Κεφάλαιο 3

Επιστήμη των Δεδομένων και Αναλυτική Μεγάλων Δεδομένων (Data Science & Big Data Analytics)

3.1. Εισαγωγή, Βασικές Έννοιες & Ορισμοί – Ιστορική Εξέλιξη

Στη συνέχεια θα ξεκινήσουμε παρουσιάζοντας τις βασικές έννοιες.

Επιστήμη των Δεδομένων

Η επιστήμη των δεδομένων είναι μια διεπιστημονική περιοχή, στην οποία συνυπάρχουν και συνεργάζονται μέθοδοι και τεχνικές από τη στατιστική, την ανάλυση δεδομένων, την επιχειρησιακή έρευνα, την πληροφορική, τη τεχνητή νοημοσύνη, κα., με στόχο την αξιοποίηση του μεγάλου όγκου ανεπεξέργαστων δεδομένων με την αποκάλυψη κρυμμένων προτύπων/μοτίβων, σχέσεων, πληροφοριών και γνώσης, προκειμένου να επιτρέψει στους αποφασίζοντες τη λήψη καλύτερων και ασφαλέστερων αποφάσεων (Minelli et al., 2013; Provost and Fawcett, 2013; Sharda et al. 2014; Davenport , 2014; Inmon and Linstedt, 2015; Dietrich et al., 2015; Buyya et al., 2016; Said and Torra, 2019; Said and Torra, 2019; Favero and Belfiore, 2019).

Συνεπώς, η επιστήμη των δεδομένων έχοντας ισχυρή σύνδεση με πολλά επιστημονικά πεδία, μπορεί να θεωρηθεί σαν ένας τρόπος για την ολοκλήρωσή τους. Τα πεδία αυτά εν συντομία είναι:

- ***Στατιστική:*** Η στατιστική έχει παρόμοιο στόχο με την ανάλυση δεδομένων και την εξαγωγή συμπερασμάτων από δεδομένα (πχ. περιγραφικές και συμπερασματικές στατιστικές).

Εικόνα 3.1. *Alan Turing*

- ***Μηχανική Μάθηση και Εξόρυξη Δεδομένων:*** Οι ορισμοί της Τεχνητής Νοημοσύνης (ΤΝ) έχουν τις ρίζες τους από τη δεκαετία του 1940 και στο ερώτημα του Turing 'αν μπορούν οι μηχανές να σκέφτονται;' (τεστ Turing). Ο Turing (Εικόνα 3.1) θεωρείται από πολλούς ως ο πατέρας της Τεχνητής Νοημοσύνης. Η μηχανική μάθηση και η εξόρυξη δεδομένων είναι το πεδίο της Τεχνητής Νοημοσύνης που επικεντρώνεται σε αυτό το στόχο. Ορισμένες μέθοδοι μηχανικής εκμάθησης μπορούν να χρησιμοποιηθούν για τη δημιουργία μοντέλων από δεδομένα (δηλαδή για τη δημιουργία επεξηγήσεων) και για την πραγματοποίηση προβλέψεων (προβλήματα ταξινόμησης και παλινδρόμησης). Μερικά από τα εργαλεία που αναπτύχθηκαν στο πλαίσιο της μηχανικής μάθησης συνδέονται στενά ή αλληλεπικαλύπτονται με στατιστικές μεθόδους.
- ***Μεγάλα Δεδομένα και Τεχνολογίες Βάσεων Δεδομένων:*** Ενώ το μέγεθος μιας βάσης δεδομένων δεν είναι απαραίτητα κρίσιμη πτυχή για την κατασκευή ενός μοντέλου και τη λήψη αποφάσεων, η συνάφεια της επιστήμης δεδομένων στην επιχείρηση έγκειται στο γεγονός ότι τα δεδομένα είναι διαδεδομένα και αποθηκεύονται σε terabytes δεδομένων με στόχο την ανάλυσή τους. Εν συνεχεία, εφαρμόζονται διάφοροι μέθοδοι από διάφορα επιστημονικά πεδία, για την ανάλυσή τους και την εξαγωγή πληροφοριών και γνώσης από αυτά. Αυτό σημαίνει ότι οι αλγόριθμοι πρέπει να προγραμματιστούν αποτελεσματικά και να οδηγούν σε λύσεις μέσα σε εύλογο χρονικό διάστημα. Για να επιτευχθεί αυτό, θα πρέπει να αξιοποιούνται οι κατάλληλες τεχνολογίες για μεγάλα δεδομένα.

Στη συνέχεια θα παρουσιάσουμε μερικά από τα επιστημονικά πεδία, μέθοδοι και τεχνικές των οποίων χρησιμοποιούνται στην επιστήμη των δεδομένων.

- ***Επιχειρησιακή Έρευνα και Βελτιστοποίηση:*** Η επιχειρησιακή έρευνα έχει κοινά σημεία και αλληλοεπιδρά με τα παραπάνω πεδία στα οποία συνεισφέρει με μεγάλο αριθμό μεθόδων από διάφορα πεδία της στην ανάλυση δεδομένων, στην απόκτηση γνώσης και τη λήψη αποφάσεων. Αρκετές μέθοδοι μοντελοποίησης μπορούν να διατυπωθούν σαν προβλήματα βελτιστοποίησης. Δηλαδή, στο πρόβλημα θα υπάρχει μια αντικειμενική συνάρτηση που θα πρέπει να μεγιστοποιηθεί (ή να ελαχιστοποιηθεί) κάτω από ένα σύνολο περιορισμών που πρέπει να ληφθούν υπόψη. Ο στόχος είναι να βρεθεί μια βέλτιστη λύση (τιμή μιας μεταβλητής) ή ένας βέλτιστος συνδυασμός τιμών μεταβλητών, που θα ικανοποιούν τους περιορισμούς του προβλήματος. Οι μέθοδοι βελτιστοποίησης μελετούν διάφορες προσεγγίσεις για την επίλυση αυτού του τύπου προβλημάτων. Οι μεθευρετικοί αλγόριθμοι είναι ένας σχετικός τομέας και αφορά την εύρεση καλών ευρετικών για την αποτελεσματική επίλυση προβλημάτων βελτιστοποίησης.

- ***Γραμμική Άλγεβρα:*** Ένα απλό μοντέλο πολυμεταβλητής γραμμικής παλινδρόμησης μπορεί να επιλυθεί καλύτερα και πιο εύκολα, χρησιμοποιώντας πίνακες και διανύσματα με τη χρήση γραμμικής άλγεβρας. Τα προβλήματα βελτιστοποίησης συνήθως διατυπώνονται χρησιμοποιώντας γραμμική άλγεβρα. Ορισμένα άλλα μοντέλα μηχανικής μάθησης και στατιστικών παρουσιάζονται και επιλύονται, τουλάχιστον για ορισμένες περιπτώσεις, χρησιμοποιώντας γραμμική άλγεβρα. Αυτή είναι η περίπτωση των μηχανών διανυσμάτων υποστήριξης (support vector machines).
- ***Θεωρία Πιθανοτήτων:*** Πολλοί τρόποι μοντελοποίησης δεδομένων βασίζονται στη θεωρία πιθανοτήτων. Τα γραφικά μοντέλα και τα Bayesian δίκτυα είναι μερικά από αυτά.
- ***Γράφοι ή Γραφήματα (Graphs):*** Ορισμένες από τις διαθέσιμες πληροφορίες αναπαρίστανται εύκολα σε γράφους/γραφήματα. Αυτή είναι η περίπτωση των κοινωνικών δικτύων. Η θεωρία γράφων (graph theory) παρέχει έννοιες και εργαλεία για την ανάλυση αυτού του τύπου δεδομένων. Τα σύνθετα δίκτυα είναι ο όρος που σημαίνει ένα δίκτυο με μη τετριμμένη τοπολογική δομή. Τα δέντρα είναι επίσης γράφοι με τον περιορισμό ότι δεν πρέπει να περιέχουν κύκλους. Επιπλέον, μερικά από τα εργαλεία μοντελοποίησης δεδομένων, όπως τα γραφικά μοντέλα, βασίζονται επίσης σε γραφήματα για την αναπαράσταση της γνώσης.
- ***Τοπολογία:*** Ο τομέας της Τοπολογικής Ανάλυσης Δεδομένων εμφανίστηκε πρόσφατα ως τρόπος εξαγωγής σχετικών χαρακτηριστικών από δεδομένα. Στην τοπολογική ανάλυση δεδομένων, τα δεδομένα αντιμετωπίζονται ως αντικείμενα μεγάλων διαστάσεων. Στόχος της είναι να ανακαλύψει σχετικές, ποιοτικές και ποσοτικές τοπολογικές δομές απευθείας από τα δεδομένα.
- ***Οπτική Αναλυτική:*** Είναι δύσκολο να κατανοήσουμε μεγάλα δεδομένα. Η οπτικοποίηση δεδομένων παρέχει εργαλεία για μια πιο αποτελεσματική κατανόηση των δεδομένων και η οπτική αναλυτική παρέχει επίσης εργαλεία για την μετατροπή (ετερογενών) δεδομένων μεγάλου όγκου και πολυπλοκότητας σε κατανοητές, οπτικοποιημένες αναπαραστάσεις που διευκολύνουν την ερμηνεία τους, την ανάλυσή τους, τη αλληλεπιδραστική εξερεύνηση τους βοηθώντας έτσι τις διαδικασίες λήψης αποφάσεων.
- ***Γλώσσες Προγραμματισμού και Λογισμικό:*** Οι κατάλληλες γλώσσες προγραμματισμού για μεγάλα δεδομένα περιλαμβάνουν τις: Python, R, Java, Scala, Julia, κα. Τα πλαίσια γλωσσών προγραμματισμού που χρησιμοποιούνται συνήθως σε αυτό το πεδίο περιλαμβάνουν μεταξύ άλλων τα: Apache Spark, MapReduce, Hadoop, Flink, κα. Ενώ τα εργαλεία οπτικοποίησης δεδομένων περιλαμβάνουν μεταξύ άλλων τα: Tableau και Spotfire.

Μεγάλα Δεδομένα

Από πού προέρχονται τα Μεγάλα Δεδομένα; Μια απλή απάντηση είναι από ‘παντού’. Οι πηγές δεδομένων που αγνοήθηκαν κατά το παρελθόν εξαιτίας τεχνικών περιορισμών αντιμετωπίζονται σήμερα σαν πηγές χρυσού. Τα μεγάλα δεδομένα αναφέρονται συνήθως σε δεδομένα που προέρχονται από πολλές διαφορετικές μορφές, δομημένες, αδόμητες ή σε ροές δεδομένων (data streams). Σημαντικές πηγές τέτοιων δεδομένων είναι τα αρχεία καταγραφής των ιστοσελίδων (Web Logs), οι αναρτήσεις στα μέσα κοινωνικής δικτύωσης όπως το Facebook ή τα δεδομένα από την κυκλοφορία, RFID, συστήματα GPS, δίκτυα αισθητήρων, κοινωνικά δίκτυα, έγγραφα κειμένων στο διαδίκτυο, ευρετήρια αναζήτησης στο διαδίκτυο, αρχεία κλήσεων, αστρονομία, μετεωρολογία, βιολογία, γενετική, πυρηνική φυσική, βιοχημικά πειράματα, ιατρικά δεδομένα, επιστημονική έρευνα, αρχεία μουσικής, φωτογραφίας, βίντεο και εφαρμογές ηλεκτρονικού εμπορίου (Sharda et al. 2014; Dietrich et al., 2015). Μια μηχανή αναζήτησης ιστού όπως η Google χρειάζεται να αναζητήσει και να δημιουργήσει ευρετήριο για δισεκατομμύρια ιστοσελίδες ώστε να μπορεί να δίνει σχετικά αποτελέσματα αναζήτησης σε κλάσμα του δευτερολέπτου. Αν και αυτό, δεν γίνεται σε πραγματικό χρόνο, η δημιουργία ενός ευρετηρίου όλων των ιστοσελίδων στο Διαδίκτυο δεν είναι εύκολη υπόθεση. Ευτυχώς για την Google, ήταν σε θέση να λύσει αυτό το πρόβλημα χρησιμοποιώντας, μεταξύ άλλων εργαλείων, αναλυτικές τεχνικές Μεγάλων Δεδομένων.

Υπάρχουν δύο θέματα στη διαχείριση των δεδομένων αυτής της κλίμακας: η αποθήκευση και η επεξεργασία τους. Αν μπορούσαμε να αγοράσουμε μια εξαιρετικά δαπανηρή λύση αποθήκευσης όλων των δεδομένων σε ένα σημείο, σε μία μονάδα, καθιστώντας αυτή την μονάδα ασφαλή αποθηκευτικό χώρο, αυτό θα απαιτούσε μια πολύ μεγάλη δαπάνη. Για την αντιμετώπιση αυτού του προβλήματος, προτάθηκε μια έξυπνη λύση που περιλάμβανε την αποθήκευση αυτών των δεδομένων σε τμήματα σε διαφορετικούς υπολογιστές που θα συνδέονται μέσω ενός δικτύου, τοποθετώντας ένα ή δύο αντίγραφα από αυτό το τμήμα σε διαφορετικές θέσεις στο δίκτυο, τόσο λογικά όσο και φυσικά. Χρησιμοποιήθηκε αρχικά στη Google (γνωστό ως Google File System) και αργότερα αναπτύχθηκε και κυκλοφόρησε σε ένα Apache έργο ως Hadoop Distributed File System (HDFS).

Ωστόσο, η αποθήκευση αυτών των δεδομένων είναι μόνο το μισό πρόβλημα. Τα δεδομένα είναι άχρηστα αν δεν παρέχουν επιχειρηματική αξία και για να προσφέρουν επιχειρηματική αξία, θα πρέπει να είναι δυνατόν να αναλυθούν. Πώς αναλύονται όμως τέτοιες τεράστιες ποσότητες δεδομένων; Η διενέργεια όλων των υπολογισμών σε έναν ισχυρό υπολογιστή δεν λειτουργεί μιας και αυτή η κλίμακα αναλύσεων θα δημιουργούσε τεράστια επιβάρυνση ακόμη και σε έναν πολύ ισχυρό υπολογιστή. Για την αντιμετώπιση και αυτού του προβλήματος, προτάθηκε μια άλλη έξυπνη λύση που συνοψίζεται στη λογική: Πιέστε τον υπολογισμό στα δεδομένα, αντί να σπρώξετε τα δεδομένα σε έναν κόμβο υπολογιστών. Αυτό ήταν ένα νέο πρότυπο και αυτό οδήγησε σε έναν εντελώς νέο τρόπο επεξεργασίας δεδομένων. Αυτό σήμερα το γνωρίζουμε ως παράδειγμα προγραμματισμού MapReduce, το οποίο έκανε την επεξεργασία των μεγάλων δεδομένων πραγματικότητα. Το MapReduce αναπτύχθηκε αρχικά στη Google ενώ μια επόμενη έκδοση κυκλοφόρησε από την Apache σαν Hadoop MapReduce.

Σήμερα, όταν μιλάμε για αποθήκευση, επεξεργασία ή ανάλυση μεγάλων δεδομένων, το HDFS και το MapReduce εμπλέκονται σε κάποιο βαθμό. Παρόλο που η κύρια εργαλειοθήκη είναι διαθέσιμη ως λογισμικό ανοιχτού κώδικα (open source), έχουν ξεκινήσει αρκετές εταιρείες για την παροχή κατάρτισης ή εξειδικευμένων υπηρεσιών αναλυτικού υλικού ή λογισμικού.

Ορισμοί

Με τον όρο Μεγάλα Δεδομένα (Big Data), άνθρωποι από διαφορετικές επιστημονικές περιοχές, χαρακτηρίζουν διαφορετικά πράγματα. Παραδοσιακά, ο όρος 'Μεγάλα Δεδομένα' έχει χρησιμοποιηθεί για να περιγράψει τους τεράστιους όγκους δεδομένων που αποκτώνται και αναλύονται από μεγάλες επιχειρήσεις όπως η Google ή επιστημονικά ερευνητικά κέντρα όπως η NASA. Τι σημαίνει όμως αυτός ο όρος για τις περισσότερες επιχειρήσεις, το μέγεθος των οποίων ποικίλει δραματικά; Η έννοια 'μεγάλος' θα πρέπει να εκτιμάται σε σχέση με το μέγεθος του κάθε οργανισμού ή της κάθε επιχείρησης καθώς και το από πού ξεκινάει κάποιος, πως προχωρά και που θέλει να φτάσει. Η ουσία είναι η ανακάλυψη νέας αξίας, νέων ιδεών και ευκαιριών.

Τα μεγάλα δεδομένα δεν είναι καινούργια. Το νέο είναι ότι ο ορισμός και η δομή των μεγάλων δεδομένων μεταβάλλονται συνεχώς. Οι εταιρείες έχουν αποθηκεύσει και αναλύουν μεγάλους όγκους δεδομένων από την εμφάνιση των αποθηκών δεδομένων (data warehouses) στις αρχές της δεκαετίας του 1990. Αν και χρησιμοποιούνταν terabytes δεδομένων σαν συνώνυμα με τις μεγάλες αποθήκες δεδομένων, τώρα είναι petabytes και ο ρυθμός αύξησης των όγκων δεδομένων συνεχίζει να αυξάνει καθώς οι επιχειρήσεις επιδιώκουν να αποθηκεύουν και να αναλύουν όλο και πιο λεπτομερή δεδομένα συναλλαγών καθώς και δεδομένα από το Web αλλά και από άλλες πηγές, προκειμένου να κατανοήσουν καλύτερα την συμπεριφορά των πελατών τους και του ανταγωνισμού.

Οι μονάδες Terabyte (TB), Petabyte (PB) κοκ., είναι ψηφιακές μονάδες μέτρησης που χρησιμοποιούνται για τη μέτρηση του μεγέθους μεγάλων ποσοτήτων πληροφοριών όπως ψηφιακών αρχείων δεδομένων, της μνήμης RAM ενός υπολογιστή, η μηνιαία ή ετήσια ανταλλαγή ψηφιακών πληροφοριών μεταξύ ορισμένων ιστότοπων ή η μέτρηση του μεγέθους κάποιων ψηφιακών αρχείων μεγάλων ιστότοπων κλπ. Είναι ενδιαφέρον ότι πρόσφατα επιστήμονες με τη θεωρία ότι ένας ανθρώπινος εγκέφαλος

μπορεί να αποθηκεύσει δεδομένα έως και 2,5 PB. Στον επόμενο πίνακα 3.1 μπορούμε να δούμε αναλυτικά τις μονάδες μέτρησης που χρησιμοποιούνται.

Πίνακας 3.1. *Μονάδες μέτρησης*

Μονάδες Μέτρησης	**Μέγεθος**	**Bytes**
Bit	Δυαδικό ψηφίο (1/0)	-
Byte	8 bits	$2^0 = 1024^0$
Kilobyte (KB)	1,024 Bytes	$2^{10} = 1024^1$
Megabyte (MB)	1,024 Kilobytes	$2^{20} = 1024^2 = 1,048,576$ bytes
Gigabyte (GB)	1,024 Megabytes	$2^{30} = 1024^3 = 1,073,741,824$
Terabyte (TB)	1,024 Gigabytes	$2^{40} = 1024^4 = 1.099.511.627.776$
Petabyte (PB)	1,024 Terabytes	$2^{50} = 1024^5 = 1,125,899,906,842,624$
Exabyte (EB)	1,024 Petabytes	$2^{60} = 1024^6 = 1,152,921,504,606,846,976$
Zettabyte (ZB)	1.024 Exabytes	$2^{70} = 1024^7 = 1,180,591,620,717,411,303,424$
Yottabyte (YB)	1,024 Zettabytes	$28^0 = 1024^8 = 1,208,925,819,614,629,174,706,176$

Σύμφωνα με την αναφορά 'Data Age 2025', μέχρι το 2025, η παγκόσμια βάση δεδομένων θα αυξηθεί σε 175 Zettabytes (από 33 το 2018), με τις συσκευές IoT να αναμένεται να δημιουργήσουν πάνω από 90 zettabytes δεδομένων (Reinsel et al., 2017). Αυτά υπό ομαλή εξέλιξη των πραγμάτων. Στην πορεία όμως, δημιουργήθηκε το ερώτημα 'Ποιες θα είναι οι επιπτώσεις από την Πανδημία στην παγκόσμια παραγωγή δεδομένων'; Τον Μάρτιο του 2021, η ερευνητική εταιρεία IDC ανακοίνωσε ότι 'ο όγκος των δεδομένων που δημιουργήθηκαν και αναπαράχθηκαν γνώρισε ασυνήθιστα υψηλή ανάπτυξη το 2020 λόγω της δραματικής αύξησης του αριθμού των ανθρώπων που εργάζονται, μαθαίνουν και διασκεδάζουν από το σπίτι'. Συνολικά, το 2020, δημιουργήθηκαν ή αντιγράφηκαν περίπου 64,2 Zettabytes δεδομένων. Σύμφωνα με την ίδια εταιρεία, προβλέπεται ότι η παγκόσμια δημιουργία και αναπαραγωγή δεδομένων θα αυξάνεται κατά 23% ετησίως κατά την περίοδο 2020-2025.

Χαρακτηριστικά Μεγάλων Δεδομένων

Δεδομένα δημιουργούνται καθημερινά και με συνεχώς αυξανόμενο ρυθμό. Πηγές παραγωγής δεδομένων όπως κινητά τηλέφωνα, μέσα κοινωνικής δικτύωσης, τεχνολογίες απεικόνισης στην ιατρική, αισθητήρες, διαδίκτυο, κα., δημιουργούν συνεχώς νέα δεδομένα τα οποία θα πρέπει να αποθηκευτούν κάπου ώστε στη συνέχεια να αξιοποιηθούν κατάλληλα. Η απλή καταγραφή και αποθήκευση αυτής της τεράστιας παραγωγής δεδομένων είναι δύσκολη, αλλά ουσιαστικά πιο δύσκολη είναι η ανάπτυξη κατάλληλων μεθόδων για την ανάλυση αυτών των τεράστιων ποσοτήτων δεδομένων για την ανακάλυψη σημαντικών προτύπων-μοτίβων και την εξαγωγή χρήσιμων πληροφοριών. Αυτές οι προσπάθειες υπόσχονται να συμβάλουν και συμβάλουν ήδη στην ανάπτυξη της επιστήμης και στην λήψη των αποφάσεων στη διοίκηση των κυβερνήσεων, των επιχειρήσεων, κα., αλλά και στην καθημερινή ζωή. Για παράδειγμα, για εταιρείες όπως το Facebook, το LinkedIn, Booking, TripAdvisor, κα., τα ίδια τα δεδομένα είναι το κύριο προϊόν τους. Οι αποφάσεις αυτών των επιχειρήσεων βασίζονται σε μεγάλο βαθμό στα δεδομένα που συλλέγουν και αποθηκεύουν, και τα οποία περιέχουν όλο και περισσότερη εγγενή αξία καθώς συνεχώς αυξάνουν (Minelli et al., 2013; Davenport, 2014; Dietrich et al., 2015; Buyya et al., 2016).

Λόγω του μεγέθους ή/και της δομής του, τα Μεγάλα Δεδομένα δεν μπορούν να αναλυθούν αποτελεσματικά χρησιμοποιώντας μόνο παραδοσιακές βάσεις δεδομένων ή μεθόδους. Τα προβλήματα των Μεγάλων Δεδομένων, απαιτούν νέα εργαλεία και τεχνολογίες για την αποθήκευση, τη διαχείριση, την ανάλυση και την απόδοση επιχειρηματικής αξίας. Αυτά τα νέα εργαλεία και τεχνολογίες επιτρέπουν τη δημιουργία, τον χειρισμό και τη διαχείριση μεγάλων συνόλων δεδομένων καθώς και τα ανάλογα περιβάλλοντα αποθήκευσης.

Τρία είναι τα καθοριστικά χαρακτηριστικά των Μεγάλων Δεδομένων:

- ***Τεράστιος Όγκος Δεδομένων:*** Αντί για χιλιάδες ή εκατομμύρια αντικείμενα (πλειάδες, περιπτώσεις, εγγραφές, γραμμές) και μερικές εκατοντάδες χαρακτηριστικά (μεταβλητές, στήλες) που μπορεί να έχουν οι κλασικές βάσεις δεδομένων, τα Μεγάλα Δεδομένα μπορεί να έχουν δισεκατομμύρια αντικείμενα και εκατομμύρια χαρακτηριστικά.
- ***Πολυπλοκότητα των Τύπων Δεδομένων και των Δομών:*** Τα Μεγάλα Δεδομένα αντικατοπτρίζουν την ποικιλία νέων πηγών δεδομένων, μορφών και δομών, συμπεριλαμβανομένων των ψηφιακών ιχνών που παραμένουν στον παγκόσμιο Ιστό και άλλων ψηφιακών αποθετηρίων για μετέπειτα ανάλυση.
- ***Ταχύτητα Δημιουργίας και Ανάπτυξης Νέων Δεδομένων:*** Τα Μεγάλα Δεδομένα μπορούν να περιγράφουν δεδομένα υψηλής ταχύτητας, με γρήγορη λήψη δεδομένων και ανάλυση σχεδόν σε πραγματικό χρόνο.

Πολλοί ακαδημαϊκοί, βιομηχανικοί αναλυτές και στελέχη επιχειρήσεων πιστεύουν ότι η ονομασία μεγάλα δεδομένα (Big Data) είναι εσφαλμένη. Αυτό για το οποίο μιλάνε κάποιοι και αυτό που εννοούν δεν είναι ακριβώς το ίδιο. Όπως και τα χαρακτηριστικά που τους αποδίδουν δεν είναι τα ίδια.

Έτσι, τα Μεγάλα Δεδομένα δεν είναι μόνο 'μεγάλα' λόγω του όγκου τους. Ο όγκος (volume) των δεδομένων είναι μόνο ένα από τα πολλά χαρακτηριστικά που συνδέονται συχνά με τα μεγάλα δεδομένα, όπως η ποικιλία (variety), η ταχύτητα (velocity), η μεταβλητότητα (variability), η αξία τους (value), κα. (Sharda et al., 2014). Αρχικά προτάθηκαν τα 3V (Laney, 2001), εν συνεχεία προστέθηκαν και άλλα για να φτάσουμε στα 6V, ενώ συνεχίζονται οι προτάσεις για περισσότερα. Βεβαίως αυτές οι συνεχείς προτάσεις, δημιουργούν σύγχυση ως προς το τι τελικά χαρακτηρίζει πραγματικά τα μεγάλα δεδομένα αλλά κυρίως για το τι περιλαμβάνει κάθε ένα από αυτά τα χαρακτηριστικά.

Ας δούμε πιο αναλυτικά τα σημαντικότερα από αυτά:

Όγκος (Volume): Ο όγκος των δομημένων και μη δομημένων δεδομένων είναι εξαιρετικά μεγάλος ενώ συνεχίζει να μεγαλώνει με ραγδαίους ρυθμούς.

Ποικιλία (Variety): Τα δεδομένα διατίθενται σήμερα σε μεγάλη ποικιλία, σε όλους τους τύπους μορφών, ξεκινώντας από παραδοσιακές βάσεις δεδομένων έως τα ιεραρχικά δεδομένα στα συστήματα OLAP, τα έγγραφα κειμένου, τα e-mails, HTML, XML, τα δεδομένα που συλλέγονται από αισθητήρες και μετρητές, τα βίντεο, τους ήχους, τα δεδομένα χρηματιστηρίων, κα.. Σύμφωνα με ορισμένες εκτιμήσεις, το 80 με 85 τοις εκατό των δεδομένων όλων των επιχειρήσεων ανήκουν σε κάποιο είδος μη δομημένης ή ημιδομημένης μορφής (δεδομένα που συλλέγονται και αφορούν αντικείμενα-πλειάδες οι οποίες όμως δεν έχουν όλες την ίδια δομή), μορφής που δεν είναι κατάλληλη για παραδοσιακές βάσεις δεδομένων. Η αξία τους όμως είναι εξαιρετικά σημαντική και ως εκ τούτου πρέπει να συμπεριλαμβάνονται στις αναλύσεις για την υποστήριξη της λήψης αποφάσεων.

Ταχύτητα (Velocity): Η ταχύτητα αναφέρεται στο πόσο γρήγορα παράγονται, συλλέγονται, αποθηκεύονται, επεξεργάζονται και αναλύονται τα δεδομένα, προκειμένου να αξιοποιηθούν για τη λήψη αποφάσεων. Η ταχύτητα είναι ίσως το πιο παραμελημένο χαρακτηριστικό των μεγάλων δεδομένων. Η άμεση αντίδραση για την αντιμετώπιση των προβλημάτων λόγω χαμηλής ταχύτητας, αποτελεί πρόκληση για τις περισσότερες επιχειρήσεις και οργανισμούς. Για τα ευαίσθητα στο χρόνο περιβάλλοντα λήψης αποφάσεων, η αξία των παραγόμενων δεδομένων υποβαθμίζεται σταδιακά με το πέρασμα του χρόνου και τελικά καθίσταται άχρηστη. Ας το σκεφτούμε αυτό για περιπτώσεις όπως στην περίπτωση της υγεία ενός ασθενούς, στην επιλογή ενός επενδυτικού χαρτοφυλακίου, κοκ.. Το συνηθισμένο είναι να εργαζόμαστε σε καταστάσεις ηρεμίας, χρησιμοποιώντας κατάλληλα λογισμικά και συστήματα για την εξόρυξη μεγάλων ποσοτήτων από ποικίλες πηγών δεδομένων. Αν και αυτό είναι εξαιρετικά σημαντικό και πολύτιμο, υπάρχει μια άλλη κατηγορία αναλυτικών που βασίζεται στη φύση της ταχύτητας των μεγάλων δεδομένων, που ονομάζεται 'streaming analytics' ή 'in-motion analytics', η οποία συνήθως παραβλέπετε. Εάν γίνει σωστά, η αναλυτική ροών δεδομένων μπορεί να είναι εξίσου πολύτιμη και σε ορισμένα επιχειρηματικά περιβάλλοντα πιο πολύτιμη από μια ανάλυση των δεδομένων σε κατάσταση ηρεμίας.

Αλήθεια - Εγκυρότητα (Veracity): Η Veracity είναι ένας όρος που χρησιμοποιείται ως το τέταρτο 'V' που περιγράφει τα Μεγάλα Δεδομένα και προτάθηκε από την IBM. Περιλαμβάνει έννοιες όπως οι: αξιοπιστία, ακρίβεια, αλήθεια, ειλικρίνεια, ποιότητα, χρησιμότητα των δεδομένων. Στο πλαίσιο των μεγάλων δεδομένων, δεν είναι μόνο η ποιότητα των δεδομένων που έχει σημασία, αλλά πόσο αξιόπιστη είναι η πηγή, ο τύπος και η επεξεργασία των δεδομένων. Για τον χειρισμό της 'αλήθειας' των Μεγάλων Δεδομένων, χρησιμοποιούνται συχνά διάφορα εργαλεία και τεχνικές, μετατρέποντας τα δεδομένα σε ποιοτικές και αξιόπιστες πληροφορίες. Λόγω των πολλών και διαφορετικών μορφών των Μεγάλων Δεδομένων, η ποιότητα και η ακρίβεια είναι λιγότερο ελεγχόμενες παράμετροι. Ας μην ξεχνάμε ότι ο κύριος στόχος της αξιοποίησης των μεγάλων δεδομένων είναι η λήψη καλύτερων αποφάσεων. Βεβαίως τίθενται διάφορα ερωτήματα που χρήζουν απάντησης προκειμένου να μειωθεί ο βαθμός αβεβαιότητας και κατά συνέπεια ο βαθμός του κινδύνου στη λήψη αποφάσεων. Ερωτήσεις όπως: Είναι τα δεδομένα μας αρκετά αξιόπιστα για να στηρίξουμε τις αποφάσεις μας πάνω σε αυτά; Μπορούμε να εμπιστευθούμε τα δεδομένα και κατά συνέπεια τα αποτελέσματα των αναλύσεών τους, ώστε να λάβουμε σημαντικές αποφάσεις; Για να αυξήσουμε την εγκυρότητα των αναλύσεων χρειάζεται να δίνουμε μεγάλη προσοχή και να αφιερώνουμε το μεγαλύτερο ποσοστό του χρόνου ενός έργου στη σωστή συγκέντρωση, διαλογή και καθαρισμό των δεδομένών μας. Αυτό θα διασφαλίζει ότι τα δεδομένα που αποτελούν τη βάση των αναλύσεων θα είναι αξιόπιστα και τα αποτελέσματα των αναλύσεων έγκυρα.

Αξία (value): Οι μεγάλες προσδοκίες από τα μεγάλα δεδομένα είναι η πρόταση αξίας τους. Διότι τι νόημα θα είχε να μπορούμε να αναλύουμε τα μεγάλα δεδομένα αν δεν μπορούμε μέσω αυτών των αναλύσεων να προσδίδουμε 'αξία' στις επιχειρήσεις; αυτός είναι ο λόγος που θεωρείται από πολλούς και το σημαντικότερο V. Μια αρχική αντίληψη για τα 'μεγάλα' δεδομένα είναι ότι περιέχει ή έχει περισσότερες πιθανότητες να περιέχει, περισσότερα μοτίβα και ενδιαφέρουσες ανωμαλίες από ότι τα 'μικρά' δεδομένα. Έτσι, αναλύοντας μεγαλύτερα δεδομένα και πιο πλούσια σε χαρακτηριστικά, οι επιχειρήσεις μπορούν να αποκτήσουν μεγαλύτερη επιχειρηματική αξία την οποία ενδέχεται να μην είχαν διαφορετικά. Ενώ οι χρήστες μπορούν να εντοπίσουν τα μοτίβα σε μικρά σύνολα δεδομένων χρησιμοποιώντας απλές μεθόδους στατιστικής και μηχανικής μάθησης ή ειδικά επί τούτω (ad-hoc) ερωτήματα και εργαλεία αναφοράς, τα 'Big Data' απαιτούν τη χρήση 'μεγάλων' αναλυτικών μεθόδων και τεχνικών. Η 'μεγάλη' αναλυτική σημαίνει μεγαλύτερη διορατικότητα και καλύτερες αποφάσεις, κάτι που κάθε επιχείρηση χρειάζεται σήμερα και θα χρειάζεται σε μεγαλύτερο βαθμό στο μέλλον για να αυξάνει την 'αξία' της.

Μεταβλητότητα (variability): Η μεταβλητότητα αναφέρεται σε δεδομένα των οποίων διάφορα χαρακτηριστικά μεταβάλλονται συνεχώς. Εκτός από τις αυξανόμενες ταχύτητες (varieties) και ποικιλίες (velocities) των δεδομένων, οι ροές δεδομένων μπορεί να είναι εξαιρετικά ασυνεπείς, με περιοδικές διακυμάνσεις. Είναι κάτι που αφορά μη αναμενόμενες επιλογές, συμπεριφορές, αντιδράσεις όπως κάτι που μπορεί να συμβεί πχ. στις διακοπές ή σε μια ανάρτηση στα μέσα κοινωνική δικτύωσης, κοκ.. Η μεταβλητότητα αφορά το βαθμό και την ταχύτητα με την οποία αλλάζει η δομή των δεδομένων μας καθώς και η συχνότητα που αλλάζει το νόημα ή το σχήμα των δεδομένων. Η καθημερινή, εποχική και οφειλόμενη σε κάποιο γεγονός αντίδραση μπορεί να είναι δύσκολη στη διαχείριση, ειδικά στα μέσα κοινωνικής δικτύωσης.

Έχουν προταθεί και άλλα Vs για να περιγράψουν τα μεγάλα δεδομένα αλλά ανεξάρτητα από το πόσα χαρακτηριστικά θα περιλάβουμε στον ορισμό μας για τα μεγάλα δεδομένα, αυτό που έχει σημασία και αποτελεί την πραγματική αξία των Μεγάλων Δεδομένων (Big Data) είναι ότι είναι εδώ και θα μείνουν μαζί μας για τη συνέχεια.

3.2. Ανάλυση Δεδομένων και Λήψη Αποφάσεων

Η επιστήμη των δεδομένων έχει γίνει πλέον ένας ευρύτατα και όλο πιο συχνά χρησιμοποιούμενος όρος (Albright et al., 2011; Favero and Belfiore, 2019). Είναι ένα ευρύ πεδίο που αντιπροσωπεύει ένα

συνδυασμό πολλών επιστημονικών κλάδων. Η επιστήμη των δεδομένων και η επιστήμη των αποφάσεων σχετίζονται άμεσα μιας και η λήψη αποφάσεων βασίζεται στην πληροφόρηση, σημαντικό μέρος της οποίας προέρχεται από την ανάλυση δεδομένων.

Οι επιχειρήσεις και οι οργανισμοί λειτουργούν στην εποχή της τεχνολογίας και των μεγάλων δεδομένων και θα πρέπει να είναι έτοιμες να αποδεχτούν τις όποιες σημαντικές (θετικές ή/και αρνητικές) επιπτώσεις συνεπάγονται αυτές. Η τεχνολογία κατέστησε δυνατή τη συλλογή και αποθήκευση τεράστιων ποσοτήτων δεδομένων. Κάθε δραστηριότητα κάθε ενέργεια μιας επιχείρησης, κάθε συναλλαγή μπορεί να καταγράφεται. Στο θέμα της καταγραφής στοιχείων έχουν τεθεί μέσω του GDRP (Γενικός κανονισμός για την προστασία δεδομένων: https://ec.europa.eu/info/law/law-topic/data-protection_el) κανόνες που πρέπει να ακολουθούνται. Η συγκέντρωση δεδομένων, έχει γίνει σχετικά εύκολη με αποτέλεσμα, οι κρατικοί φορείς, οι επιχειρήσεις κα., να έχουν στη διάθεσή τους τεράστιους όγκους από δεδομένα. Ωστόσο, είναι άλλο να γνωρίζεις ότι έχεις στη διάθεσή σου δεδομένα, άλλο να κατανοείς ότι έχεις στη διάθεσή σου κρυμμένο θησαυρό, άλλο να μπορείς να τον ανακαλύψεις, άλλο να μπορείς να αναλύσεις τα δεδομένα, άλλο να κατανοείς την αξία του και άλλο να μπορέσεις να τον αξιοποιήσεις.

Μια άλλη σημαντική επίπτωση της τεχνολογίας είναι ότι έχει δώσει, σε πολύ περισσότερους ανθρώπους, τη δυνατότητα να αναλύουν δεδομένα και να λαμβάνουν αποφάσεις βασιζόμενοι σε ποσοτικές αναλύσεις. Οι άνθρωποι που εισέρχονται στον επιχειρηματικό κόσμο έχουν κάποιας μορφής εκπαίδευση σε κάποιας μορφής λογισμικό, ιδίως λογιστικά φύλλα, γνωρίζουν να χρησιμοποιούν υπολογιστές, έχουν πρόσβαση σε επιχειρησιακά δεδομένα και είναι εξοικειωμένοι με έστω και απλές στατιστικές και άλλες ποσοτικές μεθόδους ανάλυσης δεδομένων μιας και η ποσοτική ανάλυση αποτελεί πλέον αναπόσπαστο μέρος της καθημερινής εργασίας τους.

Καθώς η ανάλυση δεδομένων έχει γίνει πολύ σημαντική για τον επιχειρηματικό κόσμο, οι εταιρείες λογισμικού ανέπτυξαν καλύτερα εργαλεία για την ανάλυση δεδομένων ή ακριβέστερα για την επιχειρηματική ευφυΐα, την επιχειρηματική αναλυτική και την Ανάλυση Μεγάλων Δεδομένων.

3.3. Μεγάλα Δεδομένα και Επιχειρήσεις

Τα μεγάλα δεδομένα από μόνα τους, ανεξάρτητα από το μέγεθος, τον τύπο ή την ταχύτητα, είναι άχρηστα, εκτός εάν οι επιχειρήσεις αποφασίσουν να τα αξιοποιήσουν προσφέροντας αξία στις ίδιες (Mineli et al., 2013; Provost and Fawcett, 2013; Davenport, 2014; Dietrich et al., 2015; Buyya et al., 2016; Favero and Belfiore, 2019). Στο σημείο αυτό εμφανίζεται η αναλυτική μεγάλων δεδομένων (big data analytics). Παρόλο που τα στελέχη των επιχειρήσεων πραγματοποιούν πάντα αναφορές και παρακολουθούν πίνακες ελέγχου (dashboards) στην οθόνη των υπολογιστών τους, οι περισσότεροι δεν έχουν ανοίξει τις αποθήκες δεδομένων για να τις εξερευνήσουν ανά περίπτωση. Αυτό οφείλεται εν μέρει στο ότι τα εργαλεία ανάλυσης είναι πολύ πολύπλοκα για τον μέσο χρήστη αλλά επειδή επίσης και οι βάσεις δεδομένων, στις οποίες αυτοί έχουν πρόσβαση, συχνά δεν περιέχουν όλα τα δεδομένα που απαιτούνται από έναν ικανό χρήστη. Αυτό όμως αρχίζει να αλλάζει με δυναμικό τρόπο, χάρη στην αναλυτική μεγάλων δεδομένων.

Με την πρόταση απόδοσης αξίας στις επιχειρήσεις, το μεγάλα δεδομένα έφεραν επίσης και μεγάλες προκλήσεις για αυτές. Τα παραδοσιακά μέσα συλλογής, αποθήκευσης και ανάλυσης δεδομένων δεν έχουν τις δυνατότητες να αντιμετωπίζουν αποτελεσματικά και αποδοτικά τις απαιτήσεις των μεγάλων δεδομένων. Οι ανάγκες αυτές δημιούργησαν τις ανάγκες για ανάπτυξη νέων τεχνολογιών σε ερευνητικό επίπεδο και την ανάγκη για επένδυση σε αυτές από την μεριά των επιχειρήσεων.

Σε κάθε επιχείρηση που εξετάζει ή θέλει να πραγματοποιήσει μια τέτοια επένδυση, τίθεται το ερώτημα της αιτιολόγησης μιας τέτοιας επιλογής. Στη συνέχεια παρατίθενται οι κρίσιμοι παράγοντες επιτυχίας, οι οποίοι θα πρέπει πληρούνται προκειμένου να διασφαλιστεί μια τέτοια επένδυση από μια επιχείρηση ή οργανισμό:

1. Να υπάρχει μια σαφής επιχειρηματική ανάγκη που θα προκύπτει από το όραμα και τη στρατηγική της επιχείρησης. Οι επιχειρηματικές επενδύσεις οφείλουν να γίνονται για το καλό της επιχείρησης και όχι για χάρη των τεχνολογικών εξελίξεων. Επομένως, ο κύριος λόγος για την ανάπτυξη αναλυτικής μεγάλων δεδομένων (Big Data Analytics) θα πρέπει να είναι οι εξυπηρέτηση των αναγκών της επιχείρησης σε οποιοδήποτε επίπεδό της (στρατηγικό, τακτικό, λειτουργίας).
2. Ισχυρή και εξασφαλισμένη χρηματοδότηση. Εάν δεν τα διαθέτετε τότε είναι εξαιρετικά δύσκολο να επιτύχετε και καλό θα είναι να μην ξεκινήσετε την προσπάθεια πριν τα διασφαλίσετε. Εάν το πεδίο εφαρμογής είναι μία ή μερικές αναλυτικές εφαρμογές, τότε η χρηματοδότηση μπορεί να είναι σε επίπεδο ενός τμήματος μιας επιχείρησης. Ωστόσο, εάν ο στόχος είναι η οργανωτική επανασχεδίαση και ο μετασχηματισμός σε επίπεδο επιχείρησης, κάτι που συμβαίνει συχνά και συνιστάται, με στόχο την αξιοποίηση των μεγάλων δεδομένων, τότε η χρηματοδότηση θα πρέπει να καλύπτει όλη την επιχείρηση και να την καλύπτει επίπεδα διοίκησής της.
3. Ευθυγράμμιση της στρατηγικής της επιχείρησης και της πληροφορικής. Είναι σημαντικό να βεβαιωθείτε ότι το έργο της αναλυτικής υποστηρίζει πάντα την επιχειρηματική στρατηγική και όχι το αντίστροφο. Η αναλυτική θα πρέπει να διαδραματίσει τον ρόλο που επιτρέπει στην επιτυχή υλοποίηση της επιχειρηματικής στρατηγικής.
4. Ανάπτυξη κουλτούρας λήψης αποφάσεων βάσει δεδομένων/γεγονότων. Σε μια κουλτούρα λήψης αποφάσεων που στηρίζεται στα δεδομένα, οι αριθμοί μάλλον παρά η διαίσθηση ή οι υποθέσεις οδηγούν στη λήψη αποφάσεων. Υπάρχει επίσης μια κουλτούρα πειραματισμού για να δείτε τι δουλεύει και τι όχι. Για να δημιουργήσει μια κουλτούρα λήψης αποφάσεων που θα στηρίζετε στα γεγονότα, η διοίκηση της επιχείρησης θα πρέπει να:
   - Αναγνωρίσει ότι ορισμένα άτομα δεν μπορούν ή δεν θέλουν να προσαρμοστούν
   - Γίνει ένθερμος υποστηρικτής
   - Τονίσει ότι οι ξεπερασμένες μέθοδοι πρέπει να σταματήσουν
   - Ζητήσει να δείτε με βάσει ποιες αναλύσεις λαμβάνονται οι αποφάσεις
   - Συνδέσει τα κίνητρα και την επιβράβευση με τις επιθυμητές συμπεριφορές
5. Μια ισχυρή υποδομή δεδομένων. Οι αποθήκες δεδομένων παρέχουν μια ισχυρή υποδομή δεδομένων για την υλοποίηση της αναλυτικής. Αυτή η υποδομή αλλάζει και βελτιώνεται στην εποχή των μεγάλων δεδομένων με νέες τεχνολογίες. Η επιτυχία απαιτεί να συνδυαστεί το παλιό με το νέο για μια ολιστική υποδομή που λειτουργεί συνεργατικά.

Επιχειρηματικά Προβλήματα που Αντιμετωπίζονται από την Αναλυτική Μεγάλων Δεδομένων

Τα κορυφαία επιχειρηματικά προβλήματα που αντιμετωπίζει η Αναλυτική Μεγάλων Δεδομένων συνολικά είναι η αποτελεσματικότητα της διαδικασίας και η μείωση του κόστους καθώς και η βελτίωση της εμπειρίας των πελατών. Βεβαίως αναλόγως της επιχείρησης, οι προτεραιότητες αυτές διαφοροποιούνται. Η αποτελεσματικότητα της διαδικασίας και η μείωση του κόστους είναι κοινά επιχειρηματικά προβλήματα που μπορούν να αντιμετωπιστούν με την Αναλυτική Μεγάλων Δεδομένων, τα οποία ίσως είναι μεταξύ των κορυφαίων προβλημάτων που μπορούν να αντιμετωπιστούν με την Αναλυτική Μεγάλων Δεδομένων στις κατασκευές, την ενέργεια και τις επιχειρήσεις κοινής ωφέλειας, τις επικοινωνίες και τα ΜΜΕ, τις μεταφορές, και στους τομείς της υγείας την κυβέρνηση, κα. Έτσι, για παράδειγμα η βελτιωμένη εμπειρία πελατών μπορεί να βρίσκεται στην κορυφή της λίστας προβλημάτων που αντιμετωπίζουν οι ασφαλιστικές εταιρείες και οι λιανοπωλητές ενώ η διαχείριση κινδύνων είναι συνήθως στην κορυφή της λίστας για τις επιχειρήσεις στο κλάδο των τραπεζών και της εκπαίδευσης.

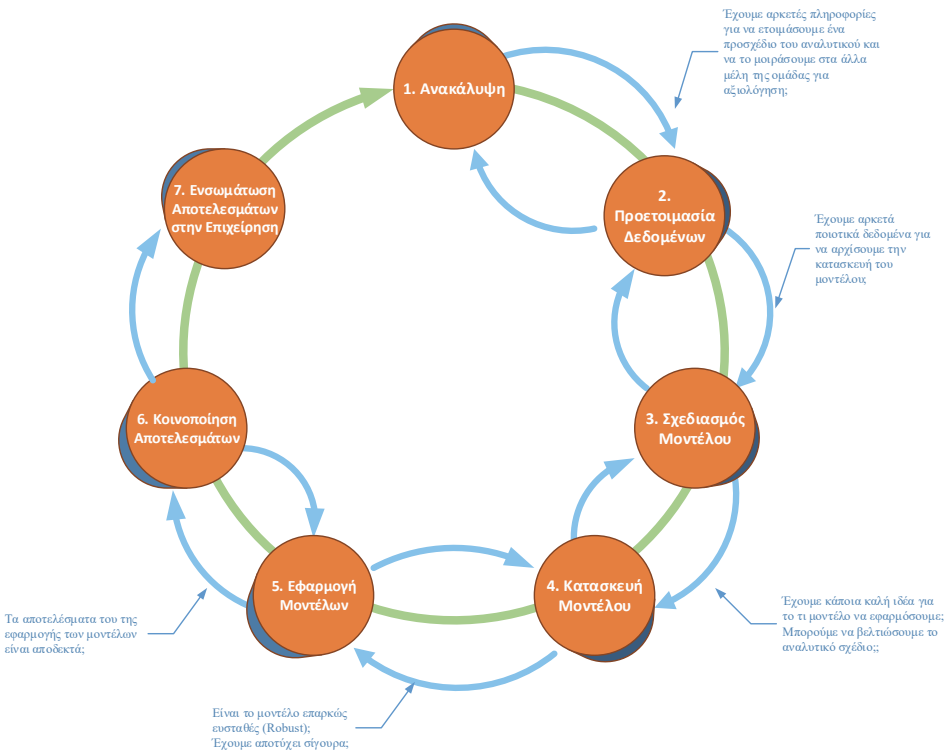
Στη συνέχεια δίνεται μια λίστα προβλημάτων που μπορούν να αντιμετωπιστούν με τη βοήθεια της Αναλυτικής Μεγάλων Δεδομένων:

- Βελτίωση της αποτελεσματικότητας των διαδικασιών και μείωση κόστους
- Διαχείριση εμπορικών σημάτων (brand name)
- Μεγιστοποίηση εσόδων, cross-selling (πώληση συμπληρωματικού προϊόντος με αυτό που έχει επιλέξει ο πελάτης) και up-selling (πώληση ακριβότερου προϊόντος από αυτό που έχει επιλέξει ο πελάτης)
- Βελτίωση της εμπειρίας πελατών
- Αναγνώριση απώλειας και προσέλκυσης πελατών
- Βελτίωση της εξυπηρέτησης πελατών
- Αναγνώριση νέων προϊόντων και ευκαιριών αγοράς
- Διαχείριση κινδύνου
- Βελτιωμένες δυνατότητες ασφάλειας

3.4. Κύκλος Ζωής Αναλυτικής Δεδομένων

Ο κύκλος ζωής της Αναλυτική δεδομένων (Data Analytics Lifecycle) έχει σχεδιαστεί ειδικά για προβλήματα Μεγάλων Δεδομένων και προγράμματα της Επιστήμης Δεδομένων. Ο κύκλος ζωής περιλαμβάνει έξι αλληλοεπιδρώσες φάσεις. Η κίνηση μεταξύ των φάσεων αλλά και εντός των φάσεων είναι συνεχής και αποσκοπεί στην πιο ακριβή απεικόνιση ενός πραγματικού έργου, στο οποίο κάποιες ενέργειες του έργου ενδέχεται να επιστρέψουν σε προηγούμενα στάδια καθώς προκύπτουν νέα δεδομένα και αναθεωρούνται διάφοροι παράμετροι και οι αναλυτές αναζητούν νέες πληροφορίες και εκτελούν νέες αναλύσεις (Davenport, 2014; Dietrich et al., 2015; Buyya et al., 2016; Favero and Belfiore, 2019).

Ο κύκλος ζωής της αναλυτικής δεδομένων καθορίζει τις βέλτιστες πρακτικές της διαδικασίας ανάλυσης που καλύπτουν διάφορες φάσεις από την ανακάλυψη έως την ολοκλήρωση του έργου (σχήμα 3.1). Ο κύκλος ζωής βασίζεται στις καθιερωμένες μεθόδους από τον τομέα της ανάλυσης δεδομένων και της επιστήμης των αποφάσεων. Αυτή η σύνθεση αναπτύχθηκε μετά τη συλλογή δεδομένων από επιστήμονες δεδομένων και λαμβάνοντας υπόψη καθιερωμένες προσεγγίσεις που συνεισφέρουν αναλόγως σε διάφορα τμήματα της όλης διαδικασίας.



Σχήμα 3.1. *Κύκλος Ζωής Αναλυτικής Δεδομένων*

Οι αναδράσεις μεταξύ των διαφόρων φάσεων, οφείλονται στο ότι τα μέλη των ομάδων εργασίας, μαθαίνουν συνήθως νέα πράγματα σε μια φάση ή διαπιστώνουν ένα πρόβλημα (πχ. έλλειψη σημαντικών ευρημάτων, μη ακριβή αποτελέσματα, κοκ) που τους αναγκάζει να επιστρέψουν σε προηγούμενες φάσεις για να βελτιώσουν εργασίες που είχαν πραγματοποιηθεί εκεί με βάση τόσο νέες πληροφορίες που τίθενται υπόψη των μελών της ομάδας όσο και πληροφορίες που ανακαλύπτονται στην πορεία. Τα διάφορα βέλη στο σχήμα (3.1), δίνουν την συνεχή κίνηση μεταξύ των διαφόρων φάσεων έως ότου τα μέλη της ομάδας έχουν επαρκείς πληροφορίες για να ολοκληρώσουν μια φάση και να μετακινηθούν στην επόμενη. Οι επεξηγήσεις περιλαμβάνουν δείγματα ερωτήσεων που τίθενται ώστε να διαπιστωθεί αν είμαστε έτοιμοι να προχωρήσουμε στην επόμενη φάση του κύκλου ζωής.

Ας δούμε εν συντομία τι περιλαμβάνει κάθε φάση:

1. ***Ανακάλυψη (Discovery):*** Στην φάση αυτή, η ομάδα αναλύει τα στοιχεία και ενημερώνεται για τον τομέα της επιχείρησης με τον οποίο θα ασχοληθεί. Σε αυτά, περιλαμβάνονται το ιστορικό υλοποίησης παρόμοιων έργων κατά το παρελθόν καθώς και των αποτελεσμάτων τους. Η ομάδα αξιολογεί επίσης τους διαθέσιμους πόρους για την υποστήριξη του έργου από άποψη ανθρώπων, τεχνολογίας, χρόνου και δεδομένων και προβαίνει στις κατάλληλες ενέργειες. Σημαντικές δραστηριότητες σε αυτήν τη φάση αποτελούν ο καθορισμός του προβλήματος της επιχείρησης-οργανισμού υπό το πρίσμα της αναλυτικής, το οποίο θα αντιμετωπιστεί στις επόμενες φάσεις, η διαμόρφωση αρχικών ερωτημάτων και υποθέσεων για διερεύνηση στη συνέχεια αλλά και η καταγραφή και μελέτη των διαθέσιμων δεδομένων.

2. ***Προετοιμασία Δεδομένων:*** Η 2η φάση, απαιτεί την παρουσία ενός αναλυτικού περιβάλλοντος δοκιμών, στο οποίο η ομάδα μπορεί να εργαστεί με δεδομένα και να εκτελέσει αναλύσεις δεδομένων για τη διάρκεια του έργου. Η ομάδα πρέπει να εκτελεί εργασίες εξαγωγής, φόρτωσης και μετασχηματισμού (Extract, Load, and Transform – ELT) ή εξαγωγής, μετατροπής και φόρτωσης (Extract, Transform and Load – ETL) δεδομένων ώστε να έχει στη διάθεσή της δεδομένα στο περιβάλλον δοκιμών. Σχεδόν πάντα τα δεδομένα μας θα πρέπει να μετασχηματιστούν έτσι ώστε η ομάδα να μπορεί να εργαστεί με αυτά και να τα αναλύσει. Σε αυτήν τη φάση, η ομάδα πρέπει επίσης να εξοικειωθεί πλήρως με τα δεδομένα και να προχωρήσει στην προεπεξεργασία τους όπως στον καθαρισμό των δεδομένων, την κανονικοποίηση των συνόλων δεδομένων και στην εκτέλεση μετασχηματισμών στα δεδομένα έτσι ώστε να είναι δυνατή η εφαρμογή μεθόδων εξόρυξης δεδομένων κα. (δες προεπεξεργασία δεδομένων §8.6).

3. ***Σχεδιασμός Μοντέλου*****:** Στη φάση αυτή γίνεται ο σχεδιασμός του μοντέλου, όπου η ομάδα εργασίας καθορίζει τις μεθόδους, τις τεχνικές και τη ροή εργασίας που σκοπεύει να ακολουθήσει στην επόμενη φάση της κατασκευής/δημιουργίας των μοντέλων. Η ομάδα εργασίας διερευνά τα δεδομένα για να μελετήσει τις σχέσεις μεταξύ των χαρακτηριστικών (μεταβλητών) και στη συνέχεια επιλέγει τις πιο σημαντικές μεταβλητές και τα πιο κατάλληλα μοντέλα.

4. ***Δημιουργία Μοντέλων*****:** Εδώ η ομάδα δημιουργεί σύνολα δεδομένων για την εκπαίδευση (training set) και τους ελέγχους (test set) των μοντέλων καθώς και για την εν συνεχεία εφαρμογή τους. Επιπλέον, σε αυτή τη φάση η ομάδα κατασκευάζει και δοκιμάζει μοντέλα με βάση τη δουλειά που έχει γίνει στην προηγούμενη φάση. Η ομάδα εξετάζει επίσης εάν το περιβάλλον εργασίας και τα υπάρχοντα εργαλεία της θα επαρκούν για τη εφαρμογή των μοντέλων στα διαθέσιμα δεδομένα ή εάν θα χρειαστεί ένα πιο ισχυρό περιβάλλον (πχ. κατανεμημένη/παράλληλη επεξεργασία, κλπ.) για την εφαρμογή των μοντέλων και των ροών εργασίας.

5. ***Εφαρμογή Μοντέλων*****:** Η ομάδα εργασίας θα πρέπει να εκτελέσει πλήρη εφαρμογή των μοντέλων σε πραγματικό περιβάλλον, επιλύοντας μέρος ή το σύνολο του προβλήματος. Εναλλακτικά θα μπορούσε να πραγματοποιήσει μια πιλοτική αλλά πλήρη εφαρμογή των μοντέλων.

6. ***Αποδοχή και Κοινοποίηση Αποτελεσμάτων*****:** Στη φάση αυτή, η ομάδα, σε συνεργασία με τους κύριους χρήστες των αποτελεσμάτων των αναλύσεων, καθορίζει εάν τα αποτελέσματα του έργου

είναι αποδεκτά ή όχι χρησιμοποιώντας και τα κριτήρια που αναπτύχθηκαν στην 1η φάση. Η ομάδα θα πρέπει να εντοπίσει βασικά ευρήματα, να ποσοτικοποιήσει την επιχειρηματική τους αξία και να ετοιμάσει μια εισήγηση που θα συνδέει και θα συνοψίζει τα ευρήματα προκειμένου να τα μεταφέρει στους χρήστες τους.

7. ***Λειτουργικότητα*:** Τέλος, η ομάδα εργασίας παραδίδει την τελική έκθεση, την τεκμηρίωση, τους ελέγχους, τον κώδικα καθώς και κάθε άλλο υλικό του συγκεκριμένου έργου.

Μόλις τα μέλη της ομάδας εφαρμόσουν τα μοντέλα και εξάγουν αποτελέσματα, είναι σημαντικό να τα παρουσιάσουν με τρόπο προσαρμοσμένο στις πραγματικές διαστάσεις του προβλήματος έτσι ώστε να είναι κατανοητά τόσο στο σύνολο των μελών της ομάδας εργασίας όσο και στα στελέχη της επιχείρησης. Επιπλέον, είναι σημαντικό να επισημαίνονται/υπενθυμίζονται τα οφέλη που θα έχει η επιχείρηση από αυτά. Εάν η ομάδα εργασίας εκτελέσει μια τεχνικά άψογη ανάλυση αλλά δεν παρουσιάσει τα αποτελέσματα σε γλώσσα κατανοητή από τη διοίκηση και τα στελέχη της επιχείρησης, τότε αυτοί δεν θα μπορέσουν να κατανοήσουν την αξία και τη χρησιμότητα που θα έχουν από αυτά με αποτέλεσμα την πιθανή μη αποδοχή ή την αδιαφορία τους για αυτά.

3.5. Τεχνολογίες Αναλυτικής Μεγάλων Δεδομένων

Η τεχνολογία αλλάζει ριζικά τον τρόπο με τον οποίο τα δεδομένα παράγονται, επεξεργάζονται, αναλύονται και χρησιμοποιούνται. Από τη μία πλευρά, η τεχνολογία βοηθά στην ανάπτυξη-παραγωγή νέων και πιο αποτελεσματικών πηγών δεδομένων. Από την άλλη, καθώς όλο και περισσότερα δεδομένα δημιουργούνται, η τεχνολογία συνεισφέρει βοηθώντας στην επεξεργασία αυτών των δεδομένων γρήγορα, αποτελεσματικά και τα απεικονίζει με παραστατικούς τρόπους με στόχο τη λήψη αποφάσεων υπό συνθήκες καλύτερης πληροφόρησης (Minelli et al., 2013; Dietrich et al., 2015; Buyya et al., 2016).

Υπάρχουν αρκετές τεχνολογίες για την επεξεργασία και την ανάλυση Μεγάλων Δεδομένων, αλλά οι περισσότερες έχουν κάποια κοινά χαρακτηριστικά (Kelly, 2012). Συγκεκριμένα, εκμεταλλεύονται τα προϊόντα υλικού (hardware) για να επιτρέψουν την οριζόντια επέκταση της υπολογιστικής ισχύος, τις παράλληλες τεχνικές επεξεργασίας, χρησιμοποιούν μη σχεσιακές δυνατότητες αποθήκευσης δεδομένων για την επεξεργασία μη δομημένων και ημιδομημένων δεδομένων, όπως και την εφαρμογή προηγμένων αναλυτικών δεδομένων και τεχνολογιών οπτικοποίησης δεδομένων σε μεγάλα δεδομένα για να παρουσιάσουν με πιο εύληπτο και παραστατικό τρόπο τα αποτελέσματα των αναλύσεων στους τελικούς χρήστες.

Υπάρχουν τρεις τεχνολογίες Μεγάλων Δεδομένων που ξεχωρίζουν και που οι περισσότεροι πιστεύουν ότι θα μεταμορφώσουν την επιχειρηματική αναλυτική και την διαχείριση δεδομένων. Αυτές είναι οι: Hadoop, MapReduce και NoSQL.

3.5.1. Hadoop

Ιστορικά Στοιχεία & Βασικές Έννοιες

Το σύστημα Hadoop είναι ένα πλαίσιο ανοιχτού κώδικα για επεξεργασία, αποθήκευση και ανάλυση τεράστιων ποσοτήτων κατανεμημένων και μη δομημένων δεδομένων (https://hadoop.apache.org/). Αρχικά δημιουργήθηκε από τον Doug Cutting (https://www.linkedin.com/in/cutting/) και τον Michael Cafarella (https://web.eecs.umich.edu/~michjc/bio.html) για την Yahoo!, οι οποίοι το εμπνεύστηκαν από το MapReduce (White, 2015; Sitto and Presser, 2015; Bengfort and Kim, 2016; Buyya et al., 2016; Jeyaraj et al., 2021). Το ίδιο το όνομα 'Hadoop' οφείλεται στον γιο του Doug Cutting, που είπε τη λέξη για ένα κίτρινο βελούδινο παιχνίδι ελέφαντα που είχε. Η Yahoo! προσέλαβε τον Cutting και επένδυσε σημαντικούς πόρους στην ανάπτυξη του έργου Hadoop, αρχικά για να αποθηκεύσει και να ευρετηριάσει