συνειδητοποίησε ότι τα δεδομένα που χρειάζονται από τη Βάση Δεδομένων για τα Ιδρύματα Ανώτερης Εκπαίδευσης (Higher Education Information Database for Institutions – HEIDI) πρέπει να εξαχθούν και να προ-επεξεργαστούν σε κατάλληλο τύπο για τη δική του στατιστική ανάλυση big data, καθώς η HEIDI δεν είναι φιλική προς το χρήστη (user-friendly). Το Πανεπιστήμιο του East London έπρεπε να εξάγει δεδομένα από την HEIDI και να αναπτύξει μια ποικιλία Macros για να αναδιαμορφώσει τα δεδομένα που είναι κατάλληλα για ανάπτυξη QlikView.Το Πανεπιστήμιο της Glasgow βελτίωσε την πρόσβαση στα δεδομένα του και τους αλγόριθμους σύγκρισης μέσω της συγκεχυμένης αναζήτησης λέξεων με αποκοπή καταλήξεων, η οποία επιτρέπει στο σύστημα να προτείνει λέξεις κλειδιά από δημοσιεύσεις και περιγραφές εργασιών.

2.2.1. Big data analytics στην ανώτερη εκπαίδευση στο Ηνωμένο Βασίλειο

Όπως είναι φανερό, η στατιστική ανάλυση των big data μπορεί να παρέχει μοναδική και πολύτιμη διορατικότητα στις τάσεις εισαγωγής των φοιτητών στο πανεπιστήμιο, στη συγκράτηση των φοιτητών και σε θέματα προόδου, στις επενδύσεις για έρευνα και την ανάλυση των αποτελεσμάτων, στη βελτίωση των υποδομών του πανεπιστημίου και σε πολλά άλλα θέματα σχετικά με την ανώτερη εκπαίδευση. Η μελλοντική έρευνα για τα big data στην ανώτερη εκπαίδευση μπορεί να εστιαστεί σε υψηλότερα επίπεδα ωριμότητας της στατιστικής ανάλυσης των big data καθώς όλο και περισσότερα πανεπιστημιακά ιδρύματα αναγνωρίζουν την αξία και αρχίζουν να χρησιμοποιούν την στατιστική ανάλυση των big data. Αυτό ίσως περιλαμβάνει την έρευνα σε προηγμένες τεχνικές στατιστικής ανάλυσης των big data και μοντέλα λήψης αποφάσεων βασισμένα σε αποδείξεις και καθοδηγούμενα από τα δεδομένα. Η στατιστική πρόβλεψη, τα περίπλοκα ερωτήματα, η συνήθης στατιστική ανάλυση, η βελτιστοποίηση, η επεξεργασία σε φυσική γλώσσα και τα προβλεπτικά μοντέλα γίνονται όλο και πιο σημαντικά για καλή λήψη αποφάσεων που βασίζεται σε αποδείξεις (evidence – based) και καθοδηγείται από τα δεδομένα (data – driven).

Καθώς τα δεδομένα σε πραγματικό χρόνο, ή κοντά σε πραγματικό χρόνο γίνονται πιο πολύτιμα για τα ιδρύματα ανώτερης εκπαίδευσης, η μελλοντική έρευνα ίσως εξετάσει την επεξεργασία και την χρήση των δεδομένων αυτών στην ανώτερη εκπαίδευση. Η μελλοντική έρευνα ίσως ακόμη πρέπει να λάβει υπόψη της και σχετικά εξωτερικά δεδομένα, όπως δεδομένα από τη Βάση Δεδομένων για τα Ιδρύματα Ανώτερης Εκπαίδευσης (Higher Education Information Database for Institutions – HEIDI), τα δεδομένα από το Research Excellence Framework (REF), τα δεδομένα για την εργασιακή απασχόληση των αποφοίτων, τα δεδομένα από το National Student Survey (NSS), και ακόμη και δεδομένα από τα social media των φοιτητών.

Η προηγμένη οπτικοποίηση των δεδομένων είναι ένα ακόμη θέμα που πρέπει να διερευνηθεί όσον αφορά την χρηστικότητα και την διαδραστικότητα με τους χρήστες, ειδικά για τους ανώτερους διευθυντές την ανώτερης εκπαίδευσης, ώστε η στατιστική ανάλυση των big data να είναι χρήσιμη και χρηστική για τους ενδιαφερόμενους φορείς. Εξαιτίας της αυξανόμενης ποσότητας δεδομένων από ποικίλες πηγές, οι ερευνητές καλούνται επίσης να ερευνήσουν τα μοντέλα και τις τεχνικές των big data σε σχέση με τις κοινωνικές, οικονομικές και γνωστικές τους διαστάσεις. Η συμπεριφοριστική στατιστική ανάλυση είναι άλλη μια περιοχή προς έρευνα καθώς οι συμπεριφορές και οι επιδόσεις ατομικής και συλλογικής μάθησης μπορούν να ελεγχθούν και να βελτιωθούν κατά τους κύκλους μαθημάτων των σπουδών. Οι θεωρίες των κοινωνικών δικτύων θα μπορούσαν επίσης να χρησιμοποιήσουν τα big data για να ερευνήσουν την δυναμική των επίσημων και ανεπίσημων δικτύων καθώς αυτά σχηματίζονται και εξελίσσονται, όπως

επίσης και για να εξετάσουν την επίδραση στη συμπεριφορά σε ατομικό επίπεδο και επίπεδο του δικτύου.

Οργανωτικά και μετασχηματιστικά θέματα στην στατιστική ανάλυση των big data θα πρέπει συνεχώς να μελετώνται, καθώς οι στρατηγικές που βρίσκουν τα ιδρύματα ανώτερης εκπαίδευσης για την διοίκηση των επιχειρήσεων και τη χρησιμοποίηση των big data αλλάζουν τον τρόπο διοίκησης των επιχειρήσεων και μεταμορφώνουν την επιχείρηση. Η συγκεκριμένη περιοχή έρευνας των big data ίσως έχει μεγάλη επιρροή στην οργάνωση και διοίκηση της ηγεσίας.

2.3. Η εφαρμογή των Big Data στον τομέα της υγείας και της περίθαλψης στις Ηνωμένες Πολιτείες της Αμερικής

Μέσα από τη βιβλιογραφική ανασκόπηση διαπιστώνει κανείς εύκολα πως πολλοί ερευνητές έχουν εστιάσει το ενδιαφέρον τους στη μελέτη της χρησιμοποίησης των big data στον τομέα της υγείας στις ΗΠΑ. Ο τομέας της υγειονομικής περίθαλψης αποτελεί έναν από τους μεγαλύτερους τομείς της οικονομίας των ΗΠΑ, καθώς παράγει περισσότερο από το 17% του ΑΕΠ (Ακαθάριστο Εγχώριο Προϊόν) και απασχολεί περίπου το 11% των εργαζομένων της χώρας (Manyika J. κ.ά., 2011). Για το λόγο αυτό, λοιπόν, η έρευνα για την χρήση των big data στον συγκεκριμένο τομέα είναι εκτεταμένη. Σε αυτή την ενότητα, θα αναφερθούμε σε αυτές που θεωρήσαμε πιο ενδεικτικές.

Σε έρευνα του McKinsey Global Institute το 2011, οι συγγραφείς ισχυρίζονται ότι η χρησιμοποίηση των big data μπορεί να επιφέρει περισσότερα από 300 δισεκατομμύρια δολάρια πρόσθετη αξία στην υγειονομική περίθαλψη των ΗΠΑ. Πιο συγκεκριμένα, σύμφωνα με την οικεία έρευνα, τα δεδομένα που αφορούν την υγεία είναι κλινικά δεδομένα, δεδομένα πληρωμών και κόστους, δεδομένα έρευνας και ανάπτυξης φαρμακευτικών και ιατρικών προϊόντων, δεδομένα συμπεριφοράς και συναισθημάτων των ασθενών. Η εφαρμογή, λοιπόν, των big data με πέντε διαφορετικούς τρόπους θα μπορούσε να μειώσει τις δαπάνες για την υγεία. Πρώτον, όσον αφορά την αποτελεσματικότητα της συγκριτικής έρευνας κατά την οποία τα ερευνητικά αποτελέσματα καθορίζουν την καλύτερη θεραπεία για κάθε ασθενή, με τα big data είναι δυνατόν να συγκρίνονται με μεγαλύτερη ακρίβεια τόσο τα δεδομένα των ασθενών, όσο και των ερευνητικών πορισμάτων, ώστε να καθορίζεται καλύτερα η αποτελεσματικότητα των ποικίλων θεραπευτικών παρεμβάσεων. Δεύτερον, χρησιμοποιώντας υποστηρικτικά συστήματα λήψης κλινικών αποφάσεων μπορεί να ενισχυθεί η αποτελεσματικότητα και η ποιότητα των χειρουργικών παρεμβάσεων. Ο τρίτος τρόπος χρήσης των κλινικών big data, είναι να αναλύονται τα δεδομένα των ιατρικών διαδικασιών και ακολούθως να δημιουργείται διαφάνεια γύρω από αυτά τα δεδομένα αφενός για να αναγνωρίζονται οι ευκαιρίες επίδοσης για τους επαγγελματίες στον τομέα της

υγείας και για τα ιδρύματα περίθαλψης, και αφετέρου για να παρέχεται βοήθεια στους ασθενείς να επιλέξουν την κατάλληλη για αυτούς περίθαλψη στην καλύτερη τιμή. Επιπλέον, η συλλογή δεδομένων για τους ασθενείς με χρόνιες παθήσεις και η ανάλυσή τους με σκοπό τον έλεγχο της υγειονομικής υποστήριξης αυτών των ασθενών, καθώς και τη μελλοντική βελτίωση των φαρμάκων και των επιλογών θεραπείας τους αποτελεί τον τέταρτο τρόπο χρήσης των κλινικών big data. Τέλος, ο πέμπτος τρόπος αφορά την εφαρμογή προηγμένων μέσων στατιστικής ανάλυσης στα προφίλ των ασθενών ώστε να αναγνωρίζονται τα άτομα που θα ωφελούνταν από την πρόληψη ή αλλαγές στον τρόπο ζωής τους. Ακόμη, όσον αφορά τις οικονομικές συναλλαγές, οι λύσεις που παρέχουν τα big data δίνουν τη δυνατότητα σε αυτοματοποιημένα συστήματα να εντοπίζουν περιπτώσεις απάτης και να ελέγχουν την ακρίβεια και τη συνέπεια στις πληρωμές. Σε σχέση με την έρευνα και την ανάπτυξη, τα big data δίνουν την δυνατότητα για χρήση μοντέλων πρόβλεψης μέσω στατιστικών εργαλείων και αλγόριθμων για να βελτιωθούν τα κλινικά σχέδια δοκιμών, να αναλύονται τα δεδομένα των κλινικών δοκιμών, των εξατομικευμένων φαρμακευτικών αγωγών και των μοτίβων των διαφόρων ασθενειών. Επίσης, δίνεται η δυνατότητα να συγκεντρώνονται και να συνθέτονται τα κλινικά αρχεία και οι πληρωμές των ασθενών και να δημιουργούνται online πλατφόρμες και κοινότητες (Manyika J. κ.ά., 2011).

Συνεχίζοντας, άλλοι ερευνητές μελέτησαν την χρησιμοποίηση της στατιστικής ανάλυσης των big data για να μειωθεί το κόστος της υγειονομικής περίθαλψης στις ΗΠΑ. Οι Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. παρουσιάζουν έξι περιπτώσεις ευκαιριών για μείωση του κόστους μέσω της χρήσης των big data: τους ασθενείς με υψηλό κόστος περίθαλψης, τις επανεισδοχές ασθενών, τη διαλογή των ασθενών, την αντιρρόπηση (τη χειροτέρευση της κατάστασης του ασθενούς), τις επιπλοκές και τη βελτίωση της θεραπείας ασθενειών που προσβάλλουν πολλαπλά όργανα (2014). Αναλυτικότερα, οι συγγραφείς δεν προτείνουν απλώς ένα τρόπο για να αναγνωρίζονται οι ασθενείς με υψηλό κόστος περίθαλψης και να τους παρέχεται φροντίδα με τέτοιο τρόπο ώστε να μειώνεται το κόστος. Λαμβάνοντας υπόψη πολλούς παράγοντες, θεωρούν πως είναι εξέχουσας σημασίας να χρησιμοποιούνται μέσα στατιστικής ανάλυσης που όχι μόνο θα αναγνωρίζουν αυτούς τους ασθενείς, αλλά και θα καθορίζουν τις ανάγκες φροντίδας και περίθαλψής τους καθώς και τα κενά στην φροντίδα τους. Επίσης, θεωρούν αναγκαία την αναγνώριση και αντιμετώπιση των προβλημάτων ψυχικής υγείας, όπως η κατάθλιψη, καθώς πολλές φορές αυτοί οι ασθενείς υποφέρουν και από αυτά. Η αντιμετώπιση των ασθενών υψηλού κόστους θα ήταν λιγότερο δαπανηρή αν ήταν εξατομικευμένη και προσαρμοσμένη στις συγκεκριμένες ανάγκες του κάθε ασθενούς, ενώ ταυτόχρονα η ανακατανομή των πόρων υπολογίζεται με αλγόριθμους και για τους ασθενείς υψηλού κόστους αλλά και για τους ασθενείς χαμηλού κόστους.

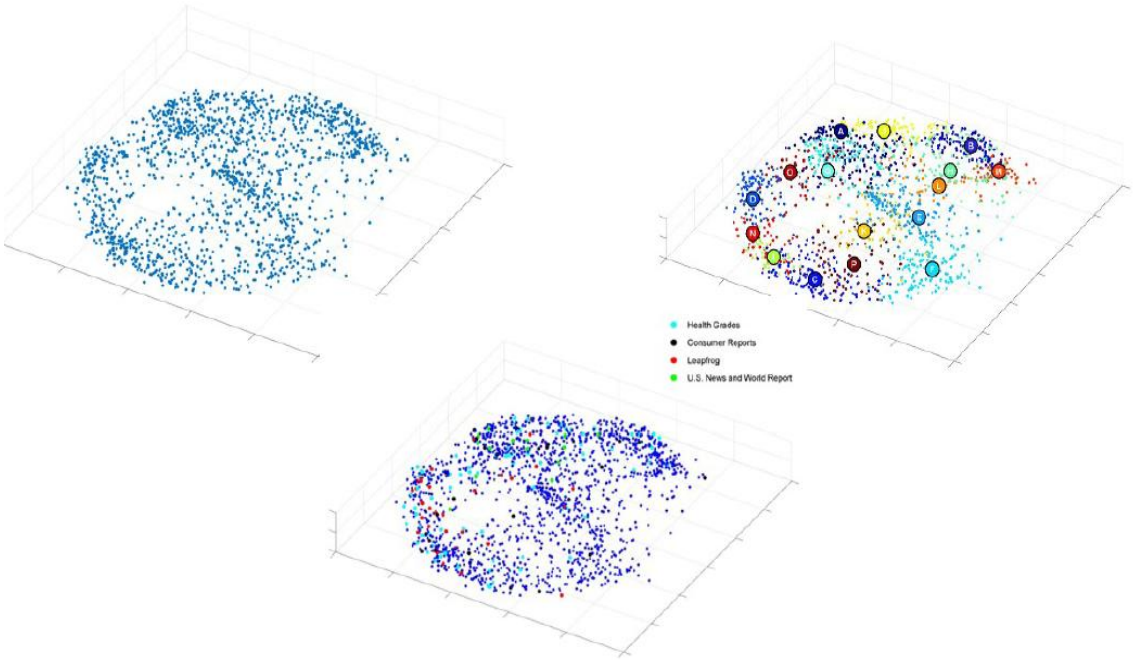
Όσον αφορά τις επανεισδοχές ασθενών στα νοσοκομεία, οι συγγραφείς της συγκεκριμένης έρευνας πιστεύουν πως με τη χρήση της στατιστικής ανάλυσης

των big data μπορούν να προβλεφθούν, αλλά και να αποφευχθούν λαμβάνοντας προληπτικά μέτρα, όπως την παροχή της κατάλληλης θεραπείας την πρώτη φορά εισαγωγής στο νοσοκομείο, την αποφυγή ιατρικών λαθών, την παρακολούθηση των ασθενών μετά την εξαγωγή τους ώστε να λαμβάνουν την κατάλληλη φαρμακευτική αγωγή ή την αντιμετώπιση ψυχολογικών προβλημάτων ιδίως των ασθενών με χρόνιες παθήσεις κ.ά. Σε σχέση με τη διαλογή των ασθενών, οι ερευνητές παρουσιάζουν 2 πιλοτικά προγράμματα στο Kaiser Permanente Northern California (KPNC), ένα ολοκληρωμένα σύστημα υγείας με περιεκτικά πληροφοριακά συστήματα. Και στα δύο προγράμματα χρησιμοποιείται ένας αλγόριθμος διαλογής ασθενών που καθορίζει τη ροή των κλινικών εργασιών με τέτοιο τρόπο ώστε να παρέχεται η καλύτερη δυνατή περίθαλψη. Οι κλινικές ομάδες χρησιμοποιούν τα εργαλεία των big data σε συνδυασμό με τα παραδοσιακά εργαλεία, όπως η κλινική εξέταση, για να αναπτύξουν τη ροή των εργασιών τους και με τον τρόπο αυτό οι ασθενείς για τους οποίους σχεδιάστηκαν τα πιλοτικά προγράμματα – νεογέννητα στο πρώτο και επείγοντα περιστατικά ενηλίκων στο δεύτερο – επωφελούνται από τις μεθόδους των big data.

Η τέταρτη περίπτωση, αυτή της αντιρρόπησης, της χειροτέρευσης της κατάστασης του ασθενούς, σύμφωνα με τους συγγραφείς, μπορεί να προβλεφθεί και κατά συνέπεια να προληφθεί με τη βοήθεια των big data. Ενώ υπάρχουν ήδη διάφορες συσκευές που ελέγχουν την κατάσταση του ασθενούς ακόμα και μετά την εξαγωγή του από το νοσοκομείο (π.χ. συσκευή ελέγχου αναπνευστικής ή καρδιακής λειτουργίας), θα μπορούσε να υπάρχει ένας συνδυασμός των δεδομένων κάθε συσκευής ώστε να προλαμβάνεται η επιδείνωση. Με τον ίδιο τρόπο μπορούν να προληφθούν και οι επιπλοκές και οι παρενέργειες των φαρμάκων που μάλιστα είναι δαπανηρό να αντιμετωπιστούν και προκαλούν νοσηρότητα και θνησιμότητα. Τέλος, η χρήση της στατιστικής ανάλυσης των big data μπορεί να επιφέρει βελτίωση και εξατομίκευση στην παροχής ιατροφαρμακευτικής περίθαλψης για ασθενείς με παθήσεις που προσβάλλουν πολλαπλά όργανα ή ομάδες οργάνων (όπως η ρευματοειδής αρθρίτιδα ή ο ερυθηματώδης λύκος) και παράλληλα να μειώσει το κόστος της θεραπείας τους.

Η επόμενη έρευνα στην οποία θα αναφερθούμε είναι των Downing N.S., Cloninger A., Venkatesh A.K., Hsieh A., Drye E.E., Coifman R.R., κ.α. (2017) και αφορά την επίδοση των νοσοκομείων στις ΗΠΑ. Οι συγγραφείς χρησιμοποίησαν στατιστική ανάλυση big data για να περιγράψουν τις επιδόσεις των νοσοκομειακών ιδρυμάτων. Παρόλο που υπάρχουν πολλά εργαλεία μέτρησης της ποιότητας, εκείνοι ανέπτυξαν μια καινούρια προσέγγιση που υπογραμμίζει τις ομοιότητες και τις διαφορές ανάμεσα στα νοσοκομεία και που αναγνωρίζει συνήθη μοτίβα στην επίδοσή τους. Συγκεκριμένα, ο αλγόριθμος που έφτιαξαν συνδύασε τα υπάρχοντα διαθέσιμα εργαλεία μέτρησης της ποιότητας για 1614 νοσοκομεία στις ΗΠΑ για να περιγράψουν την επίδοσή τους γραφικά και ποσοτικά. Στην οπτικοποίηση των αποτελεσμάτων φαίνεται πως υπάρχουν συστάδες νοσοκομείων με συγκεκριμένα προφίλ επίδοσης, ενώ άλλα προφίλ επίδοσης είναι πιο σπάνια.

Ακόμη, διάφορα κορυφαία νοσοκομεία που έχουν παρόμοια επίδοση σύμφωνα με τα παραδοσιακά εργαλεία μέτρησης, στην οπτικοποίηση των ερευνητών φαίνεται πως διαφέρει ο τρόπος με τον οποίο διαπρέπουν, δηλαδή η κορυφαία ποιότητά τους είναι σε πολλούς διαφορετικούς τομείς. Τα big data έδωσαν την δυνατότητα για μια κατάταξη των νοσοκομείων που φωτίζει στοιχεία για την ποιότητα της νοσοκομειακής περίθαλψης τα οποία δεν λαμβάνονταν υπόψη από τους υπάρχοντες τρόπους αξιολόγησης των νοσοκομείων, όπως με ποιους ακριβώς τομείς το κάθε νοσοκομείο ξεχωρίζει, σε ποιους συγκεκριμένους τομείς παρουσιάζει χαμηλή ποιότητα κ.ά, ενώ παράλληλα συνυπολογίζονται δημογραφικά και γεωγραφικά δεδομένα. Η εικόνα “Χάρτης διάχυσης” παρουσιάζει το χάρτη διάχυσης (diffusion map) όπως παρουσιάζεται στην οικεία μελέτη.

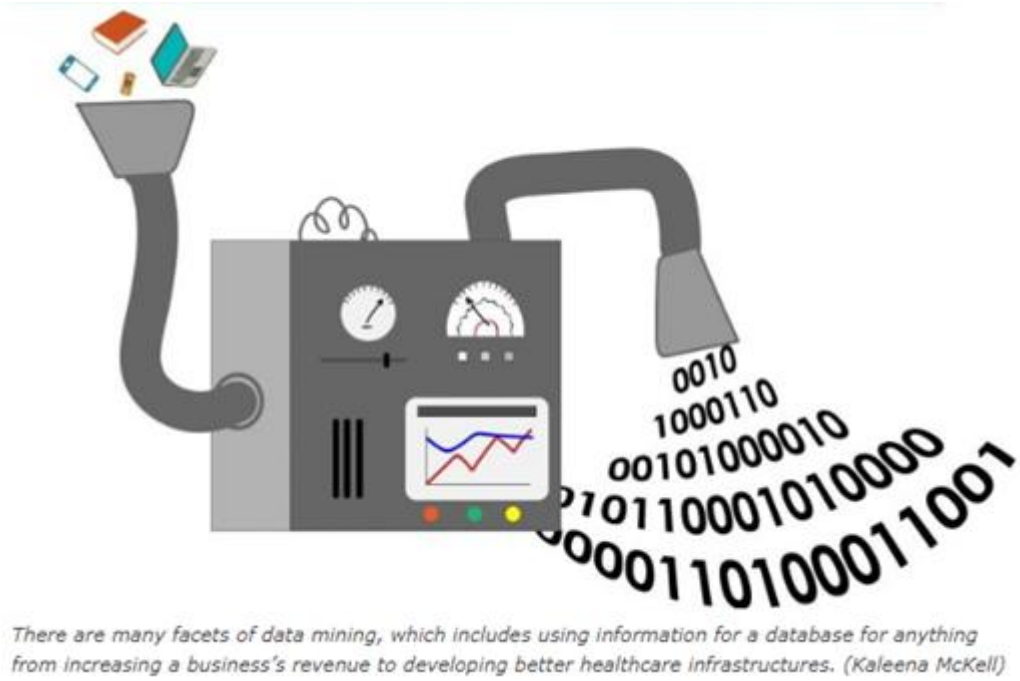


Εικόνα 5: Χάρτης διάχυσης

3. Εξόρυξη γνώσης βασισμένη στα Big Data (Data Mining)

Είναι γεγονός πως παρατηρείται ραγδαία αύξηση του όγκου των δεδομένων τα οποία συλλέγονται τόσο στην επιχειρησιακή όσο και στην επιστημονική έρευνα, για να πραγματοποιηθούν οι ανάλογες μελέτες. Πιο συγκεκριμένα, διάφορα επιστημονικά πεδία όπως η Αστρονομία ή η Ιατρική χρησιμοποιούν ένα μεγάλο όγκο δεδομένων τα οποία αντλούν από ογκώδεις βάσεις δεδομένων. Παράλληλα το ίδιο συμβαίνει και στον επιχειρησιακό τομέα, με τις πολυεθνικές εταιρείες, οι οποίες συλλέγουν terabytes δεδομένων από της πολυάριθμες καθημερινές συναλλαγές που πραγματοποιούνται. Το αξιοσημείωτο είναι ότι τα δεδομένα από τα παραπάνω παραδείγματα συλλέγονται με τρομερά υψηλές συχνότητες, η οποία αγγίζει τα gigabyte ανά ώρα.

Αν οι επεξεργασία όλου αυτού του όγκου δεδομένων σταματούσε στην συλλογή των δεδομένων είναι αυτονόητο ότι θα αποτελούσαν μια ανούσια πληροφορία. Ως εκ τούτου προκειμένου τα δεδομένα να παράγουν χρήσιμες πληροφορίες απαραίτητη είναι η κατάλληλη επεξεργασία τους. Στο σημείο αυτό να επισημανθεί ότι λόγω του τεράστιου μεγέθους των δεδομένων, η σωστή ανάλυση αποτελεί μια επιστημονική πρόκληση. Παράλληλα, μια ακόμα δυσκολία αποτελεί η αλληλεξάρτηση των δεδομένων, η οποία είναι εξαιρετικά πολύπλοκη. Όλα αυτά καθιστούν απαραίτητη την εφαρμογή ενός καινούριου επιστημονικού καθώς και ερευνητικού πεδίου. Το νέο αυτό πεδίο καλείται να επεξεργαστεί αλλά και να εξάγει συμπεράσματα τα οποία βασίζονται σε ένα μεγάλο όγκο των δεδομένων. Το καινούργιο επιστημονικό πλαίσιο ονομάζεται Εξόρυξη Γνώσης από Δεδομένα (Data Mining).



Εικόνα 6: Data Mining (*https://universe.byu.edu/2018/03/27/data-mining-1/*)

3.1. Ορισμός

Με τον όρο εξόρυξη γνώσης από δεδομένα ή αλλιώς εξόρυξη δεδομένων (data mining), αναφερόμαστε σε ένα κλάδο της τεχνολογίας όπου με δυναμικό τρόπο αναπτύσσονται τεχνικές με σκοπό να εστιάσουμε σε πληροφορίες που βρίσκεται μέσα στις αντίστοιχες αποθήκες δεδομένων (data warehouses). Πιο συγκεκριμένα, υλοποιούνται τεχνικές οι οποίες έχουν τη δυνατότητα να αναζητήσουν αλλά και να εντοπίσουν σε γρήγορο χρονικό διάστημα δεδομένα σε βάσεις με σκοπό την εξόρυξη σημαντικών πληροφοριών καθώς και την αναζήτηση κρυμμένων προτύπων (patterns). Ως εκ τούτου, καταλήγουμε στο να ορίσουμε συνοπτικά την εξόρυξη δεδομένων ως μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων.

Κατά την πάροδο του χρόνου, ο όρος εξόρυξη δεδομένων (data mining) έχει ως επί το πλείστων χρησιμοποιηθεί για να βοηθήσει στην μελέτη στατιστικών στοιχείων, από αναλυτές δεδομένων, και από συστήματα διαχείρισης πληροφοριών (management information systems- MIS). Παράλληλα, ο όρος είναι αρκετά δημοφιλής στο επιστημονικό πεδίο των βάσεων δεδομένων. Η πρώτη φορά που χρησιμοποιήθηκε ο όρος ήταν το 1989 σε ένα εργαστήριο (Piatetsky-Shapiro 1991) όπου υπογραμμίζεται ότι η γνώση είναι το τελικό προϊόν από μια ανακάλυψη με γνώμονα τα δεδομένα. Από τότε έως και σήμερα έχει υπάρξει μια αξιόλογη πρόοδο σε πολλά επιστημονικά

πεδία όπως για παράδειγμα η μηχανικής μάθησης. Με βάση τους Piatetsky-Shapiro και Frawley, το 1991 διατυπώθηκε ο ορισμός που ακολουθεί :

> «Η εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής γνώσης, μέσω της σύνδεσης χρήσιμης γνώσης υπό τη μορφή συσχετίσεων προτύπων και τάσεων, η οποία επιτυγχάνεται με την ανάλυση δεδομένων από τις διαθέσιμες πηγές (δομημένων η μη) και τη χρήση τεχνικών από τους τομείς της μηχανικής μάθησης, της αναγνώρισης προτύπων, της στατιστικής και της οπτικοποίησης».

Είναι κοινός αποδεκτό ότι η εξόρυξης δεδομένων έχει ως σκοπό να συλλέξει χρήσιμες και καινούριες πληροφορίες, επεξεργάζοντας τα δεδομένα που αποθηκεύονται στις βάσεις. Από την άλλη όμως, τα εργαλεία τα οποία χρησιμοποιούνται για την επίτευξη αυτού του σκοπού ποικίλουν και παρουσιάζεται μια ευρεία γκάμα. Πιο συγκεκριμένα, η εξόρυξη δεδομένων αποτελείται από ένα ευρύ πεδίο υπολογιστικών μεθόδων. Μερικά παραδείγματα υπολογιστικών μεθόδων αποτελούν τα δέντρα αποφάσεων (decision trees), η γραφική οπτικοποίηση (graphic visualization), τα νευρωνικά δίκτυα (neural networks), η στατιστική ανάλυση (statistical analysis), καθώς και η εξαγωγή κανόνων (rule induction). Οι μέθοδοι που προαναφέρθηκαν χρησιμοποιούνται με σκοπό να εντοπίσουν συσχετίσεις μεταξύ των δεδομένων, αλλά και τον εντοπισμό προτύπων και δομών οι οποίες υπάρχουν στις δυναμικά αναπτυσσόμενες βάσεις δεδομένων.

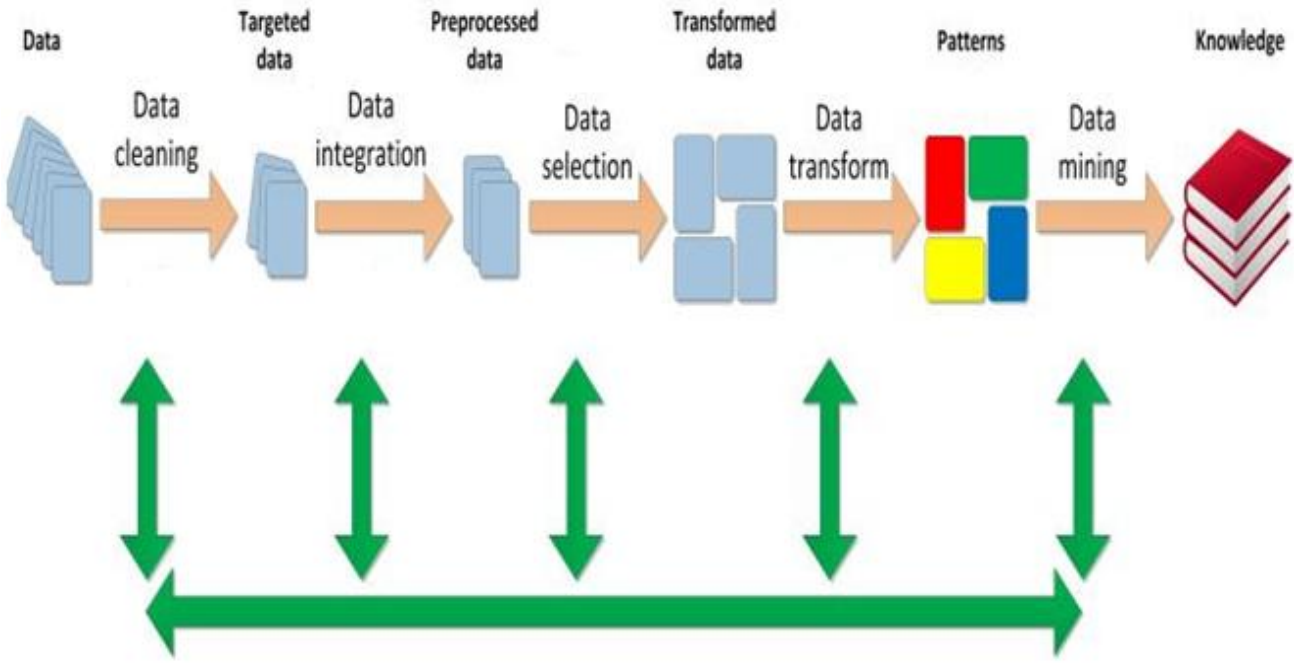
3.2. Ανάκτηση γνώσης από εξόρυξη δεδομένων

Τα τελευταία χρόνια ο όγκος των δεδομένων ο οποίος συλλέγεται από τις εταιρίες καθημερινά είναι μεγάλος. Αυτό καθιστά αναγκαίο το να αναπτυχθούν διάφοροι τρόποι οι οποίοι αποσκοπούν στη σωστή διαχείριση των δεδομένων, με σκοπό να αποτελέσουν εργαλείο για την εταιρία που θα βοηθήσει στην ανάπτυξή της. Στο σημείο αυτό εφαρμόζεται η εξόρυξη δεδομένων, η οποία καθιστά εφικτό για τις σύγχρονες εταιρείες να επικεντρωθούν στα δεδομένα των βάσεων και να λάβουν χρήσιμες πληροφορίες κάνοντας βιώσιμη και αποτελεσματική τη λειτουργία τους. Ουσιαστικά, με τη βοήθεια της εξόρυξης δεδομένων από μεγάλες βάσεις επιτυγχάνεται η ανάκτηση χρήσιμων πληροφοριών που βοηθούν στην καλύτερη ανάπτυξη της εταιρείας.

Για τη σύγχρονη επιχείρηση, η γνώση αποτελεί πολύτιμο συστατικό και η Εξόρυξη Δεδομένων είναι το εργαλείο για την ανάκτηση της. Η Εξόρυξη Δεδομένων εφαρμόζεται καθημερινά σε ένα μεγάλο ποσοστό επιχειρηματικών διαδικασιών. Ένα παράδειγμα εφαρμογής της αποτελεί η χρήση της στις διαφημίσεις αλλά και στις πωλήσεις, όπου χρησιμοποιείται με σκοπό να μελετηθεί η συμπεριφορά των

καταναλωτών. Εφόσον μελετηθεί ο τρόπος με τον οποίο επιλέγουν οι καταναλωτές να επενδύσουν τα λεφτά τους, οι εταιρείες λαμβάνουν σοβαρά υπόψη τους τα αποτελέσματα που προέκυψαν, βοηθώντας έτσι τον σχεδιασμό των προϊόντων ή της υπηρεσίας που προσφέρουν. Εν συνεχεία προσαρμόζουν την διαφήμιση στα θέλω του καταναλωτικού κοινού με αποτέλεσμα να υποβοηθούνται οι πωλήσεις. Παράλληλα, να τονιστεί ότι με το να γνωρίζει η εταιρία το προφίλ του πελάτη στον οποίο απευθύνεται, διευκολύνει το εισερχόμενο μάρκετινγκ καθώς και το καθοδηγούμενο από γεγονότα μάρκετινγκ. Ένα ακόμα σημαντικό παράδειγμα αποτελούν οι τράπεζες, στις οποίες η εξόρυξη δεδομένων εκτός από τη διαφήμιση και την πώληση εφαρμόζεται και για να γίνεται σωστή διαχείριση του ρίσκου ή για την αποφυγή απάτης. Όπως μπορούμε να καταλάβουμε η γνώση η οποία θα εξαχθεί από τις βάσεις δεδομένων μπορεί να αποφέρει πολλά οφέλη σε μία επιχείρηση. Αρχικά πολύ σημαντικό είναι ότι μια εταιρία με την εξόρυξη δεδομένων μπορεί να προβλέψει μελλοντικές συμπεριφορές. Κατά δεύτερον μπορούν να δημιουργηθούν διάφορα πρότυπα και να αποτυπωθούν οι συνήθειες των καταναλωτών. Τέλος, οι εταιρείες κινούνται με μεγαλύτερη ασφάλεια καθώς αποκτούν εμπειρία βασισμένη στα θέλω και τις ανάγκες των καταναλωτών.

Σκοπός της παρούσας ενότητας είναι να αναλυθεί το πως μπορούμε να αποκτήσουμε διάφορες γνώσεις, όπως τα παραδείγματα που προαναφέραμε με τη βοήθεια της εξόρυξη δεδομένων από μεγάλες βάσεις δεδομένων. Αρχικά είναι σημαντικό να γίνει αντιληπτή η έννοια τόσο της εξόρυξη γνώσης (Data mining), όσο και της ανεύρεση γνώσης στις βάσεις δεδομένων (Knowledge discovery in data bases, KDD). Στη διεθνή βιβλιογραφία, λόγω του ότι οι δύο όροι παρουσιάζουν μια συνάφεια, υπάρχει μια γενικότερη σύγχυση. Η σχέση που συνδέει τους δύο αυτούς όρους πολλές φορές μπορεί να ταυτίζεται και κάποιες άλλες η εξόρυξη δεδομένων να περιγράφεται ως υποσύνολο της ανάκτησης γνώσεων. Έχοντας ως στόχο να γίνει αποσαφήνιση και πλήρη κατανόηση του όρου παρακάτω αναπτύσσεται ο τρόπος που ακολουθείται κατά την ανάκτηση γνώσης μέσω της εξόρυξης δεδομένων. Προκειμένου να ανακτήσουμε γνώσεις ακολουθεί μια επαναληπτική διαδικασία η οποία δομείται με βάση μια ακολουθία. Ακολουθώντας την σειρά αυτή γίνεται εφικτή η συλλογή δεδομένων και η επεξεργασία τους με σκοπό να μπορεί ο χρήστης να λάβει σημαντικές πληροφορίες.



Εικόνα 7: Βήματα επεξεργασίας (https://www.researchgate.net/figure/Steps-in-processes-of-knowledge-discovery-Data-cleaning-removes-data-samples-containing_fig1_305481273)

1. Καθαρισμός δεδομένων (Data cleaning): Κατά το πρώτο βήμα της επεξεργασίας γίνεται απομάκρυνση τον δεδομένων που δεν μπορούν να βοηθήσουν στο τελικό αποτέλεσμα

2. Ενσωμάτωση δεδομένων (Data integration): Στο δεύτερο βήμα εφαρμόζεται μια λύση ενοποίησης των δεδομένων

3. Επιλογή δεδομένων (Data selection): Στη συνέχεια γίνεται επιλογή όσων δεδομένων είναι συναφή με την έρευνα που θα ακολουθηθεί

4. Τροποποίηση δεδομένων (Data transformation): Ύστερα από την επιλογή μόνο των χρήσιμων δεδομένων, προχωράμε στην επιθυμητή τροποποίηση των δεδομένων. Αυτό γίνεται με σκοπό να είναι έτοιμα τα δεδομένα για να γίνει η διαδικασία της εξόρυξης γνώσης

5. Εξόρυξη δεδομένων (Data mining): Όπως μπορούμε να καταλάβουμε και από τον τίτλο, το βήμα αυτό είναι το κρισιμότερο για την συγκεκριμένη διαδικασία. Στο βήμα αυτό με την εφαρμογή διαφόρων τεχνικών προκύπτουν τα πρότυπα με δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις

6. Αξιολόγηση προτύπων (Pattern evaluation): Σε αυτό το στάδιο αναλύονται τα πρότυπα που προέκυψαν με βάση επιστημονικές μετρικές αξιολόγησης

7. Αναπαράσταση γνώσης (Knowledge representation): Τέλος, εφόσον έχει παραχθεί η τελική γνώση ακολουθώντας πιστά τα παραπάνω βήματα, γίνεται αναπαράσταση της γνώσης αποτυπώνοντας κατανοητά τα συμπεράσματα που προέκυψαν.

Με βάση τα παραπάνω καταλήγουμε στο συμπέρασμα ότι βασικό στοιχείο στην ανάκτησης γνώσης αποτελεί η εξόρυξη δεδομένων. Η εξόρυξη δεδομένων στην διαδικασία που προαναφέρθηκε αποτελεί μόνο ένα βήμα στην επίτευξη της εξαγωγής γνώσεις από ογκώδεις βάσεις δεδομένων. Παρά το γεγονός αυτό, αποτελεί το σημαντικότερο βήμα στην επίτευξη του στόχου αλλά και το πιο πολύπλοκο.

3.3. Στόχος της εξόρυξης δεδομένων μέσω ανάκτηση γνώσης

Βασικός σκοπός της ανάκτηση γνώσης, όπως έχει αναφερθεί ήδη είναι η εξόρυξη πληροφοριών μέσω ενός μεγάλου όγκου δεδομένων προκειμένου να βοηθήσει τη λειτουργία των επιχειρήσεων καθώς και των υπηρεσιών. Μέσω της γνώσης που θα παραχθεί η εκάστοτε επιχείρηση ή ο οργανισμός έχει ως στόχο να συγκεντρώσει πληροφορίες που θα την κάνουν βιώσιμη στο μέλλον. Με βάση τα δεδομένα που παράγονται οι εταιρίες αναπτύσσουν διάφορα πρότυπα τα οποία μοντελοποιούν με σκοπό να προβλέψουν μελλοντικές κινήσεις τον καταναλωτών αποκομίζοντας το μεγαλύτερο κέρδος.

Πιο συγκεκριμένα, ας υποθέσουμε ότι ένα μεγάλο ποσοστό καταναλωτών που αγοράζει κρεβάτι, μετά αγοράζει σαλόνι. Αυτή την πληροφορία θα μπορούσε να την χρησιμοποιήσει η εκάστοτε εταιρεία και εν συνεχεία να προσφέρει στους συγκεκριμένους καταναλωτές ένα πακέτο ή κάποια προσφορά με σκοπό όλα τα προϊόντα που θα χρειαστεί για την επίπλωση του σπιτιού του να τα αγοράσει από την συγκεκριμένη επιχείρηση. Παράλληλα ένα άλλο πρότυπο που μπορεί να αναπτυχθεί είναι μια ομάδα καταναλωτών οι οποίοι επιλέγουν να αγοράσουν πολλές φορές ένα προϊόν κατά την εκπτωτική περίοδο ή κάποιος που αγοράζει έναν υπολογιστή πιθανών να αγοράσει και άλλες ηλεκτρονικές συσκευές.

Όλες αυτές οι πληροφορίες αποτελούν για την εταιρεία σημαντικούς παράγοντες για την ανάληψη αποφάσεων ως προς την λειτουργία της. Παραδείγματα των αποφάσεων αυτών θα μπορούσε να είναι το πως θα δομηθεί το πρόγραμμα τον εργαζομένων, το πως θα κυμανθούν τα ποσοστά των εκπτώσεων αλλά και το πως θα ταξινομηθούν στο χώρο της επιχείρησης τα προϊόντα. Τέλος πολύ βασικό είναι να αναφερθεί ότι οι πληροφορίες που ανακτώνται αποσκοπούν στο να βοηθήσουν στην διαδικασία του μάρκετινγκ της

επιχείρησης. Ακολουθούν οι στόχοι της ανάκτηση γνώσης μέσω της εξόρυξης δεδομένων.

- Πρόβλεψη: Μέσω της εξόρυξης δεδομένων γίνεται εφικτή η ανάδειξη μελλοντικών συμπεριφορών με βάση ορισμένα γνωρίσματα που αποτυπώνονται από τα δεδομένα που ανακτώνται από τις ογκώδεις βάσεις δεδομένων.

- Ταυτοποίηση: Βάση της ταυτοποίησης δίνεται η δυνατότητα ώστε οι μορφές των δεδομένων να χρησιμοποιηθούν με σκοπό να προσδιοριστεί η ύπαρξη ενός προϊόντος, ενός γεγονότος, η μιας δραστηριότητας

- Περιγραφή: Σκοπός της συγκεκριμένης διαδικασία είναι να αναπτυχθούν διάφορα πρότυπα και να αποτυπωθούν τα δεδομένα με τον καλύτερο δυνατό τρόπο προκειμένου να είναι πιο κατανοητά και χρήσιμα.

- Ταξινόμηση: Με τη βοήθεια της ταξινόμησης, η εξόρυξη δεδομένων μπορεί να τμηματοποιήσει τα δεδομένα ώστε να μπορούν να αναπτυχθούν διαφορετικές κατηγορίες συνδυάζοντας κατάλληλα την παραμετροποίηση.

- Βελτιστοποίηση: Τέλος ένας βασικός στόχος της εξόρυξης δεδομένων που δίνει την δυνατότητα στις εταιρείες να υπάρξουν καλύτερα αποτελέσματα και αποδοτικότερη ανάκτηση γνώσεων είναι η βελτιστοποίηση. Επομένως, η βελτιστοποίηση αποσκοπεί στο να επιτευχθεί γρηγορότερη ανάκτηση γνώσεων, να επέλθει μείωση των εξόδων που δαπανούνται ή και να μεγιστοποιηθούν τα έξοδα όπως για παράδειγμα οι πωλήσεις ή τα κέρδη με βάση τους περιορισμούς που επικρατούν.

Στο σημείο αυτό να αναφερθεί ότι οι εμπορικές εφαρμογές οι οποίες αποσκοπούν στην απόκτηση γνώσεων μέσω της εξόρυξης δεδομένων, στοχεύουν στο να αναπτυχθεί όχι μόνο σε υψηλού επιπέδου προγραμματιστικά περιβάλλοντα αλλά και τον μέσο χρήστη ο οποίος θέλει να κατανοήσει την επιστήμη της στατιστικής, να εκπαιδεύσει το δικό του σύστημα ή να πειραματιστεί με την τεχνητή νοημοσύνη. Οι απαιτήσεις ποικίλουν, τόσο για μια μεγάλη επιχείρηση η οποία θέλει να επεξεργαστεί στατιστικά δεδομένα με σκοπό τις μελλοντικές της κινήσεις όσο και ένας χρήστης που χρησιμοποιεί ένα σύστημα μηχανικής μάθησης. Στοχεύοντας σε διαφορετικά πεδία, μπορούν να χρησιμοποιούν λογισμικό εξόρυξης γνώσης με σκοπό να λάβουν τα αποτελέσματα που έχουν ανάγκη.

Με βάση τα παραπάνω, καταλήγουμε στο συμπέρασμα πως οι βασικοί στόχοι που έχει η εξόρυξη δεδομένων είναι να εφαρμόσει διάφορες τεχνικές πρόβλεψης της συμπεριφοράς των χρηστών, την

ταυτοποίηση και την περιγραφή μεγάλου όγκου δεδομένων, καθώς και την ταξινόμηση, αλλά και την βελτιστοποίηση των πόρων της. Όλοι αυτοί οι στόχοι είναι πολύ σημαντικοί, βοηθώντας ο καθένας στην αποδοτικότερη και ορθότερη απόκτηση γνώσεων. Τέλος να τονίσουμε την σημαντικότητα του τελευταίου στόχου, ο οποίος κάθε φορά αποδίδει όλο και καλύτερα αποτελέσματα βοηθώντας την επιχειρησιακή έρευνα και συμβάλει στην βελτίωση τον υπόλοιπων στόχων.

3.4. Εφαρμογές εξόρυξη γνώσης

Στο πλαίσιο αυτής της ενότητας μελετώνται μερικές από τις σημαντικότερες εφαρμογές της επιστήμης της εξόρυξης γνώσεων. Τα παραδείγματα εξόρυξης γνώσης από ογκώδεις βάσεις δεδομένων είναι πάρα πολλά και ποικίλουν με βάση τον επιστημονικό και επιχειρησιακό τομέα στον οποίο αναπτύσσονται. Πιο συγκεκριμένα, οι εφαρμογές της εξόρυξης γνώσεις μπορούν να κυμαίνονται από την στατιστική ανάλυση των πληροφοριών που συλλέγονται από τις βάσεις δεδομένων ως και το να υλοποιηθούν καμπάνιες προώθησης για την διευκόλυνση του μάρκετινγκ. Παράλληλα, αναλύονται οι τομείς που έχουν άμεση σχέση με την εξόρυξη γνώσεων.

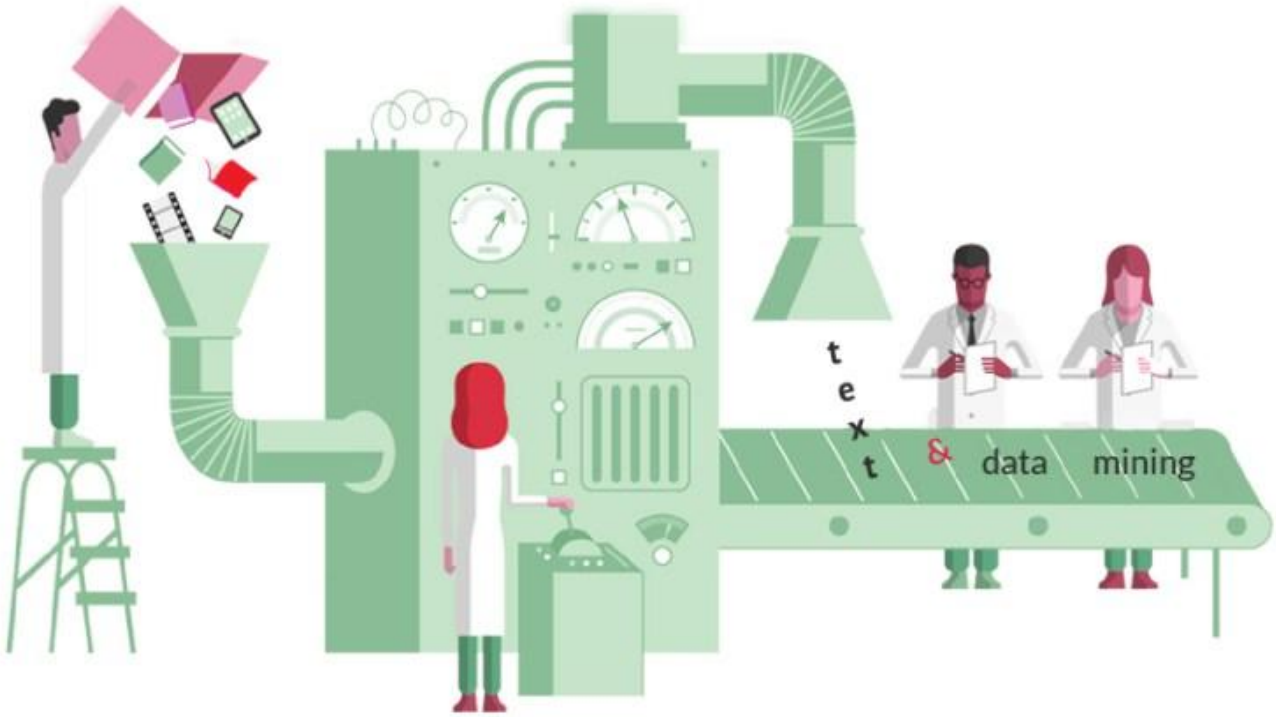
Είναι πλέον γνωστό πως καθημερινά παρατηρείτε μια ολοένα και αυξανόμενη ροή δεδομένων κάτι που οδηγεί στο να γίνεται ολοένα και πιο περίπλοκη η επεξεργασία τους. Παρόλα αυτά, πλέον με βάση διάφορες τεχνικές που εφαρμόζονται, η πρόσβαση αλλά και η επεξεργασία αυτού του μεγάλου όγκου δεδομένων με σκοπό την εξόρυξη γνώσης γίνεται εφικτή. Ως εκ τούτου η μεγάλη ανάγκη που δημιουργείται, οδήγησε στην ανάπτυξη εφαρμογών εξόρυξης γνώσης.

Όπως αναφέρθηκε και σε παραπάνω ενότητα όταν αναφερόμαστε στον όρο ανάλυση δεδομένων, απαραίτητη διεργασία που πρέπει να γίνει είναι η αποτελεσματική επεξεργασία των δεδομένων. Η σωστή επεξεργασία αποσκοπεί στην οργάνωση των πληροφοριών που εξάγονται βοηθώντας τον χρήστη στην τελική ερευνά. Είναι κατανοητό ότι μια σωστά επεξεργασμένη και οργανωμένη πληροφορία θα οδηγήσει στην εξαγωγή καλύτερων αποτελεσμάτων με αποτέλεσμα η λήψη των αποφάσεων να χαρακτηρίζεται από ορθολογικότητα.

Αν μπούμε στη διαδικασία να σκεφτούμε ορισμένα παραδείγματα εφαρμογών εξόρυξης δεδομένων πολύ τομείς θα μας έρθουν στο μυαλό. Ορισμένοι από τους σημαντικότερους τομείς οι οποίοι παρουσιάζουν ιδιαίτερο ενδιαφέρον είναι οι ακόλουθοι:

- ο τομέας των τηλεπικοινωνιών,
- ο χρηματοοικονομικός τομέας,
- ο τομέας της βιομηχανίας και της έρευνας,
- ο τομέας της υγείας και

- ο τομέας της εκπαίδευσης



Πηγή: https://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/

Στη συνέχεια θα αναλύσουμε τους τομείς που προαναφέρθηκαν με σκοπό να γίνει αντιληπτό το πόσο σημαντική είναι η εφαρμογή της εξόρυξης γνώσεις για να αναπτυχθούν αυτοί οι τομείς. Αρχικά, ως αναφορά τον τομέα των τηλεπικοινωνιών, έχει γίνει κατανοητό ότι η ανάλυση του μεγάλου όγκου δεδομένων που συνεχώς συλλέγεται, βοηθά τις εταιρίες να αναπτύξουν νέες μεθόδους για να εδραιωθούν έναντι των ανταγωνιστών τους αλλά και να κατανοήσουν το συνεχώς μεταβαλλόμενο περιβάλλον του τομέα τον τηλεπικοινωνιών. Οι τεχνικές που εφαρμόζονται μεταξύ άλλων είναι το να κατηγοριοποιήσω τα θέλω των πελατών με σκοπό την παροχή κατάλληλων υπηρεσιών, να μελετήσουν τα χαρακτηριστικά των χρηστών έτσι ώστε να προσελκύσουν νέους πελάτες αλλά και το να προβλέπεται πότε ο πάροχος θα πρέπει να προβεί σε διακοπή της σύνδεσης.

Από την άλλη στον χρηματοοικονομικό τομέα παρατηρείται μεγάλη αύξηση του ανταγωνισμού γεγονός που απορρέει τόσο από τον μεγάλο αριθμό τραπεζικών ιδρυμάτων όσο και από την οικονομική κατάσταση που επικρατεί στην χώρα. Επομένως με την εφαρμογή τεχνικών εξόρυξης επιτυγχάνεται το να παρέχονται κίνητρα στους πελάτες ώστε να μην αλλάζουν τραπεζικό ίδρυμα αλλά και να παρέχονται δελεαστικά κίνητρα ώστε να προσελκύσουν καινούριους ανεβάζοντας έτσι τα κέρδη. Παράλληλα με την εφαρμογή τεχνικών

προβλέψεις αποφεύγονται κινήσεις που μπορούν να ζημιώσουν το εν λόγω τραπεζικό ίδρυμα όπως για παράδειγμα η ανάλυση του πιστωτικού κινδύνου που μπορεί να προέλθει από κάποιον πελάτη.

Ένας άλλος τομέας στον οποίο βρίσκει εφαρμογή η εξόρυξη δεδομένων είναι ο τομέας της βιομηχανίας και της έρευνας. Οι διαρκώς μεταβαλλόμενες προτιμήσεις των πελατών αλλά και ο μεγάλος όγκος δεδομένων που συλλέγονται καθημερινά από τα προϊόντα που επιλέγουν, καθιστά άκρως βοηθητική την σωστή επεξεργασία των δεδομένων έχοντας ως στόχο την εφαρμογή των γνώσεων που προκύπτουν. Με την βοήθεια των γνώσεων που απορρέουν από όλα τα δεδομένα τα οποία συλλέγονται καθημερινά, δημιουργούνται διάφορα πρότυπα. Εν συνεχεία, έχοντας τα πρότυπα αυτά στη διάθεσή τους επιτυγχάνεται εξατομικευμένη σχέση απέναντι σε κάθε καταναλωτή αποσκοπώντας σε μια μακροχρόνια συνεργασία και αύξηση των εσόδων.

Στο σημείο αυτό, θα ήταν αδύνατο να μην αναφερθούμε στον τομέα της υγείας, ο οποίος αποτελεί έναν από τους σημαντικότερους τομείς και είναι προς όφελος όλων μας οι γνώσεις που εξάγονται. Πιο συγκεκριμένα η εφαρμογή της εξόρυξης δεδομένων, μεταξύ άλλων αφορά την έρευνα των αποτελεσμάτων των φαρμάκων κατηγοριοποιώντας και μελετώντας σωστά τα δεδομένα. Παράλληλα μελετάτε και προβλέπεται η συμπεριφορά που έχουν διάφορες ασθένειες, συμβάλλοντας έτσι στην έγκυρη και έγκαιρη αντιμετώπιση τους. Τέλος, ο τομέας της εκπαίδευσης βελτιώνεται με την εφαρμογή της εξόρυξης γνώσεις κάνοντας πιο εύκολες τις έρευνες που αναπτύσσονται σε ακαδημαϊκό επίπεδο αλλά και βελτιώνοντας τον τρόπο διδασκαλίας.

Παραπάνω εξετάστηκαν οι κυριότερες εφαρμογές της εξόρυξης γνώσης κάνοντας έτσι κατανοητό ότι οι χρήσεις της εξόρυξης γνώσης δεν περιορίζονται μόνο στην απλή στατιστική, αλλά παρεμβαίνουν δυναμικά, προβλέποντας μελλοντικά αποτελέσματα και μελετώντας τις κινήσεις των χρηστών. Επομένως, συμπεραίνουμε ότι η εξόρυξη γνώσης αποτελεί αναπόσπαστο κομμάτι σε αρκετούς τομείς της κοινωνίας μας. Επίσης η ολοένα και αυξανόμενη συλλογή δεδομένων καθιστά τις μέχρι τώρα υπάρχουσες τεχνικές αναποτελεσματικές και χρονοβόρες. Πλέον η επεξεργασία των πληροφοριών γίνεται σε πολύ πιο γρήγορο χρόνο και βελτιώνεται διαρκώς με την βοήθεια του ιστορικού που αναλύεται.

3.5. Συμπεράσματα

Στο παρόν κεφάλαιο έγινε μια προσπάθεια να αναλύσουμε τον ρόλο της εξόρυξης γνώσεις από μεγάλες βάσεις δεδομένων. Όπως γίνετε κατανοητό, η συλλογή δεδομένων στις ογκώδεις βάσεις δεν θα είχε κανέναν νόημα αν δεν τις επεξεργαζόμαστε, αξιοποιώντας τα δεδομένα με σκοπό να ανακτήσουμε διάφορες γνώσεις. Όταν

αναφερόμαστε στον όρο μεγάλα δεδομένα (big data), εννοούμε τις πληροφορίες που συγκεντρώνονται στο διαδίκτυο και μπορεί να περιλαμβάνουν από κείμενα έως και φωτογραφίες και video.

Τα τελευταία χρόνια έχουμε μια διαρκή ροή τέτοιων πληροφοριών, στις ήδη ογκώδεις βάσεις δεδομένων. Ύστερα από αυτό καταλήγουμε στο συμπέρασμα ότι η διαδικασία που συντελείται με σκοπό την εξόρυξη γνώσης, αποτελεί μια πολύπλοκη αλλά και χρονοβόρα διαδικασία. Η πολυπλοκότητα της εξόρυξης γνώσεις οφείλεται κατά κύριο λόγο στην διαρκώς μεταβαλλόμενη μορφή των δεδομένων αλλά και στον διαρκώς αυξανόμενο όγκο τους. Όλο αυτό καθιστά αναγκαίο να αναπτυχθούν διάφορες αυτοματοποιημένες τεχνικές επεξεργασίας, οι οποίες έχουν ως στόχο την εξόρυξη γνώσης από τις ογκώδεις βάσεις δεδομένων.

Η συνεισφορά της εξόρυξης δεδομένων, όπως αναφέρθηκε αποτελεί σημαντικό πυλώνα για την καλύτερη λειτουργία πολλών τομέων, όπως ο χρηματοοικονομικός ή της εκπαίδευσης. Πιο συγκεκριμένα βρίσκει εφαρμογή κατά την συλλογή, την κατανόηση καθώς και την βελτίωση των πληροφοριών. Εν συνεχεία η σωστή ανάλυση των πληροφοριών που εξάγονται βοηθά έτσι ώστε η εκάστοτε εταιρία ή ο οργανισμός να λάβει σωστότερες αποφάσεις. Παράλληλα, τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων αρκετές φορές μπορούν να χρησιμοποιηθούν από οικονομικούς αναλυτές για την λήψη στρατηγικών αποφάσεων στο ανάλογο οικονομικό πεδίο που ενδιαφέρει κάθε φορά. Τέλος να τονίσουμε ότι απαραίτητη προϋπόθεση είναι να υπάρχει το κατάλληλο υπόβαθρο από ιστορικά δεδομένα ώστε να είναι δυνατή η σωστή επεξεργασία και εν τέλει η ανάλυση των πληροφοριών που ανακτώνται από τις ογκώδεις βάσεις δεδομένων.

Εν κατακλείδι, το συμπέρασμα είναι ότι στις μέρες μας, είναι εφικτό με την συμβολή της τεχνολογίας οι χρήσεις της εξόρυξης γνώσης να μην περιορίζονται στην στατιστική, αλλά να παρεμβαίνουν δυναμικά, προβλέποντας μελλοντικές συμπεριφορές και καθαρίζοντας τις κινήσεις με την εφαρμογή διάφορων μοντέλων.

4. Apache Spark

Σε αυτή την ενότητα θα γίνει μία συνοπτική περιγραφή του Apache Spark. Το Apache Spark είναι μια ισχυρή μηχανή επεξεργασίας ανοιχτού κώδικα που βασίζεται στην ταχύτητα, την ευκολία χρήσης και τα εξελιγμένα αναλυτικά στοιχεία. Αρχικά αναπτύχθηκε στο UC Berkeley το 2009. Το Databricks [databricks] ιδρύθηκε από τους δημιουργούς του Spark το 2013.

Το engine Spark τρέχει πάνω από ένα μεγάλο σύνολο περιβαλλόντων, από τις cloud υπηρεσίες μέχρι Hadoop ή Mesos [mesos] clusters. Χρησιμοποιείται για την εκτέλεση ETL διαδικασιών, διαδραστικά ερωτήματα (SQL), προηγμένες αναλύσεις (π.χ. μηχανική μάθηση) και streaming πάνω σε μεγάλα σύνολα δεδομένων σε ένα ευρύ φάσμα αποθηκευμένων δεδομένων (π.χ. HDFS, HBase [hbase]). Το Spark υποστηρίζει μια ποικιλία από δημοφιλείς γλώσσες προγραμματισμού, όπως είναι η Java, η Python και η Scala.

Από τη μέρα κυκλοφορίας του, το Spark έχει δει ταχεία υιοθέτηση του από τις επιχειρήσεις σε ένα ευρύ φάσμα βιομηχανιών. Ταυτόχρονα, έγινε πολύ γρήγορα η μεγαλύτερη κοινότητα ανοιχτού κώδικα σε μεγάλα δεδομένα, με περισσότερους από 500 συνεργάτες από 100+ οργανισμούς.

4.1. Spark έναντι άλλων Big Data frameworks

Το Spark έχει αρκετά πλεονεκτήματα σε σύγκριση με άλλα frameworks για τη διαχείριση μεγάλων δεδομένων και MapReduce τεχνολογίες όπως είναι το Hadoop και το Storm [storm]. Πρώτα απ' όλα, το Spark παρέχει ένα ολοκληρωμένο και ενοποιημένο πλαίσιο για τη διαχείριση και την επεξεργασία μεγάλων δεδομένων που προέρχονται από διαφορετικές μορφές αναπαράστασης δεδομένων, όπως είναι τα δεδομένα κειμένου, τα δεδομένα γραφημάτων, κλπ. Το Spark είναι σε θέση να εξάγει δεδομένα από μία μεγάλη ποικιλία πηγών, όπως είναι η συλλογή δεδομένων μνήμης, αρχεία κειμένου και πολλά κατανεμημένα συστήματα αρχείων (Cassandra, HDFS κ.λ.π.).

Επιπρόσθετα, το Spark υποστηρίζει την επεξεργασία και την ανάλυση αναλυμένων δεδομένων εντός της μνήμης, γεγονός που βελτιώνει την απόδοση επαναληπτικών εργασιών. Για παράδειγμα, το Spark επιτρέπει σε εφαρμογές σε Hadoop clusters να τρέχουν έως και 100 φορές ταχύτερα στη μνήμη σε σύγκριση με το Map-Reduce [Matei Zaharia κ.α., 2012]. Επιπλέον, το Spark διαθέτει ένα ενσωματωμένο σύνολο από πάνω από 80 τελεστές υψηλού επιπέδου. Ένα μεγάλο πλεονέκτημα του Spark έναντι άλλων frameworks που αποθηκεύουν δεδομένα στη μνήμη είναι ότι προσφέρει fault tolerance με έναν υπολογιστικά φθηνό τρόπο, καθώς κρατά ως πληροφορία το lineage και δεν χρησιμοποιεί τεχνικές όπως είναι το data replication ή τα checkpoint.

4.2. Βασικά Components

Όσον αφορά την αρχιτεκτονική, το Apache Spark βασίζεται σε δύο βασικές έννοιες:

- Resilient Distributed Datasets (RDDs) [Matei Zaharia κ.α., 2012]
- DAG execution engine

Όσον αφορά τα σύνολα δεδομένων, το Spark υποστηρίζει δύο τύπους RDD: παράλληλες συλλογές δεδομένων, που βασίζονται σε υπάρχουσες συλλογές δεδομένων της Scala και σύνολα δεδομένων Hadoop που δημιουργούνται από τα αρχεία που είναι αποθηκευμένα στο HDFS. Τα RDDs υποστηρίζουν δύο είδη πράξεων:

- Transformations
- Actions

Τα transformations δημιουργούν νέα σύνολα δεδομένων από την είσοδο (π.χ. λειτουργίες map και filter), ενώ τα actions επιστρέφουν μια τιμή μετά από την εκτέλεση υπολογισμών πάνω στο σύνολο δεδομένων (π.χ., reduce και count).

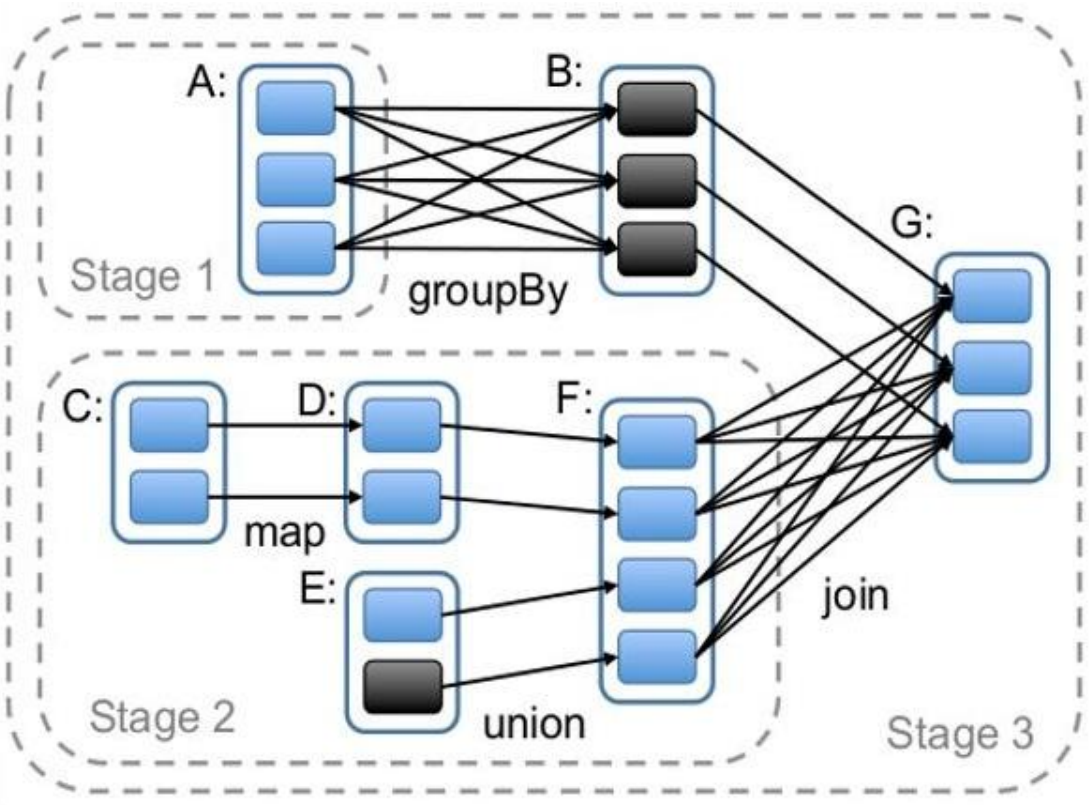
Το DAG engine βοηθά στην εξάλειψη του μοντέλου εκτέλεσης πολλαπλών σταδίων του MapReduce και προσφέρει σημαντικές βελτιώσεις στην απόδοση. Κάθε Spark job δημιουργεί ένα DAG των σταδίων των tasks που πρέπει να εκτελεστούν στο σύμπλεγμα. Σε σύγκριση με το MapReduce, το οποίο δημιουργεί ένα DAG με δύο προκαθορισμένα στάδια - Map and Reduce, τα DAG που δημιουργούνται από το Spark μπορούν να περιέχουν οποιοδήποτε αριθμό από στάδια. Αυτό επιτρέπει ορισμένες εργασίες να ολοκληρωθούν γρηγορότερα από ό, τι στο MapReduce, τόσο με απλές εργασίες που μπορεί να ολοκληρωθούν μετά από ένα μόνο στάδιο όσο και πιο πολύπλοκες εργασίες που συμπληρώνουν σε ένα μόνο spark Job πολλαπλών σταδίων, αντί να χρειάζεται να χωριστούν σε πολλαπλά jobs. Λόγω της lazy υλοποίησης και με βάση το DAG, το Spark μπορεί να εκτελέσει διαδοχικούς 1-1 τελεστές με τη τεχνική του pipeline.

4.3. Πλάνο εκτέλεσης

Το spark διαχωρίζει τα transformations σε δύο τύπους:

- Ευρείς εξαρτήσεις (Wide dependencies): Μετασχηματισμοί στους οποίους ένα γονικό partition θα χρησιμοποιηθεί από πολλαπλά partitions παιδιών. Οι ευρείς εξαρτήσεις απαιτούν το shuffling δεδομένων από όλα τα partitions του πατέρα στα παιδιά. Εάν ένα RDD partition ενός παιδιού χαθεί, τότε ολόκληρο το RDD του πατέρα θα πρέπει να υπολογιστεί ξανά. Για το λόγο αυτό, όταν βρεθεί μια ευρεία εξάρτηση, η υλοποίηση του μετασχηματισμού θα υπολογιστεί αμέσως.

- Στενές εξαρτήσεις (Narrow dependencies): Μετασχηματισμοί στους οποίους χρησιμοποιείται ένα γονικό partition το πολύ από ένα παιδί partition. Αυτοί οι μετασχηματισμοί δεν υλοποιούνται αμέσως, αλλά η υλοποίηση τους λαμβάνει χώρα όταν θα εμφανιστεί μια ευρεία εξάρτηση (ή ένας action τελεστής). Οι διαδοχικές στενές εξαρτήσεις μπορούν να διοχετευθούν. Εκτός από αυτό, μια στενή εξάρτηση στην οποία ένα διαμέρισμα παιδιού θα χρειαστεί το πολύ ένα διαμέρισμα πατέρα (π.χ. map, filter, union), μπορεί να υπολογιστεί τοπικά, χωρίς την ανάγκη μεταφοράς δεδομένων μεταξύ των όλων των workers.

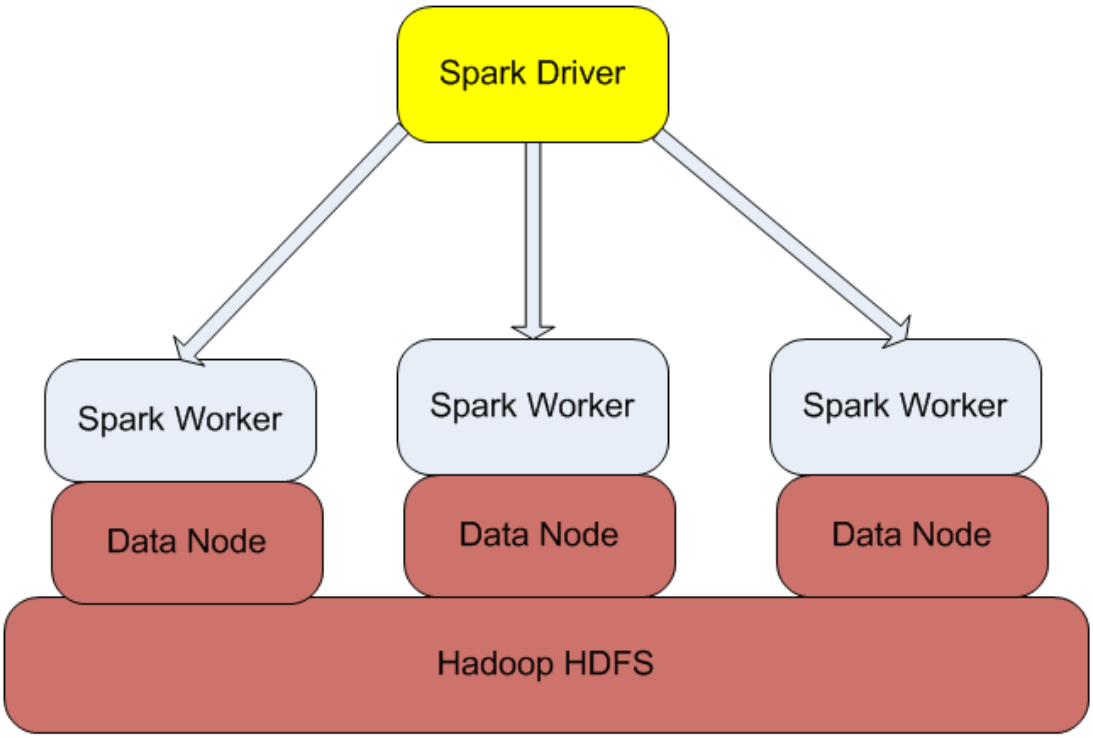


Εικόνα 8: Παράδειγμα πλάνου εκτέλεσης του Spark

Το Spark χρησιμοποιεί τη τεχνική της lazy υλοποίησης, αποφεύγοντας να υλοποιήσει μετασχηματισμούς, μέχρι να εμφανιστεί ένας action τελεστής. Σε αυτή την περίπτωση, το Spark χρησιμοποιεί το DAG των μετασχηματισμών και χωρίζει το σχέδιο σε πολλαπλά στάδια, με βάση τις ευρείες εξαρτήσεις. Πιο συγκεκριμένα, όταν θα βρεθεί μια ευρεία εξάρτηση, τότε θα ολοκληρωθεί ένα στάδιο (Εικόνα: Παράδειγμα πλάνου εκτέλεσης του Spark). Οι στενές εξαρτήσεις εντός ενός σταδίου μπορούν να εκτελεστούν με τη τεχνική της διοχέτευσης. Επιπρόσθετα, το Spark επιλέγει join αλγόριθμους και άλλους ευρείς key-value μετασχηματισμούς (π.χ. reduceByKey) με βάση το

partitioning. Επομένως, παρόλο που αυτοί οι μετασχηματισμοί είναι ευρείς, μερικές φορές, αν το partition είναι κατάλληλο, οι μετασχηματισμοί αυτοί μπορούν να γίνουν είτε στενοί (π.χ., hash join) είτε υβριδικοί (px το join στην εικόνα παραδείγματος του πλάνου εκτέλεσης του Spark). Τέλος, το Spark αποφεύγει των υπολογισμό partitions που έχουν αποθηκευτεί στη cache προηγουμένως.

4.4. Αρχιτεκτονική



Εικόνα *9*: Αρχιτεκτονική του Spark

Το Spark απαιτεί ένα cluster manager και ένα distributed storage system. Όσον αφορά το cluster manager, το Spark υποστηρίζει standalone [12], Hadoop YARN [yarn] ή Apache Mesos. Όσον αφορά το κατανεμημένο σύστημα αποθήκευσης, το Spark μπορεί να λειτουργήσει πάνω από το HDFS, τη Cassandra, το S3 και πολλά άλλα.

Όσον αφορά την εκτέλεση ενός προγράμματος από το χρήστη, στο Spark υπάρχουν δύο τύποι κόμβων, όπως φαίνεται και στην εικόνα Αρχιτεκτονική του Spark:

- Spark Driver: Ο spark Driver αποτελεί το συντονιστή μίας διεργασίας. Οι χρήστες γράφουν ένα driver program, το οποίο με ένα σύμπλεγμα από κόμβους εργάτες. Ο οδηγός ορίζει ένα ή περισσότερα RDD και

εκτελεί ενέργειες πάνω σε αυτά. Ο Driver Κόμβος είναι υπεύθυνος για την παρακολούθηση του lineage των RDDs [Matei Zaharia κ.α., 2012].

- Workers: Οι κόμβοι εργαζόμενοι είναι long-lived διαδικασίες που αποθηκεύουν τα δεδομένων των RDD partitions και εκτελούν ενέργειες πάνω σε αυτά με παράλληλο τρόπο.

4.5. Απλό Παράδειγμα

Σε αυτή την ενότητα, θα περιγράψουμε το σχέδιο εκτέλεσης ενός απλού παραδείγματος. Ας πούμε ότι έχουμε αποθηκεύσει τις γραμμές του παρακάνω κειμένου (το κείμενο είναι κομμένο, οπότε δεν βγάζει εννοιολογικό νόημα) σε τρεις φυσικές μηχανές:

to be or not to be
in the mind to suffer
the slings and the arrows
or to take arms
and by opposing end them
to say we end

Ο στόχος είναι να βρεθεί η λέξη με την υψηλότερη συχνότητα εμφάνισης, χωρίς ωστόσο να ληφθούν υπόψη οι λέξεις (the, a και an). Στον παρακάτω αλγόριθμο μπορούμε να δούμε τον κώδικα Spark, ο οποίος είναι υπεύθυνος να βρει τη λέξη με την υψηλότερη συχνότητα

Αλγόριθμος

Input: $source$: data set; $stopList$: stop-word list

Output: $higherFrequencyWord$: the word with the higher frequency

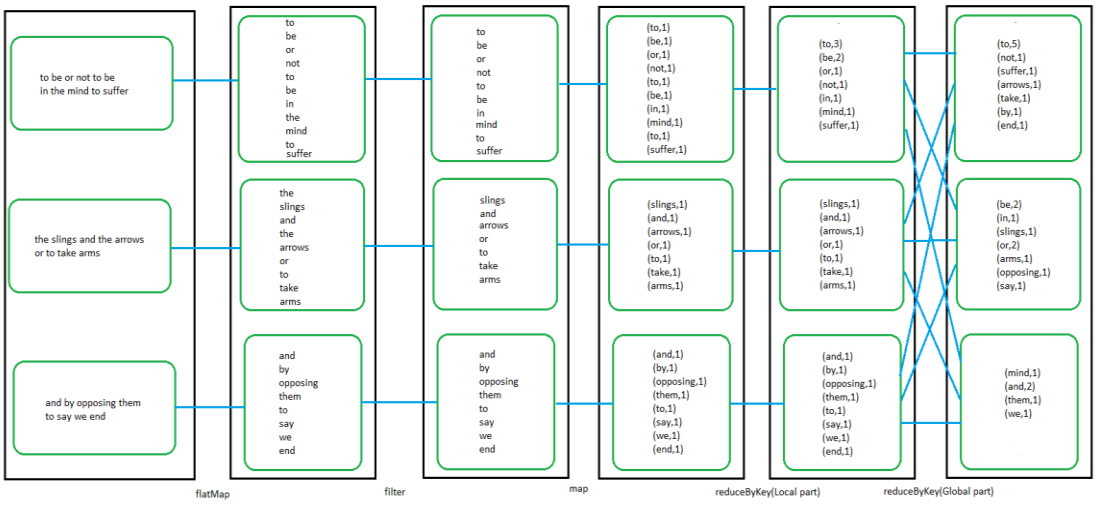
```
data = source.cache()
splitWords = data.flatmap(x.split(" "))
usefullWords = data.filter(filterStopWords)
initFrequency = usefullWords.map(word => (word,1))
getFrequency = initFrequency.reduceByKey((x,y) => x+y)
getLocalHigherFrequency = getFrequency.mapPartitions(higherFrequency)
candidateWords = getLocalHigherFrequency.collectAsMap()
```

higherFrequencyWord = higherFrequency(candidateWords)

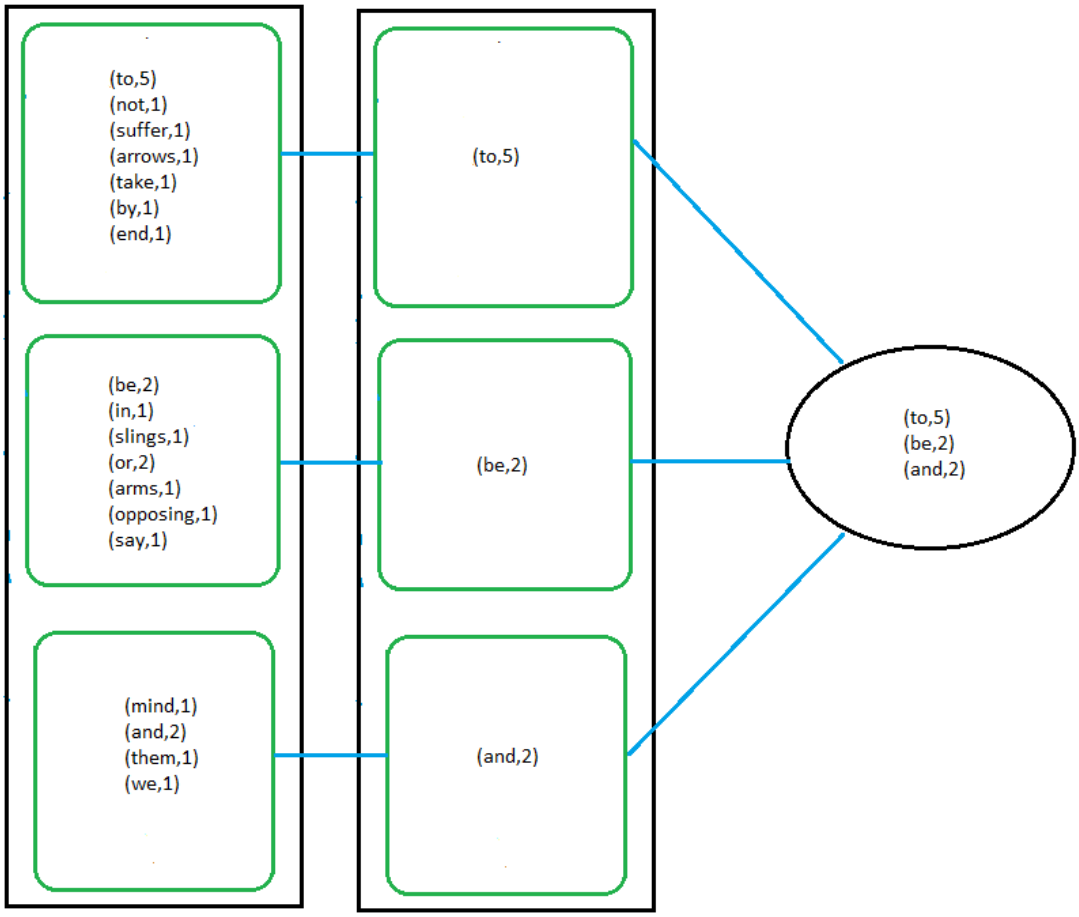
return higherFrequencyWord

Στον παραπάνω αλγόριθμο, αρχικά, χωρίζουμε κάθε γραμμή σε μία ή περισσότερες λέξεις, βάση τον delimiter «κενό». Στη συνέχεια, φιλτράρουμε τις λέξεις, αφαιρώντας τις λέξεις the, a και an. Οι υπόλοιπες λέξεις χαρτογραφούνται σε μια δισδιάστατη πλειάδα, όπου το πρώτο στοιχείο είναι η ίδια η λέξη, ενώ το δεύτερο στοιχείο είναι ο αριθμός 1. Στη συνέχεια, γίνετε μία τοπική συνάθροιση, όπου κάθε μηχάνημα συναθροίζει τις εγγραφές που αντιστοιχούν στην ίδια λέξη, υπολογίζοντας με αυτόν τον τρόπο, πόσες φορές εμφανίζεται μία λέξη στα τοπικά δεδομένα. Έπειτα τα δεδομένα ανακατεύονται μεταξύ των μηχανών, με τον περιορισμό ότι όλα τα διαθέσιμα δεδομένα για οποιαδήποτε συγκεκριμένη λέξη θα πάνε στο ίδιο μηχάνημα και εν συνεχεία γίνεται η τελική συνάθροιση. Μετά, κάθε partition υπολογίζει τη λέξη με τη μεγαλύτερη συχνότητα, από αυτές που έχει διαθέσιμες. Τέλος τα δεδομένα μαζεύονται στον κόμβο οδηγό, ο οποίος και υπολογίζει τη τελική λέξη με τη μεγαλύτερη συχνότητα εμφάνισης.

Στις επόμενες δύο εικόνες παρουσιάζεται μία εικονική αναπαράσταση των παραπάνω ενεργειών:



Εικόνα 10: Υπολογισμός λέξης με μεγαλύτερη συχνότητα (1)



Εικόνα 11: Υπολογισμός λέξης με μεγαλύτερη συχνότητα (2)

4.6. Βασικότερα Spark Transformation

Στην υποενότητα αυτή θα γίνει περιγραφή των βασικότερων Spark μετασχηματισμών και ενεργειών:

- **map**: Πρόκειται για ένα στενό μετασχηματισμό, ο οποίος μετατρέπει κάθε εγγραφή του γονέα RDD σε μια νέα εγγραφή για το παιδικό RDD, με βάση μια συνάρτηση που ορίζει ο χρήστης. Αυτός ο μετασχηματισμός εκτελείται τοπικά σε κάθε partition και το παιδί RDD έχει τα ίδια partitions και τις προτιμώμενες τοποθεσίες, όπως και ο πατέρας RDD.

- **flatmap**: Είναι ένας στενός μετασχηματισμός, ο οποίος λαμβάνει ως είσοδο κάθε εγγραφή του γονέα RDD και παράγει πολλαπλές εγγραφές εξόδου, βάση της συνάρτησης που έχει ορίσει ο χρήστης. Αυτός ο

μετασχηματισμός εκτελείται τοπικά και το παιδικό RDD έχει τα ίδια partitions και προτιμώμενες τοποθεσίες με το γονέα RDD.

- **mapPartition:** Πρόκειται για ένα στενό μετασχηματισμό, ο οποίος εκτελείται σε κάθε partition ξεχωριστά. Σε κάθε διαμέρισμα, η είσοδος είναι ένας iterator πάνω από όλες τις εγγραφές του αντίστοιχου γονικού διαμερίσματος και η έξοδος είναι ένας iterator με δεδομένα που ορίζει ο χρήστης. Αυτός ο μετασχηματισμός μπορεί να είναι πολύ χρήσιμος όταν οι χρήστες θέλουν να συναθροίσουν δεδομένα ανά διαμέρισμα ξεχωριστά και όχι σε καθολικό επίπεδο. Αυτός ο μετασχηματισμός εκτελείται τοπικά σε κάθε διαμέρισμα και το παιδικό RDD έχει τα ίδια διαμερίσματα και προτιμώμενες τοποθεσίες όπως το γονικό RDD. Επειδή αυτός ο μετασχηματισμός τροποποιεί κάθε εγγραφή, ο χρήστης είναι σε θέση να δηλώσει εάν το παιδί RDD διατηρεί το σχήμα διαμέρισης του γονικού RDD του ή όχι.

- **filter:** Πρόκειται επίσης για ένα στενό μετασχηματισμό, ο οποίος φιλτράρει τις εγγραφές του γονικού RDD, βάση μίας συνάρτησης ορισμένης από το χρήστη. Αυτός ο μετασχηματισμός εκτελείται τοπικά σε κάθε partition και το RDD παιδί έχει τα ίδια partitions, προτιμώμενες τοποθεσίες και το σχήμα διαίρεσης του γονέα RDD.

- **reducebyKey:** Είναι ένας μετασχηματισμός ευρείας κλίμακας, ο οποίος μπορεί να εφαρμοστεί μόνο σε pair RDDs (key-value). Ο μετασχηματισμός αυτός συγκεντρώνει τις τιμές των εγγραφών με το ίδιο κλειδί, βάση μιας συνάρτησης που έχει ορίσει ο χρήστης. Αυτός ο μετασχηματισμός εκτελείται αρχικά σε τοπικό επίπεδο, συγκεντρώνοντας τις εγγραφές με το ίδιο κλειδί τοπικά και στη συνέχεια συγκεντρώνει τα δεδομένα σε καθολικό επίπεδο. Οι χρήστες είναι σε θέση να ορίσουν τη συνάρτηση κατανομής, βάσει της οποίας τα δεδομένα θα ανακατευθύνονται μεταξύ των εργαζομένων, καθώς και τον αριθμό των κατατμήσεων εξόδου. Η προεπιλεγμένη έξοδος βασίζεται σε μια συνάρτηση κατακερματισμού, όπου ο αριθμός των διαμερισμάτων της εξόδου είναι ο ίδιος με τον αριθμό των διαμερισμάτων της εισόδου

- **groupyByKey:** Πρόκειται για έναν ευρύ μετασχηματισμό, ο οποίος μπορεί να εφαρμοστεί μόνο σε ένα pair RDDs και ομαδοποιεί τις τιμές των εγγραφών που έχουν το ίδιο κλειδί. Όπως και στο reduceByKey, στον μετασχηματισμό groupByKey, οι χρήστες είναι σε θέση να ορίσουν τη συνάρτηση κατανομής, με βάση την οποία τα δεδομένα θα ανακατευθύνονται μεταξύ των εργαζομένων, όπως και τον αριθμό των κατατμήσεων εξόδου

- **reduce:** Είναι μία ενέργεια, η οποία συγκεντρώνει τα δεδομένα ενός RDD, με βάση μίας προκαθορισμένης από το χρήστη συνάρτηση. Η ενέργεια αυτή λειτουργεί με κατανεμημένο τρόπο και όχι κεντρικοποιημένο.

- **takeSample**: Πρόκειται για μια ενέργεια, η οποία επιστρέφει ένα δείγμα του RDD στον κύριο κόμβο με βάση ένα προκαθορισμένο seed. Κάθε κόμβος εργάτης διατάσσεται να επιστρέψει ένα σύνολο εγγραφών στον κόμβο οδηγό. Μετά από αυτό, ο κόμβος οδηγός θα ενώσει τα αρχεία που επιστράφηκαν από όλους τους κόμβους και θα επιστρέψει το τελικό αποτέλεσμα στο χρήστη. Σε περίπτωση που ο χρήστης επιθυμεί το δείγμα να αποτελείται από μοναδικά στοιχεία, τότε κάθε εργαζόμενος επιστρέφει ένα δείγμα βασισμένο στη συνάρτηση Bernouli, διαφορετικά χρησιμοποιεί τη συνάρτηση Poisson. Σε περίπτωση που ο χρήστης επιθυμεί την επιστροφή μόνο μοναδικών εγγραφών, τότε ο κόμβος οδηγός είναι υπεύθυνος να μην επιστρέψει περισσότερες από μία ίδιες εγγραφές. Εάν τα στοιχεία του τελικού συνόλου εγγραφών είναι μικρότερα από το ζητούμενο αριθμό συνόλου, τότε η ίδια διαδικασία εκτελείται ακόμη μία φορά με τη χρήση διαφορετικού seed και στο τέλος ο κόμβος οδηγός μπορεί να ενώσει τα αποτελέσματα της πρώτης και της δεύτερης εκτέλεσης, προκειμένου να επιστραφεί η επιθυμητή δειγματοληψία στο χρήστη.

- **count:** Είναι μια ενέργεια, η οποία επιστρέφει τον αριθμό των εγγραφών του RDD στον κόμβο οδηγό. Αυτή η ενέργεια δεν χρησιμοποιεί πληροφορίες μεταδεδομένων για να υπολογίσει τον αριθμό των εγγραφών, αλλά πρέπει να μετρήσει τις εγγραφές, κάθε φορά που θα καλείται. Πιο συγκεκριμένα, ο κόμβος οδηγός διατάζει τους κόμβους των εργαζομένων να επιστρέψουν τον αριθμό των εγγραφών που είναι αποθηκευμένα σε αυτά. Ο κόμβος οδηγός συγκεντρώνει το μέγεθος από κάθε κόμβο και τα συνοψίζει για να υπολογίσει το συνολικό μέγεθος του RDD

- **collect:** Είναι μια ενέργεια, η οποία επιστρέφει τις εγγραφές του RDD στον στο χρήστη, ως μια σειρά αντικειμένων. Κάθε κόμβος επιστρέφει τις διαθέσιμες εγγραφές στον κόμβο οδηγό. Στο τέλος, ο κόμβος οδηγός συγκεντρώνει τις εγγραφές και το τελικό αποτέλεσμα επιστρέφεται στον πελάτη ως μια σειρά αντικειμένων

- **collectAsMap:** Πρόκειται για μια ενέργεια, η οποία μπορεί να καλείται μόνο σε pair RDDs και επιστρέφει τις εγγραφές του RDD ως ένα map στον χρήστη. Κάθε διαμέρισμα επιστρέφει τις διαθέσιμες εγγραφές στον κόμβο οδηγό ως μια σειρά αντικειμένων. Ο κόμβος οδηγός αναλαμβάνει να συναθροίσει τη συστοιχία αντικειμένων από κάθε διαμέρισμα και να δημιουργήσει το τελικό map. Σε περίπτωση που περισσότερες από μία εγγραφές έχουν το ίδιο κλειδί, τότε το

επιστρεφόμενο map θα περιέχει ως value, για το συγκεκριμένο κλειδί, μόνο μία από τις αντίστοιχες εγγραφές.

5. Μεγάλα Δεδομένα και Ιδιωτικότητα

Όταν το διαδίκτυο σχεδιάστηκε, πολλοί άνθρωποι πίστευαν ότι πρόκειται για το αποκορύφωμα όσον αφορά τις ψηφιακές επικοινωνίες. Μέσω αυτού, δόθηκε η δυνατότητα στους χρήστες να μοιράζονται πληροφορίες, ακόμη και αν βρίσκονται σε δύο αντιδιαμετρικές πλευρές στον πλανήτη. Καθώς ο ηλεκτρονικός χώρος αποθήκευσης μη δομημένων δεδομένων αυξήθηκε ραγδαία σε ευρεία κλίμακα, οι “πρωτοπόροι” της τεχνολογίας άρχισαν να ενώνουν τα κομμάτια και μετέφεραν τη ψηφιακή ανταλλαγή πληροφοριών σε ένα εντελώς νέο επίπεδο.

Τη σήμερον ημέρα, τα μεγάλα δεδομένα έχουν γίνει μια από τις πιο ελπιδοφόρες ιδέες στον χώρο της τεχνολογίας. Μπορούμε να τα βρούμε παντού - από υπηρεσίες ροής μουσικής έως νοσοκομεία που αποθηκεύουν ψηφιακά ιατρικά αρχεία. Οι μεγάλες αναλύσεις δεδομένων επιτρέπουν επίσης στις επιχειρήσεις να βελτιώσουν τις στρατηγικές τους και ταυτόχρονα άρχισαν να εκτελούν εκστρατείες μάρκετινγκ που βασίζονται σε δεδομένα.

Αλλά ποια είναι τα αποτελέσματα για έναν καθημερινό χρήστη που ποτέ δεν έχει ακούσει για τα μεγάλα δεδομένα; Τι γίνεται αν ένας απλός υπάλληλος δεν είναι σε θέση να ανησυχεί για τα μεγάλα στατιστικά στοιχεία; Για παράδειγμα, είναι πιθανό να εξετάσει το ενδεχόμενο να διαθέσει ορισμένες από τις προσωπικές του πληροφορίες σε διαδικτυακές εφαρμογές με αντάλλαγμα εξατομικευμένες υπηρεσίες και εμπειρίες. Αυτό, ωστόσο, ταυτόχρονα σημαίνει πως δημιουργούνται κενά όσον αφορά την ασφάλεια και την ιδιωτικότητα. Εξάλλου, ο Παγκόσμιος Ιστός δεν είναι ξένος προς τις δυσάρεστες παραβιάσεις δεδομένων που θέτουν σε κίνδυνο τις πληροφορίες χρηστών.

Το παραπάνω φαινόμενο δημιουργεί τις ακόλουθες προκλήσεις όσων αφορά την ασφάλεια και την ιδιωτικότητα [challenges-in-privacy]:

- Προστασία των έντυπων συναλλαγών και των δεδομένων

  Τα δεδομένα που είναι αποθηκευμένα σε ένα μέσο αποθήκευσης, όπως είναι τα αρχεία καταγραφής συναλλαγών και άλλες ευαίσθητες πληροφορίες, μπορεί να έχουν διαφορετικά επίπεδα ασφάλειας, αλλά αυτό από μόνο του δεν αρκεί. Για παράδειγμα, η μεταφορά των δεδομένων μεταξύ αυτών των επιπέδων επιτρέπει στον διαχειριστή του infrastructure (IT) να διαβάσει πληροφορίες σχετικά με τα δεδομένα που μετακινούνται. Όσο το μέγεθος των δεδομένων αυξάνεται ραγδαία συνεχώς, η κλιμάκωση και η διαθεσιμότητα καθιστούν την αυτόματη αντιστοίχιση απαραίτητη για τη διαχείριση των μεγάλων αποθηκευτικών δεδομένων. Ωστόσο, δημιουργούνται νέες προκλήσεις για την αποθήκευση των μεγάλων δεδομένων, καθώς η μέθοδος της αυτόματης αντιστάθμισης δεν παρακολουθεί τη θέση αποθήκευσης των δεδομένων.

- Επικύρωση και φίλτρα end-to-end

  Οι τελικές συσκευές είναι οι κύριοι παράγοντες για τη διατήρηση μεγάλων δεδομένων. Η αποθήκευση, η επεξεργασία και άλλες απαραίτητες εργασίες εκτελούνται με τη βοήθεια δεδομένων εισόδου, τα οποία παρέχονται από τα τελικά σημεία. Ως εκ τούτου, ένας οργανισμός θα πρέπει να επιβεβαιώνει ότι γίνεται χρήση αυθεντικών και νόμιμων τελικών συσκευών.

- Κρυπτογράφηση δεδομένων

  Η κρυπτογράφηση δεδομένων αποτελεί σημαντικό κομμάτι για τη διατήρηση της ιδιωτικότητας, ωστόσο η κρυπτογράφηση πρέπει να γίνεται τόσο στο κομμάτι της μεταφοράς, όσο και στις συσκευές που επεξεργάζονται και συντηρούν τα δεδομένα.

- Ασφάλεια στα frameworks μεγάλων δεδομένων

  Η ασφάλεια που προσφέρουν τα κατανεμημένα frameworks, όπως είναι το MapReduce του Hadoop, έχουν ως επί το πλείστον έλλειψη προστασίας. Οι δύο βασικοί τρόποι για την προφύλαξη απέναντι σε αυτό είναι η παροχή προστασίας για τους mappers και η προστασία των δεδομένων απέναντι στην παρουσία ενός μη εξουσιοδοτημένου mapper.

- Προστασία δεδομένων σε πραγματικό χρόνο

Λόγω της δημιουργίας μεγάλων ποσοτήτων δεδομένων, οι περισσότερες οργανώσεις δεν είναι σε θέση να τηρούν τακτικούς ελέγχους. Ωστόσο, είναι πολύ ωφέλιμο να πραγματοποιούνται έλεγχοι ασφαλείας και παρακολούθησης σε πραγματικό χρόνο ή σχεδόν πραγματικό χρόνο.

5.1. Σκάνδαλο με Facebook & Cambridge Analytica

Φέτος, το 2018, ένα τεράστιο σκάνδαλο που αφορά στην παραβίαση της ιδιωτικότητας και των προσωπικών δεδομένων ξεσκεπάστηκε και έφτασε να εξετάζεται στο Κογκρέσο των ΗΠΑ. Πιο συγκεκριμένα, τα προσωπικά δεδομένα 87 εκατομμυρίων χρηστών του Facebook διέρρευσαν καταλήγοντας στα χέρια της εταιρείας Cambridge Analytica, η οποία είχε αναλάβει την προεκλογική εκστρατεία του Donald Trump το 2016. Η εταιρεία κατηγορήθηκε για απόπειρα πολιτικής χειραγώγησης και επηρεασμού του εκλογικού αποτελέσματος, ενώ οι αμερικανοί κατηγορούν και τη Ρωσία ως ηθικό αυτουργό της όλης επιχείρησης.

Ο Mark Juckerberg, ο ιδρυτής του ιστότοπου κοινωνικής δικτύωσης Facebook, που κλήθηκε να καταθέσει, παραδέχτηκε πως ακόμη και ο ίδιος είχε πέσει θύμα παραβίασης των προσωπικών του δεδομένων.

Τα δεδομένα των χρηστών του facebook εξατομικεύουν τις ιστοσελίδες που τους προτείνονται, τις διαφημίσεις που τους προβάλλονται και όλα όσα έχουν να κάνουν με τον κόσμο του internet. Σε λάθος χέρια, λοιπόν, τα δεδομένα αυτά γίνονται όπλο χειραγώγησης και προπαγάνδας παντός είδους. Η Cambridge Analytica συνέλεγε τα στοιχεία εκατομμυρίων χρηστών του Facebook και χρησιμοποιούσε τις πληροφορίες αυτές για να φτιάξει τα «πολιτικά και ψυχολογικά» προφίλ τους ώστε να επηρεάσει την ψήφο τους με τα κατάλληλα πολιτικά μηνύματα.

Ποτέ δεν θα μάθουμε με ακρίβεια αν όντως η διαρροή αυτών των δεδομένων επηρέασε πράγματι τις εκλογές στις ΗΠΑ υπέρ του D. Trump ή ακόμα και το δημοψήφισμα στη Μ. Βρετανία υπέρ του Brexit, όπως φημολογείται. Όποια και αν είναι η αλήθεια, είναι γεγονός πως το συγκεκριμένο σκάνδαλο αποτέλεσε αφορμή για ευαισθητοποίηση της κοινής γνώμης σε σχέση με την ιδιωτικότητα (privacy) και τα προσωπικά δεδομένα, αλλά και των κρατών, των κοινοβουλίων και των νομοθετών. Άλλωστε, ο νέος νόμος που τέθηκε σε εφαρμογή την 25^{η} Μαΐου 2018 για την προστασία των προσωπικών δεδομένων – ευρωπαϊκός Γενικός Κανονισμός για την Προστασία Δεδομένων - αποτελεί τη μεγαλύτερη μεταρρύθμιση εδώ και 20 χρόνια. Οι εταιρείες που δραστηριοποιούνται στην Ευρωπαϊκή Ένωση θα έχουν να κάνουν με νέους κανόνες ως προς το πώς διαχειρίζονται τα δεδομένα των πολιτών, ενώ επιβάλλονται νέες, αυστηρότερες ποινές για την παραβίαση του νόμου.

Επίλογος

Η διαχείριση και η επεξεργασία των μεγάλων δεδομένων είναι ένα από τα περισσότερο σημαντικά ανοιχτά ζητήματα στις μέρες μας. Τα δεδομένα που παράγονται από την ολοένα και αυξανόμενη χρήση των διαδικτυακών εφαρμογών και υπηρεσιών τόσο στην καθημερινή ζωή όσο και στην επαγγελματική, συνδράμουν στην παραγωγή ενός τεράστιου όγκου δεδομένων. Αυτό έχει ως αποτέλεσμα αρκετές από τις τεχνικές που χρησιμοποιούνταν στο παρελθόν να μην μπορούν να ικανοποιήσουν τις σημερινές ανάγκες.

Στη βιβλιογραφία μπορούν να βρεθούν αρκετοί ορισμοί σχετικά με τον όρο Big Data, ωστόσο όλοι οι ορισμοί συμφωνούν στα «3 V»:

- **Volume:** Μεγάλος και πολύπλοκος όγκος δεδομένων
- **Velocity:** Ανάγκη για ταχύτητα στην επεξεργασία και στη διαχείριση των δεδομένων
- **Variety:** Μεγάλη ποικιλία ως προς την μορφή των δεδομένων(δομημένα, ημι-δομημένα και αδόμητα δεδομένα – εικόνες, video, ήχος κ.α.)

Στα πλαίσια της πτυχιακής εργασίας γίνεται ανάλυση της έννοιας των μεγάλων δεδομένων και της σημασίας τους για τη κοινωνία. Επιπρόσθετα, γίνεται αναφορά σε πραγματικά παραδείγματα αξιοποίησης των μεγάλων δεδομένων από επιχειρήσεις και οργανισμούς.

Μια δεύτερη πρόκληση, επιπρόσθετα του τρόπου αξιοποίησης των μεγάλων δεδομένων, είναι η δυνατότητα επεξεργασίας αυτών. Ο τεράστιος και πολύπλοκος όγκος των Big Data οδήγησε στην ανάγκη δημιουργίας εξατομικευμένων frameworks για την επεξεργασία αυτών. Σημαντικό κομμάτι της εργασίας αποτελεί η περιγραφή του δημοφιλέστερου Big Data framework σήμερα, το Spark και η περιγραφή χρήσης του για την επίλυση του προβλήματος της εύρεση της λέξης με τη μεγαλύτερη συχνότητα μέσα σε ένα μεγάλο κείμενο.

Ωστόσο, πέρα από τα οφέλη που μας προσφέρουν τα μεγάλα δεδομένα έρχονται και ορισμένοι κίνδυνοι, με σημαντικότερο το θέμα της παραβίασης της ιδιωτικότητας. Στη σημερινή εποχή, οι χρήστες προσφέρουν όλο και περισσότερα προσωπικά δεδομένα στις διαδικτυακές εφαρμογές και αν αυτές οι πληροφορίες πέσουν σε ξένα χέρια, τότε αυτό αυτόματα σημαίνει πως υπάρχει παραβίαση της ιδιωτικότητας. Το αποτέλεσμα μίας τέτοιας ενέργειας μπορεί να έχει πολύ σημαντικές συνέπειες στη ζωή των ανθρώπων. Χαρακτηριστικό παράδειγμα αποτελεί η φημολογία ότι η διαρροή των πληροφοριών του facebook, μέσω της Cambridge Analytica**,** μπορεί να επηρέασε τις εκλογές στις ΗΠΑ.

Πίνακας Ορολογίας

Ξενόγλωσσος όρος	Ελληνικός Όρος
Actions	Ενέργειες
Analytics	Ανάλυση
Big Data	Μεγάλα δεδομένα
Business Intelligence	Επιχειρηματική Ευφυΐα
Cache	Συνιστώσα προσωρινής αποθήκευσης δεδομένων
Checkpoint	Σημείο ελέγχου
Cluster	Σύμπλεγμα
Cluster manager	Διαχειριστή συμπλέγματος
Components	Συνιστώσες περιβάλλοντος
Data	Δεδομένα
Data Replication	Δημιουργία αντιγράφων
Delimiter	Διαχωρίζει δεδομένα
Distributed storage system	Κατανεμημένο σύστημα αποθήκευσης
Driver	Οδηγός
Engine	Μηχανή
Execution Engine	Μηχανή εκτέλεσης
Fault Tolerance	Παροχή ανοχής σε σφάλματα
Filter	Φίλτρο
Framework	Πλαίσια για την υποστήριξη ανάπτυξης εφαρμογών
Internet	Διαδίκτυο
Iterator	Επαναληπτικός δρομέας πάνω από τα δεδομένα
Join	Ένωση εγγραφών με βάση ένα κλειδί
Key-Value	Ζεύγος κλειδιού-τιμής
Lazy	Υπολογισμός παράσταση στο σημείο που θα γίνει χρήση
long-lived	Μακρόχρονες
Map	Λεξικό
Market Intelligence	Αγορά Νοημοσύνης
Metadata	Μεταδεδομένα
Online Learning	Ηλεκτρονική Μάθηση
Open source	Ανοιχτός Κώδικας
Pair	Ζεύγος
Partition	Διαμέρισμα

Pipeline	Διοχέτευση
Program	Πρόγραμμα
Project	Έργο
Repositories	Αποθετήρια
Seed	Σπόρος για την παροχή ντετερμινιστικής παραγωγής τυχαίων αριθμών
Shuffling	Ανακατανομή δεδομένων σε μηχανήματα
Smart health	«Έξυπνη» Υγεία
Standalone	Αυτόνομο
Streaming	Ροή δεδομένων
Tables	Πίνακας
Tasks	Καθήκοντα
Transformations	Μετασχηματισμοί
Union	Ένωση
Value	Τιμή
Variety	Ποικιλία
Velocity	Ταχύτητα
Veracity	Αλήθεία
Volume	Όγκος
Workers	Εργάτες

Συντμήσεις – Αρκτικόλεξα – Ακρώνυμα

AHP	Analytic Hierarchy Process
BI	Business intelligence
DAG	Directly Acyclic Graph
DLs	Digital libraries
EMS	Estates Management Statistics
ETL	Extract, Transformation, Load
GDP	Gross Domestic Product
GPS	Global Positioning System
HEIDI	Higher Education Information Database for Institutions
IoT	Internet of Things
KPIs	Key Performance Indicators
NSS	National Student Survey
RDD	Resilient Distributed Datasets
REF	Research Excellence Framework
SQL	Structured Query Language
UKBA	UK Border Agency
VLDLs	Very large digital libraries

Βιβλιογραφία

1. "Big data analytics – Advanced analytics in Oracle database", an Oracle White Paper, March 2013
2. Anderson, C. (2007), "The end of theory: the data deluge makes the scientific method obsolete", Wired, Vol. 16 No. 7, available at: www.wired.com/2008/06/pb-theory/ (accessed 27 January 2016)
3. apache, https://httpd.apache.org/
4. apache-spark, https://spark.apache.org/
5. Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. Hive: A warehousing solution over a map-reduce framework. Proc. VLDB Endow., 2(2): 1626–1629, August 2009. ISSN 2150-8097. doi: 10.14778/1687553.1687609. URL http://dx.doi.org/10.14778/1687553.1687609
6. Askitas, N. and Zimmermann, K.F. (2009), "Google econometrics and unemployment forecasting", Applied Economics Quarterly, Vol. 55 No. 2, pp. 107-120, available at: http://doi.org/10.3790/aeq.55.2.107
7. Atzori, L., Iera, A. and Morabito, G. (2010), "The internet of things: a survey", Computer Networks, Vol. 54 No. 15, pp. 2787-2805
8. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123-1131
9. Beyer, M.A. and Laney, D. (2012), The Importance of "Big Data": A Definition, Gartner, Stamford, CT
10. Boyd, D. and Crawford, K. (2012), "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon", Information, Communication & Society, Vol. 15 No. 5, pp. 662-679, available at: http://doi.org/10.1080/1369118X.2012.678878

11. Buhl, H.U., Röglinger, M., Moser, F. and Heidemann, J. (2013), “Big data”, Business & Information Systems Engineering, Vol. 5 No. 2, pp. 65-69, available at: http://doi.org/10.1007/s12599-013-0249-5

12. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C. and Schuldt, H. (2007), “The DELOS digital library reference model”, Foundations for Digital Libraries

13. Candela, L., Manghi, P. and Ioannidis, Y. (2012), “Fourth workshop on very large digital libraries: on the marriage between very large digital libraries and very large data archives”, ACM SIGMOD Record, Vol. 40 No. 4, pp. 61-64, available at: http://doi.org/10.1145/2094114. 2094130

14. challenges-in-privacy, https://dataconomy.com/2017/07/10-challenges-big-data-security-privacy

15. Chen, H., Chiang, R. and Storey, V. (2012), “Business intelligence and analytics: from big data to big impact”, MIS Quarterly, Vol. 36 No. 4, pp. 1165-1188

16. Cricelli, L. and Grimaldi, M. (2008), “A dynamic view of knowledge and information: a stock and flow based methodology”, International Journal of Management and Decision Making, Vol. 9 No. 6, pp. 686-698, available at: http://doi.org/10.1504/IJMDM.2008.021221

17. databricks, https://databricks.com/

18. datafication, https://en.wikipedia.org/wiki/Datafication

19. Davenport, T.H. and Patil, D.J. (2012), “Data scientist: the sexiest job of the 21st century”, Harvard Business Review, Vol. 90 No. 10, pp. 70-76

20. Dean, J. and Ghemawat, S. (2008), “MapReduce: simplified data processing on large clusters”, Communications of the ACM, Vol. 51 No. 1, pp. 1-13

21. Dean, J. and Ghemawat, S. (2008), “MapReduce: simplified data processing on large clusters”, Communications of the ACM, Vol. 51 No. 1, pp. 1-13

22. Downing NS, Cloninger A, Venkatesh AK, Hsieh A, Drye EE, Coifman RR, et al. (2017) Describing the performance of U.S.

hospitals by applying big data analytics.PLOS ONE 12(6):e0179603. https://doi.org/10.1371/journal.pone.0179603

23. Dumbill, E. (2013), “Making sense of big data”, Big Data, Vol. 1 No. 1, pp. 1-2
24. Estrin, D., Culler, D., Pister, K. and Sukhatme, G. (2002), “Connecting the physical world with pervasive networks”, IEEE Pervasive Computing, Vol. 1 No. 1, pp. 59-69, available at: http://doi.org/10.1109/MPRV.2002.993145
25. Evans, D. (2011), The Internet of Things – How the Next Evolution of the Internet is Changing Everything, Cisco Systems, San Jose, CA
26. flink, https://flink.apache.org/
27. Gartner (2014), “Gartner says the internet of things will transform the data center”, available at: www.gartner.com/newsroom/id/2684616
28. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), “Detecting influenza epidemics using search engine query data”, Nature, Vol. 457 No. 7232, pp. 1012-1014, available at: http://doi.org/10.1038/nature07634
29. google-books-library, https://books.google.com/googlebooks/library/
30. Guzman, G. (2011), “Internet search behavior as an economic forecasting tool: the case of inflation expectations”, Journal of Economic and Social Measurement, Vol. 36 No. 3, pp. 119-167
31. H. Chen, R. Chiang, and V. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” MIS Quarterly, vol. 36 (4), pp. 1165-1188, 2012
32. hadoop, http://hadoop.apache.org/
33. hbase, https://hbase.apache.org/
34. hdfs, https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
35. Hilbert, M. and López, P. (2011), “The world’s technological capacity to store, communicate, and compute information”, Science, Vol. 332 No. 6025, pp. 60-65, available at: http://doi.org/10.1126/science.1200970
36. https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131 - (Piatetsky-Shapiro & Frawley, 1991). Wray Buntine. Stratifying samples to improve learning. In G. PiatetskyShapiro

and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 305{314. AAAI/MIT Press, Cambridge, MA, 1991

37. J. Bichsel, “Analytics in Higher Education - Benefits, Barriers, Progress, and Recommendations,” Research Report, EDUCAUSE Center for Applied Research, August 2012, available from http://www.educause.edu/ecar

38. Jansen, W., Barbera, R., Drescher, M., Fresa, A., Hemmje, M., Ioannidis, Y. and Stanchev, P. (2013), “e-infrastructures for digital libraries...the future”, Lecture Notes in Computer Science, Vol. 8092, pp. 480-481, available at: http://doi.org/10.1007/978-3-642-40501-3

39. Manovich, L. (2012), “Trending: the promises and the challenges of big social data”, in Gold, M.K. (Ed.), Debates in the Digital Humanities, University of Minnesota Press, Minneapolis, MN, pp. 460-475

40. Manyika J., Chui M.,Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A. (2011) Big data: the next frontier for innovation, competition and productivity, McKiney global report, May 2011

41. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, New York, NY

42. map-reduce, https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

43. Matei Zaharia, Mosharaf Chowdhury, Tathagata Dasm Ankur Dave, Justen Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, NSDI’12, pages 2-2, Berkley, CA, USA, 2012. USENIX Association, https://dl.acm.org/citation.cfm?id=2228298.2228301

44. Mayer-Schönberger, V. and Cukier, K. (2013), Big Data: A Revolution That Will Transform How We Live, Work and Think, John Murray, London

45. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D. and Barton, D. (2012), "Big data: the management revolution", Harvard Business Review, Vol. 90 No. 10, pp. 61-67

46. mesos, http://mesos.apache.org/

47. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P. and Aiden, E.L. (2011), "Quantitative analysis of culture using millions of digitized books", Science, Vol. 331 No. 6014, pp. 176-182, available at: http://doi.org/10.1126/science.1199644

48. Michelino, F., Bianco, F. and Caputo, M. (2008), "Internet and supply chain management: adoption modalities for Italian firms", Management Research News, Vol. 31 No. 5, pp. 359-374

49. Moore, G.E. (2006), "Cramming more components onto integrated circuits, reprinted from Electronics", Vol. 38 No. 8, April 19, 1965, pp. 114, ff, IEEE Solid-State Circuits Newsletter, Vol. 11 No. 5, pp. 33-35, available at: http://doi.org/10.1109/N-SSC.2006.4785860

50. Ong, V. K. (2016). Business intelligence and big data analytics for higher education: Cases from UK Higher Education Institutions. Information Engineering Express, 2(1), 65-75

51. P. B. Goes, "Big Data and IS Research," MIQ Quarterly, vol. 38, no. 3, September 2014, pp. iii-viii

52. Pearson, T. and Wegener, R. (2013), Big Data: The Organizational Challenge, Bain & Company, available at: www.bain.com/images/bain_brief_big_data_the_organizational_challenge.pdf

53. Powell, J. (2012), "'At scale' author name matching with Hadoop/MapReduce", Library Hi Tech News, Vol. 29 No. 4, pp. 6-12, available at: http://doi.org/10.1108/07419051211249455

54. Prescott, A. (2013), "Bibliographic records as humanities big data", Proceedings – 2013 IEEE International Conference on Big Data, IEEE, pp. 55-58, available at: http://doi.org/10.1109/ BigData.,2013.6691670

55. Ronda-Pupo, G.A. and Guerras-Martin, L.A. (2012), "Dynamics of the evolution of the strategy concept 1962-2008: A co-word analysis",

Strategic Management Journal, Vol. 33 No. 2, pp. 162-188, available at: http://doi.org/10.1002/smj.948

56. Rowley, J. (2007), "The wisdom hierarchy: representations of the DIKW hierarchy", Journal of Information Science, Vol. 33 No. 2, pp. 163-180

57. Russom, P. (2011), "Big data analytics", TDWI Best Practices Report, Fourth Quarter, pp. 1-35, available at: www.tableau.com/sites/default/files/whitepapers/tdwi_bpreport_q411_big_data_analytics_tableau.pdf

58. Shvachko, K., Kuang, H., Radia, S. and Chansler, R. (2010), "The Hadoop distributed file system",IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010, IEEE, pp. 1-10

59. storm, http://storm.apache.org/

60. U. Kelly, I. McNicoll and J. White, "The Impact of Universities on the UK Economy", Uni-versities UK (ISBN: 978-1-84036-304-3)

61. Wang, J., Zhang, Y., Gao, Y. and Xing, C. (2013), " A new plug-in system supporting very large digital library", in Urs, S.R., Na, J.C. and Buchanan, G. (Eds), 15th International Conference on Asia-Pacific Digital Libraries, ICADL 2013, Bangalore, 9-11 December, (Lecture No, Vol. 8279, pp. 45-52), Springer International Publishing, Bangalore, available at: http://doi.org/10.1007/978-3-319-03599-4

62. Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y. and Xu, C. (2013), "A characterization of big data benchmarks", Proceedings – 2013 IEEE International Conference on Big Data, Big Data 2013, pp. 118-125, available at: http://doi.org/10.1109/BigData.,2013.6691707

63. yarn, https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html