ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «ΨΗΦΙΑΚΕΣ ΕΠΙΚΟΙΝΩΝΙΕΣ ΚΑΙ ΔΙΚΤΥΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Ανάλυσης Μεγάλων Δεδομένων για Υποστήριξη Αποφάσεων

ΦΟΙΤΗΤΗΣ

ΖΑΝΝΗΣ ΛΕΜΟΣ

Α.Μ: ΜΕ 1554

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

ΔΗΜΟΣΘΕΝΗΣ ΚΥΡΙΑΖΗΣ

ΑΘΗΝΑ

ΜΑΙΟΣ 2017

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Δημοσθένη Κυριαζή για τις γνώσεις και τα ερεθίσματα που μου προσέφερε ώστε να ανακαλύψω τον χώρο των Big Data, καθώς και την πολύτιμη καθοδήγηση του κατά την διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Επίσης θα ήθελα να ευχαριστήσω το Ίδρυμα Ωνάση για την οικονομική στήριξη που μου παρείχε μέσω υποτροφίας καθ'όλη την διάρκεια του μεταπτυχιακού μου.

ΠΕΡΙΛΗΨΗ

Η παραγωγή τεράστιου όγκου δεδομένων και πληροφοριών στο σύνολο σχεδόν της επιχειρηματικότητας και των οργανισμών ανεξαρτήτου κλάδου, οδήγησε στην ανάγκη αξιοποίησης και εκμετάλλευσης των αχρησιμοποίητων παραγώμενων πληροφοριών καθώς η επεξεργασία τους παράγει τεράστια κέρδη. Αυτό οδήγησε στην δημιουργία νέων τεχνικών και τεχνολογιών ανάλυσης δεδομένων και γέννησε τον γενικότερο πεδίο των Big Data.

Στην παρούσα εργασία παρουσιάζουμε τις δυνατότητες που μας παρέχουν οι νέες τεχνολογίες στον χώρο των Big Data που μας βοηθούν στην ανάλυση μεγάλων datasets και την εξαγωγή αποτελεσμάτων βάση αυτών.

Ποιο συγκεκριμένα ‘μαθαίνουμε’ στον υπολογιστή μέσω της επεξεργασίας αρχείων που αφορούν την συμπεριφορά χρηστών ως προς τις προτιμήσεις τους σε ταινίες (βάση των βαθμολογιών τους), να εξάγει προτεινόμενες ταινίες βάση της ανάλυσης αυτής, δηλαδή βάση των προτιμήσεων του κάθε χρήστη.

ABSTRACT

In our days, the production of huge amounts of data and information in every aspect of the economy as a whole, has led many industries and organizations to exploit this unused information, because this exploitation generated enormous insight and profits. This led to the creation of new data analysis techniques and technologies and gave birth to the broader field of Big Data.

In this paper, we present the capabilities of new Big Data technologies that help us analyze huge datasets and export helpful results based on them.

More specifically, we make our computer 'learn' by processing files that present past user preferences in movies (based on their ratings), suggesting, based on this analysis, movies that we may like.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1^{ο} : ΕΙΣΑΓΩΓΗ ΣΤΑ BIG DATA

Οι περισσότερες και ποιο καθοριστικές εφευρέσεις σε όλη την ανθρώπινη ιστορία, είτε αυτό αφορούσε την δημιουργία της γλώσσας και του λεξιλογίου ή την εφεύρεση των υπολογιστών, είχαν ως σκοπό την παραγωγή και χρησιμοποίηση δεδομένων και πληροφορίας. Στην εποχή μας, έχουμε βιώσει μια τρομερή αύξηση στην ποσότητα των δεδομένων. Τα Big Data όπως ονομάζονται έχουν αρχίσει να αποκτούν κρίσιμο ρόλο σε κάθε τομέα της οικονομίας, και αυτό οφείλεται εν μέρη στη ταχεία ανάπτυξη των τεχνολογιών που επεξεργάζονται την ψηφιακή πληροφορία.

Τα δεδομένα έχουν κατακλύσει κάθε πτυχή και κάθε τομέα της παγκόσμιας οικονομίας. Οι εταιρίες διαχειρίζονται τεράστιες ποσότητες δεδομένων που αφορούν τις συναλλαγές τους με τους πελάτες τους αλλά και με τους προμηθευτές τους, την χρηματιστηριακή τους αξία, τα project που αναλαμβάνουν κτλπ. Εκατομμύρια αισθητήρες έχουν ενσωματωθεί στον φυσικό μας κόσμο όπως σε συσκευές σαν τα κινητά μας τηλέφωνα, στα αυτοκίνητα ακόμα και στα ψυγεία που παράγουν τεράστια ποσότητα πληροφοριών. Εκατομμύρια άνθρωποι σε όλο τον κόσμο συμβάλλουν στην εκθετική αύξηση των δεδομένων. Για παράδειγμα κάθε δευτερόλεπτο ενός high-definition video παράγει περισσότερο από 2000 φορές τα bytes που χρειάζονται για την αποθήκευση μιας σελίδας ενός κειμένου. Σε έναν ψηφιακό κόσμο όπως τον δικό μας, ο κάθε άνθρωπος που περνάει την μέρα του κάνοντας απλές καθημερινές δραστηριότητες όπως να μιλάει στο κινητό, να σερφάρει στο internet, να αγοράζει προϊόντα και υπηρεσίες online, δημιουργεί το δικό του ψηφιακό μονοπάτι αφήνοντας πίσω τεράστιες ποσότητες δεδομένων.

Αυτήν την στιγμή, ύστερα από έρευνες που πραγματοποιήθηκαν τα τελευταία χρόνια εκτιμάται ότι η ποσότητα των δεδομένων που είναι αποθηκευμένα σε επιχειρήσεις ή σε διαφόρους φορείς γενικότερα, υπερβαίνουν τα 7 exabytes και επιπλέον νέα δεδομένα τα οποία αποθηκεύουν οι καθημερινοί χρήστες του διαδικτύου ή των κινητών επικοινωνιών αφορούν άλλα 6 exabytes. Για να δώσω ένα παράδειγμα και να γίνει κατανοητό το μέγεθος το οποίο προανέφερα, τα 13 exabytes είναι ισοδύναμο με 60.000 εθνικές βιβλιοθήκες. Επιπλέον, σκεφτείτε ότι αν όλες οι λέξεις που ειπώθηκαν από το ανθρώπινο είδος είχαν ψηφιοποιηθεί σε κείμενο, θα προσέγγιζαν περίπου τα 5 exabytes, αισθητά λιγότερο από όσο αποθηκεύουν οι καθημερινοί χρήστες σε έναν χρόνο! Παράγουμε τόσα πολλά δεδομένα στις μέρες μας που είναι φυσικός αδύνατο να

αποθηκευτούν όλα. Για παράδειγμα στον τομέα της υγείας απορρίπτεται το 90% των δεδομένων που παράγονται (τα οποία συνήθως είναι real-time χειρουργικές επεμβάσεις). Η συνεχιζόμενη αύξηση του όγκου των δεδομένων που αποθηκεύουν οι εταιρίες σε συνδυασμό με την διαρκή άνοδο της χρήσης πολυμέσων και των social media όπως επίσης και η ενσωμάτωση στην ζωή μας του IoT (Internet of Things) θα αυξήσουν εκθετικά τον όγκο των διαχειρίσιμων δεδομένων στο εγγύς μέλλον.

Δεν υπάρχει καμία αμφιβολία ότι το τεράστιο μέγεθος των big data αλλά και η προοπτική που υπάρχει για την περαιτέρω εξάπλωσή τους, βρίσκονται στο επίκεντρο των περισσότερων ερευνών στις μέρες μας. Ένα από τα βασικά ερωτήματα που καλούμαστε να απαντήσουμε είναι τι αντίκτυπο έχει αυτό το tsunami δεδομένων στην οικονομία ή και στις ζωές των ανθρώπων γενικότερα. Πολλοί άνθρωποι είναι καχύποπτοι σχετικά με την ποσότητα των δεδομένων που συλλέγονται για κάθε πτυχή της ζωής τους, από το πως και που ψωνίζουν, τι ταινίες τους αρέσουν, που πάνε για διακοπές, μέχρι για την υγεία τους και το εισόδημά τους, όλες αυτές οι πληροφορίες που συλλέγονται μέσω επεξεργασίας δεδομένων μπορούν να αναλυθούν και να παρέχουν σε εταιρίες στοχευμένο marketing για τον κάθε πολίτη ξεχωριστά. Αυτό μας δείχνει πως τα Big Data έχουν διεισδύσει στις καθημερινές ζωές μας, αλλά και πως από αυτό στην πραγματικότητα μπορούμε να ωφεληθούμε ως άτομα ξεχωριστά αλλά και ως κοινωνία γενικότερα. Τα δεδομένα μπορούν πράγματι να δημιουργήσουν μια σημαντική οικονομική παγκόσμια ανάπτυξη, ενισχύοντας την παραγωγικότητα και την ανταγωνιστικότητα μεταξύ των επιχειρήσεων και την δημιουργία ενός σημαντικού οικονομικού πλεονάσματος για τις κυβερνήσεις και τους πολίτες. (Peter Lyman, 2003)

1.1) Η ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ ΤΩΝ BIG DATA

Το ανθρώπινο είδος προσπαθούσε να αποθηκεύσει και να διαχειριστεί δεδομένα χιλιάδες χρόνια πριν την εμφάνιση των υπολογιστών και της τεχνολογίας γενικότερα όπως την αντιλαμβανόμαστε στις μέρες μας. Η ικανότητα μας να αποθηκεύουμε και να αναλύουμε δεδομένα εξελίχθηκε παράλληλα με την εξέλιξη των ανθρώπων, αν και αυτή επιταχύνθηκε πάρα πολύ τον τελευταίο αιώνα, με την εφεύρεση μονάδων ψηφιακής αποθήκευσης, του internet κτλπ. Με τα Big Data να βρίσκονται στο επίκεντρο των ερευνών στις μέρες μας, ας ρίξουμε μια σύντομη ματιά πως φτάσαμε ως εδώ.

1.1.1) Τα δεδομένα στην αρχαία εποχή

Τα πρώτα παραδείγματα στην ανθρώπινη ιστορία που αφορούν την αποθήκευση και την ανάλυση δεδομένων μας πάνε χιλιάδες χρόνια προ Χριστού όπου παλαιολιθικές φυλές σχημάτιζαν εγκοπές σε μπαστούνια ή οστά για να καταγράφουν τις προμήθειές τους ή τις εμπορικές συναλλαγές τους με τις άλλες φυλές. Αρκετές χιλιάδες χρόνια μετά γύρω στο 2.500 π.Χ. οι Βαβυλώνιοι ανακάλυψαν τον άβακα, η οποία ήταν η πρώτη συσκευή που κατασκευάστηκε ειδικά για την εκτέλεση υπολογισμών. Επίσης την ίδια εποχή δημιουργήθηκαν και οι πρώτες βιβλιοθήκες στην πρώτη μας απόπειρα να αποθηκεύσουμε μαζικά δεδομένα και πληροφορίες. Η βιβλιοθήκη της Αλεξάνδρειας (την εποχή του Μ. Αλεξάνδρου) είναι η μεγαλύτερη συλλογή δεδομένων που πραγματοποιήθηκε στα χρόνια του αρχαίου κόσμου, στεγάζοντας ως και μισό εκατομμύριο παπύρους οι οποίοι αφορούσαν δεδομένα που κάλυπταν τα πάντα που ήταν γνωστά ως εκείνη την εποχή της ιστορίας. Ο αρχαιότερος υπολογιστής που έχει ανακαλυφθεί ονομάζεται μηχανισμός των Αντικυθήρων και θεωρείται ότι σχεδιάστηκε για την παρακολούθηση των άστρων καθώς και για την διατήρηση του χρονικού κύκλου των Ολυμπιακών αγώνων.

1.1.2) Η ανακάλυψη της στατιστικής

Στο Λονδίνο το 1663 ο επιστήμονας John Graunt πραγματοποιεί το πρώτο καταγεγραμμένο πείραμα στην στατιστική ανάλυση των δεδομένων. Με την καταγραφή δεδομένων πάνω στην θνησιμότητα, διατύπωσε την θεωρία ότι μπορεί να σχεδιάσει ένα σύστημα έγκαιρης προειδοποίησης για την βουβωνική πανώλη που μάστιζε την Ευρώπη εκείνη την εποχή. Το 1865 ο όρος “business intelligence” χρησιμοποιείται για πρώτη φορά από τον Richard Devens, ο οποίος περιέγραφε έναν τραπεζίτη ονόματι Henry Furnese, ο οποίος κατάφερε να αποκτήσει πλεονέκτημα έναντι των ανταγωνιστών του με την συλλογή και ανάλυση πληροφοριών σχετικά με τις επιχειρηματικές του δραστηριότητες. Αυτό θεωρείται ότι είναι η πρώτη μελέτη μιας επιχείρησης που χρησιμοποιήθηκε η ανάλυση δεδομένων για εμπορικούς σκοπούς. Το 1880 η υπηρεσία απογραφής των Η.Π.Α. αντιμετώπιζε ένα πρόβλημα. Υπολόγισαν ότι θα χρειαζόντουσαν τουλάχιστον 8 χρόνια για να αναλύσουν όλα τα δεδομένα που είχαν συλλέξει στην απογραφή και είχαν υπολογίσει ότι τα δεδομένα που θα έπαιρναν το 1890 θα χρειαζόντουσαν πάνω από 10 χρόνια, που σήμαινε ότι δεν θα είχαν επεξεργαστεί τα δεδομένα πριν την αρχή της

νέας απογραφής το 1900. Σε αυτό το πρόβλημα έδωσε λύση το 1881 ένας νεαρός μηχανικός που είχε προσληφθεί από την υπηρεσία. Ο Herman Hollerith, όπως λεγόταν δημιούργησε την Hollerith Tabulating μηχανή. Χρησιμοποιώντας διάτρητες κάρτες κατάφερε να μειώσει μια δουλειά που θα χρειαζόταν 10 και πλέον χρόνια σε μόλις 3 μήνες και θεωρείται ο πατέρας του σύγχρονου αυτοματοποιημένου υπολογισμού. Η εταιρία που ίδρυσε θα γινόταν αργότερα η γνωστή σε όλους μας IBM.

1.1.3) Η αρχή της αποθήκευσης δεδομένων

Η πρώτη απόπειρα για αποθήκευση δεδομένων στην σύγχρονη εποχή πραγματοποιήθηκε από τον Γερμανό εφευρέτη Fritz Pfleumer όπου ανακάλυψε μια μέθοδο για να αποθηκεύει δεδομένα μαγνητικά σε ταινία. Οι βασικές αρχές που ανέπτυξε χρησιμοποιούνται ακόμα και σήμερα με την μεγάλη πλειοψηφία των ψηφιακών δεδομένων να αποθηκεύονται μαγνητικά σε σκληρούς δίσκους. Ο William Dersch, μηχανικός της IBM, κατασκευάζει μια συσκευή που μπορεί να καταλάβει 16 λέξεις στα αγγλικά και αριθμούς και να τα μετατρέπει σε ψηφιακές πληροφορίες. Είναι η πρώτη απόπειρα δημιουργίας μηχανής για αναγνώριση φωνής. Το 1965 η Η.Π.Α. δημιουργούν το πρώτο μεγάλο κέντρο για την αποθήκευση δεδομένων με σκοπό να καταχωρήσουν 742 εκατομμύρια φορολογικές δηλώσεις και 175 εκατομμύρια δακτυλικά αποτυπώματα σε μαγνητικές ταινίες. Λίγα χρόνια μετά παρουσιάζεται ένα μοντέλο ‘σχεσιακής βάσης δεδομένων’ για την αποθήκευση πληροφοριών σε ιεραρχική μορφή, την οποία μπορεί να προσπελάσει οποιοσδήποτε που ξέρει τι ψάχνει. Το μοντέλο αυτό έβαλε τις βάσεις για τον τρόπο λειτουργίας πολλών βάσεων δεδομένων στις μέρες μας. Πριν την δημιουργία αυτού του μοντέλου, για να μπορούσες να έχεις πρόσβαση σε δεδομένα θα έπρεπε να είχες την βοήθεια ενός ειδικού.

1.1.4) Το internet και οι πρώτες σκέψεις για Big Data

Το 1991 ο Tim Bernes ανακοινώνει την δημιουργία του παγκόσμιου ιστού, γνωστού και ως World Wide Web, όπου τα δεδομένα θα είναι διαθέσιμα σε όλους από όπου και αν βρίσκονται. Το 1999 ο όρος Big Data κάνει την εμφάνιση του στο βιβλίο ‘Visually exploring gigabyte datasets in real time’ όπου επισημαίνεται ότι δεν έχουμε κανέναν τρόπο να αναλύσουμε μεγάλο όγκο δεδομένων και ότι ο σκοπός των υπολογιστών είναι να μας παρέχει διορατικότητα πάνω στα δεδομένα που δεν μπορούμε να επεξεργαστούμε και όχι να κάνουν αριθμητικές πράξεις. Το 2000

ο Hal Varian, ο οποίος είναι τώρα επικεφαλής του τμήματος των οικονομικών της Google, προσπάθησε για πρώτη φορά να υπολογίσει την ποσότητα της ψηφιακής πληροφορίας αλλά και τον ρυθμό ανάπτυξής της. Κατέληξε στο συμπέρασμα ότι για τα δεδομένα που παράγονται ανά χρόνο αντιστοιχούν 250 megabytes ανά άνθρωπο. Έναν χρόνο μετά διατυπώνονται 3 όροι που στις μέρες μας είναι κοινός αποδεκτοί για τον προσδιορισμό των χαρακτηριστικών των Big Data. Velocity, Volume, Variety. Το 2005 δημιουργείται το Hadoop, το οποίο δημιουργήθηκε ειδικά για την αποθήκευση και την ανάλυση των datasets που αφορούν τα Big Data. Γίνεται ευρέως γνωστό για την ικανότητά του να διαχειρίζεται αδόμητα δεδομένα όπως φωνή,video,κείμενα κλπ. τα οποία γνώριζαν ιδιαίτερη άνθιση εκείνη την εποχή.

1.1.5) <u>Τα Big Data στις μέρες μας</u>

ΕΙΚΟΝΑ 1) <u>Τα δεδομένα που παράγονται κάθε λεπτό</u>

Το άρθρο ‘The end of theory: The data deluge makes the scientific model obsolete’ εισάγει για πρώτη φορά την έννοια των Big Data στο ευρύ κοινό. Τα δεδομένα πλέον βομβαρδίζουν την καθημερινότητά μας και παράγονται σε εκθετικό βαθμό. Υπολογίζεται ότι μια μέση αμερικάνικη εταιρία με χίλιους ή παραπάνω υπαλλήλους αποθηκεύει 200 terabytes δεδομένων. Το 2010 ο πρόεδρος της Google Eric Schmidt, λέει σε ένα συνέδριο ότι ανά δύο μέρες δημιουργούνται τόσα δεδομένα, όσα από την αρχή της ανθρωπότητας ως το 2003. Τέλος, το 2011 σε έκθεση της McKinsey προβλέπεται ότι το 2018 η Η.Π.Α. θα έχουν έλλειμμα επαγγελματιών επιστημόνων με τις γνώσεις που χρειάζεται για να προσληφθούν ως Data Scientists-Analysts της τάξης των 150.000 ανθρώπων. (Rijmenam, 2016)

1.2) ΤΙ ΕΝΝΟΟΥΜΕ ΛΕΓΟΝΤΑΣ BIG DATA

Τι εννοούμε όμως όταν αναφερόμαστε στον όρο Big Data? Όταν λέμε Big Data εννοούμε τα datasets εκείνα στα οποία το σύνολο των δεδομένων είναι τόσο μεγάλο, όπου συμβατές τεχνολογίες ανάλυσης δεδομένων δεν μπορούν να τα αποθηκεύσουν και να τα αναλύσουν. Ο όρος Big Data είναι εσκεμμένα ασαφής και υπόκειται σε υποκειμενική κρίση καθώς δεν υπάρχει ας πούμε κάποιο όριο που αν το dataset το ξεπεράσει τότε θα θεωρηθεί ότι ανήκει στην κατηγορία Big Data. Και αυτό γιατί γνωρίζοντας ότι η τεχνολογία κάνει άλματα σε πολύ μικρά χρονικά διαστήματα, έτσι και τα δεδομένα θα πολλαπλασιάζονται και κάθε φορά ο όγκος ενός dataset που θα ανήκει στην κατηγορία των Big Data θα πρέπει να αυξάνεται. Επίσης ο ορισμός των Big Data έχει διαφορετική ερμηνεία ανά τομέα ανάλογα με το τι είδος λογισμικού είναι διαθέσιμο και ποια datasets είναι ίδια σε μια συγκεκριμένη βιομηχανία. Έχοντας υπόψιν μας αυτές τις παραμέτρους τα Big Data κυμαίνονται από λίγα Gigabytes σε δεκάδες Petabytes.

Για να κατανοήσουμε καλύτερα τι ακριβώς αντιπροσωπεύει η έννοια των Big Data τα αναλύουμε σε 5 κατηγορίες:

1) **ΟΓΚΟΣ**: Οι οργανισμοί και οι εταιρίες μαζεύουν πλέον δεδομένα από ένα μεγάλο αριθμό πηγών όπως από επιχειρηματικές συναλλαγές, τα social media (Facebook, Twitter κλτπ.) και επίσης από αισθητήρες που ολοένα και περισσότερο χρησιμοποιούνται για machine-to-machine επικοινωνία οδηγώντας μας στην IoT

εποχή. Στο παρελθόν, τόσο μεγάλος όγκος δεδομένων θα ήταν εξαιρετικά δύσκολο έως αδύνατο να αναλυθεί όμως στις μέρες μας έχουν δημιουργηθεί τεχνολογίες όπως το Hadoop και η Spark που έχουν βοηθήσει πολύ στην εξοικονόμηση χρόνου από την ανάλυση των δεδομένων.

2) **ΤΑΧΥΤΗΤΑ**: Πλέον οι συνεχής ροές δεδομένων σε πραγματικό χρόνο δημιούργησαν την ανάγκη για γρήγορη επεξεργασία και άντληση δεδομένων και παροχή αποτελεσμάτων σε έγκαιρο χρόνο**.**

3) **ΠΟΙΚΙΛΙΑ ΔΕΔΟΜΕΝΩΝ**: Τα δεδομένα στις μέρες μας έρχονται με κάθε δυνατό τρόπο σε οποιαδήποτε μορφή. Αυτά τα δεδομένα μπορεί να είναι δομημένα, αριθμητικά, αδόμητα έγγραφα κειμένων, e-mail, video, ήχος, συναλλαγές και πολλά άλλα. Άρα υπάρχει η ανάγκη για άμεση και γρήγορη ανάλυση των δεδομένων σε οποιαδήποτε μορφή αυτά ληφθούν.

4) **ΜΕΤΑΒΛΗΤΟΤΗΤΑ**: Εκτός από το γεγονός ότι έχουμε να αντιμετωπίσουμε τις αυξανόμενες ταχύτητες και τους διάφορους τύπους δεδομένων, έχουμε επίσης να αντιμετωπίσουμε την ασυνέπεια στην ροή των δεδομένων. Π.χ. σε παγκόσμιες γιορτές ή σε κάποιο νέο trend που μπορεί να σαρώνει τα social media παρατηρούνται οι λεγόμενες περίοδοι αιχμής όπου ο καταιγισμός από δεδομένα είναι δύσκολο να αντιμετωπιστεί και ειδικότερα στις περιπτώσεις που τα δεδομένα είναι σε αδόμητη μορφή.

5) **ΠΟΛΥΠΛΟΚΟΤΗΤΑ**: Η κίνηση των δεδομένων απλώνεται πάνω σε πολλές διαφορετικές ασύνδετες πηγές και γι αυτό είναι αναγκαία η συσχέτιση των σχέσεων μεταξύ των διαφόρων πηγών αλλιώς μπορεί πολύ γρήγορα να χαθεί ο έλεγχος. (Inc., 2016)

ΕΙΚΟΝΑ 2) ‘Ρεαλιστική’ απεικόνιση των Big Data

1.2.1) Γιατί τα Big Data είναι σημαντικά?

Τα τελευταία 50 χρόνια, η κοινωνία μας από εκεί που η πλειοψηφία των ανθρώπων ήταν σχετικά ελεύθερη από δεδομένα και πληροφορίες, ξαφνικά σ’ ένα τόσο σύντομο κομμάτι της ιστορίας μας, ζούμε σε έναν κόσμο όπου τα πάντα (κυριολεκτικά) που αφορούν τις κινήσεις μας καταγράφονται, συλλέγονται και αναλύονται. Ένα απλό παράδειγμα είναι ότι στην εποχή των

γιαγιάδων και των παππούδων μας, οι πληροφορίες που αφορούσαν τις ζωές τους βρισκόντουσαν σε χαρτί που κατέγραφε την ημερομηνία γέννησης, τους γάμους ή τις συλλήψεις που δημοσιευόντουσαν σε εφημερίδες, την ιδιοκτησία η οποία ήταν καταγεγραμμένη και φυλασσόταν σε κυβερνητικά κτίρια και τέλος η καταγραφή των θανάτων πάλι από εφημερίδες. Αυτά τα δεδομένα, μεταφράζονται σε 1 δευτερόλεπτο στην δικιά μας ζωή κατ' αντιστοιχία. Για να γίνει κατανοητό, έστω ότι αγοράζετε ένα εισιτήριο από το κινητό σας για να δείτε ένα έργο στον κινηματογράφο. Αυτό γίνεται άμεσα γνωστό και καταγράφεται από:

1) **Την τηλεπικοινωνιακή εταιρία στην οποία ανήκετε**

2) **Την εταιρία που έφτιαξε την εφαρμογή από την οποία αγοράσατε το εισιτήριο**

3) **Την εταιρία που παρέχει το λογισμικό στο κινητό σας (Apple, Google, BlackBerry, Microsoft)**

4) **Τον πωλητή των εισιτηρίων**

5) **Την τράπεζα από όπου θα παρθούν τα χρήματα**

6) **Την τράπεζα του πωλητή των εισιτηρίων που θα κατατεθούν τα χρήματα**

Πέρα από αυτά τα στοιχεία που κάθε εταιρία συλλέγει για να τα αξιοποιήσει μέσω των Big Data, υπάρχουν και άλλα στοιχεία που συλλέγουν γιατί είναι άμεσα ή έμμεσα χρήσιμα για την εξαγωγή συμπερασμάτων στις συνήθειες του κάθε πελάτη ξεχωριστά. Αυτά αφορούν:

1) **Το προϊόν που αγοράστηκε**

2) **Την ακριβή ώρα αγοράς του προϊόντος**

3) **Την τοποθεσία της αγοράς**

4) **Το γεγονός ότι η αγορά έγινε μέσω τηλεφώνου**

5) **Τον τύπο του τηλεφώνου και το λογισμικό που έχει εγκατεστημένο**

6) **Το ποσό που καταβλήθηκε και την μέθοδο πληρωμής**

7) **Που βρίσκεται το cinema για το οποίο κλείσατε εισιτήρια**

8) **Πόσα άτομα θα φέρεται μαζί σας**

Όλα αυτά τα στοιχεία καταγράφονται από μια απλή αγορά εισιτηρίου και είτε οι εταιρίες τα εκμεταλλεύονται οι ίδιες, είτε τα πουλάνε σε εταιρίες που έχουν κέρδος από αυτές τις πληροφορίες. Από αυτά τα στοιχεία προέρχεται το στοχευμένο marketing αρκετών εταιριών όπου μπορεί να λάβετε ένα e-mail με ένα καλό εστιατόριο κοντά στο cinema πριν το έργο, ή κάποια διαφήμιση για κάποιο κοντινό κατάστημα πωλήσεων. Αυτή η ανάλυση των δεδομένων αποφέρει τεράστια κέρδη στις εταιρίες και τους οργανισμούς που έχουν επενδύσει στην real-time ανάλυση των Big Data και αυτός είναι ο λόγος που κάθε εταιρία κατέχει τεράστιο όγκο δεδομένων.

Τα ψηφιακά δεδομένα βρίσκονται παντού πλέον, διέπουν την καθημερινότητά μας και αποτελούν σημαντικό, μερικές φορές και αναπόσπαστο κομμάτι των επιχειρήσεων, των χρηστών ψηφιακής τεχνολογίας και γενικότερα της οικονομίας. Αν και αυτό το θέμα κάποτε θα αφορούσε μόνο λίγους εξειδικευμένους επιστήμονες, στις μέρες τα Big Data έχουν πρωταγωνιστικό ρόλο σε πάρα πολλούς τομείς και οι καταναλωτές των προϊόντων και υπηρεσιών θα ωφεληθούν πάρα πολύ από την υλοποίηση τους. Η ικανότητα που έχουμε στις μέρες μας για αποθήκευση, συμψηφισμό και συνδυασμό στοιχείων για να εκτελέσουμε πολύπλοκες αναλύσεις ογκωδέστατων αρχείων, έχει γίνει ευκολότερη από την ανάπτυξη τεχνολογιών όπως το cloud computing, τα οποία μας παρέχουν αυτήν την δυνατότητα. Για λιγότερο από 1000 euro, ο καθένας μας μπορεί να αποθηκεύσει πλέον σε έναν καλό σκληρό δίσκο την μουσική ολόκληρου του κόσμου. Τα μέσα που μας παρέχονται για να εξάγουμε την γνώση από τα αρχεία, βοηθιούνται αρκετά και από το όλο και καλύτερο software και τις νέες εξελιγμένες τεχνικές που δημιουργούνται πάνω στην ανάλυση δεδομένων, καθώς επίσης και από την ολοένα και αυξανόμενη υπολογιστική δύναμη που μπορεί πλέον να έχει ο καθένας μας, ακόμα και μέσα στο σπίτι του.

Βέβαια η ικανότητά μας για να μπορούμε ακόμα και να χρησιμοποιούμε την έννοια των Big Data προήλθε από το γεγονός ότι τα δεδομένα πλέον παράγονται σε εξωφρενικά γρήγορο βαθμό. Σε έρευνα που πραγματοποιήθηκε το 2014, περισσότεροι από 4 δισεκατομμύρια άνθρωποι

ή το 60% του παγκόσμιου πληθυσμού, χρησιμοποιούσαν κινητό τηλέφωνο, εκ των οποίων το 44% από αυτά ήταν smart phone και υπολογίζεται ότι το ποσοστό τους θα αυξάνεται 20% κάθε χρόνο. Επιπλέον υπάρχουν περισσότεροι από 30 εκατομμύρια δικτυωμένοι κόμβοι αισθητήρων που παρέχουν μεγάλες ποσότητες δεδομένων σε πάρα πολλούς κλάδους όπως των μεταφορών, της αυτοκινητοβιομηχανίας, των πωλήσεις κλπ. Ο αριθμός των αισθητήρων υπολογίζεται ότι αυξάνεται με ρυθμό 30% κάθε χρόνο. (Gantz, 2007)

Όταν οι εταιρίες μπορούν να αναλύουν αποτελεσματικά τα δεδομένα που παράγονται, τότε είναι σε θέση να κατανοούν καλύτερα την αποδοτικότητα της επιχείρησης τους όπως επίσης και τους πελάτες τους, τους ανταγωνιστές τους, τα προϊόντα που πουλάνε κτλπ., το οποίο μπορεί να οδηγήσει σε βελτίωση της αποδοτικότητας, σε αυξημένες πωλήσεις, σε μείωση του κόστους και την καλύτερη εξυπηρέτηση των πελατών. Ας δώσουμε κάποια παραδείγματα για να γίνει ποιο κατανοητή η χρησιμότητα τους:

- Κατασκευαστικές εταιρίες αναπτύσσουν αισθητήρες τους οποίος ενσωματώνουν στα προϊόντα τους, και οι οποίοι παρέχουν real-time αποτελέσματα από τις καταναλωτικές συνήθειες των πελατών μέχρι και τα μέρη που βγαίνουν το βράδυ. Έτσι αναλύοντας αυτά τα δεδομένα προκύπτουν πρότυπα αγορών και ενδιαφερόντων για τον κάθε άνθρωπο ξεχωριστά και αυτό προσφέρει στις εταιρίες την ευκαιρία για βελτίωση των προϊόντων.

- Ο πολλαπλασιασμός των smart phones που περιέχουν ενσωματωμένα GPS προσφέρει την δυνατότητα στους διαφημιστές να στοχεύσουν τους καταναλωτές όταν βρίσκονται κοντά σε ένα εστιατόριο ή κατάστημα που ενδέχεται να τους αρέσει βάση της ανάλυσης των δεδομένων από προηγούμενες επιλογές. Αυτό ανοίγει νέα πηγή εσόδων για τους παρόχους τέτοιων υπηρεσιών και προσφέρει παράλληλα σε πολλές επιχειρήσεις την ευκαιρία να προσελκύσουν νέους πελάτες.

- Οι πωλητές συνήθως ξέρουν ποιοι αγόρασαν τα προϊόντα τους. Η χρήση των social media αλλά και τα αρχεία που αποθηκεύονται στο διαδίκτυο από την χρήση του κάθε καταναλωτή τους δίνει την δυνατότητα να καταλάβουν ποιος δεν αγόρασε

έναν προϊόν και γιατί. Αυτό επιτρέπει την δημιουργία στοχευμένου marketing και σε παράλληλη βελτίωση των προϊόντων τους.

Πολλοί επιστήμονες έχουν καταλήξει στο συμπέρασμα ότι τα Big Data θα προκαλέσουν μια έκρηξη ανάλογη με το Internet το 1990, και ότι ακόμα βρισκόμαστε στο 10% στο να κατανοήσουμε την δύναμη τους και τον αντίκτυπο που θα έχουν στην παγκόσμια οικονομία. Θα αλλάξει δραματικά ο τρόπος με τον οποίων θα λειτουργούν οι επιχειρήσεις και θα επηρεάσουν σε μεγάλο βαθμό την καθημερινότητα των ανθρώπων.

ΚΕΦΑΛΑΙΟ $2^{ο}$: ΠΩΣ ΤΑ BIG DATA ΑΛΛΑΖΟΥΝ ΤΟΝ ΚΟΣΜΟ ΜΑΣ

Στις μέρες μας, όλοι οι μεγάλοι τομείς της οικονομίας έχουν αρχίσει να εντάσσουν την επεξεργασία των δεδομένων στην καθημερινότητά τους καθώς αυτό τους αποφέρει χρόνο και χρήμα. Υπάρχουν αρκετές τεχνικές που αν εφαρμοστούν θα αποφέρουν μεγάλα κέρδη σε εταιρίες και οργανισμούς και θα αλλάξουν τον τρόπο με τον οποίο μια επιχείρηση θα σχεδιάζεται.

Αρχικά, θα πρέπει μια εταιρία ή ένας οργανισμός να δίνει τα δεδομένα που παράγονται ευκολότερα στους ενδιαφερόμενους φορείς χωρίς χρονοτριβές, γιατί αυτό μπορεί να δημιουργήσει τεράστια αξία. Για παράδειγμα αν σε μια επιχείρηση τα δεδομένα διαμοιράζονταν εγκαίρως σε όλα τα τμήματα της επιχείρησης ανεξαρτήτως τμήματος (marketing, IT, logistics κλπ.) τότε αυτό θα μείωνε δραστικά τον χρόνο που θα σπαταλούσαν οι εργαζόμενοι για την αναζήτηση και την επεξεργασία των δεδομένων.

Επιπλέον, καθώς οι εταιρίες αποθηκεύουν και διαχειρίζονται όλο και περισσότερα δεδομένα σε ψηφιακή μορφή, αυτό τους επιτρέπει μέσω της επεξεργασίας τους να έχουν μια ποιο ακριβή και λεπτομερή εικόνα της απόδοσης του εκάστοτε προϊόντος ή υπηρεσίας σε πραγματικό χρόνο. Το IT κάθε εταιρίας αναλύει την μεταβλητότητα της απόδοσης, και προσπαθεί μέσω διαφόρων πειραμάτων και τεχνικών να παρουσιάσει τα αίτια αυτής της μεταβλητότητας. Αυτό επιτρέπει στα ανώτερη στελέχη να διαχειρίζονται καλύτερα και να αυξάνουν τις επιδόσεις των προϊόντων ή των υπηρεσιών τους αντίστοιχα.

Ένας ακόμη τρόπος για να επωφεληθούν οι εταιρίες από την ανάλυση των Big Data είναι η τμηματοποίηση του πληθυσμού ανάλογα με την γεωγραφική περιοχή, τα ήθη και τα έθιμα των ανθρώπων σε κάθε χώρα ή σε κάθε πόλη ειδικότερα, ο μέσος όρος ηλικίας ανά πόλη κλπ. Τα Big Data επιτρέπουν στους οργανισμούς να δημιουργήσουν στοχευμένο marketing και στοχευμένες διαφημίσεις συλλέγοντας τα παραπάνω δεδομένα για να τροποποιήσουν τα προϊόντα τους ανάλογα με την ζήτηση που υπάρχει ανά γεωγραφική περιοχή. Ακόμα και εταιρίες που ήδη χρησιμοποιούσαν αυτήν την μέθοδο, εισάγουν την επιστήμη των Big Data, καθώς αυτά προσφέρουν ακόμα καλύτερα αποτελέσματα στις εταιρίες.

Ένα άλλος τρόπος που έχει δεχτεί κριτική κυρίως επειδή αντικαθιστά τις ανθρώπινες αποφάσεις και κρίσεις, αλλά παράγει πολύ καλύτερα αποτελέσματα, είναι η αυτοματοποιημένη

λήψη αποφάσεων πάνω στα αποτελέσματα συγκεκριμένων αλγορίθμων. Οι εταιρίες είτε έχουν αντικαταστήσει τελείως την ανθρώπινη απόφαση από αυτοματοποιημένους αλγορίθμους είτε κάποιες άλλες, χρησιμοποιούν αυτούς τους αλγορίθμους για να ενισχύσουν μια ανθρώπινη απόφαση. Ο λόγος που έχει συμβεί αυτό είναι ότι με την βαθύτερη, γρηγορότερη και ποιο εύστοχη ανάλυση που παρέχουν οι αλγόριθμοι, βελτιώνεται σημαντικά η ικανότητα των εκτελεστικών διευθυντών για λήψη σημαντικών αποφάσεων ως προς την στρατηγική που θα ακολουθήσει ένας οργανισμός, και ταυτόχρονα αυτό συμβαίνει ελαχιστοποιώντας τους κινδύνους και τα ρίσκα που βρίσκονται σε κάθε απόφαση. Τέτοιες υλοποιήσεις εφαρμόζονται ή μπορούν να εφαρμοστούν σε φοροτεχνικούς οργανισμούς όπου χρησιμοποιώντας αλγορίθμους, μπορούν να βρουν γρηγορότερα και αποτελεσματικότερα τους παραβάτες, ή σε εταιρίες τύπου amazon ή γενικότερα σε οποιαδήποτε εταιρία διαθέτει τα προϊόντα της προς πώληση όπου διαμορφώνουν τις τιμές τους βάση της real-time απογραφής των διαθέσιμων προϊόντων, καθώς και της real-time ζήτησης που έχουν. Είναι απολύτως λογικό οι εταιρίες να προτιμούν αποφάσεις που πάρθηκαν από λεπτομερή ανάλυση των Big Data παρά από τις χρονοβόρες αναλύσεις των ανθρώπων σε φύλλα χαρτιού. Η λήψη των αποφάσεων, ιδιαίτερα όσο οδεύουμε προς το μέλλον, δεν θα είναι ποτέ ξανά η ίδια. Μερικές επιχειρήσεις ήδη έχουν τεράστια οφέλη αναλύοντας τεράστια datasets από πελάτες, υπαλλήλους ή ακόμη και από σένσορες ενσωματωμένους στα προιόντα.

Τα Big Data επιτρέπουν στις εταιρίες να δημιουργήσουν νέα προϊόντα και υπηρεσίες, να ενισχύσουν τις υπάρχουσες, ή και να δημιουργούν εντελώς νέα επιχειρηματικά μοντέλα. Στις κατασκευαστικές εταιρίες χρησιμοποιούν τα δεδομένα από την χρήση των τωρινών προϊόντων τους για την ανάπτυξη της επόμενης γενιάς προϊόντων και την δημιουργία καινοτόμων υπηρεσιών μετά την πώληση των προϊόντων. Επιπλέον η εμφάνιση και η ανάλυση των real-time δεδομένων έχει δημιουργήσει ένα νέο σύνολο υπηρεσιών βάση της τοποθεσίας που βρίσκεται ο κάθε άνθρωπος, από πλοήγηση έως την τιμή της ασφάλισης των αυτοκινήτων με βάση το πως οι άνθρωποι οδηγούν τα αυτοκίνητα τους ανά γεωγραφική περιοχή. (desjardins, 2017)

2.1) ΤΑ BIG DATA ΕΛΕΓΧΟΥΝ ΤΟ ΜΕΛΛΟΝ

Αναμφίβολα, βάση μελετών και ερευνών είναι κοινώς αποδεκτό πλέον ότι βρισκόμαστε στην γενιά των δεδομένων και της πληροφορίας. Τα Big Data αυξάνονται εκθετικά. Αυτή η ανάπτυξη όμως των δεδομένων συμβαίνει μόνο σε συγκεκριμένους τομείς της οικονομίας?

Η απάντηση είναι όχι. Η ανάπτυξη των Big Data είναι ένα φαινόμενο που παρουσιάζεται σε όλους τους τομείς της οικονομίας και κατ' επέκταση τους επηρεάζει άμεσα. Ποιο συγκεκριμένα ο μέσος αριθμός δεδομένων που αποθηκεύεται ανά εταιρία, σε όλους τους τομείς της οικονομίας είναι αρκετός για να χρησιμοποιηθούν τεχνικές αναλύσεις για Big Data, σε άλλους τομείς της οικονομίας περισσότερο σε άλλους λιγότερο, πάντως δεν υπάρχει κάποιος τομέας που δεν χρειάζεται ή δεν θα χρειαστεί στο μέλλον Data Scientists-Analysts. Οι επιχειρηματικοί ηγέτες αρχίσουν να προσαρμόζονται στην εποχή των Big Data και προσπαθούν να βρουν τρόπους να πάρουν όσο γίνεται καλύτερα αποτελέσματα από την αξιοποίησή τους.

ΕΙΚΟΝΑ 3) <u>Η εκθετική αύξηση των Data</u>

Όπως είναι προφανές σε κάποιους κλάδους υπάρχει μεγαλύτερη ανάγκη για επεξεργασία των δεδομένων καθώς επεξεργάζονται μεγαλύτερους όγκους από αυτά. Αυτό σημαίνει ότι για αυτούς τους κλάδους είναι ακόμη σημαντικότερη η άμεση αξιοποίηση των δεδομένων που διαθέτουν καθώς θα τους αποφέρει άμεσα κέρδη. Για παράδειγμα οι οικονομικοί κλάδοι, είτε πρόκειται για επενδυτικά funds είτε για τράπεζες, διαθέτουν τα περισσότερα ψηφιακά δεδομένα προς αξιοποίηση σε σχέση με τους υπόλοιπους κλάδους. Τα δεδομένα που δέχονται ανά

δευτερόλεπτα είναι τεράστια (real-time συναλλαγές στο χρηματιστήριο, οικονομικές συναλλαγές κλπ.) και χρειάζονται άμεση ανάλυση καθώς τα κέρδη από τις εκάστοτε αποφάσεις μπορούν να αλλάξουν ανά λεπτό που περνάει. Άλλοι τομείς που επηρεάζονται άμεσα είναι οι επικοινωνίες και τα ΜΜΕ καθώς αυτές οι υπηρεσίες λαμβάνουν υψηλό όγκο δεδομένων από την τηλεπικοινωνιακή κίνηση αλλά και από τα social media.

Επιπλέον τα δεδομένα που δέχεται ο κάθε τομέας είναι διαφορετικά και χρειάζονται διαφορετική ανάλυση. Για παράδειγμα στους οικονομικούς κλάδους ή σε κλάδους πωλήσεων παράγονται κυρίως αριθμητικά δεδομένα που αφορούν τις συναλλαγές με τους πελάτες, μαθηματικά μοντέλα πωλήσεων κλπ. Στον τομέα της υγείας τα δεδομένα που παράγονται είναι εικόνες από ακτινογραφίες ή βίντεο από χειρουργικές επεμβάσεις. Στις επικοινωνίες και στα ΜΜΕ παράγονται κυρίως HD φωτογραφίες και ακουστικό υλικό. Όλες αυτές οι διαφοροποιήσεις των δεδομένων ανάλογα των τομέα απαιτεί την προσαρμογή των Data Analysts ανάλογα με τα δεδομένα που δέχονται ώστε να επιτυγχάνεται το καλύτερο δυνατό αποτέλεσμα. Δεν υπάρχει δηλαδή ένας 'Big Data αλγόριθμος' που εφαρμόζεται παντού. Γι' αυτό και υπάρχει τόσο μεγάλη και αυξανόμενη ανάγκη για περισσότερους αναλυτές και επιστήμονες στον χώρο των Big Data. Αυτήν την χρονική στιγμή το 70% των παγκόσμιων δεδομένων αναλύεται και αποθηκεύεται σε Η.Π.Α. και Ευρώπη, αλλά θα πρέπει να έχουμε υπόψιν μας ότι πολλά από τα δεδομένα μπορούν να παράγονται σε μια χώρα και να αποθηκεύονται και να αναλύονται σε μια άλλη.

2.1.1) <u>Η εκθετική ανάπτυξη των Big Data</u>

Σε όλο τον παγκοσμιοποιημένο-διασυνδεδεμένο κόσμο, διάφορες τάσεις και μόδες θα συνεχίσουν να αυξάνουν εκθετικά τα δεδομένα που παράγονται, αποθηκεύονται και αναλύονται. Αυτές οι τάσεις περιλαμβάνουν τις παραδοσιακές συναλλαγές, την συνεχή χρήση και ανάπτυξη στον χώρο των πολυμέσων, την ολοένα και αυξανόμενη χρήση των social media, όπως και επίσης την ενσωμάτωση στην καθημερινότητα μας χιλιάδων αισθητήρων στον νέο κόσμο του IoT. Οι επιχειρήσεις πλέον δεν αφήνουν κανενός είδους πληροφορία να πάει χαμένη, κάθε συναλλαγή με πελάτες κάθε σχόλιο στα social media, κάθε νέα τάση που γίνεται viral, καταγράφεται, αποθηκεύεται και αναλύεται. Σκοπός ιδιαίτερα των μεγάλων πολυεθνικών που δραστηριοποιούνται ανά τον κόσμο σε τελείως διαφορετικές κουλτούρες και σε διαφορετικά ήθη και έθιμα, είναι η συλλογή προσωπικών δεδομένων και προσωπικών συνηθειών σε διαφορετικά

γεωγραφικά περιβάλλοντα. Αυτό αυξάνει την ανάγκη για περισσότερο διαθέσιμο χώρο αποθήκευσης και αναλυτική ικανότητα στην επεξεργασία των δεδομένων. Για παράδειγμα η amazon αυτήν την στιγμή διαχειρίζεται περισσότερα από 5 Petabytes δεδομένων και ο αριθμός αυτός αυξάνεται δραματικά κάθε χρόνο.

Η αυξανόμενη χρήση των πολυμέσων σε κλάδους όπως της υγείας και των βιομηχανιών πώλησης προϊόντων ή υπηρεσιών, συνέβαλε σημαντικά στην ανάπτυξη των Big Data και θα συνεχίσει με ακόμα μεγαλύτερη ένταση στο μέλλον. Τα βίντεο δημιουργούν έναν τεράστιο όγκο δεδομένων. Στην ιατρική για παράδειγμα, τα High Definition βίντεο στις χειρουργικές επεμβάσεις παράγουν 25 φορές περισσότερα δεδομένα ανά λεπτό ακόμα και από την υψηλότερης ανάλυσης εικόνα από τον αξονικό τομογράφο. Για να καταλάβετε, η κάθε εικόνα από την αξονική τομογραφία απαιτεί πολλές χιλιάδες περισσότερα bytes από αριθμητικά δεδομένα ή από μια σελίδα κειμένου. Περισσότερα από το 95% των δεδομένων που παράγεται στον κλάδο της υγειονομικής περίθαλψης είναι βίντεο. Δεδομένα που προέρχονται από οπτικοακουστικό υλικό αποτελούν περισσότερο από το 50% της κίνησης που δημιουργείται στο internet και αυτό το ποσοστό αναμένεται να αυξηθεί. Επιπλέον η ολοένα και μεγαλύτερη αύξηση στην χρήση των social media δημιουργεί τεράστιες ποσότητες δεδομένων. Παλιότερα η πλειοψηφία των χρηστών αφορούσε τους νέους ανθρώπους αλλά χρόνο με τον χρόνο, περισσότεροι άνθρωποι που ανήκουν στις ηλικίες 40+ δημιούργησαν λογαριασμούς στα social media. Για να γίνει κατανοητό το μέγεθος των δεδομένων που παράγεται από τα social media, οι 700 εκατομμύρια ενεργοί χρήστες του Facebook ξοδεύουν περισσότερες από 9.3 δισεκατομμύρια ώρες τον μήνα!!, στο site όπου ο μέσος χρήστης δημιουργεί 90 κομμάτια περιεχομένου είτε αυτά είναι φωτογραφίες, βίντεο, μουσική ή κείμενο οπότε είναι ευκόλως κατανοητό ότι δημιουργείται καθημερινά ένα tsunami δεδομένων.

Η χρησιμότητα των Big Data έχει ήδη ωθήσει πολλές εταιρίες να φτιάξουν ειδικό τμήμα ανάλυσης των Big Data και έχουν καταφέρει να έχουν θεαματικά αποτελέσματα. Για παράδειγμα, τα ΙΚΕΑ διαχειρίζονται ένα τεράστιο αριθμό δεδομένων από συναλλαγές και προτιμήσεις των πελατών που χωρίζεται ανά χώρα, πόλη ακόμα και ηλικία. Ο σκοπός της εταιρίας είναι να φτιάξει ένα προφίλ για στοχευμένο marketing αντί για ένα ενιαίο και γενικότερο, όπως για παράδειγμα προώθηση συγκεκριμένων προϊόντων ανάλογα με τις προτιμήσεις που έχουν δείξει να έχουν οι πελάτες ανά γεωγραφική περιοχή. Αυτή η εφαρμογή τους έχει αποφέρει

τεράστια κέρδη και μάλιστα θα ενισχύσουν την ροή των δεδομένων που δέχονται με σένσορες που θα εγκατασταθούν σε περιοχές των προϊόντων τους και θα παρέχουν real-time στοιχεία για τους πελάτες. Επιπλέον εταιρίες όπως η Amazon, ή το IMDB, εφαρμόζουν αλγορίθμους που ανάλογα με το προϊόν ή την ταινία αντίστοιχα στην οποία έχει προτίμηση ο κάθε χρήστης, αυτοί παρέχουν συγκεκριμένες προτάσεις όπως 'ο χρήστης που αγόρασε το χ προϊόν αγόρασε και το…'. Η American express χρησιμοποιεί τα Big Data για να αναλύσει και να προβλέψει την συμπεριφορά των πελατών. Βλέποντας τις ιστορικές συναλλαγές και ενσωματώνοντας στους αλγορίθμους πάνω από 100 μεταβλητές, η εταιρία παράγει ένα πρωτοποριακό μοντέλο που της επιτρέπει να έχει μια καλή αντίληψη πάνω στους πελάτες της. Η εταιρία υποστηρίζει ότι στην Αυστραλία μπορούν να προβλέψουν το 25% των λογαριασμών που θα κλείσουν μέσα σε τέσσερις μήνες. Το Miniclip που είναι site όπου μπορούν όσοι ασχολούνται με την δημιουργία παιχνιδιών να τα ανεβάζουν και να γίνονται διαθέσιμα σε όλους, χρησιμοποιεί τα Big Data για την παρακολούθηση και την βελτίωση της εμπειρίας του χρήστη. Ο πρωταρχικός τους σκοπός είναι να διατηρήσουν την μάζα των gamers που διαθέτουν καθώς έτσι θα παράγουν περισσότερα λεφτά και θα είναι κίνητρο για τους game designers να προτιμούν αυτό το site για τα παιχνίδια που αναπτύσσουν. Έτσι, μέσω πειραμάτων και χρήση του machine learning (θα αναληθεί στο επόμενο κεφάλαιο), η εταιρία μπορεί να βελτιώσει την εμπειρία για τους χρήστες και να παρέχει μια ποιο ελκυστική εμπειρία. Τέλος, αν έχετε ποτέ αναρωτηθεί πως τα γνωστά σε όλους μας Starbucks μπορούν να ανοίξουν 3 καταστήματα στον ίδιο δρόμο χωρίς να υπάρχει επίπτωση στον τζίρο τους, αυτό οφείλεται στην χρήση των Big Data που σε μεγάλο ποσοστό προβλέπει την πιθανή επιτυχία των νέων καταστημάτων, παίρνοντας ως παραμέτρους την κίνηση, την περιοχή, την συμπεριφορά των πελατών και πολλά ακόμη. (O'Neill, 2016)

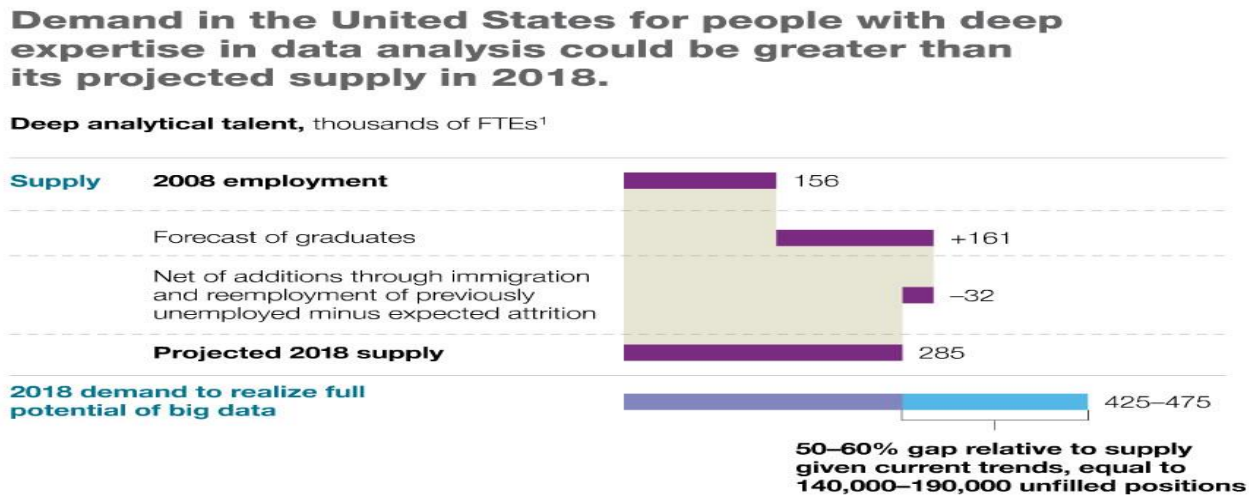
2.2) <u>ΠΡΟΚΛΗΣΕΙΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ</u>

Όπως σε κάθε νέα τεχνολογία και γενικότερα σε κάθε άλμα της τεχνολογίας ανεξαρτήτως κλάδου ή χρησιμότητας, μαζί με τα πολλά θετικά έρχονται και νέες προκλήσεις και εμπόδια που πρέπει να αντιμετωπιστούν.

2.2.1) Έλλειψη σε ικανό προσωπικό

Μία από αυτές τις προκλήσεις που επηρεάζουν άμεσα τις εταιρίες ή τους οργανισμούς είναι η έλλειψη επιστημόνων που ασχολούνται με τα Big Data. Αυτό δεν επιτρέπει στις επιχειρήσεις να έχουν το μέγιστο δυνατό κέρδος από την ανάλυση και την διαχείριση τους γι' αυτό και είναι μια σημαντική και επείγουσα πρόκληση που πρέπει να αντιμετωπιστεί στο άμεσο μέλλον. Συγκεκριμένα στις Η.Π.Α. έχει υπολογιστεί ότι το 2018 θα υπάρχουν 200.000 κενές θέσεις εργασίας με σημαντικές αναλυτικές ικανότητες καθώς και 1,5 εκατομμύρια κενές θέσεις από αναλυτές και διαχειριστές των αποτελεσμάτων που προκύπτουν από την ανάλυση των Big Data. Επιπλέον το να γίνεις ένας ικανός αναλυτής στον χώρο των Big Data απαιτεί χρόνια εκπαίδευσης ειδικά άμα ο εργαζόμενος προέρχεται από ένα καθαρά μαθηματικό περιβάλλον χωρίς καμία εμπειρία σε προγραμματισμό. Τα ανώτερα στελέχη πολλών επιχειρήσεων αδυνατούν να κατανοήσουν ή δεν έχουν κατανοήσει πλήρως την αξία των Big Data και πως να την εκμεταλλευτούν. Αυτό μπορεί να κρίνει ακόμα και την βιωσιμότητα τις επιχείρησης στο άμεσο μέλλον, καθώς εταιρίες που έχουν ξεκινήσει να ξεκλειδώνουν την δύναμη των Big Data, έχουν προβάδισμα έναντι ανταγωνιστικών εταιριών. Και όπως ανέφερα και ποιο πάνω, πολλές εταιρίες δεν έχουν ή δεν βρίσκουν τους εργαζόμενους με τις ικανότητες που χρειάζονται για την αξιοποίηση των Big Data.

ΕΙΚΟΝΑ 4) Το 2018, οι θέσεις εργασίας θα είναι περισσότερες από το διαθέσιμο εργατικό δυναμικό



[1] Deep analytical talent are people who have advanced training with statistics or machine learning.
FTE = full-time equivalent.
Source: Dun & Bradstreet; company interviews; US Bureau of Labor Statistics; US Census Bureau; McKinsey Global Institute analysis

2.2.2) Προστασία των προσωπικών δεδομένων

Ταυτόχρονα με την άνθιση των Big Data υπήρξαν πολλές αντιδράσεις και πολλές δημοσιεύσεις για την καταπάτηση της ιδιωτικότητας των ανθρώπων. Καθώς ο όγκος των δεδομένων που ψηφιοποιείται αυξάνεται δραματικά και μεταδίδεται οπουδήποτε στον κόσμο, δημιουργείται μια σειρά ζητημάτων τα οποία πρέπει να αντιμετωπιστούν, όπως στην προστασία της ιδιωτικότητας, στην ασφάλεια, στην πνευματική ιδιοκτησία κλπ. Σίγουρα η προστασία της ιδιωτικότητας είναι ένα θέμα του οποίου η σημασία, ιδιαίτερα για τους καταναλωτές, αυξάνεται παράλληλα με την άνθιση των Big Data. Προσωπικά δεδομένα όπως η υγεία και οι οικονομίες που διαθέτει ο κάθε άνθρωπος μπορούν αν αξιοποιηθούν σωστά να προσφέρουν σημαντικά οφέλη στην ζωή του κάθε ανθρώπου όπως συγκεκριμένη θεραπευτική αγωγή, εξειδικευμένη στον κάθε ασθενή ή την προώθηση των κατάλληλων προϊόντων ανάλογα με τις οικονομικές τους δυνατότητες. Ωστόσο, πολλοί άνθρωποι βλέπουν αυτές τις κατηγορίες ως τις ποιο ευαίσθητες και προσωπικές για την ζωή του κάθε ανθρώπου και πιστεύουν ότι αυτές πρέπει να μένουν ιδιωτικές και είναι ανήθικο να τις εκμεταλλεύονται. Γίνεται λοιπόν σαφές ότι στις μελλοντικές κοινωνίες όπου τα Big Data θα έχουν περίοπτη θέση, η γραμμή μεταξύ προστασίας δεδομένων και χρησιμότητας από την ανάλυση τους θα είναι πολύ λεπτή και μερικές φορές μη διακριτή.

Ένα άλλο θέμα που πρέπει να αντιμετωπιστεί και έχει σχέση με τα προσωπικά δεδομένα είναι η ασφάλεια των δεδομένων δηλαδή πως θα προστατεύονται δεδομένα εταιριών και ανθρώπων που αφορούν κάποια εφεύρεση ή κάποια καινοτομία. Τα τελευταία χρόνια έχουμε παρατηρήσει πολλά παραδείγματα όπου οι παραβιάσεις των δεδομένων δεν αφορούσαν μόνο προσωπικά δεδομένα καταναλωτών, αλλά και εμπιστευτικές εταιρικές πληροφορίες και ακόμα και απόρρητα έγγραφα και πληροφορίες για την ασφάλεια κρατών. Και καθώς αυτές οι υποκλοπές δεδομένων αυξάνονται όσο περνάει ο καιρός, είναι αναγκαία η ασφάλεια των δεδομένων είτε μέσω της τεχνολογίας, είτε θεσπίζοντας νόμους που θα τα θέτουν υπό αυστηρότερη προστασία.

Η αυξανόμενη οικονομική σημασία που προσδίδουν τα Big Data στις επιχειρήσεις έρχεται να αντιμετωπίσει μια σειρά από νομικά ζητήματα όπως το πότε θεωρείται νόμιμο να έχει κάποιος στοιχεία η κάποια λίστα από στοιχεία για μια μάζα ανθρώπων ή μια άλλη εταιρία?, ή πότε θεωρείται ότι έχει ξεπεραστεί το όριο της νομιμότητας και ποια κριτήρια ισχύουν?. Επίσης αν μια ανάλυση σε ένα dataset των Big Data οδηγήσει σε μη θεμιτά αποτελέσματα, ποιος θα έχει την

ευθύνη γι'αυτό?. Τέτοιου είδους ζητήματα χρειάζονται διευκρίνηση ώστε να μπορούν και τα προσωπικά δεδομένα να προστατευτούν και παράλληλα να εξάγεται το μέγιστο δυνατό κέρδος από την επεξεργασία τους.

Σε μια έρευνα που έγινε το 2011 στην Νέα Υόρκη, πολυεθνικές λιανικής πώλησης, κατέληξαν στο συμπέρασμα ότι οι καταναλωτές ψωνίζουν περισσότερο στο διάστημα μετά την γέννηση ενός παιδιού. Επειδή η γέννηση ενός παιδιού γίνεται δημοσίως γνωστή, οι εταιρίες προσπαθούν να μάθουν πότε οι καταναλωτές τους αποκτούν παιδί και έτσι να στοχεύσουν το marketing στους νέους γονείς. Έτσι προσέλαβαν στατιστικολόγους και αναλυτές, που κατέληξαν στο συμπέρασμα μέσα από την ανάλυση προηγούμενων συναλλαγών ότι η πλειοψηφία των νέων γονιών αγοράζει βιταμίνες, τσάντες αρκετά μεγάλες ώστε να χωράνε και οι πάνες και πολλά άλλα. Έχοντας λοιπόν αποκομίσει την συγκεκριμένη πληροφορία, στέλνανε στις μητέρες κουπόνια με εκπτώσεις στα συγκεκριμένα προϊόντα και κατάφεραν να εκτοξεύσουν τις συγκεκριμένες πωλήσεις. Αυτό ήταν απλά ένα παράδειγμα πως οι εταιρίες ή οι κυβερνήσεις μπορούν να εκμεταλλευτούν γεγονότα στην ζωή ενός ανθρώπου και να επηρεάσουν τις μελλοντικές τους αποφάσεις είτε αυτό αφορά στην πώληση προϊόντων, είτε ακόμη και στην διαμόρφωση πολιτικής γνώμης μέσω στοχευμένης προπαγάνδας. Αυτά τα γεγονότα έχουν προκαλέσει θύελλα αντιδράσεων και πολλές χώρες όπως το Ισραήλ, η Ελβετία και η Ιαπωνία έχουν θεσπίσει νόμους όπου η προστασία των προσωπικών δεδομένων προστατεύεται ως ανθρώπινο δικαίωμα και υπάρχουν σοβαρές ποινικές ευθύνες σε ανθρώπους, εταιρίες ή οργανισμούς που προσπαθούν να τα εκμεταλλευτούν. (Theresa Payton, 2014)

ΕΙΚΟΝΑ 5) <u>Big Data παντού!</u>

2.2.3) <u>Τεχνολογία και πρόσβαση στα Big Data</u>

Όπως είναι προφανές, για να αξιοποιηθούν πλήρως τα Big Data και να αποφέρουν τα μεγαλύτερα κέρδη στις εταιρίες που θα τα χρησιμοποιήσουν, θα πρέπει να υπάρχει η τεχνολογική ικανότητα για γρήγορή και αποτελεσματική προσπέλαση και ανάλυση και επίσης να έχουν οι εταιρίες πρόσβαση σε ικανοποιητικό αριθμό δεδομένων αλλιώς οι αλγόριθμοι θα τους οδηγήσουν σε λάθος συμπεράσματα.

Για να μην υπάρχουν λανθασμένες εκτιμήσεις θα πρέπει οι ενδιαφερόμενες εταιρίες ή οργανισμοί να συλλέγουν δεδομένα από πολλές διαφορετικές πηγές. Για να επιτευχθεί αυτό, μερικές φορές οι επιχειρήσεις αγοράζουν δεδομένα ώστε να καλύψουν κενά που προκύπτουν άμα δεν υπάρχουν επαρκείς δεδομένα για την ασφαλέστερη εξαγωγή συμπερασμάτων. Ωστόσο, δεν είναι πάντα εύκολο να αποκτήσει μια εταιρία πρόσβαση σε δεδομένα τρίτων. Πολλές φορές οι εταιρίες που έχουν αυτά τα δεδομένα ζητάνε πολλά λεφτά για να τα μοιραστούν, και μερικές φορές

δεν θέλουν να τα μοιραστούν ειδικά αν πρόκειται για ανταγωνιστική εταιρία ή οργανισμό. Για παράδειγμα μια εταιρία που ασχολείται με τις επενδύσεις και έχει συλλέξει δεδομένα πολύ σημαντικά που η αξιοποίηση τους θα της αποφέρει τεράστιο κέρδος, δεν έχει κάποιο οικονομικό κίνητρο να τα πουλήσει γιατί θα χάσει το ανταγωνιστικό πλεονέκτημα που έχει έναντι άλλων εταιριών.

Βέβαια, μερικές φορές ακόμα και να έχει μια εταιρία όλα τα δεδομένα που χρειάζεται για να έχει κέρδη, χρειάζεται επίσης να έχει και την απαραίτητη τεχνολογία και τεχνογνωσία για να έχει αποτέλεσμα. Σε αυτόν τον τομέα, έχει επιτευχθεί τεράστια πρόοδος και ξοδεύεται τεράστιος αριθμός χρημάτων για την δημιουργία προγραμμάτων και υπηρεσιών που θα δίνουν την δυνατότητα για ποιο ακριβής και γρηγορότερη ανάλυση καθώς και περισσότερο αποθηκευτικό χώρο για τα δεδομένα. Επιπλέον, αν κάποιος οργανισμός δεν έχει συμβαδίσει με τις νέες τεχνολογίες που γεννιούνται συνέχεια, και έχει μείνει σε παλιά πρότυπα ανάλυσης, τότε ακόμα και το ποιο εξειδικευμένο προσωπικό δεν θα μπορεί να παράγει τα επιθυμητά αποτελέσματα. Οι αναβαθμίσεις στο τεχνολογικό υπόβαθρο πρέπει να είναι συνεχείς ώστε τα νέα προγράμματα ανάλυσης και αποθήκευσης να μπορούν να αξιοποιούν τα κέρδη από την ανάλυση των Big Data.

ΚΕΦΑΛΑΙΟ $3^{ο}$: ΤΕΧΝΙΚΕΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΤΩΝ BIG DATA

Όπως ανέφερα σε προηγούμενα κεφάλαια, η αύξηση των δεδομένων που παράγονται κάθε χρόνο είναι τεράστια και οι εταιρίες προσπαθούν να συλλέξουν όσα περισσότερα μπορούν για την καλύτερη και αποτελεσματικότερη ανάλυσή τους. Όμως αυτός ο όγκος των δεδομένων έφερε στο προσκήνιο αδυναμία στην ανάλυσή τους καθώς δεν υπήρχαν οι τεχνικές εκείνες που θα μπορούσαν να τα αναλύσουν αποτελεσματικά, ούτε η υπολογιστική ισχύς και ακόμη και ο διαθέσιμος αποθηκευτικός χώρος δεν επαρκούσε. Έτσι από την πρώτη δεκαετία της νέας χιλιετίας, πολλές εταιρίες ασχολήθηκαν με την δημιουργία νέων τεχνικών και τεχνολογιών που θα μπορούσαν να αντιμετωπίσουν αυτήν την πρόκληση που προκύπτει από την τεράστια αύξηση των διαθέσιμων δεδομένων.

Οι τεχνικές και οι τεχνολογίες που αναπτύχθηκαν και αναπτύσσονται έχουν ως σκοπό τους την συγκέντρωση των δεδομένων από διαφορετικές πηγές, την ανάλυση των δεδομένων όσο γίνεται ταχύτερα και αποτελεσματικότερα και τέλος την οπτικοποίηση τους και την εξαγωγή συμπερασμάτων προς υλοποίηση. Αυτές οι τεχνολογίες και τεχνικές αναπτύχθηκαν από διάφορους κλάδους όπως αυτούς της στατιστικής, της επιστήμης των υπολογιστών, των εφαρμοσμένων μαθηματικών και των οικονομικών. Από αυτό προκύπτει ότι μια εταιρία για να έχει το μέγιστο κέρδος από την ανάλυση των Big Data θα πρέπει να έχει εργαζομένους τόσο με προγραμματιστικές ικανότητες, όσο και με την ικανότητα να αντιλαμβάνονται στατιστικά και οικονομικά μοντέλα και όρους που θα προκύπτουν από την ανάλυση των datasets. Μερικές από αυτές τις τεχνικές και τεχνολογίες αναπτύχθηκαν από ακαδημαϊκούς, και άλλες από επιχειρήσεις που είχαν άμεσο συμφέρον από την εκμετάλλευση των Big Data, όπως εταιρίες με online συναλλαγές σαν την amazon. Κάποιες από αυτές τις τεχνικές και τις τεχνολογίες που αναπτύχθηκαν τα τελευταία χρόνια, αναπτύχθηκαν σε μια εποχή όπου υπήρχε πολύ μικρότερη πρόσβαση σε μεγάλες ποσότητες δεδομένων. Πολλές από αυτές όμως εξελίχθηκαν και έχουν προσαρμοστεί στα δεδομένα της εποχής με μεγάλη επιτυχία και μπορούν να εφαρμοστούν σε πολύ μεγάλα datasets.

Πολλά από τα εργαλεία που είναι πλέον στην διάθεση μας για να αναλύσουμε τα Big Data είναι εξειδικευμένα σε κλάδους και υπηρεσίες. Δηλαδή πολλές φορές οι τεχνικές πρέπει να παραμετροποιούνται ανάλογα με την εταιρία που τις χρησιμοποιεί και τον στόχο που έχει από την

ανάλυση των Big Data. Σε αυτό το κεφάλαιο θα αναλύσω μερικές τεχνικές και τεχνολογίες που εφαρμόζονται στις μέρες μας, αν και αυτές εξελίσσονται ή δημιουργούνται νέες που αντικαθιστούν τις υπάρχουσες καθώς γράφω αυτήν την εργασία. Η ιστορία των Big Data γράφεται στις μέρες μας και οι καινοτομίες σε νέες τεχνικές και τεχνολογίες εμφανίζονται συνεχώς. Ίσως μερικές από αυτές που περιγράφω τώρα να εφαρμόζονται μερικώς ή ακόμα και να έχουν πλήρως αντικατασταθεί ως το 2020 όπως συμβαίνει αυτήν την στιγμή με την Apache Spark που αντικαθιστά σιγά σιγά το Hadoop λόγω της γρηγορότερης ανάλυσης που προσφέρει σε τεραστίους όγκους δεδομένων (θα αναλυθούν και τα δύο στην συνέχεια). Στην συνέχεια θα αναλύσω τις σημαντικότερες τεχνικές και τεχνολογίες που υπάρχουν στις μέρες μας για την ανάλυση των Big Data.

ΕΙΚΟΝΑ 6) <u>Τα αποτελέσματα που παράγουν οι διάφορες τεχνικές ανάλυσης των Big Data</u>

Οι λόγοι που αναπτύχθηκαν τεχνικές και για την καλύτερη αξιοποίηση των Big Data είναι:

- Για την παραγωγή αξιόπιστων αποτελεσμάτων από την ανάλυση ογκωδέστατων datasets

- Για την γρηγορότερη ανάλυση τους

- Για την πρόβλεψη μοτίβων συμπεριφορών και την αποφυγή λανθασμένων στρατηγικών

Στην συνέχεια θα αναλύσω ορισμένες από αυτές τις τεχνικές που χρησιμοποιούνται στις μέρες μας. Πολλές από αυτές τις τεχνικές είναι αποτέλεσμα εξέλιξης προηγουμένων τεχνικών που χρησιμοποιούσαν παλιότερα οι εταιρίες όταν ο όγκος των δεδομένων ήταν μικρότερος, και προσαρμόστηκαν στις απαιτήσεις των Big Data. Καθώς γράφω αυτές τις γραμμές, ολοένα και περισσότερες και περισσότερο αξιόπιστες τεχνικές αναδύονται καθώς έχουν δοθεί πολλά κονδύλια για έρευνα στο συγκεκριμένο τομέα και νέες βελτιωμένες τεχνικές παράγονται συνεχώς. Οι περισσότερες από αυτές έχουν πολλά στοιχεία από την επιστήμη των υπολογιστών σε συνδυασμό με την επιστήμη της στατιστικής. Οι ποιο σημαντικές από αυτές είναι:

3.1) DATA MINING

Το Data Mining περιλαμβάνει μια σειρά από στατιστικά στοιχεία καθώς και την δυνατότητα να μάθει στον υπολογιστή να ξεχωρίζει μοτίβα μελετώντας υπάρχων datasets και μετά την ανάλυση μας παρέχει με πληροφορίες που θα ήταν αδύνατον να βρούνε οι εργαζόμενοι μόνοι τους χωρίς την ανάλυση. Στην ιδανική περίπτωση, το Data Mining προβλέπει μοτίβα πελατών και παρέχει πληροφορίες προς αξιοποίηση στις ενδιαφερόμενες εταιρίες. Ουσιαστικά αυτό που μας παρέχει σαν γνώση είναι όχι ‘ποια είναι η σχέση μεταξύ διαφημίσεων και πωλήσεων’ αλλά ‘ποια συγκεκριμένη διαφήμιση, ή συγκεκριμένο προϊόν πρέπει να δείξω σε έναν καταναλωτή που ψωνίζει στο διαδίκτυο εκείνη την στιγμή’. Ένα άλλο ενδιαφέρον στοιχείο είναι ότι πέρα από ατομικές προβλέψεις, κατηγοριοποιεί τους καταναλωτές ανάλογα με τις καταναλωτικές συναλλαγές που είχαν στο παρελθόν, και έχοντας αυτήν την πληροφορία μια εταιρία προσαρμόζει σε κάθε group καταναλωτών διαφορετική στρατηγική στο marketing.

Με το Data Mining εφαρμόζουμε μια τεχνική που χρησιμοποιούσαν ήδη οι οικονομολόγοι, οι στατιστικολόγοι, οι μετεωρολόγοι και αυτή αφορούσε την ιδέα ότι μοτίβα

δεδομένων μπορούν να προκύψουν, αν αναλυθούν τα δεδομένα. Το Data Mining μας δίνει την δυνατότητα αυτά τα μοτίβα να προκύπτουν από ανάλυση ογκωδέστατων datasets που μεγαλώνουν καθημερινά όπως οι καταναλωτικές συνήθειες των πελατών μιας αλυσίδας μαγαζιών. Αυτή η δυνατότητα να προκύπτουν μοτίβα πρόβλεψης συνηθειών των ανθρώπων βάζει το Data Mining στην πρώτη γραμμή επιλογής για εταιρίες πωλήσεων ή εταιρίες διαφημίσεων κυρίως αλλά μπορεί να εφαρμοστεί σχεδόν παντού. Έχει εκτιμηθεί ότι η ποσότητα των δεδομένων που αποθηκεύονται σε βάσεις δεδομένων σε όλο τον κόσμο, διπλασιάζεται κάθε 20 μήνες, οπότε γίνεται κατανοητό ότι σε αυτήν την τρομακτική αύξηση των δεδομένων, το Data Mining γίνεται αναπόσπαστο κομμάτι για την ανάλυση των Big Data. Το θετικό είναι ότι παρόλη την τεράστια αύξηση των δεδομένων, πλέον υπάρχουν ακόμα και κατ' οίκον τα μηχανήματα εκείνα τα οποία διαθέτουν την υπολογιστική ισχύ για ανάλυση και εξαγωγή αποτελεσμάτων, οπότε περισσότερος κόσμος έχει την δυνατότητα να ωφεληθεί από το Data Mining. Καθώς τα δεδομένα παράγονται σε εξωφρενικό ρυθμό και με μεγάλη πολυπλοκότητα, το Data Mining είναι το σημαντικότερο όπλο που διαθέτουμε για την ανακάλυψη και την κατανόηση κρυμμένων μοτίβων στον ωκεανό των νέων δεδομένων, τα οποία οδηγούν σε νέες επιχειρηματικές ιδέες και σε εμπορικά πλεονεκτήματα έναντι των ανταγωνιστών. Το Data Mining ουσιαστικά σχετίζεται με την επίλυση προβλημάτων από την ανάλυση δεδομένων που βρίσκονται ήδη σε βάσεις δεδομένων ανεξαρτήτου μεγέθους.

ΕΙΚΟΝΑ 7) <u>Η σειρά διαδικασιών στο data mining</u>

Για να δώσω ένα παράδειγμα και να γίνει ποιο κατανοητή η χρησιμότητα του, ας υποθέσουμε ότι μια εταιρία θέλει να μάθει κατά πόσο οι πελάτες της ψωνίζουν από την δικιά της αλυσίδα καταστημάτων και με τη συχνότητα ψωνίζουν από άλλες. Αφορά δηλαδή την πίστη που έχουν οι πελάτες της στο δικό της brand name σε μια άκρως ανταγωνιστική αγορά. Το κλειδί για την επίλυση του συγκεκριμένου προβλήματος βρίσκεται σε βάσεις δεδομένων που έχουν στοιχεία με τις προγενέστερες επιλογές του πελάτη σε συνδυασμό με τα προφίλ των πελατών. Μπορούμε να αναλύσουμε μοτίβα συμπεριφοράς πελατών που είχαν αγοράσει στο παρελθόν προϊόντα ή υπηρεσίες από την επιχείρηση μας, και μέσω της ανάλυσης να δούμε κατά πόσο προτιμήσαν να ξαναγοράσουν από την επιχείρηση μας ή αγόρασαν προϊόντα ανταγωνιστικών εταιριών. Μόλις ολοκληρωθεί η ανάλυση μέσω του Data Mining και βρεθούν τα χαρακτηριστικά που ψάχνουμε, μπορούμε να τα εφαρμόσουμε σε τωρινούς πελάτες και να εξακριβώσουμε την πιθανότητα να προτιμήσουν κάποιο άλλο προϊόν. Αυτό το group των πελατών στην συνέχεια στοχεύετε από την επιχείρηση μας με συγκεκριμένη στρατηγική στο marketing, πράγμα που θα ήταν πολύ δαπανηρό άμα γινόταν στο σύνολο των πελατών μας. Η στρατηγική αυτή μπορεί να αφορά την έκπτωση σε ένα προϊόν που δείχνουν να μην απολαμβάνουν ή την προσφορά ενός άλλου προϊόντος μαζί με αυτό για να γίνει ποιο ελκυστικό. Έτσι μεγαλώνει η πιθανότητα οι πελάτες αυτοί να το αγοράσουν το προϊόν ενώ θα ήταν σχεδόν βέβαιο ότι οι περισσότεροι από αυτούς θα είχαν στραφεί σε άλλη επιχείρηση αν δεν είχε προηγηθεί ανάλυση των δεδομένων. Αυτό προσφέρει τεράστια κέρδη στις επιχειρήσεις και αυτό το απλό παράδειγμα καταδεικνύει την χρησιμότητα της ανάλυσης των Big Data στην σημερινή, πελατοκεντρική κοινωνία μας, και για ποιον λόγο έχει προκληθεί τόσος θόρυβος γύρο από αυτά.

Παρακάτω θα αναλύσω μερικές από τις Data Mining τεχνικές που χρησιμοποιούνται στις μέρες μας όπως: Association rule learning, Cluster analysis και Classification. (Galit Shmueli, 2017)

3.1.1) Association rule learning

Αυτή η τεχνική μας δίνει την δυνατότητα να ανακαλύψουμε ζευγάρια μεταβλητών που μπορεί να σχετίζονται με ένα αποτέλεσμα. Αυτό συμβαίνει με διαδοχικά τεστ και παραμετροποιήσεις στους αλγορίθμους μέχρι να βγει ένα επιθυμητό μοτίβο. Όταν λέμε ένα επιθυμητό μοτίβο αυτό σημαίνει ο αλγόριθμός μας θα έχει κάποια κατώτερα όρια, τα οποία αν δεν

ικανοποιηθούν δεν επιστρέφουν αποτέλεσμα. Αυτό μας βοηθάει να διώχνουμε τις περιττές μεταβλητές που θα επηρέαζαν την ευστοχία της πρόβλεψης του αλγορίθμου. Αυτή η τεχνική χρησιμοποιείται κατά κόρων στο Data Mining από εταιρίες πωλήσεων προϊόντων όπως η Amazon, από supermarkets και γενικότερα από εταιρίες πωλήσεων που θέλουν να βρουν μοτίβα πίσω από τις καταναλωτικές συνήθειες των πελατών.

Ένα παράδειγμα χρησιμότητας της συγκεκριμένης τεχνικής είναι ότι μια αλυσίδα supermarket χρησιμοποιώντας την συγκεκριμένη τεχνική κατέληξε στο συμπέρασμα ότι άντρες καταναλωτές που έγιναν γονείς πρόσφατα, αγόραζαν πάνες για το μωρό τους μαζί με μπύρες για την προσωπική τους διασκέδαση. Έτσι τα supermarket βρίσκοντας τέτοια μοτίβα στους καταναλωτές τους, είτε δημιουργούν μια προσφορά μεταξύ δυο προϊόντων, είτε φέρνουν τα προϊόντα στο supermarket σε παραπλήσιους διαδρόμους.

3.1.2) Clustering

Αυτή η τεχνική χρησιμοποιείται συνήθως όταν έχουμε έναν τεράστιο αριθμό από δεδομένα τα οποία προσπαθούμε να ομαδοποιήσουμε βάση κοινών χαρακτηριστικών και γνωρισμάτων τα οποία δεν ξέραμε προηγουμένως έτσι ώστε να πετύχουμε συγκεκριμένα αποτελέσματα.

Ένα παράδειγμα που είχα αναφέρει προηγουμένως είναι ότι από όλο το πελατολόγιο μιας εταιρίας πωλήσεων, ο αλγόριθμος χωρίζει τους καταναλωτές σε groups βάση των προτιμήσεών τους ώστε να μπορεί η εταιρία να προσαρμόσει το marketing της, στις συνήθειες του κάθε group.

Ένα άλλο παράδειγμα που έφτιαξα για να δω από πρώτο χέρι την λειτουργικότητα του Clustering, είναι ότι είχα τα στοιχεία, που είναι διαθέσιμα για 1309 επιβάτες που ταξίδεψαν με τον τιτανικό, σ’ ένα αρχείο.

ΕΙΚΟΝΑ 8) <u>Στοιχεία επιβατών στον τιτανικό</u>

survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	######	B5	S	2		St Louis, MO
1	Allison, Master. Hudson Trevor	male	0.917	1	2	113781	######	C22 C26	S	11		Montreal, PQ / Chesterville, ON
0	Allison, Miss. Helen Loraine	female	2	1	2	113781	######	C22 C26	S			Montreal, PQ / Chesterville, ON
0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	######	C22 C26	S		135	Montreal, PQ / Chesterville, ON
0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	######	C22 C26	S			Montreal, PQ / Chesterville, ON
1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
0	Astor, Col. John Jacob	male	47	1	0	PC 17757	######	C62 C64	C		124	New York, NY
1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18	1	0	PC 17757	######	C62 C64	C	4		New York, NY
1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
1	Barber, Miss. Ellen "Nellie"	female	26	0	0	19877	78.8500		S	6		
1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
0	Baumann, Mr. John D	male		0	0	PC 17318	25.9250		S			New York, NY
0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	######	B58 B60	C			Montreal, PQ
1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50	0	1	PC 17558	######	B58 B60	C	6		Montreal, PQ
1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917	D15	C	8		
0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
1	Behr, Mr. Karl Howell	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
1	Bidois, Miss. Rosalie	female	42	0	0	PC 17757	######		C	4		
1	Bird, Miss. Ellen	female	29	0	0	PC 17483	######	C97	S	8		
0	Birnbaum, Mr. Jakob	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
1	Bishop, Mr. Dickinson H	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
1	Bishop, Mrs. Dickinson H (Helen Walton)	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
1	Bissette, Miss. Amelia	female	35	0	0	PC 17760	######	C99	S	8		
1	Bjornstrom-Steffansson, Mr. Mauritz Hakan	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Washington, DC

και χρησιμοποιώντας τον αλγόριθμο Mean Shift, ο οποίος χρησιμοποιείται για Data Mining/Clustering, τροποποιώντας κάποια από τα στοιχεία που θα επηρέαζαν αρνητικά την αποδοτικότητα του αλγορίθμου όπως π.χ. αφαιρέθηκε η στήλη με το όνομα, με τον αριθμό εισιτηρίου και η στήλη με το λιμάνι προορισμού, καθώς δεν θα βοηθούσε τον υπολογιστή, όπως δεν θα βοηθούσε και εμάς, να κατηγοριοποιήσουμε τα groups των ανθρώπων που επέβαιναν στο πλοίο, επέστρεψε τα εξής αποτελέσματα:

```
{0: 0.37782982045277125, 1: 0.7692307692307693, 2: 0.1, 3: 1.0}
```

Χώρισε με τα στοιχεία που του παρείχαμε τους επιβάτες σε 4 groups βάση του ποσοστού επιβιωσημότητας. Αναλύοντας 2 από τα groups το 0 με ποσοστό επιβιωσημότητας 37.7% και το 1 με ποσοστό επιβιωσημότητας 76.92% βλέπουμε ότι:

```
{0: 0.37782982045277125, 1: 0.7692307692307693, 2: 0.1, 3: 1.0}
            pclass     survived          age        sibsp        parch  \
count  1281.000000  1281.000000  1020.000000  1281.000000  1281.000000
mean      2.307572     0.377830    29.669281     0.486339     0.322404
std       0.830482     0.485034    14.376403     1.039084     0.670731
min       1.000000     0.000000     0.166700     0.000000     0.000000
25%       2.000000     0.000000    21.000000     0.000000     0.000000
50%       3.000000     0.000000    28.000000     0.000000     0.000000
75%       3.000000     1.000000    38.000000     1.000000     0.000000
max       3.000000     1.000000    80.000000     8.000000     4.000000

              fare        body  cluster_group
count  1280.000000  119.000000         1281.0
mean     29.228115  159.571429            0.0
std      37.820193   97.302914            0.0
min       0.000000    1.000000            0.0
25%       7.895800   71.000000            0.0
50%      13.900000  155.000000            0.0
75%      30.000000  255.500000            0.0
max     247.520800  328.000000            0.0
       pclass   survived        age      sibsp      parch        fare  body  \
count    13.0  13.000000  13.000000  13.000000  13.000000   13.000000   0.0
mean      1.0   0.769231  36.384615   1.692308   2.307692  301.117954   NaN
std       0.0   0.438529  19.050540   1.109400   0.947331   93.738703   NaN
min       1.0   0.000000  13.000000   0.000000   1.000000  262.375000   NaN
25%       1.0   1.000000  21.000000   1.000000   2.000000  262.375000   NaN
50%       1.0   1.000000  28.000000   2.000000   2.000000  263.000000   NaN
75%       1.0   1.000000  58.000000   3.000000   3.000000  263.000000   NaN
max       1.0   1.000000  64.000000   3.000000   4.000000  512.329200   NaN

       cluster_group
count           13.0
mean             1.0
std              0.0
min              1.0
25%              1.0
50%              1.0
75%              1.0
max              1.0
```

Στο group 0 που έβαλε την πλειοψηφία των επιβατών υπάρχει διασπορά εισιτηρίων δηλαδή και από τις τρεις κατηγορίες με μέσο όρο τιμής 37.82 λίρες ενώ στο group 1 όλοι οι επιβάτες είναι πρώτης κατηγορίας εισιτηρίου με μέση τιμή 301 λίρες και πολύ μεγαλύτερο ποσοστό επιβιωσημότητας. Αυτό δείχνει ότι ο αλγόριθμος έμαθε βάση των στοιχείων που του έδωσα, ότι αυτοί που είχαν πληρώσει ακριβότερο εισιτήριο είχαν μεγαλύτερη πιθανότητα να ζήσουν (βρισκόντουσαν στο πάνω μέρος του πλοίου με γρηγορότερη πρόσβαση στις σωσίβιες λέμβους

και έτυχαν γενικότερα καλύτερης αντιμετώπισης από το πλήρωμα). Για να ενισχύσω το συγκεκριμένο επιχείρημα, απομόνωσα από το cluster 0 που περιείχε επιβάτες και των τριών κατηγοριών, μόνο αυτούς που είχαν αγοράσει εισιτήριο πρώτης κατηγορίας και είχαμε τα εξής αποτελέσματα:

```
{0: 0.37782982045277125, 1: 0.7692307692307693, 2: 0.1, 3: 1.0}
       pclass    survived         age       sibsp       parch        fare
count   305.0  305.000000  266.000000  305.000000  305.000000  305.000000
mean      1.0    0.606557   39.276003    0.390164    0.285246   73.947350
std       0.0    0.489316   14.438003    0.521165    0.579767   53.848165
min       1.0    0.000000    0.916700    0.000000    0.000000    0.000000
25%       1.0    0.000000   28.625000    0.000000    0.000000   30.500000
50%       1.0    1.000000   39.000000    0.000000    0.000000   57.750000
75%       1.0    1.000000   49.750000    1.000000    0.000000   90.000000
max       1.0    1.000000   80.000000    2.000000    2.000000  247.520800
```

Οι 305 από τους 1280 επιβάτες του group 0 που είχαν αγοράσει εισιτήριο πρώτης κατηγορίας (με μέση τιμή 73,9 λίρες σε σχέση με τις 29.22 λίρες που είχε το group ως μέσο όρο) είχαν ποσοστό επιβιωσημότητας 60.6% σε σχέση με το 37.7% που είχε το συνολικό group.

Βλέπουμε ότι με έναν απλό αλγόριθμο Data Mining με την τεχνική του Clustering ο υπολογιστής έχει την δυνατότητα να βρίσκει μοτίβα και με αυτά να χωρίζει τους ανθρώπους σε groups. Με την ίδια ακριβώς τεχνική, με πολύ καλύτερους αλγόριθμους, και με datasets χιλιάδων καταναλωτών, οι εταιρίες χωρίζουν τους ανθρώπους βάση αγορών και συνηθειών και εξάγουν τα δικά τους αποτελέσματα προς υλοποίηση.

3.1.3) Classification

Η τεχνική του Classification, σε αντίθεση με το Clustering, χρησιμοποιείται όταν έχουμε συγκεκριμένα αποτελέσματα από ένα dataset, τα groups είναι ήδη χωρισμένα βάση των αποτελεσμάτων, και βάση αυτών των στοιχείων όταν έρχεται ένα νέο data point ο αλγόριθμος το κατατάσσει στο group με τα περισσότερα κοινά χαρακτηριστικά. Αυτή η τεχνική ονομάζεται και supervised learning καθώς σε αντίθεση με το clustering (unsupervised learning) που δεν υπήρχαν διαμορφωμένα groups, εδώ τα groups υπάρχουν και μαθαίνουμε τον αλγόριθμο να σκέφτεται παρέχοντας του ένα κομμάτι των αποτελεσμάτων από το dataset.

Για να γίνει πιο κατανοητό, έστω ότι έχουμε ένα dataset το οποίο στις στήλες του έχει τα στοιχεία που χρειάζονται για να αποφανθεί ένας γιατρός αν μια γυναίκα έχει εμφανίσει κακοήθη ή καλοήθη όγκο στον μαστό. Από αυτά τα στοιχεία τα οποία υπάρχουν στο dataset, μαθαίνουμε στον αλγόριθμο μέσω της τεχνικής του Classification να βρίσκει τα μοτίβα εκείνα μεταξύ των στοιχείων που υποδεικνύουν αν ένας όγκος είναι καλοήθης ή κακοήθης, και όταν θα εισάγουμε τα στοιχεία μιας νέας ασθενούς, ο αλγόριθμος έχοντας δημιουργήσει μοτίβα από τα στοιχεία που τα παρείχαμε, κατατάσσει την ασθενή σε ένα από τα δύο groups. Ουσιαστικά με αυτήν την τεχνική δημιουργούνται αποτελέσματα πρόβλεψης, τα οποία θα έπαιρναν πάρα πολύ χρόνο στους γιατρούς χωρίς την τεχνική αυτή. Τα αποτελέσματα αυτά, συνοδεύονται από ένα ποσοστό πρόβλεψης δηλαδή ότι κατά 98,7% η ασθενής Χ ανήκει στο group με τους καλοήθεις όγκους. Αυτή η πρόβλεψη βοηθάει το προσωπικό να γλιτώσει τον χρόνο ψαξίματος και να ψάξει στοχευμένα για το αν η πρόβλεψη ήταν αληθής η όχι. (το 98.7% θα ήταν ένα υπέροχο νούμερο σε πρόβλεψη μετοχής στο χρηματιστήριο, αλλά σε ένα τόσο ευαίσθητο θέμα οι γιατροί πρέπει να είναι 100% σίγουροι πριν παραδώσουν τα αποτελέσματα).

3.2) MACHINE LEARNING AND STATISTICS

Όπως προδίδει και το όνομα του το Machine Learning, που οι τεχνικές του συνδυάζονται πολλές φορές με την στατιστική και το Data Mining, είναι ένα είδος τεχνητής νοημοσύνης (AI) το οποίο δίνει την δυνατότητα στους ηλεκτρονικούς υπολογιστές να μάθουν μοτίβα και συνήθειες χωρίς να έχουν προγραμματιστεί γι' αυτόν τον σκοπό. Ο Data Analyst που θα χρησιμοποιήσει την τεχνική του Machine Learning, ουσιαστικά εστιάζει στην δημιουργία προγραμμάτων τα οποία μπορούν να αλλάξουν και να προσαρμόσουν τα αποτελέσματα τους κάθε φορά που δέχονται νέα δεδομένα.

Σαν έννοια το Machine Learning δεν είναι καινούργιο, αλλά ουσιαστικά τώρα με την υπάρχουσα τεχνολογία γίνεται χρήσιμο. Η ιδέα υπήρχε από το 1970, ότι δηλαδή οι υπολογιστές μπορούν να μάθουν να κάνουν συγκεκριμένα πράγματα χωρίς να έχουν προγραμματιστεί γι' αυτά (πρώιμη θεωρία τεχνητής νοημοσύνης). Η εφαρμογή του στις μέρες μας, όπως γενικότερα η άνοδος των Big Data, προέρχεται από την δυνατότητα που έχουμε πλέον να χρησιμοποιούμε πολύ δυνατούς υπολογιστές με πολύ μικρό κόστος (για την δουλειά που παράγουν) σε συνδυασμό με τις τεράστιες ποσότητες δεδομένων που έχουμε διαθέσιμες.

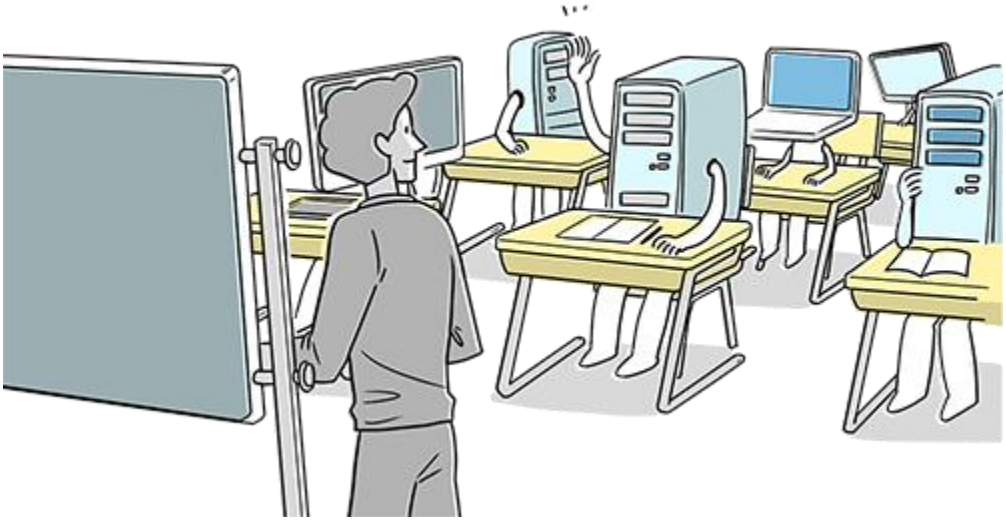
Η τεχνική του Machine Learning μοιάζει πολύ με εκείνη του Data Mining και πολλές φορές ο διαχωρισμός τους δεν είναι εύκολα διακριτός. Και οι δύο οι τεχνικές ψάχνουν στα δεδομένα που τους έχουμε τροφοδοτήσει και προσπαθούν να βρουν συσχετίσεις. Ωστόσο, σε αντίθεση με το Data Mining που εξάγει δεδομένα για να βγουν αποτελέσματα μη διακριτά σε ανθρώπους λόγω του τεράστιου όγκου των δεδομένων, το Machine Learning χρησιμοποιεί αυτά τα δεδομένα για να εντοπίσει μοτίβα και προσαρμόζει τα αποτελέσματα του προγράμματος ανάλογα.

Ένα απλό παράδειγμα χρήσης Machine Learning, που πολλοί από εσάς θα το έχετε καταλάβει άμα χρησιμοποιείται το Facebook, είναι ότι χρησιμοποιούν οι αναλυτές της εταιρίας αλγορίθμους που εξατομικεύουν για κάθε χρήστη την ροή των ειδήσεων ή των φίλων που θα βλέπει όταν θα ανοίγει την σελίδα. Αν ας πούμε ένας χρήστης της εφαρμογής σταματάει συχνά να διαβάσει μια δημοσίευση ή να πατήσει like σ' ένα συγκεκριμένο φίλο ή φίλη, τότε ο αλγόριθμος συσχετίζει αυτήν την κίνηση με τον χρήστη και την επόμενη φορά που θα ανοίξει την σελίδα, αν ο φίλος του έχει δημοσιεύσει κάτι καινούργιο, αυτό θα είναι στην κορυφή της σελίδας ώστε να το δει πρώτο. Το λογισμικό πολύ απλά χρησιμοποιεί στατιστική ανάλυση των κινήσεων κάθε ατόμου και όταν έχει συλλέξει αρκετά δεδομένα τότε μπορεί βάση των προηγούμενων κινήσεων, να προβλέπει τι ενδιαφέρει τον κάθε χρήστη συγκεκριμένα και να προσαρμόζει ανάλογα την ροή των δημοσιεύσεων. Αν στην περίπτωση που σταματήσει ο χρήστης μας να διαβάζει ή να πατάει like στον συγκεκριμένο φίλο, τότε αυτά τα νέα δεδομένα προσαρμόζονται σαν νέο μοτίβο στον αλγόριθμό και ξανά αλλάζει την ροή των δημοσιεύσεων που θα βλέπει. (rouse, 2016)

Το Machine Learning χρησιμοποιείται παντού επειδή πλέον πολλές εταιρίες και οργανισμοί έχουν στην διάθεση τους τεράστιες ποσότητες αχρησιμοποίητων δεδομένων που αν αναλυθούν σωστά θα τους αποφέρουν τεράστια κέρδη.

Παρακάτω θα αναλύσω μερικές τεχνικές που βασίζονται στο Machine Learning και στην στατιστική όπως Regression, Natural Language Processing, A/B testing και Spatial analysis.

ΕΙΚΟΝΑ 9) <u>Machine Learning!</u>



3.2.1) <u>Regression</u>

Το Regression αφορά μια σειρά τεχνικών που είχαν χρησιμοποιηθεί αρχικώς στην στατιστική και αφορούσαν την μεταβολή μια μεταβλητής (label) όταν άλλαζαν οι τιμές συσχετιζόμενων μεταβλητών. Αυτό υιοθετήθηκε από τους αναλυτές που χρησιμοποιούν το Machine Learning, μαθαίνοντας στον υπολογιστή πως η label επιθυμητή μεταβλητή αλλάζει όταν τις παρέχουμε συνεχώς νέα δεδομένα τα οποία μπορεί να είναι και real-time (μεταβολή μιας μετοχής).

Συνήθως χρησιμοποιείται στο κλάδο των οικονομικών από εταιρίες και ανθρώπους που ασχολούνται με το χρηματιστήριο, τις επενδύσεις, τις τιμές των ακινήτων και πως αυτές επηρεάζονται σε καθημερινή βάση. Ένα παράδειγμα λειτουργίας είναι ότι αναλυτές που ασχολούνται με την διακύμανση των μετοχών, έχουν αλγορίθμους στους οποίους έχουν περάσει τα δεδομένα της διακύμανσης μιας μετοχής του τελευταίου διμήνου, μαθαίνοντας στον υπολογιστή να συσχετίζει την άνοδο ή την πτώση μιας μετοχής συγκρίνοντας την με μια σειρά από μεταβλητές, και έτσι σε κάθε νέα συνεδρίαση όπου νέα real-time δεδομένα παρέχονται στον αλγόριθμο, βάση των συσχετίσεων που είχε κάνει από τα αποτελέσματα που του είχαμε δώσει για

το τελευταίο δίμηνο, δίνει μια πρόβλεψη για την διακύμανση της μετοχής κατά την διάρκεια της ημέρας. (investopedia, 2016)

3.2.2) Natural Language Processing

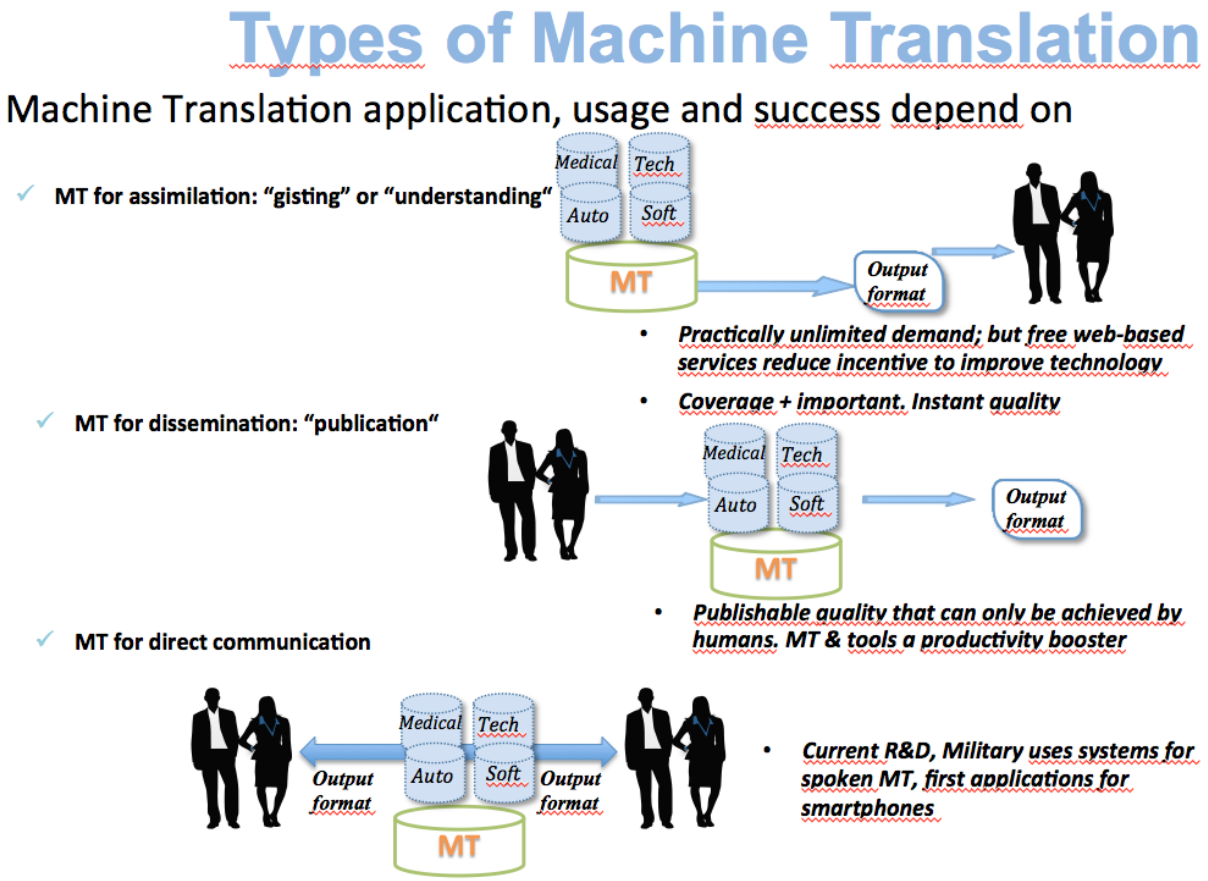
Αυτό είναι ένα γενικότερο πεδίο που έχει στοιχεία από την επιστήμη των υπολογιστών, την τεχνητή νοημοσύνη, και της υπολογιστικής γλωσσολογίας και πραγματεύεται τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας και ειδικότερα με την κατανόηση από τους υπολογιστές μέσω προγραμμάτων της δομής, της σύνταξης και ειδοποιών διαφορών κάθε γλώσσας, έτσι ώστε να μπορεί να αντιλαμβάνεται ο υπολογιστής τις εντολές που του δίνουμε όταν λαμβάνει σαν δεδομένα κείμενα από κάθε γλώσσα.

Ένα τέτοιο παράδειγμα χρησιμοποιώντας μια τεχνική που ανήκει στο πεδίο του Natural Language Processing είναι το **Sentiment analysis** όπου αυτό χρησιμοποιείται για τον εντοπισμό και την εξαγωγή πληροφοριών από διάφορα κείμενα. Βασικός σκοπός της ανάλυσης αυτής είναι να προσδιορίζει σ' ένα κείμενο αν ο τόνος των σχολίων είναι θετικός ή αρνητικός. Αυτό χρησιμοποιείται κατά κόρων από εταιρίες πωλήσεων όπως η Amazon, όπου χρησιμοποιώντας αυτήν την τεχνική κατηγοριοποιούν τις κριτικές σε θετικές ή αρνητικές, χωρίς να χρειαστεί να διαβαστούν αυτοτελώς οι κριτικές. Αυτό έχει ως αποτέλεσμα την εξαγωγή συμπερασμάτων πάνω στα υπό πώληση προϊόντα, δηλαδή κατά πόσο ανταποκρίθηκαν θετικά ή αρνητικά στις προσδοκίες των καταναλωτών (οι αλγόριθμοι είναι φτιαγμένοι έτσι ώστε ο υπολογιστής να μπορεί να κατανοεί ακόμα και το αν ένα σχόλιο ήταν μετρίως αρνητικό ή απόλυτα αρνητικό και θετικό αντίστοιχα), αλλά επίσης και σε μεγάλη εξοικονόμηση χρόνου, καθώς αν δεν έκανε ο αλγόριθμος την συγκεκριμένη δουλεία, θα χρειαζόντουσαν δεκάδες εργατοώρες για την προσπέλαση όλων των σχολίων σε όλα τα προϊόντα.

Μια άλλη τεχνική ονομάζεται **word segmentation** η οποία μπορεί να διαχωρίζει ένα αρχείο συνεχούς κειμένου σε ξεχωριστές λέξεις. Για μια γλώσσα σαν τα αγγλικά αυτό είναι αρκετά ασήμαντο δεδομένου ότι οι λέξεις είναι συνήθως χωρισμένες με κενά. Ωστόσο, ορισμένες γλώσσες όπως τα Κινέζικα ή τα Ιαπωνικά, στα οποία δεν τμηματοποιούνται οι λέξεις όπως σε μια ευρωπαϊκή γλώσσα, οπότε η διαχώριση των λέξεων είναι σημαντική και απαιτεί την γνώση και τις πολλές ερμηνείες της κάθε λέξης καθώς και την μορφολογία της γλώσσας.

Μια από τις ποιο δύσκολες εφαρμογές είναι το **machine translation** όπου δεν είναι τίποτα άλλο από μετάφραση από την μια γλώσσα στην άλλη. Ακούγεται πολύ εύκολο και είναι σε μερικές περιπτώσεις όταν οι γλώσσες είναι φτιαγμένες από τον ίδιο κορμό (Γαλλικά, Ιταλικά, Ισπανικά), αλλά όταν οι γλώσσες διαφέρουν πάρα πολύ (Αγγλικά σε σχέση με Ελληνικά ή Κινέζικα), οι διαφορές στην γραμματική, στο συντακτικό, στην σήμανση των λέξεων, στους ιδιωματισμούς κλπ.) καθιστά πολύ δύσκολη την διαδικασία 'εκμάθησης' του υπολογιστή. Παράδειγμα είναι το google translate όπου όπως έχετε δει η μετάφραση από τα αγγλικά στα ελληνικά είναι μέτρια στην καλύτερη περίπτωση.

Εικόνα 10) <u>Types of machine translation</u>



Τέλος μια ακόμη δύσκολη, και υπό διαρκής εξέλιξη τεχνική του Natural Language Processing, είναι το **speech recognition,** δηλαδή η δυνατότητα που έχουμε να μιλάμε στον υπολογιστή μέσω μικροφώνου και αυτό να το γράφει σε κείμενο. Οι δυσκολίες εδώ πέρα βρίσκονται στο γεγονός ότι στον προφορικό μας λόγο ο κάθε ανθρωπος έχει τις δικές του

ιδιαιτερότητες (χροιά φωνής, καθαρότητα λόγου, ταχύτητα λόγου), όπου ο υπολογιστής πρέπει σε κάθε περίπτωση να επεξεργαστεί με τον ίδιο τρόπο, και επίσης είναι δύσκολο να καταλάβει πότε σταματάμε μια πρόταση επειδή συνήθως στην ομιλία μας δεν υπάρχουν σχεδόν καθόλου παύσεις μεταξύ των διαδοχικών λέξεων. Επίσης σε κάθε γλώσσα, οι ήχοι που αντιπροσωπεύουν τα συνεχόμενα γράμματα δεν είναι ίδιοι (π.χ. Γερμανικά σε σχέση με Αγγλικά) οπότε η μετατροπή σε χαρακτήρες γίνεται μια πολύ δύσκολη διαδικασία (στην συγκεκριμένη τεχνική έχουν επενδύσει αρκετά κονδύλια πολύ μεγάλες εταιρίες και θεωρείται το νέο μεγάλο επίτευγμα στον χώρο της τεχνητής νοημοσύνης). (Wikipedia, 2017)

3.2.3) Statistics

Όπως και το Natural Language Processing, και η στατιστική είναι ένα γενικότερα πεδίο με πολλές τεχνικές στο Machine Learning. Οι περισσότερες από τις τεχνικές αυτές προϋπήρχαν πριν την χρήση των υπολογιστών για ανάλυση δεδομένων, και εξελίχθηκαν για να μπορούν να υλοποιηθούν σε ένα υπολογιστικό περιβάλλον. Ουσιαστικά η επιστήμη αυτή αφορά την συλλογή, οργάνωση και ερμηνεία των δεδομένων καθώς επίσης τον σχεδιασμό πειραμάτων για την αποτελεσματικότερη ανάλυση των δεδομένων. Οι περισσότερες από τις τεχνικές που χρησιμοποιούνται προσπαθούν να καθορίσουν αν η σχέση μεταξύ δύο μεταβλητών σ' ένα dataset είναι τυχαία, ή αν υπάρχει κάποια σχέση μεταξύ των μεταβλητών και αν ναι, από ποιους παράγοντες προκλήθηκε και ποιοι παράγοντες την επηρεάζουν.

Μια πρώτη τεχνική ονομάζεται **A/B testing.** Με αυτήν, συγκρίνουμε ένα group δεδομένων που έχουμε δημιουργήσει εμείς, με διάφορα test groups, έτσι ώστε να καθορίσουμε τις αλλαγές που χρειάζεται να γίνουν ώστε να βελτιώσουμε το αρχικό μας group. Για παράδειγμα αυτή η τεχνική χρησιμοποιείται συνεχώς από web developers ώστε να κάνουν την ιστοσελίδα ποιο προσιτή στον κόσμο και να την επιλέγουν και βάση της αισθητικής. Πραγματοποιούνε διάφορα πειράματα όπως, τι περίγραμμα είναι περισσότερο θεμιτό στο κοινό, τι εικόνες προτιμώνται, τι χρώμα να είναι το φόντο κλπ. Τα Big Data μας επιτρέπουν να τρέχουμε χιλιάδες πειράματα ταυτόχρονα και να αναλύουν τα αποτελέσματα ώστε να βρίσκουν το βέλτιστο αποτέλεσμα.

Μια άλλη τεχνική με αρκετά στοιχεία από το πεδίο της στατιστικής είναι η **Spatial analysis,** όπου χρησιμοποιείται για να γίνει ανάλυση μια γεωγραφικής περιοχής, ώστε να μπορέσουν οι υποψήφιοι επενδυτές που θέλουν να επενδύσουν εκεί πέρα, να ξέρουν (κατά ένα

βαθμό) αν θα έχει επιθυμητά αποτελέσματα η επένδυσή τους. Γι' αυτή την ανάλυση χρησιμοποιούνται δεκάδες μεταβλητές όπως, το γεωγραφικό προφίλ της περιοχής, το καταναλωτικό κοινό, ακόμα και τον καιρό που έχει η συγκεκριμένη περιοχή και γενικότερα οτιδήποτε μπορεί να επηρεάσει θετικά ή αρνητικά την απόφαση των καταναλωτών να αγοράσουν ένα προϊόν από ένα συγκεκριμένο κατάστημα.

3.3) ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΕΙΤΑΙ ΤΟ MACHINE LEARNING

Οι περισσότερες εταιρίες που διαθέτουν μεγάλο όγκο δεδομένων έχουν καταλάβει ότι η αξιοποίησή τους θα τους αποφέρει τεράστια κέρδη. Χρησιμοποιώντας το Machine Learning προσπαθούν να αποκομίσουν στοιχεία από αυτά τα δεδομένα, και πολλές φορές σε πραγματικό χρόνο, έτσι ώστε να έχουν πλεονέκτημα έναντι των ανταγωνιστών τους αλλά και για να μειώσουν το κόστος. Πέντε μεγάλοι κλάδοι τις οικονομίας έχουν αρχίσει να υλοποιούν τις διάφορες τεχνικές του Machine Learning προς όφελος τους. Αυτοί οι πέντε κλάδοι αντιπροσωπεύουν ένα μεγάλο μέρος της παγκόσμιας οικονομίας και διαφέρουν ως προς τα αποτελέσματα που θέλουν από την χρήση των Big Data και γι' αυτό τον λόγω οι τεχνικές που χρησιμοποιούνται διαφέρουν ανά κλάδο.

ΕΙΚΟΝΑ 11) Πως αξιοποιούνται τα Big Data σε κάθε κλάδο

3.3.1) <u>Υγεία</u>

Στον τομέα της υγειονομικής περίθαλψης συνεργάζονται επιχειρηματικοί κλάδοι από τις φαρμακευτικές εταιρίες, τις εταιρίες παραγωγής ιατρικού εξοπλισμού, παρόχους, ασθενείς κλπ. και κάθε μια από αυτές τις συνδεόμενες μεταβλητές παράγει τεράστια ποσότητα δεδομένων. Κάθε μια από αυτές έχει διαφορετικά συμφέροντα και επιχειρηματικά κίνητρα, και στην πλειοψηφία τους τα δεδομένα που παράγονται δεν αναλύονται. Επίσης πολλά από τα κλινικά δεδομένα δεν έχουν ψηφιοποιηθεί ακόμα οπότε είναι αδύνατη η εκμετάλλευσή τους. Επιπλέον, παρατηρείται μια αύξηση 5% τον χρόνο στα έξοδα που αφορούν τον συγκεκριμένο κλάδο, και αυτό επιβαρύνει τον προϋπολογισμό των κυβερνήσεων και διογκώνει το δημόσιο χρέος. Σε αυτό έρχεται να προστεθεί και η παγκόσμια γήρανση του πληθυσμού όπως και επίσης νέες, ποιο δαπανηρές θεραπείες. Γενικότερα ο κλάδος της υγείας έχει μείνει αρκετά πίσω σε σχέση με άλλους κλάδους στην βελτίωση των λειτουργικών εξόδων, όπως επίσης και στην ενσωμάτωση τεχνολογιών για την βελτίωση τους. Αυτά τα προβλήματα δημιουργούν μια τεράστια ευκαιρία από την εκμετάλλευση των δεδομένων αυτών για την δημιουργία κέρδους από την ανάλυση τους, αν αυτά τα δεδομένα ψηφιοποιηθούν, συνδεθούν μεταξύ τους και χρησιμοποιηθούν σε αυτά κατάλληλες τεχνικές.

Τα Big Data που μπορούμε να αναλύσουμε στον τομέα της υγείας χωρίζονται σε τέσσερις κατηγορίες:

- **Φαρμακευτικές έρευνες και δαπάνες** (κλινικές μελέτες φαρμάκων, κόστος παρασκευής και πώλησης, αποτελεσματικότητα κόστους/θεραπείας κλπ.)

- **Κλινικά δεδομένα ασθενών** (ηλεκτρονική απεικόνιση ιστορικού υγείας, ψηφιακό αρχείο από ακτινογραφίες)

- **Υγειονομικές δαπάνες των ιατρείων αλλά και των ασθενών**

- **Δεδομένα που αφορούν την συμπεριφορά των ασθενών και τις προτιμήσεις τους κατά την διαρκεια της νοσηλείας**

Όπως ανέφερα, πολλά από αυτά τα πολύ σημαντικά δεδομένα δεν έχουν ψηφιοποιηθεί ακόμη και δεν υπάρχει εκμετάλλευση των δεδομένων. Αυτά τα δεδομένα χρειάζεται να αναλυθούν καθώς θα παρέχουν άμεσα αποτελέσματα. Μερικά παραδείγματα από το κέρδος της ανάλυσης τους είναι:

- Στοχευμένη ανάλυση ιστορικού υγείας για κάθε περίπτωση ασθενούς, ώστε να παρέχεται η καλύτερη εξατομικευμένη θεραπεία.

- Real-time συλλογή δεδομένων από ασθενείς με χρόνιες παθήσεις, όπου παρακολουθείται αν οι ασθενείς έχουν κάνει αυτά που προβλεπόντουσαν, και άμεση αναπροσαρμογή θεραπείας ή φαρμάκων όπου κρίνεται απαραίτητο, ανάλογα με την πρόοδο του ασθενούς.

- Ανάλυση του προφίλ των ασθενών, για τον εντοπισμό ατόμων που θα ωφεληθούν από προληπτική φροντίδα ή αλλαγές στις διατροφικές συνήθειες και στον τρόπο ζωής, αν βρεθεί ότι υπάρχουν αρκετές πιθανότητες (μέσω αλγορίθμων πρόβλεψης) να αναπτύξουν στο μέλλον μια ασθένεια.

- Αποτελεσματικότητα θεραπειών/φαρμάκων βάση απόδοσης/κόστους, και μείωση τιμών όπου τα αποτελέσματα δεν είναι τα θεμιτά. (Adamson, 2016)

3.3.2) <u>Χρηματοπιστωτικές υπηρεσίες</u>

Από το 2015 οι μεγάλες τράπεζες εφάρμοσαν σε μεγάλη κλίμακα την ανάλυση των αχρησιμοποίητων δεδομένων, και τα αποτελέσματα που είδαν θεωρήθηκαν τόσο σημαντικά, ώστε το 2015 να θεωρείται χρονιά σταθμός στον τραπεζικό τομέα και στις χρηματοπιστωτικές αγορές. Στις μέρες μας καταβάλλεται μεγάλη προσπάθεια για το μέγιστο κέρδος από τα Big Data, καθώς συνεχίζεται η αναδιοργάνωση των διαδικασιών, που τόσα χρόνια παρέμεναν ίδιες. Βλέπουμε ότι οι τράπεζες και όσοι ασχολούνται με επενδύσεις και χρηματιστήριο προσπαθούν να προσαρμοστούν στην νέα εποχή για να παραμείνουν ανταγωνιστικοί.

Για τις τράπεζες, τα κίνητρα για να χρησιμοποιήσουν Big Data βρίσκονται κυρίως στην καλύτερη χαρτογράφηση των πελατών της και την μείωση του επιχειρησιακού κινδύνου. Για τις εταιρίες ή τους ανθρώπους που ασχολούνται με το χρηματιστήριο ή τις επενδύσεις, τα αποτελέσματα που θέλουν χρειάζεται να είναι άμεσα, καθώς το κέρδος βρίσκεται στην λήψη γρήγορων αποφάσεων κατά την διάρκεια της ημέρας. Οι ευκαιρίες που παρουσιάζονται από την επεξεργασία και την ανάλυση των δεδομένων είναι:

- Machine Learning αλγόριθμοι θα επεξεργάζονται Real-Time τα εισερχόμενα δεδομένα και θα είναι σε θέση να ενημερώνουν άμεσα για οικονομικές απάτες.

- Η επεξεργασία των δεδομένων θα προσφέρει στοχευμένες υπηρεσίες για τραπεζικές συναλλαγές ή δάνεια ανάλογα με το προφίλ του κάθε καταναλωτή.

- Θα παρέχουν σε επενδυτές μια ποιο ασφαλή εικόνα για τις μελλοντικές τους επενδύσεις. (O'Dowd, 2016)

3.3.3) <u>Δημόσια Διοίκηση</u>

Οι περισσότερες κυβερνήσεις ανά τον κόσμο βρίσκονται υπό αυξανόμενη πίεση για να δοθεί ώθηση στην παραγωγικότητα, δηλαδή να προσφέρουν περισσότερα με το μικρότερο δυνατό κόστος. Αυτή η ανάγκη γίνεται ιδιαιτέρως επιτακτική ύστερα από τους κλυδωνισμούς που επέφερε η παγκόσμια οικονομική κρίση. Πολλές κυβερνήσεις προσπαθούν να παρέχουν όσο τον δυνατόν καλύτερες δημόσιες υπηρεσίες σε μια εποχή όπου υπάρχουν σημαντικοί δημοσιονομικοί περιορισμοί για να μειωθούν τα δημοσιονομικά ελλείμματα και το δημόσιο χρέος. Επιπλέον πολλές χώρες αντιμετωπίζουν και την γύρναση του πληθυσμού, που θα αυξήσει σημαντικά τη ζήτηση σε ιατρικές και κοινωνικές υπηρεσίες, άρα και τα έξοδα.

Για να καταφέρουν οι κυβερνήσεις να είναι μέσα στους δημοσιονομικούς τους στόχους, αλλά και για να δώσουν ώθηση στο δημόσιο τομέα θα πρέπει να αυξήσουν την παραγωγικότητά τους. Έρευνες έχουν δείξει ότι τα τελευταία χρόνια ο ιδιωτικός τομέας σε πολλά

κράτη έχει ξεπεράσει τον δημόσιο. Πως όμως μπορούν τα Big Data να βοηθήσουν στην αύξηση της παραγωγικότητας της δημόσιας διοίκησης? Τα δεδομένα που παράγονται στον δημόσιο τομέα είναι κυρίως δεδομένα κειμένου ή αριθμητικά δεδομένα, και σε σχέση με τον τομέα της υγείας όπου τα δεδομένα είναι εικόνες υψηλής ευκρίνειας ή βίντεο από εγχειρήσεις, διαπιστώνουμε ότι στο δημόσιο τα δεδομένα είναι λιγότερα σε μέγεθος, αλλά εξίσου σημαντικά. Η σημαντική διαφορά όμως είναι ότι τα δεδομένα που παράγονται είναι κατά 90% ψηφιοποιημένα καθώς στις περισσότερες χώρες λειτουργούν μέσω ηλεκτρονικών υπηρεσιών για την διευκόλυνση των πολιτών. Τα σημαντικά πλεονεκτήματα που μπορούν να προσφέρουν τα Big Data αν ενσωματωθούν πλήρως στην δημόσια διοίκηση είναι:

- **Εξοικονόμηση χρόνου.** Τα στοιχεία των πολιτών είναι αποθηκευμένα σε datasets τα οποία δεν αξιοποιούνται και τους ζητούνται να ξαναγράφουν τα στοιχεία τους, ή τις φορολογικές τους ενημερότητες από την αρχή.

- **Καταπολέμηση φοροδιαφυγής.** Μέσω των Machine Learning αλγορίθμων θα μπορούν να ελέγχονται Real-Time τα στοιχεία που δηλώνουν οι πολίτες και να ενημερώνουν για πιθανές παραβατικές υποθέσεις ελέγχοντας παράλληλα ογκωδέστατα αρχεία που θα χρειαζόντουσαν δεκάδες εργατοώρες από το ανθρώπινο προσωπικό.

- **Παραγωγή πλούτου και μείωση του κόστους.** Νέα επιχειρηματικά μοντέλα, προϊόντα και υπηρεσίες γίνονται διαθέσιμα με τα Big Data, μειώνοντας το κόστος λειτουργίας του δημόσιου τομέα και βελτιώνοντας τις δημοσιονομικές επιδόσεις της οικονομίας του κάθε κράτους.

3.3.4) <u>Marketing και Πωλήσεις</u>

Αν και δεν μπορεί να υπάρξει άμεση σύγκριση μεταξύ κλάδων, θεωρείται από πολλούς ότι το marketing και οι πωλήσεις έχουν να παρουσιάσουν τα μεγαλύτερα κέρδη από την χρήση των Big Data. Υπολογίζεται ότι με την πλήρη ενσωμάτωση των Big Data τεχνικών στους συγκεκριμένους κλάδους, θα υπάρχει αύξηση της παραγωγικότητας κατά 0.5% ανά έτος μέχρι το

2020, χωρίς να υπολογίζεται σε αυτό η συνεχιζόμενη εμφάνιση τεχνικών και τεχνολογιών που δίνουν την δυνατότητα για ακόμη καλύτερη αξιοποίηση των Big Data.

Ο λόγος που υπάρχει αυτή η δυνατότητα άμεσου κέρδους στον συγκεκριμένο κλάδο βρίσκεται στην χρήση των ψηφιακών δεδομένων. Τα ψηφιακά δεδομένα αποκτούν καίριο ρόλο στον κλάδο καθώς οι καταναλωτές αναζητούν, ερευνούν, συγκρίνουν και τελικά αγοράζουν προϊόντα διαδικτυακά. Αυτό αφήνει ένα ψηφιακό μονοπάτι για τον κάθε καταναλωτή, το οποίο η κάθε εταιρία μπορεί να το αναλύσει προς όφελός της. Αλλά τα οφέλη θα είναι αρκετά και στους καταναλωτές καθώς προβλέπεται ότι με τις νέες τεχνολογίες που αναπτύσσονται στον χώρο των Big Data, οι τιμές θα πέσουν αισθητά. Μια εφαρμογή που ήδη κυκλοφορεί ονομάζεται RedLaser και επιτρέπει στους καταναλωτές να σκανάρουν το bar-code του προϊόντος που επιθυμούν να αγοράσουν σε ένα κατάστημα από τα κινητά τους, και αυτό τους επιστρέφει άμεσα συγκρίσεις με τιμές σε άλλα καταστήματα καθώς και σε παρεμφερή προϊόντα. Επιπλέον με την αύξηση των αγορών από το διαδίκτυο γίνεται ευκολότερο για τους καταναλωτές να επιλέξουν την καλύτερη τιμή και ποιότητα σε ένα προϊόν που επιθυμούν. Αυτοί οι νέοι τρόποι αγορών θα επιφέρουν τεράστια κέρδη στους καταναλωτές. Παρακάτω θα επισημάνω τα σημαντικότερα κέρδη που αποκομίζονται από την χρήση των Big Data τεχνικών στον κλάδο.

- **Προτεινόμενα προϊόντα.** Τα κέρδη από την ανάλυση των καταναλωτικών προτιμήσεων του πελάτη, το ιστορικό των αγορών του, την τοποθεσία του κλπ. μέσω των προτεινόμενων προϊόντων είναι τεράστια. Η Amazon ανέφερε ότι το 30% των πωλήσεων της είναι πλέον από τα προτεινόμενα προϊόντα και από προτεινόμενα παρεμφερή προϊόντα που μπορούν να φανούν χρήσιμα από την προηγούμενη αγορά.

- **Στοχευμένο marketing.** Τα Big Data δίνουν την δυνατότητα για στοχευμένο marketing πάνω στις ανάγκες των ανθρώπων ανα χώρα, πόλη ακόμα και βάση γεωγραφικών περιοχών μέσα στην πόλη, αναλύοντας τις οικονομικές δυνατότητες των καταναλωτών, τις προτιμήσεις τους, τον μέσο όρο ηλικίας, τα ήθη και έθιμα κλπ.

- **Real-time 'ανάλυση' πελατών.** Πολλές εταιρίες αναλύουν real-time την συμπεριφορά των καταναλωτών όταν βρίσκονται μέσα στο κατάστημα

τους μέσα από σένσορες, από στοιχεία που βρίσκουν online, ακόμα και από το gps του κινητού που ενημερώνει τους ενδιαφερόμενους πόση ώρα αφιερώνουν ανά περιοχή του καταστήματος. Αυτό επιτρέπει στους πωλητές να έχουν μια καλύτερη εικόνα για τις προτιμήσεις των καταναλωτών και να υπάρχει η αντίστοιχη εξατομικευμένη προσφορά.

3.3.5) <u>Τοποθεσία και μεταφορά</u>

Μέσω του GPS που πλέον βρίσκεται ακόμα και στα κινητά μας ή στα αυτοκίνητα, είναι πολύ εύκολο να ξέρουν οι ενδιαφερόμενοι την ακριβή τοποθεσία του κάθε καταναλωτή. Τα προσωπικά στοιχεία χρησιμοποιούνται για να δημιουργήσουν ένα προφίλ του καταναλωτή σε σχέση με τα μέρη από τα οποία περνάει συνήθως, τα καταστήματα τα οποία επισκέπτεται κλπ. Επίσης εταιρίες που χρησιμοποιούν εφαρμογές για να παρέχουν βέλτιστες διαδρομές στους πελάτες της, μαζεύουν και αναλύουν ογκωδέστατα δεδομένα που προκύπτουν από την τοποθεσία των ανθρώπων κάθε χρονική στιγμή. Πολλές εφαρμογές ήδη χρησιμοποιούνται και άλλες βρίσκονται σε εξέλιξη και θα τις παραθέσω παρακάτω.

- **Βέλτιστη Διαδρομή.** Εφαρμογές σαν το γνωστό σε όλους μας google maps που αναλύοντας την κίνηση στους δρόμους καθορίζει την βέλτιστη διαδρομή από το σημείο Α στο Β. Η βελτίωση και η ακρίβεια που παρέχεται συνεχώς βελτιώνεται όσο οι real-time τεχνικές ανάλυσης των δεδομένων γίνονται ολοένα και αποτελεσματικότερες.

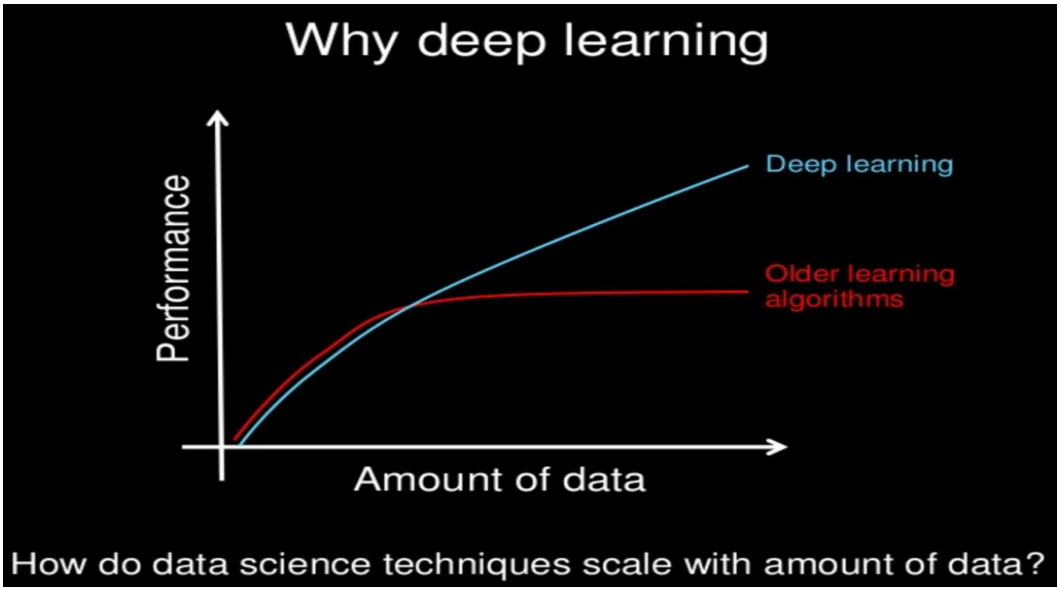
- **Έξυπνα αυτοκίνητα.** Υπάρχουν κάποια πρωτότυπα αυτοκίνητα (τα TESLA) τα οποία επιτρέπουν μέσα από μια σειρά από σένσορες στο αμάξι να κινείται αυτόνομα, αναλύοντας real-time σε microseconds τα δεδομένα από την άμεση περιοχή γύρο του. Αυτή η τεχνολογία αναπτύσσεται συνεχώς και βασίζεται στην πολύ γρήγορη ανάλυση των δεδομένων. Αναμένεται τα επόμενα χρόνια τα αυτοκίνητα στους δρόμους να είναι τελείως αυτόνομα.

- **Μεγαλύτερη ασφάλεια.** Από τα δεδομένα τοποθεσίας είναι ευκολότερο να εντοπιστεί ένας άνθρωπος με πρόβλημα υγείας ή να εντοπιστεί από τα όργανα ασφαλείας σε περίπτωση που διατρέχει κίνδυνο.

3.4) DEEP LEARNING

Το Deep Learning είναι νέα πρωτοποριακή τεχνολογία έχοντας πολλά κοινά στοιχεία με το Machine Learning, για την ακρίβεια θεωρείται η επιτομή της τεχνολογικής εξέλιξης του, και δημιουργήθηκε έχοντας ως πρότυπο, την λειτουργία του εγκεφάλου δηλαδή πως μεταφέρουν οι νευρώνες στον εγκέφαλο μας πληροφορίες μέσω συνάψεων. Το Deep Learning δημιουργήθηκε πάνω σε προϋπάρχουσες ιδέες για νευρωνικά δίκτυα (neural networks) στους υπολογιστές, και αυτό έχει καταστεί εφικτό από την υπολογιστική δύναμη που έχουμε σήμερα στην διάθεσή μας, και από τον τεράστιο όγκο των δεδομένων που έχουμε διαθέσιμο. Για την ακρίβεια η διαφορά του Deep Learning σε σχέση με τεχνικές στο πεδίο του Machine Learning βρίσκεται στην επεξεργασία ογκωδέστατων datasets. Έχει παρατηρηθεί ότι οι επιδόσεις Deep Learning αλγορίθμων με μικρά datasets είναι αισθητά χειρότερες από αυτές που επιτυγχάνονται με τους τωρινούς Machine Learning αλγορίθμους, αλλά όταν το dataset είναι μεγάλο, το Deep Learning ξεχωρίζει. (brownlee, 2016)

ΕΙΚΟΝΑ 12) Deep Learning



Υπάρχει η αντίληψη ότι το Deep Learning είναι κάτι που θα εφαρμοστεί στο μέλλον και ότι είναι σε περίοδο ερευνών αυτήν την στιγμή. Για την ακρίβεια όμως το Deep Learning χρησιμοποιείται ήδη σε πολλές εφαρμογές που χρειάζονται οι επιχειρήσεις στις μέρες μας. Μερικά από τα πεδία στα οποία εφαρμόζεται είναι:

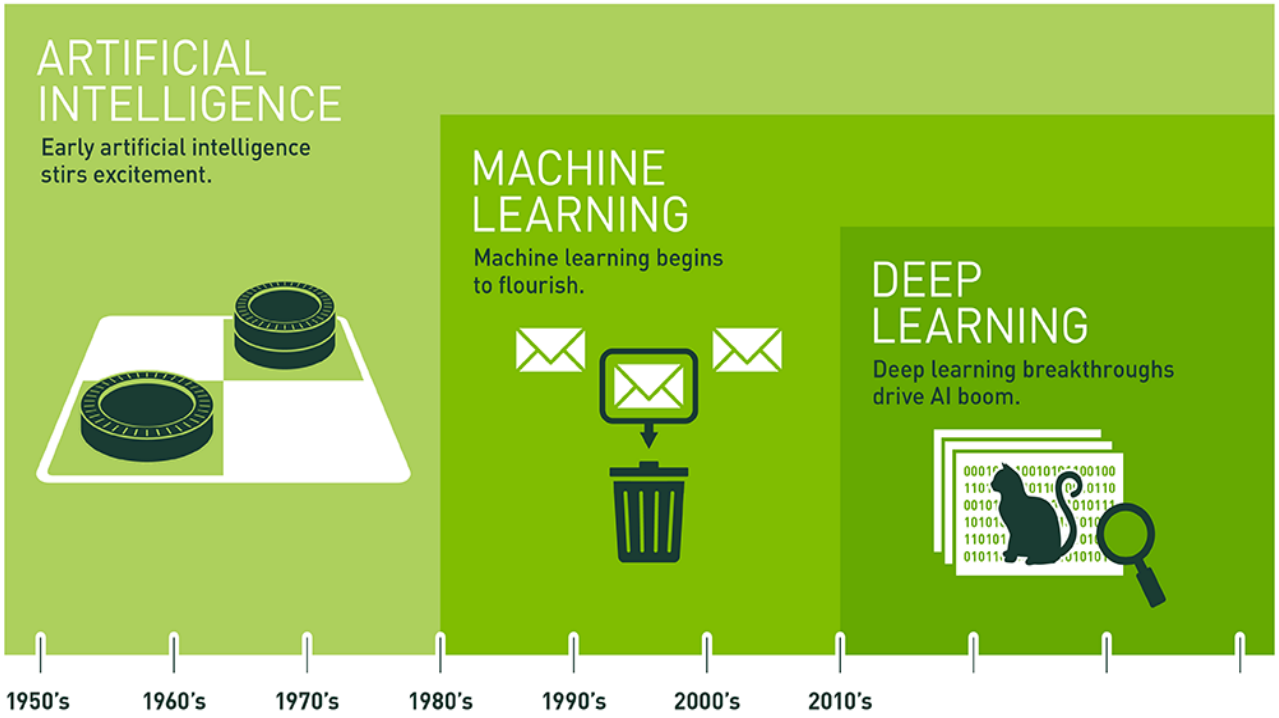
- **Αναγνώριση ομιλίας.** Τεχνικές και αλγόριθμοι προϋπήρχαν πριν την εφαρμογή του Deep Learning, η διαφορά όμως βρίσκεται στην πολύ καλύτερη αποτελεσματικότητα που παρέχει στην αναγνώριση της ανθρώπινης φωνής και των φωνητικών διαφορών ανά γλώσσα. Μερικές από της εταιρίες που το χρησιμοποιούν είναι το Skype, η Google, η Microsoft (Xbox), και η Apple (Siri).

- **Αναγνώριση εικόνας.** Οι αλγόριθμοι είναι σε θέση να αναγνωρίζουν τις εικόνες που προβάλλονται σε μια φωτογραφία, ακόμη και να περιγράφουν το τοπίο της. Αυτό ήδη έχει τρομερή ζήτηση από τις εγκληματολογικές υπηρεσίες για την ανάλυση φωτογραφιών από κάμερες και περαστικούς στον τόπο του εγκλήματος. Επίσης αυτό εφαρμόζεται μερικώς και στα πλήρως αυτόνομα αυτοκίνητα μέσω χρήσης καμερών με οπτικό πεδίο 360 μοιρών.

- **Ανάλυση φυσικής γλώσσας.** Αυτή η τεχνική χρησιμοποιείται χρόνια μέσω Machine Learning αλγορίθμων για την καταγραφή και ανάλυση παραπόνων από καταναλωτές, από άρθρα σε blogs, από τα social media κλπ. και παρατηρήθηκε μεγαλύτερη αποτελεσματικότητα με την τεχνική του Deep Learning

- **Συστήματα προτεινόμενων προϊόντων ή ταινιών.** Αυτό είναι το πείραμα στην εργασία μου και αφορά την σύσταση προϊόντων ή ταινιών, σειρών κλπ. βάση ανάλυσης του ιστορικού προτιμήσεων του κάθε καταναλωτή-χρήστη, σε σχέση με άλλους ανθρώπους που είχαν τις ίδιες προτιμήσεις στο παρελθόν. Με το Deep Learning αυτό πάει ένα βήμα παραπέρα, καθώς θα χρησιμοποιείται και σε ποιο πολύπλοκα περιβάλλοντα όπως να προτείνει μουσική και ρουχισμό.

ΕΙΚΟΝΑ 13) <u>Η εξέλιξη της τεχνητής νοημοσύνης</u>



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

3.4.1) <u>Neural Networks</u>

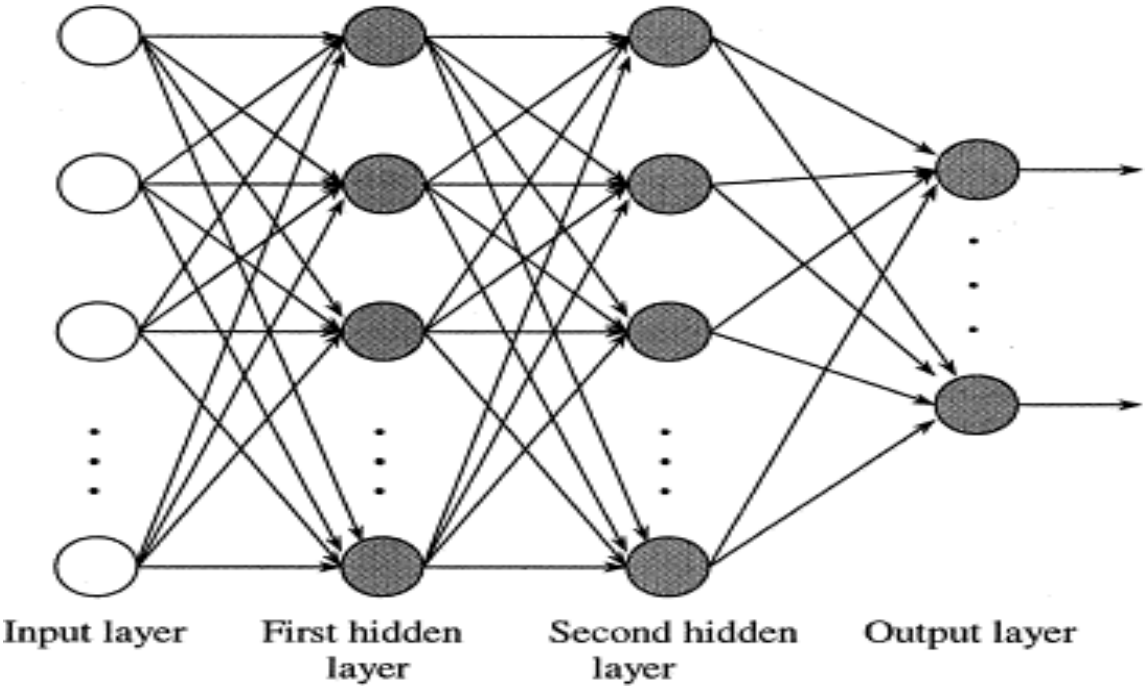
Τα νευρωνικά δίκτυα είναι η τεχνική στην οποία βασίζεται το Deep Learning. Ο τρόπος με τον οποίον λειτουργεί ένα νευρωνικό δίκτυο μοιάζει πολύ με τον τρόπο που λειτουργουν οι νευρώνες του εγκεφάλου, για την ακρίβεια αυτό αποτέλεσε την έμπνευση για να δημιουργηθεί κάτι παρόμοιο στους υπολογιστές.

Σαν ιδέα υπήρχε από την δημιουργία των πρώτων υπολογιστών, όμως η χαμηλή υπολογιστική ισχύς αλλά και τα ελάχιστα δεδομένα (σε σχέση με την εποχή μας) που παραγόντουσαν τότε, ήταν αποτρεπτικό από τον να λειτουργήσει σε πρακτικό επίπεδο.

Τα τελευταία όμως χρόνια με την αύξηση της δύναμης των υπολογιστών (ακόμη και των προσωπικών) και των διαθέσιμων ογκωδέστατων δεδομένων, η θεωρία έγινε πράξη, και τα νευρωνικά δίκτυα αποτελούν μια σειρά από τεχνικές που ανήκουν στην κατηγορία του Deep

Learning. Ο τρόπος που λειτουργούνε μιμείται τον τρόπο λειτουργίας των νευρώνων του εγκεφάλου δηλαδή κάθε νευρώνας δέχεται τα δεδομένα, τα επεξεργάζεται και στέλνει τα δεδομένα μέσω συνάψεων στην επόμενη συστάδα νευρώνων όπου ακολουθείται η ίδια διαδικασία. Τα layers νευρώνων, όπως αποκαλούνται, μπορούν να είναι 'άπειρα' αλλά κάθε layer έχει μεγάλη επίδραση στην δυνατότητα επεξεργασίας από τον υπολογιστή.

ΕΙΚΟΝΑ 14) <u>Τρόπος λειτουργίας των Neural Networks</u>



Τα Neural Networks και το Deep Learning προσπαθεί να αλλάξει τον τρόπο που δίνουμε στον υπολογιστή τις εντολές για να λύσει ένα πρόβλημα. Ο τρόπος υπολογισμού κάθε προβλήματος δεν πηγαίνει γραμμή-γραμμή όπως σε ένα καθιερωμένο κώδικα, αλλά επιλύονται όλα μαζί τα κομμάτια του προβλήματος, γι αυτό τον λόγο υπερέχουν σε περιοχές προβλημάτων όπου η λύση ενός προβλήματος είναι δύσκολο να εκφραστεί μέσα από ένα παραδοσιακό πρόγραμμα υπολογιστή. Με λίγα λόγια ένα νευρωνικό δίκτυο προσπαθεί να λύσει ένα πρόβλημα με τον ίδιο τρόπο που λύνει ένα πρόβλημα ο ανθρώπινος εγκέφαλος. Οι εφαρμογές στις οποίες χρησιμοποιούνται και παράγουν καλύτερα αποτελέσματα βασίζονται κυρίως στην επαρκή ποσότητα δεδομένων και σε πολύ γρήγορους υπολογισμούς. Μερικές από αυτές τις εφαρμογές είναι: (Wikepedia, 12)

- **Έλεγχος αυτόνομων συστημάτων**

- **Κβαντική χημεία**

- **Παιχνίδια με πολλαπλούς πιθανούς συνδυασμούς.** (σκάκι, τάβλι, πόκερ)

- **Αναγνώριση προτύπων.** (ραντάρ, αναγνώριση εικόνας, αναγνώριση προσώπου, ίριδας ματιού, δαχτυλικών αποτυπωμάτων κλπ.)

- **Ιατρικές διαγνώσεις**

- **Data Mining** (σε πολύ μεγάλα dataset >1T φαίνεται η διαφορά)

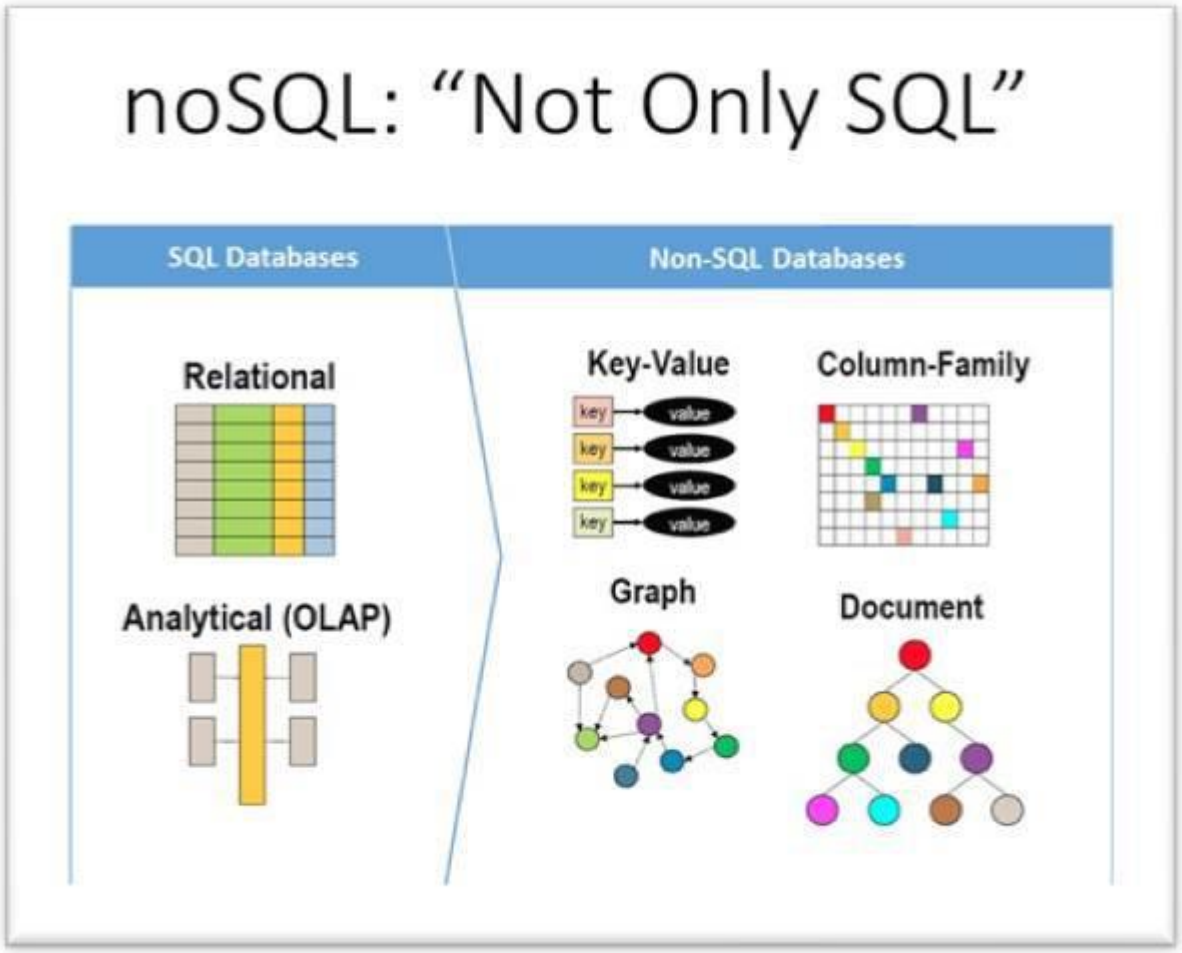
ΚΕΦΑΛΑΙΟ $4^{ο}$: Η ΤΕΧΝΟΛΟΓΙΑ ΣΤΗΝ ΔΙΑΘΕΣΗ ΤΩΝ BIG DATA

Τα Big Data δεν θα μπορούσαν να υπάρχουν αν δεν είχαν δημιουργηθεί τεχνολογίες που θα έκαναν γρήγορη και εύχρηστη την προσπέλαση τους. Οι τεχνολογίες των Big Data μπορούν να χωριστούν σε δύο μεγάλες κατηγορίες: τα συστήματα που παρέχουν επιχειρησιακές δυνατότητες σε πραγματικό χρόνο, όπου τα δεδομένα δεσμεύονται και αποθηκεύονται και συστήματα τα οποία παρέχουν την δυνατότητα για πολύπλοκες αναλύσεις δεδομένων τα οποία μπορούν να προσπελάσουν σχεδόν το σύνολο των δεδομένων. Αυτές οι δύο κατηγορίες τεχνολογίας είναι συμπληρωματικές και συνήθως αναπτύσσονται μαζί. Αξίζει να σημειωθεί ότι δαπανώνται τεράστια ποσά σε έρευνες για νέες τεχνολογίες και τεχνικές στων χώρο των Big Data, οπότε αυτές που έχω αναφέρει τώρα μπορεί να αντικατασταθούν από άλλες σε 3-4-5 χρόνια. (όπως έγινε με Hadoop και Apache Spark που θα αναλύσω στην συνέχεια.)

4.1) BIG DATA ΚΑΙ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Για να μπορέσουμε να αξιοποιήσουμε τις δυνατότητες που μας παρέχουν οι τεχνικές που αναπτύσσονται για τα Big Data, θα πρέπει να έχουμε ογκωδέστατα αρχεία για προσπέλαση. Αυτό από μόνο του δεν αρκεί καθώς πέρα από την χωρητικότητα που θα πρέπει να έχουμε στην διάθεση μας, θα πρέπει να έχουμε και τεχνολογίες που να μας επιτρέπουν να αυξάνουμε την ταχύτητα προσπέλασης των αρχείων καθώς και τον αποθηκευτικό χώρο, επειδή τα δεδομένα αυξάνονται δραματικά κάθε μέρα που περνάει. Εκεί έγκειται το πρόβλημα με τις SQL databases (MySQL, Oracle, SQLite, Postgres) . Οι SQL Databases είναι φτιαγμένες ώστε να είναι κάθετα επεκτάσιμες, σε σχέση με τις NoSQL (MongoDB, BigTable, Redis, RavenDB, Cassandra, HBase, Neo4j, CouchDB) που επεκτείνονται οριζόντια. Αυτό έχει ως αποτέλεσμα οι SQL βάσεις δεδομένων να χρειάζονται περισσότερη υπολογιστική ισχύς για την παραγωγή αποτελεσμάτων, ανάλογη με την αύξηση του όγκου των δεδομένων. Σε αντίθεση με την χρησιμοποίηση NoSQL βάσεων, απλά αυξάνουμε τους servers ώστε να διαμοιραστεί ο φόρτος εργασίας και να αυξηθεί η ταχύτητα προσπέλασης και παραγωγής αποτελεσμάτων, χωρίς την παράλληλη αύξηση της υπολογιστικής δύναμης. (ISSAC, 2014)

ΕΙΚΟΝΑ 15) <u>SQL vs NoSQL</u>



Τα συστήματα που χρησιμοποιούν NoSQL για Big Data έχουν σχεδιαστεί έτσι ώστε να εκμεταλλευτούν το cloud computing (θα αναλυθεί ποιο κάτω), το οποίο επιτρέπει μαζικούς πολύπλοκους υπολογισμούς χωρίς την αύξηση της υπολογιστικής δύναμης με γρήγορα αποτελέσματα. Αυτό μειώνει τον χρόνο και τον φόρτο εργασίας σε έναν Data Scientist, και έτσι μπορεί να διαχειριστεί καλύτερα και ταχύτερα περισσότερα δεδομένα σε λιγότερο χρόνο.

Εκτός από την προσπέλαση των δεδομένων από τον χρήστη για την παραγωγή αποτελεσμάτων, τα περισσότερα συστήματα έχουν φτιαχτεί έτσι ώστε να παρέχουν αποτελέσματα σε δεδομένα που δέχονται σε πραγματικό χρόνο. Για παράδειγμα χρησιμοποιώντας μια εφαρμογή για το χρηματιστήριο, τα συστήματα καθώς δέχονται real-time δεδομένα, ενημερώνουν τον χρήστη για τις επόμενες κινήσεις που πρέπει να κάνει. Κάποια NoSQL συστήματα μπορούν να

παρέχουν γνώσεις σχετικά με πρότυπα και τάσεις στο άμεσο μέλλον με απλούς κώδικες και χωρίς παρεμβάσεις από Data Scientists.

4.1.1) MongoDB

Μια NoSQL βάση δεδομένων που προτιμάτε αρκετά από τους Data Scientists (και εμένα προσωπικά) είναι η Mongo DataBase. Πως όμως ξέρουμε πότε θα προτιμήσουμε μια NoSQL βάση δεδομένων? Αυτό εξαρτάται καθαρά από το είδος των δεδομένων, τον όγκο, και τι ακριβώς θέλουμε να κάνουμε με αυτά.

Οι NoSQL βάσεις δεδομένων είναι ιδανικές όταν πρόκειται για μη δομημένα δεδομένα, όπως real-time δεδομένα από το χρηματιστήριο ή τα social media. Όπως έχω αναφέρει και προηγουμένως, τα δεδομένα στην εποχή μας είναι πολλά, συνήθως μη δομημένα και προσπαθούμε από αυτά να πάρουμε το καλύτερο δυνατό αποτέλεσμα. Αυτό οδήγησε στην δημιουργία των NoSQL βάσεων. Η δυνατότητα που παρέχουν για μεγαλύτερο αποθηκευτικό χώρο και ευκολότερη ανάλυση και διαχείριση των δεδομένων τις καθιστά απαραίτητες για κάθε Data scientist.

Από όλες τις διαθέσιμές NoSQL βάσεις δεδομένων, η Mongo θεωρείται από πολλούς η καλύτερη για μια σειρά από λόγους. Είναι open-source, εύχρηστη, και συμβατή με όλες τις Big Data τεχνολογίες ανάλυσης όπως Hadoop και Spark. Επίσης έχει φτιαχτεί με τέτοιο τρόπο ώστε να προσφέρει στις επιχειρήσεις γρήγορη και ευέλικτη πρόσβαση στα δεδομένα τους, είτε πρόκειται για real-time δεδομένα είτε δομημένα. Μερικές από τις πολυεθνικές που την χρησιμοποιούν είναι Bosch, MetLife, Expedia και άλλες. Αναλύτικα τα χαρακτηριστικά που την κάνουν προτιμητέα σε σχέση με άλλες SQL ή NoSQL βάσεις δεδομένων είναι:

- **Δυνατότητα αποθήκευσης μεγάλου όγκου δεδομένων, συνήθως μη δομημένων.** Οι συνηθισμένες SQL βάσεις δεδομένων αποθηκεύουν δομημένα δεδομένα όπως τα στοιχεία των πελατών μιας εταιρίας. Αλλά για ογκωδέστατα αδόμητα δεδομένα, όπως για τις προτιμήσεις του πελάτη, την τοποθεσία του, τις προηγούμενες αγορές του, τα πράγματα που του αρέσουν στο Facebook κλπ. μια NoSQL βάση επιτρέπει την προσθήκη διαφόρων ειδών δεδομένων. Και επειδή η Mongo είναι αρκετά προσιτή και εύκολη στην χρήση, επιτρέπει την αποθήκευση

των JSON αρχείων σ' ένα μέρος χωρίς να χρειάζεται να δηλωθούν οι τύποι των διαφορετικών δεδομένων εκ των προτέρων.

- **Αξιοποιεί στο έπακρο το cloud computing και storage.** Η αποθήκευση δεδομένων στο cloud έχει αποδειχθεί ότι είναι μια εξαιρετικά οικονομική λύση. Η Mongo δίνει την δυνατότητα να φορτώσει μεγάλους όγκους δεδομένων σε cloud και παρέχει την δυνατότητα για εύκολη πρόσβαση και ανάλυση τους σε μικρό χρονικό διάστημα καθώς μπορούμε να βάλουμε τα δεδομένα μας σε πολλούς servers και να διαμοιράσουμε έτσι την δουλεία μειώνοντας τον χρόνο ανάλυσης.

- **Επέκταση των βάσεων αποτελεσματικά και ανέξοδα.** Με την Mongo μπορούμε να μεταφέρουμε τα δεδομένα μας σε άλλους υπολογιστές στο δίκτυο μας, ή σε cloud, χωρίς να χρειάζεται επιπλέον λογισμικό.

Ποιο συγκεκριμένα η Mongo έχει στην διάθεση της εργαλεία που βοηθούν στην συλλογή και ανάλυση δεδομένων και μπορεί να ανταποκριθεί στις προκλήσεις των Big Data.

- **Ανάλυση δεδομένων βάση τοποθεσίας.** Αν ο Data Scientist χρειάζεται δεδομένα από συγκεκριμένες τοποθεσίες, η Mongo διαθέτει ενσωματωμένα εργαλεία που προσφέρουν την δυνατότητα για την συλλογή αυτών των δεδομένων από συγκεκριμένες τοποθεσίες χωρίς πολύπλοκες διαδικασίες εξαγωγής τους.

- **Εύκολη 'συνεργασία' με το Internet of Things (IoT).** Έχουμε την δυνατότητα να αναλύσουμε των τεράστιο όγκο δεδομένων που παράγουν οι σένσορες και οι διαδικτιακά συνδεδεμένες συσκευές, ανεξαρτήτου είδους, μέσα στην βάση.

- **Παρέχει ανάλυση σε πραγματικό χρόνο βάση της συμπεριφοράς των καταναλωτών.** Μπορούμε να προσωποποιήσουμε τα δεδομένα βάση της τοποθεσίας των καταναλωτών, τα μέρη που προτιμούν, τις προηγούμενες αγορές κτλπ. και να προβλέψουμε τις μελλοντικές τους κινήσεις. Η διαχείριση

τόσο πολλών διαφορετικών δεδομένων δεν πραγματοποιείται εύκολα σε SQL βάση δεδομένων. (MongoDB, 2017)

4.2) ΑΝΑΛΥΣΗ ΤΩΝ BIG DATA

Προφανώς δεν θα μπορούσαμε να συζητάμε για Big Data αν δεν υπήρχαν οι τεχνολογίες εκείνες οι οποίες θα μας βοηθούσαν να τα αναλύσουμε. Όταν λέμε για ανάλυση των Big Data εννοούμε την διαδικασία προσπέλασης των δεδομένων, για να αποκαλυφθούν κρυμμένα μοτίβα ή άγνωστοι συσχετισμοί μεταξύ μεταβλητών όπως και επίσης να μας βοηθήσουνε να κατανοήσουμε καλύτερα την τάση της αγοράς, τις προτιμήσεις των πελατών και άλλες χρήσιμες επαγγελματικές πληροφορίες. Τα ευρήματα αυτά μπορούν να οδηγήσουν σε ποιο αποτελεσματικό και στοχευμένο marketing, νέες επιχειρηματικές ευκαιρίες προς αξιοποίηση για την αύξηση των εσόδων, την καλύτερη εξυπηρέτηση των πελατών, την βελτίωση της ανταγωνιστικότητας έναντι άλλων εταιριών και καλύτερη και αποτελεσματικότερη εσωτερική λειτουργία της επιχείρησης.

Ο πρωταρχικός στόχος από την ανάλυση των Big Data είναι να βοηθήσει τις εταιρίες να λαμβάνουν πιο σωστές επιχειρηματικές αποφάσεις, μέσα από την ανάλυση τους από τους Data Scientists-Analysts, ανακαλύπτοντας πληροφορίες που δεν είναι δυνατό να βρεθούν από τα τωρινά προγράμματα των επιχειρήσεων. Αυτό, και ανάλογα τον τομέα κάθε εταιρίας, θα μπορούσε να σημαίνει, ανάλυση του περιεχομένου των κοινωνικών δικτύων, αγοραστικές προτιμήσεις καταναλωτών, δεδομένα από σένσορες, δεδομένα περιήγησης στο Internet, δεδομένα κινητής τηλεφωνίας και πολλά άλλα.

Όπως ανέφερα και στο προηγούμενο κεφάλαιο, τα ημι-δομημένα ή τα αδόμητα δεδομένα είναι πολύ δύσκολο να τα χειριστούμε με τις παραδοσιακές SQL βάσεις δεδομένων και αυτό επειδή δεν έχουν την δυνατότητα να επεξεργάζονται γρήγορα και αποτελεσματικά τις απαιτήσεις των Big Data που πρέπει να ενημερώνονται αρκετά συχνά ή ακόμη και συνεχώς αν τα δεδομένα είναι real-time όπως για παράδειγμα οι διακυμάνσεις των μετοχών στο χρηματιστήριο, νέες δημοσιεύσεις στα social media κλπ. Αυτό έχει ως αποτέλεσμα οι εταιρίες να στρέφονται για την συλλογή και την ανάλυση των Big Data σε νέες τεχνολογίες που είναι ικανές να ανταποκριθούν στις απαιτήσεις τους όπως YARN, MapReduce, Hive, Pig, NoSQL βάσεις δεδομένων και στην αιχμή του δόρατος βρίσκεται η Apache Spark, μια νέα τεχνολογία που έχει αποφέρει επαναστατικούς τρόπους ανάλυσεις τεραστίων datasets. Οι τεχνολογίες αυτές

αποτελούν τον πυρήνα ανοιχτού κώδικα λογισμικών που μπορούν να αντεπεξέλθουν στην επεξεργάσια των Big Data.

ΕΙΚΟΝΑ 16) Big Data τεχνολογίες

Overview of Operational vs. Analytical Systems

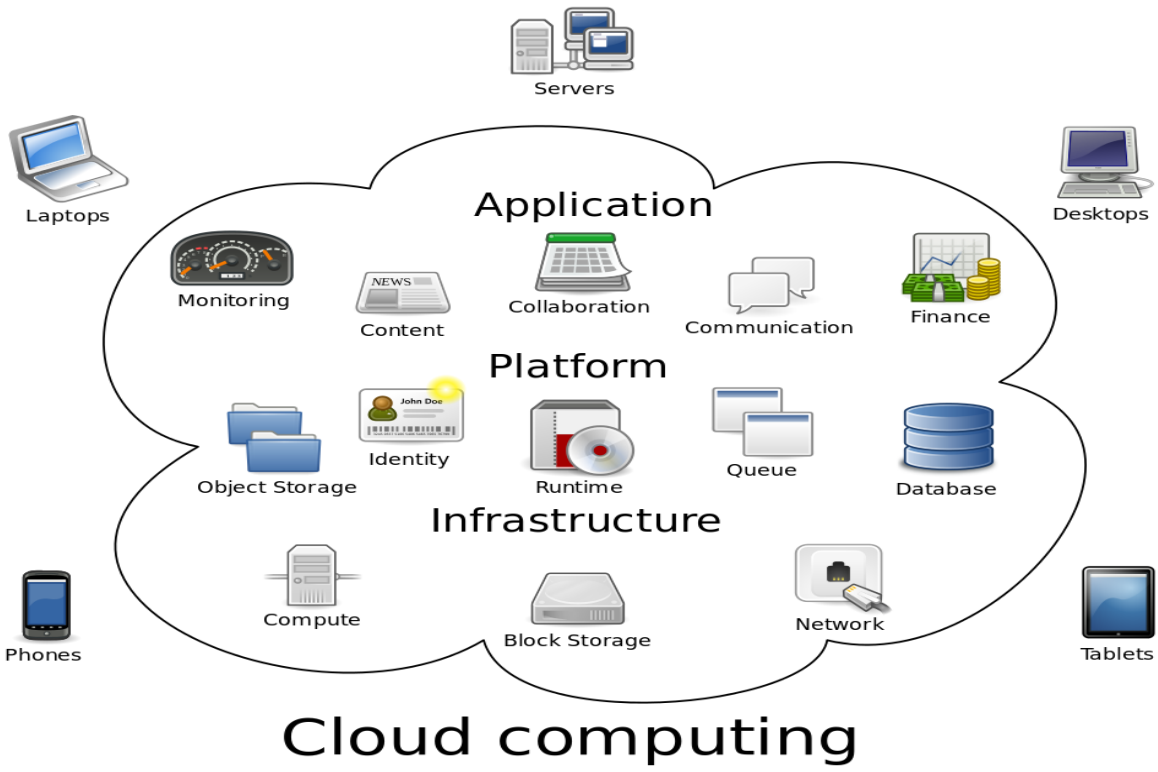
	Operational	Analytical
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	Hadoop, Apache Spark (sas, 2017)

4.2.1) Cloud Computing

Με το cloud computing αναφερόμαστε σε ένα ευρύ σύνολο υπολογιστών το οποίο παρέχεται σε εμάς σαν υπηρεσία ώστε να μπορούμε να εκτελέσουμε υπολογισμούς που θα ήταν αδύνατοι ή εξαιρετικά χρονοβόροι αν χρησιμοποιούσαμε μόνο τον προσωπικό μας υπολογιστή ή ακόμα και το δίκτυο υπολογιστών μιας εταιρίας. Ο χρήστης που θέλει να χρησιμοποιήσει τις δυνατότητες του cloud computing μπορεί να δεσμεύσει όση υπολογιστική δύναμη, μνήμη και χώρο αποθήκευσης χρειάζεται ώστε να πραγματοποιήσει την δουλειά του. Αυτό που κάνει το cloud computing να ξεχωρίζει είναι ότι ουσιαστικά δεν υπάρχει περιορισμός (επΙ πληρωμή) στο

πόσο χώρο ή υπολογιστική δύναμη μπορεί κάποιος να δεσμεύσει. Έτσι μπορούν να επιλυθούν ακόμα και τα ποιο απαιτητικά προβλήματα.

ΕΙΚΟΝΑ 17) <u>Cloud Computing</u>



Για να θεωρηθεί μια εργασία ως Cloud computing, χρειάζεται να παρέχουμε τα δεδομένα μας ή τα προγράμματα μας στο Internet. Ο χρήστης που χρησιμοποιεί το cloud δεν έχει την δυνατότητα να ξέρει τι συμβαίνει ακριβώς στην άλλη πλευρά, δηλαδή που βρίσκονται οι servers πως είναι η δομή του cloud κλπ. Όμως αυτό δεν χρειάζεται καθώς το αποτέλεσμα είναι το ίδιο αρκεί να υπάρχει μια σύνδεση στο Internet ανεξαρτήτου τοποθεσίας.

Αυτά ισχύουν για μεμονωμένους χρήστες ή μικρές επιχειρήσεις. Η χρήση του cloud από μεγάλες επιχειρήσεις και οργανισμούς είναι κάτι διαφορετικό. Συνήθως οι μεγάλες επιχειρήσεις χρησιμοποιούν το cloud σαν υπηρεσία δεσμεύοντας πολλούς πόρους (και πληρώνοντας πολλά λεφτά γι' αυτό). Επίσης μπορούν να δημιουργήσουν εφαρμογές στο Cloud και να είναι από εκεί προσβάσιμες σε όλους στην εταιρία. Τέλος πολύ μεγάλες εταιρίας σαν την Amazon έχουν φτιάξει το δικό τους cloud το οποίο νοικιάζουν σε άλλες μεγάλες εταιρίες (η

Netflix χρησιμοποιεί σαν πελάτης τις υπηρεσίες που προσφέρει η Amazon στο cloud). Το cloud computing ακούγεται νέο, αλλά πρόκειται για μια επιχειρηματική δραστηριότητα που παράγει περίπου 127 δισεκατομμύρια δολάρια το 2017. Μερικές από τις ποιο διαδεδομένες cloud υπηρεσίες είναι το Google Drive, I Cloud, Amazon Cloud Drive. (Griffith, 2016)

4.2.2) Hadoop

Με το Hadoop ήρθε η πρώτη μεγάλη επανάσταση στον τομέα της ανάλυσης των Big Data καθώς παρέχει όλα τα εργαλεία που χρειάζονται για την πραγματοποίηση γρήγορων και πολύπλοκων υπολογισμών είτε τοπικά (local mode, περισσότερος χρόνος προσπέλασης και υπολογισμού) είτε μέσω του cluster mode όπου ο κάθε ενδιαφερόμενος χρήστης ή εταιρία μπορεί να δημιουργήσει όσα nodes χρειάζεται ώστε να γίνουν οι υπολογισμοί γρηγορότερα (πολλές φορές η υπηρεσία αυτή είναι επί πληρωμή όπως της Amazon).

Ποιο συγκεκριμένα το Hadoop είναι ένα open-source, βασισμένο στην Java, λογισμικό, που επιτρέπει την επεξεργασία και την αποθήκευση πολύ μεγάλων αρχείων μέσω ενός κατανεμημένου υπολογιστικού περιβάλλοντος.

Με την χρήση του Hadoop δημιουργήθηκε για πρώτη φορά η δυνατότητα εκτέλεσης εφαρμογών σε συστήματα που χρησιμοποιούσαν χιλιάδες κόμβους για τις υπολογιστικές πράξεις, τα οποία είναι ικανά να επεξεργαστούν χιλιάδες Terabytes δεδομένων σε εύλογο (για το μέγεθος του dataset) χρονικό διάστημα. Το κατανεμημένο σύστημα αρχείων που διαθέτει το Hadoop διευκολύνει την γρήγορη μεταφορά τους μεταξύ των κόμβων και επιτρέπει στο σύστημα να συνεχίσει την λειτουργία του, ακόμη και αν χαθεί λόγω προβλήματος ένας ή περισσότεροι κόμβοι. Αυτή η προσέγγιση από τους αρχιτέκτονες του Hadoop μειώνει τον κίνδυνο καταστροφικής βλάβης του συστήματος, το οποίο θα οδηγούσε σε απώλεια κρίσιμων δεδομένων, ακόμη και αν μείνουν εκτός λειτουργίας αρκετοί κόμβοι του συστήματος. Αυτά τα πλεονεκτήματα που έχει η συγκεκριμένη τεχνολογία, την ανέδειξαν ως βασικό όπλο για την αποτελεσματική ανάλυση και εξαγωγή αποτελεσμάτων στα Big Data.

ΕΙΚΟΝΑ 18) <u>Hadoop</u>

Οι Data Scientists που χρησιμοποιούν το Hadoop, μπορούν να το χρησιμοποιήσουν στο τοπικό δίκτυο της εταιρίας τους ή στον προσωπικό τους υπολογιστή. Πρέπει όμως να επισημάνουμε ότι τα περισσότερα big data projects βασίζονται στην γρήγορη χρήση μεγάλης υπολογιστικής δύναμης. Γι' αυτό τον λόγο η πλειοψηφία των Data Scientists χρησιμοποιεί cloud εφαρμογές ώστε να τους παρέχονται ο χώρος και η υπολογιστική δύναμη που χρειάζονται για να τρέξουν ένα project. Μερικές από αυτές τις υπηρεσίες που είναι φτιαγμένες κυρίως για Big Data εργασίες είναι της Microsoft το Azure HDinsight, της Amazon το Elastic MapReduce και της Google το Cloud Dataproc. Το Hadoop έχει στην διάθεση του πολλά ενσωματωμένα εργαλεία που βοηθούν τους data scientists να εξοικονομούν χρόνο έχοντας όλα τα εργαλεία που χρειάζονται σε ένα μόνο interface. Μερικά από τα ποιο σημαντικά εργαλεία είναι: (technologies, 2014)

4.2.2.1) Hadoop Distributed File System (HDFS)

Το HDFS είναι η βασική μονάδα όπου αποθηκεύονται τα αρχεία από τις εφαρμογές μας στο Hadoop. Είναι ουσιαστικά ένα κατανεμημένο σύστημα αποθήκευσης αρχείων που παρέχει υψηλή αποδοτικότητα στην πρόσβαση των αρχείων σε όλα τα clusters που έχουμε δημιουργήσει για την ανάλυση των datasets. Το HDFS θεωρείται η βασική τεχνολογία για την διαχείριση των Big Data αρχείων και για την υποστήριξη εφαρμογών για την ανάλυση τους.

Το HDFS είναι σχεδιασμένο έτσι ώστε να μπορούν να το χρησιμοποιήσουν ακόμη και σε προσωπικούς υπολογιστές οι Data Scientists χωρίς δηλαδή να χρειάζονται τρομερές απαιτήσεις χωρητικότητας και υπολογιστικής δύναμης. Όπως ανέφερα και προηγουμένως, έχει την δυνατότητα να λειτουργεί και να μεταφέρει με μεγάλη ταχύτητα από τον ένα κόμβο στον άλλον ακόμα και αν αρκετοί κόμβοι από το cluster αχρηστευτούν.

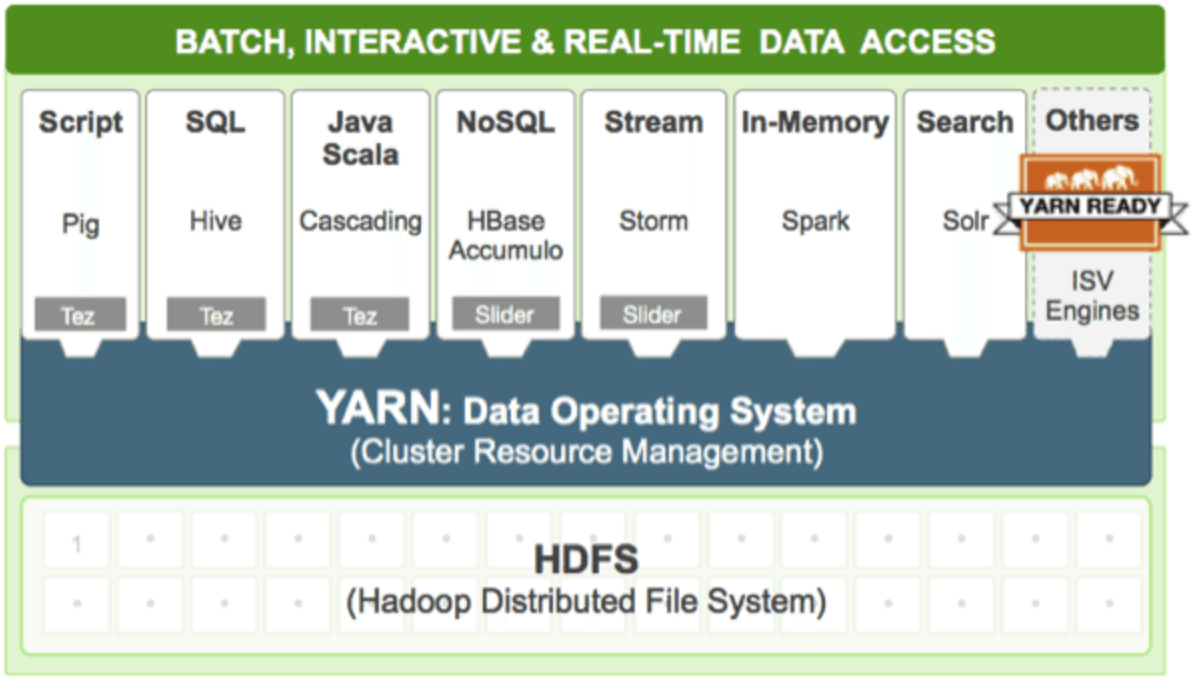
Όταν φορτώνουμε ένα αρχείο στο HDFS, αυτό διαμοιράζει τις πληροφορίες σε πολλά ατομικά κομμάτια και τις προωθεί σε όλους τους διαθέσιμους κόμβους του cluster που έχουμε δημιουργήσει συνήθως από μια cloud εφαρμογή. Αυτό επιτρέπει την παράλληλη προσπέλαση και επεξεργασία των αρχείων, μειώνοντας σημαντικά των χρόνο που χρειάζεται για την παραγωγή αποτελεσμάτων. Επίσης αντιγράφει τις πληροφορίες που περιέχει το κάθε node και τις διαμοιράζει σε άλλα nodes ώστε αν ένα σταματήσει να λειτουργεί, η πληροφορία θα έχει σωθεί και θα συνεχιστεί η επεξεργασία της σε άλλο node.

Το HDFS έχει φτιαχτεί έτσι ώστε να υποστηρίζει αρχεία πολλών Terabytes. Χρησιμοποιεί την αρχιτεκτονική master/slave που με λίγα λόγια σε κάθε εργασία που θα τρέχει θα υπάρχει ένας master και σε αυτόν φτιάχνουμε όσα slaves χρειάζονται για την μείωση του χρόνου επεξεργασίας. Στο κάθε slave διαμοιράζονται εργασίες οι οποίες τρέχουν παράλληλα και επιστρέφει η κάθε μια το αποτέλεσμα της όταν τελειώσει η επεξεργασία.

4.2.2.2) Apache Hadoop YARN (Yet Another Resource Negotiator)

Το YARN είναι μια τεχνολογία διαχείρισης των clusters. Είναι τεχνολογία κορμού για το Hadoop καθώς ελέγχει την ομαλή λειτουργία των συστημάτων που αναλύουν τα αρχεία. Αποτελείται από έναν διαχειριστή ο οποίος δέχεται πληροφορίες για την λειτουργία των clusters, μέχρι το τελευταίο node, και ελέγχει κατά πόσο οι κόμβοι σε κάθε cluster είναι αποτελεσματικοί (δηλαδή πόσα nodes λειτουργούν σε κάθε cluster) ώστε να ελέγχει την ταχύτητα επεξεργασίας των αρχείων και να διαμοιράζει το φορτίο από clusters με λιγότερα ενεργά nodes σε clusters με περισσότερα.

ΕΙΚΟΝΑ 19) Η αρχιτεκτονική του Hadoop



4.2.2.3) Map Reduce

Η τελευταία βασικότατη τεχνολογία που υπάρχει στο Hadoop είναι το Map Reduce. Όπως ανέφερα προηγουμένως το Hadoop επιτρέπει την κατανεμημένη επεξεργασία δομημένων ή αδόμητων δεδομένων τα οποία διαμοιράζει σε πολλά clusters στα οποία κάθε κόμβος του cluster

περιέχει τον δικό του αποθηκευτικό χώρο. Το Map Reduce έρχεται να προσθέσει δύο πολύ βασικά στοιχεία στην λειτουργία του.

Το πρώτο είναι ο διαμοιρασμός των εργασιών σε διαφόρους κόμβους των clusters και ο δεύτερος είναι η οργάνωση και η διαχείριση των αποτελεσμάτων που παράγεται από κάθε κόμβο, και κατ' επέκταση από κάθε cluster, σε μια κατανοητή ως προς τον υπολογιστή και τον προγραμματιστή απάντηση.

Αποτελείται από διάφορα εργαλεία όπως:

- **JobTracker.** Είναι το κύριο (master) node, το οποίο αναλαμβάνει την διαχείριση όλων των εργασιών σε ένα cluster.

- **TaskTrackers.** Είναι agents που βρίσκονται σε κάθε κόμβο, σε κάθε cluster που έχουμε δημιουργήσει για την εργασία, και αναλαμβάνουν την επεξεργασία και την αποστολή των αποτελεσμάτων από κάθε κόμβο.

- **JobHistoryServer.** Ένα εργαλείο το οποίο βρίσκει τους κόμβους στους οποίους έχουν ολοκληρωθεί οι εργασίες που τους ανατέθηκαν.

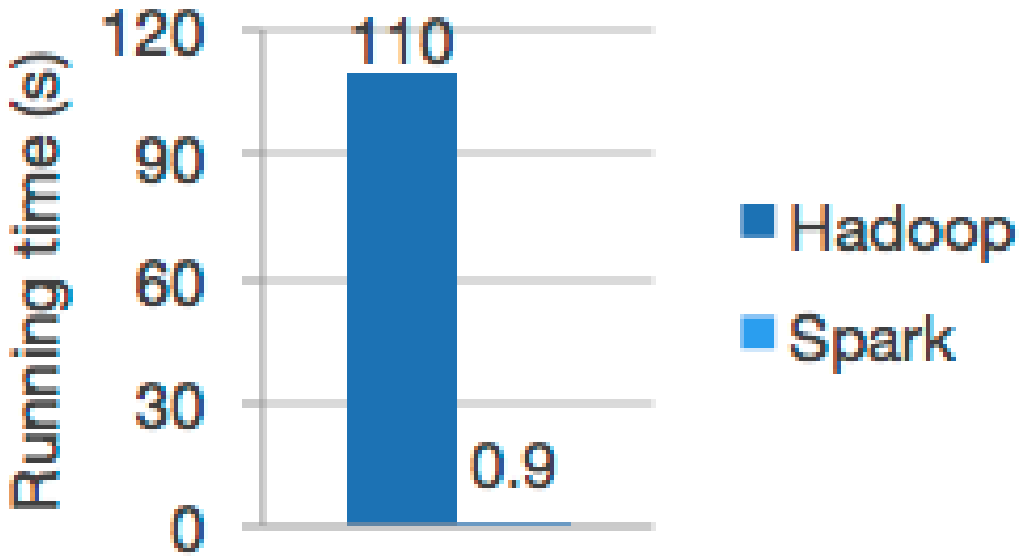
Για τον διαμοιρασμό των δεδομένων και την συλλογή αποτελεσμάτων, το MapReduce λειτουργεί παράλληλα σε όσα clusters χρησιμοποιούμαι. Από την στιγμή που το μέγεθος των clusters δεν επηρεάζει το τελικό αποτέλεσμα που θα παραχθεί, οι εργασίες μπορούν να διαμοιράζονται σε όσους αριθμούς από servers θέλουμε. Έτσι το MapReduce και κατ' επέκταση το Hadoop απλοποιεί πάρα πολύ την ανάπτυξη του λογισμικού. Το MapReduce είναι διαθέσιμο σε μια σειρά από προγραμματιστικές γλώσσες όπως C, C++, Java, Ruby, Perl και Python. Οι προγραμματιστές μπορούν να χρησιμοποιήσουν τις βιβλιοθήκες του MapReduce χωρίς να χρειάζεται να ασχοληθούν με την επικοινωνία και την συνεργασία μεταξύ των κόμβων.

4.2.3) Apache Spark

Η Apache Spark είναι η τελευταία λέξη της τεχνολογίας στον τομέα της ανάλυσης των Big Data (και η τεχνολογία που χρησιμοποίησα στην εργασία μου). Είναι και αυτή, όπως και το Hadoop ένα open source λογισμικό που διακρίνεται για την ταχύτητα του, την ευκολία στην χρήση του αλλά και την αποτελεσματική ανάλυση των Big Data.

Η Spark έχει πολλά πλεονεκτήματα σε σχέση με άλλες μεγάλες τεχνολογίες όπως Hadoop και Storm. Αρχικώς, η Spark δίνει ένα ολοκληρωμένο, ενιαίο πλαίσιο για την διαχείριση των τεράστιων απαιτήσεων επεξεργασίας και ανάλυσης των Big Data, σε όλο το φάσμα των δεδομένων (δηλαδή δεδομένα κειμένου, δεδομένα γραφημάτων, real-time δεδομένα κλπ.). Έχει αποδειχθεί ότι οι εφαρμογές που τρέχουν με Spark σε clusters είναι έως 100 φορές γρηγορότερες στην μνήμη και 10 φορές γρηγορότερες ακόμα και αν τρέχουν στον δίσκο. (spark, 2017)

ΕΙΚΟΝΑ 20) Spark vs Hadoop



Με την Spark έχουμε την δυνατότητα να φτιάξουμε τους αλγορίθμους μας σε μια σειρά από γλώσσες όπως Java, Python, Scala, R. Επιπλέον υποστηρίζει εντολές από SQL, live streaming δεδομένα, Machine Learning (διαθέτει και δική της ML βιβλιοθήκη), καθώς και επεξεργασία δεδομένων από γραφήματα. Οι Data Scientists μπορούν να χρησιμοποιήσουν αυτές τις εφαρμογές είτε αυτόνομα ή σε συνδυασμό ανάλογα με το project που τρέχουν.

Η Spark είναι σχεδιασμένη έτσι ώστε οι εργασίες Map και Reduce να πραγματοποιούνται με όσο το δυνατόν λιγότερες προσπελάσεις στο εκάστοτε αρχείο, μειώνοντας έτσι εκθετικά τον χρόνο επεξεργασίας και ανάλυσης. Επίσης στην γρήγορη απόδοση που διαθέτει η Spark βοηθάνε και στοιχεία όπως η δυνατότητα που έχουμε για αποθήκευση δεδομένων στην μνήμη και σχεδόν real-time επεξεργασία δεδομένων. Άλλο ένα σημαντικό στοιχείο που διαθέτει η Spark είναι ότι μπορεί να αποθηκεύσει αποτελέσματα στην μνήμη αντί να τα αποθηκεύει όλα στον δίσκο, και αυτό είναι πολύ χρήσιμο όταν χρειάζεται να δουλέψουμε στο ίδιο dataset πολλές φορές. Δηλαδή είναι έτσι σχεδιασμένη ώστε να επεξεργάζεται τα στοιχεία που της παρέχουμε χρησιμοποιώντας παράλληλα και την μνήμη και τον δίσκο. Ο τρόπος που είναι σχεδιασμένη να λειτουργεί είναι να αποθηκεύει όσο γίνεται περισσότερα δεδομένα στην μνήμη και τα υπόλοιπα στον δίσκο.

Πέρα από τον βασικό της σχεδιασμό και τα ενσωματωμένα χαρακτηριστικά που διαθέτει, υπάρχουν επιπλέον βιβλιοθήκες που μπορούν να χρησιμοποιηθούν από την Spark και προσφέρουν πολύ σημαντικές δυνατότητες (όπως θα δείτε και στο πείραμα που έχω κάνει) στην ανάλυση των Big Data και σε Machine Learning αλγορίθμους. Αυτές οι βιβλιοθήκες περιλαμβάνουν:

- **Spark streaming.** Τεχνολογία που επιτρέπει την επεξεργασία δεδομένων σε real-time χρόνο, χωρίς να υπάρχει περιορισμός στο μέγεθος των εισερχόμενων δεδομένων (αρκεί να έχουμε φτιάξει ένα αρκετά μεγάλο cluster).

- **Spark SQL.** Δίνει την δυνατότητα χρησιμοποίησης εντολών SQL σε Spark datasets. Μπορούμε επίσης να μετατρέψουμε τα δεδομένα από διαφορετικά formats ώστε να ανταποκρίνονται στην συγκεκριμένη τεχνολογία.

- **Spark MLlib.** Η Spark διαθέτει την δική της βιβλιοθήκη για Machine Learning χρησιμοποιώντας αλγορίθμους και υπηρεσίες που ήδη υπάρχουν, μετατρέποντας τες σε Spark. Οι αλγόριθμοι αφορούν μια γκάμα Machine Learning κατηγοριών όπως Classification, Clustering, Regression, Collaborative Filtering, Optimization algorithms κ.α.

- **Spark GraphX.** Το API της Spark για γραφήματα. Παρέχει αλγορίθμους για παράλληλη επεξεργασία γραφημάτων, με τον ίδιο τρόπο που πραγματοποιούνται οι αναλύσεις των datasets, οι οποίοι βοηθούν στην ανάλυση των γράφων.

4.2.3.1) Resilient Distributed Datasets (RDDs)

Τα RDDs αποτελούν τα βασικό κορμό της Spark τεχνολογίας. Τα RDDs είναι σαν πίνακες σε μια βάση δεδομένων και μπορούν να αποθηκεύσουν όλους τους τύπους των εισερχόμενων δεδομένων καθώς τα RDDs αποθηκεύονται σε διαφορετικά μέρη στην μνήμη. Με αυτό τον τρόπο πραγματοποιείται η παράλληλη επεξεργασία των δεδομένων καθώς κάθε RDD εκτελείται αυτόνομα. Τα στοιχεία που διαθέτουν τα RDDs βρίσκονται στο όνομα και είναι:

- **Resilient.** Δηλαδή δεν μπορούν να σταματήσουν την διαδικασία επεξεργασίας λόγο προβλημάτων. Σε περίπτωση που υπάρχει βλάβη σε κάποιο node έχουν την δυνατότητα να ξαναϋπολογίσουν τα χαμένα δεδομένα.

- **Distributed.** Δηλαδή μοιράζουν τα εισερχόμενα δεδομένα σε πολλαπλούς κόμβους σε ένα cluster.

- **Dataset.** Μια συλλογή διαμοιρασμένων δεδομένων με συγκεκριμένες τιμές

Τα RDDs υποστηρίζουν δύο τύπους εφαρμογών: **Transformations** και **Actions**. Τα Transformations αφορούν εντολές που δεν επιστρέφουν μια συγκεκριμένη τιμή, αλλά ένα νέο RDD. Με άλλα λόγια τίποτα δεν υπολογίζεται όταν καλούμε μια transformation συνάρτηση, απλά χρησιμοποιούμε ένα υπάρχων RDD και επιστρέφει ένα καινούργιο RDD. Μερικές από τις συναρτήσεις αυτές είναι: filter, flatMap, reduceByKey, pipe και πολλές άλλες.

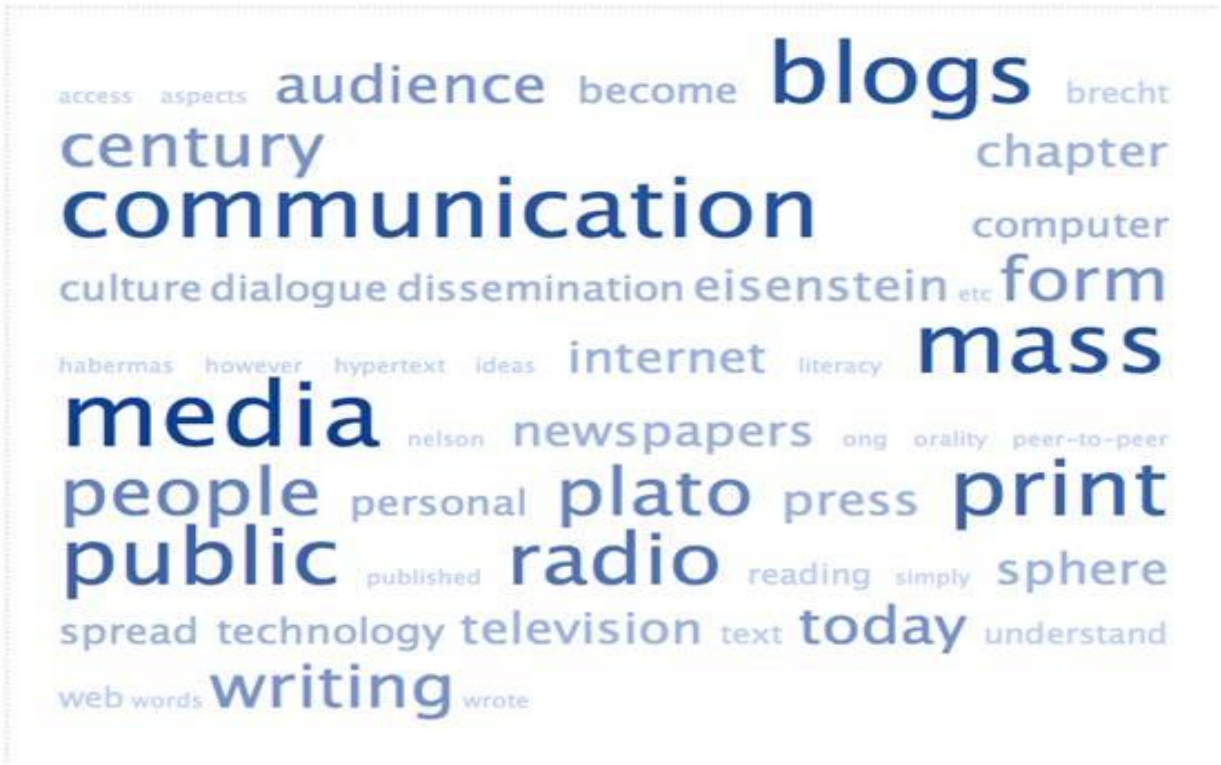
Τα actions αφορούν τις συναρτήσεις εκείνες που επιστρέφουν ύστερα από ανάλυση μια τιμή από ένα RDD. Μερικές από αυτές τις συναρτήσεις είναι: reduce, collect, count, first, take και άλλες.

4.3) VISUALIZATION

Ένα άλλο σημαντικό κομμάτι των Big Data είναι το visualization ή με άλλα λόγια η απεικόνιση των δεδομένων αφού έχουν επεξεργαστεί και αναλυθεί από τους αλγορίθμους μας. Η παρουσίαση των πληροφοριών που αποκομίζουμε από την ανάλυση των Big Data είναι σημαντικό να γίνεται όσο ποιο κατανοητή γίνεται στους ανθρώπους που χρειάζονται αυτά τα δεδομένα. Γι' αυτό τον λόγο αυτήν την στιγμή επενδύονται τεράστια ποσά και εκατοντάδες ώρες έρευνας στον τομέα της παρουσίασης των δεδομένων δηλαδή τεχνικές και τεχνολογίες για την δημιουργία εικόνων και γραφημάτων, ικανά να απεικονίσουν όσο καλύτερα γίνεται την ανάλυση εκατοντάδων Terabytes. Μερικές από αυτές τις τεχνικές είναι:

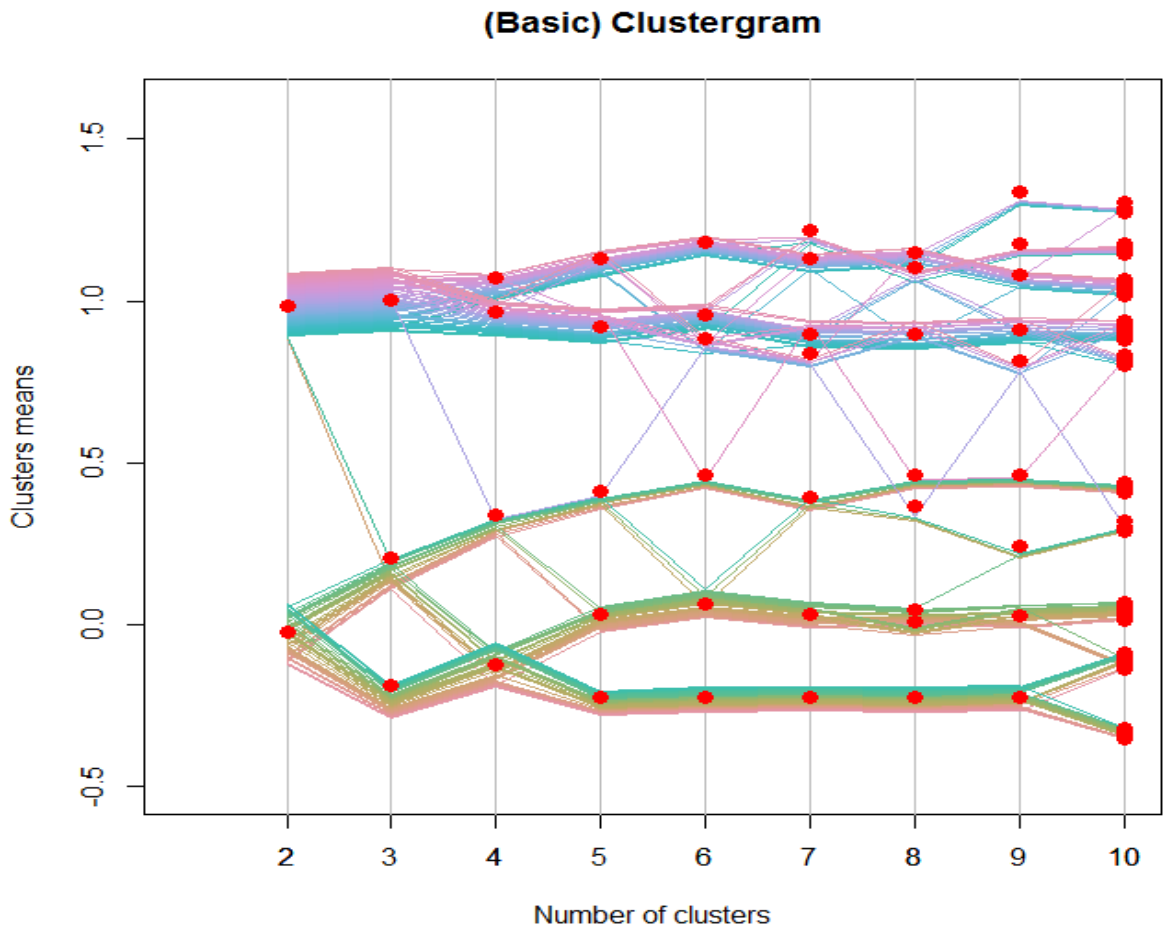
Tag cloud: Αυτή η τεχνική χρησιμοποιείται ύστερα από την ανάλυση κειμένων στα οποία μας ενδιαφέρει να μάθουμε ποιες λέξεις εμφανίζονται με την μεγαλύτερη συχνότητα. Οπότε η συγκεκριμένη τεχνική μας δείχνει σε μια εικόνα με έντονα και μεγαλύτερα γράμματα τις λέξεις αυτές. Έτσι δίνεται στον ενδιαφερόμενο να καταλάβει εύκολα και γρήγορα το τι πραγματεύεται ένα μεγάλο κομμάτι των κειμένων.

ΕΙΚΟΝΑ 21) Tag cloud



Clustergram: Το clustergram μας επιτρέπει να παρουσιάσουμε τα data points που ανήκουν σ' ένα cluster, όσο ο αριθμός των clusters αυξάνεται. Στα Big Data όταν αναλύουμε ένα αρχείο δεδομένων με την τεχνική του clustering, είναι πολύ σημαντικός ο αριθμός των clusters που χρησιμοποιούμε και πόσα data points αντιστοιχούν στο καθένα, οπότε η οπτικοποίηση του βοηθάει τα μέγιστα στην παραγωγή καλύτερων αποτελεσμάτων από των Data Scientist αλλά και στην ευκολότερη παρουσίαση των αποτελεσμάτων.

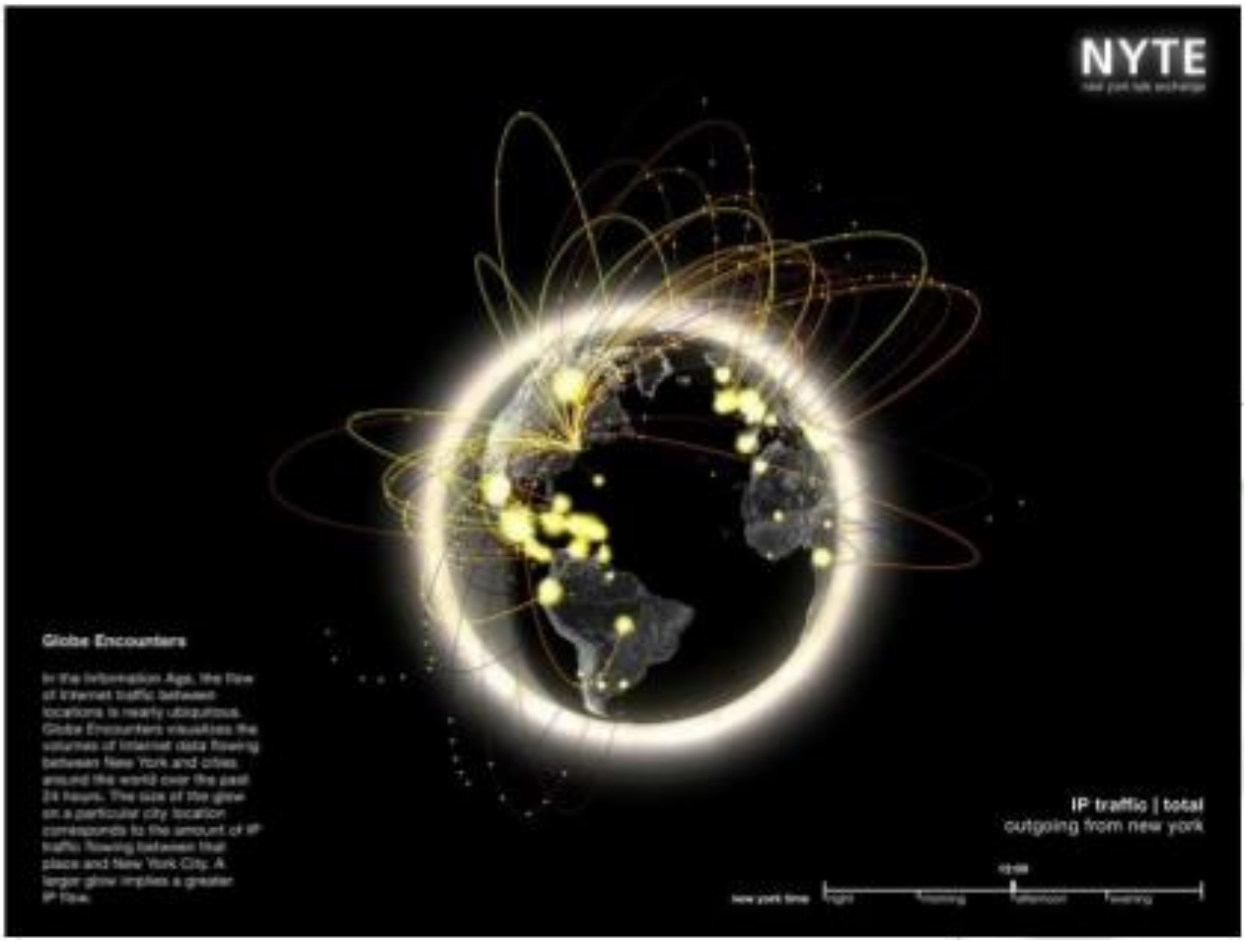
ΕΙΚΟΝΑ 22) Clustergram



Spatial Information Flow: Αυτή η τεχνική χρησιμεύει για την απεικόνιση πληροφοριών που σχετίζονται με την χωρική κατανομή τους. Για παράδειγμα στην παρακάτω εικόνα δείχνει την ποσότητα των data που 'ταξιδεύουν' σε άλλες χώρες και πόλεις με επίκεντρο την Νέα Υόρκη τις τελευταίες 24 ώρες. Όσο ποιο έντονη και φωτεινή είναι η κουκίδα στις πόλεις, σημαίνει ότι πραγματοποιήθηκε μεγαλύτερη ανταλλαγή δεδομένων. Αυτή η οπτικοποίηση των δεδομένων μας

επιτρέπει να καταλάβουμε εύκολα ποιες πόλεις είναι περισσότερο συνδεδεμένες με την Νέα Υόρκη στο πλαίσιο της επικοινωνίας και ανταλλαγής δεδομένων.

ΕΙΚΟΝΑ 23) Spatial information flow



4.4) ΓΛΩΣΣΕΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΓΙΑ BIG DATA

Όσο επεκτείνονται τα Big Data και γίνονται όλο και ποιο χρήσιμα στην καθημερινότητα των επιχειρήσεων, ακόμα και οι ποιο μικρές λεπτομέρειες παίζουν σημαντικό ρόλο στην ταχύτητα και στην αποτελεσματικότητα της επεξεργασίας τους. Οι κύριες γλώσσες που χρησιμοποιούνται αυτήν την στιγμή από Data Scientists είναι η Python, R, Scala, Java.

R

Η R είναι μια γλώσσα προγραμματισμού βασισμένη στην στατιστική και οι Data Scientists μπορούν να την μάθουν αν έχουν λίγο χρόνο για εκπαίδευση, ειδικότερα αν έχουν ήδη

εμπειρία με το Matlab, ή με SAS. Η R είναι ίσως η ποιο ισχυρή γλώσσα στον τομέα της ανάλυσης των δεδομένων, αλλά μόνο εκεί, δηλαδή σαν μια γλώσσα γενικού περιεχομένου μένει ποιο πίσω σε σχέση με τις άλλες γλώσσες. Αυτό που κάνουν αρκετοί Data Scientists είναι να φτιάχνουν ένα καλό μοντέλο ανάλυσης σε R και μετά να το μετατρέπουν σε Python η Scala πριν τρέξουν το πρόγραμμα. Η R δεν είναι αρκετά καλή για την παραγωγή κώδικα σε εταιρίες που χρησιμοποιούν μεγάλα συστήματα από clusters καθώς το debugging σε αυτήν την γλώσσα είναι τρομερά χρονοβόρο.

PYTHON

Η γλώσσα αυτή κερδίζει χρόνο με τον χρόνο όλο και μεγαλύτερο μερίδιο στην προτίμηση των Data Scientists (και προσωπικά η αγαπημένη μου γλώσσα για Big Data), και πάρα πολλοί Data Scientists είναι εξοικειωμένοι με αυτήν. Ο λόγος είναι ότι είναι πολύ εύκολη να την μάθει κάποιος, (ενδείκνυται από πολλά πανεπιστήμια για πρώτη γλώσσα εκμάθησης σε αρχάριους), και είναι πολύ ευκολότερο να διαβάσεις τον κώδικα της από οποιαδήποτε άλλη γλώσσα. Έχει μεγάλη υποστήριξη από έτοιμες βιβλιοθήκες για στατιστική ανάλυση ή για αλγορίθμους Machine Learning και Deep Learning, στο οποίο θεωρείται ιδανικότερη, και επιπλέον είναι η backup γλώσσα πίσω από την Scala για την Apache Spark.

SCALA

Η Scala μπαίνει στην λίστα κυρίως επειδή η νέα καινοτόμος τεχνολογία Apache Spark είναι γραμμένη σε αυτήν. Μοιάζει αρκετά στην σύνταξή της με την Python οπότε η εκμάθηση της δεν απαιτεί πολύ καιρό. Είναι ιδιαίτερα προτιμητέα στον χρηματοπιστωτικό κλάδο όπου οι εταιρίες χρειάζονται να δουλεύουν με τεράστια datasets διαφορετικών μεταβλητών. Στα αρνητικά της είναι ότι πολλές φορές ο κώδικας της Scala απλώνεται επειδή μπορείς να κάνεις το ίδιο πράγμα με πολλούς τρόπους, και πολλοί Data Scientists χρησιμοποιούν δύο φορές το μέγεθος του κώδικα τον οποίον χρειάζεται ένας Data Scientist που χρησιμοποιεί Python. Επίσης είναι ελάχιστα ποιο αργή από τις υπόλοιπες γλώσσες (σε datasets εκατοντάδων Terabytes μόνο φαίνεται η διαφορά).

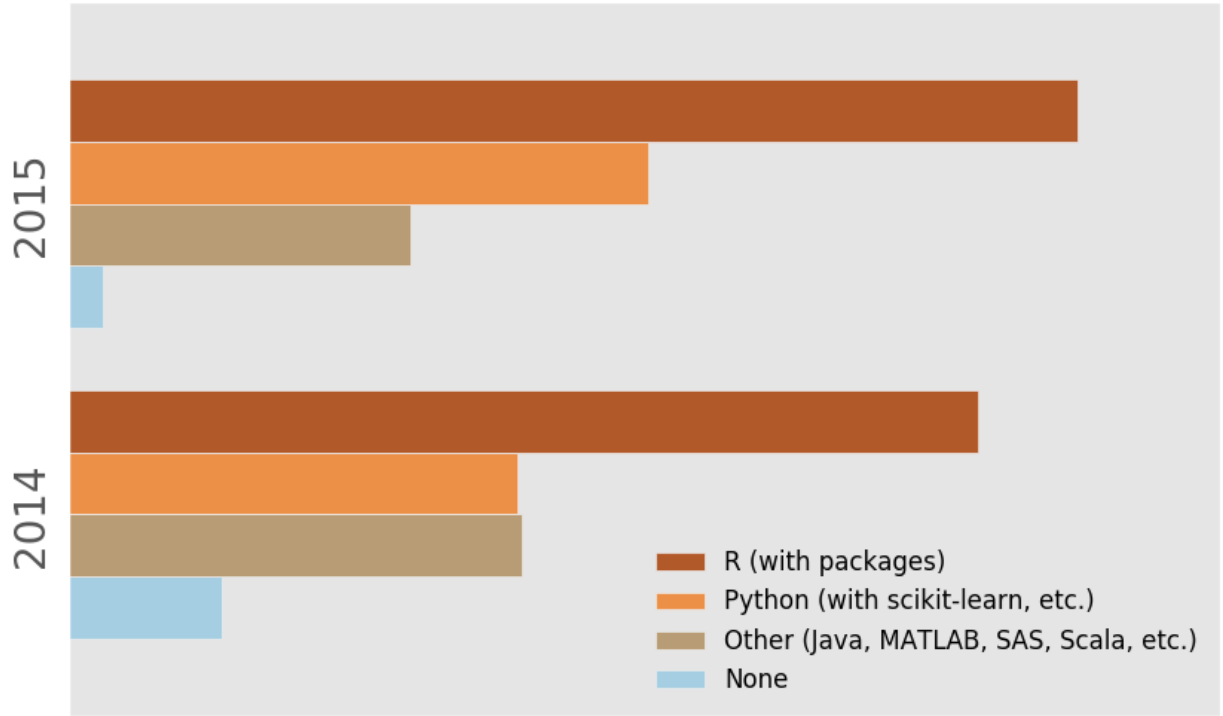
JAVA

Ο λόγος που η Java μπαίνει στην λίστα είναι εκτός ότι προφανώς μπορεί να χρησιμοποιηθεί από Spark και Hadoop, είναι η γλώσσα που υπάρχει εκεί έξω περισσότερο από οποιαδήποτε άλλη, την γνωρίζει πάρα πολύς κόσμος και είναι ευκολότερο να βρεθεί ή να ‘μετατραπεί’ ένας

προγραμματιστής σε Data Scientist λόγω της εκ τον προτέρων γνώσης της γλώσσας από πολλούς. Η Java έχει τεσταριστεί χρόνια, έχει υποστεί αναρίθμητες βελτιώσεις και έχει την μεγαλύτερη υποστήριξη από οποιαδήποτε άλλη γλώσσα. Στα αρνητικά της είναι ότι ένας κώδικας φτιαγμένος σε Java είναι μακράν μεγαλύτερος από έναν σε οποιαδήποτε άλλη Big Data γλώσσα και ποιο δύσκολο να διαβαστεί από τους ανθρώπους.

Οπότε και για να απαντήσω στο ερώτημα του υποκεφαλαίου. Υπάρχει κάποια γλώσσα που να προτιμάτε από τις υπόλοιπες? Η απάντηση είναι εξαρτάται ποιο είναι το project που θέλει να παράξει ο Data Scientist. Αν αυτό αφορά ανωτέρου επιπέδου στατιστικής ανάλυσης, χρησιμοποιώντας πολύπλοκα μοντέλα, τότε η R είναι η γλώσσα επιλογής. Αν χρειάζεται κάποιος την υλοποίηση αλγορίθμων Machine Learning και κυρίως Deep Learning με πολλαπλά νευρωνικά δίκτυα, τότε η Python είναι η προφανής επιλογή. Τέλος για real-time δεδομένα πολύ επιλέγουν μια εκ των Scala και Java ανάλογα με ποια γλώσσα έχουν την μεγαλύτερη εξοικείωση. Τα περισσότερα IT στελέχη μεγάλων εταιριών προτιμούν μια εκ των R και Python και πολλές φορές τον συνδυασμό τους. (Chakraborty, 2016)

ΕΙΚΟΝΑ 24) <u>Στοιχεία από το 2014-2015</u>



ΚΕΦΑΛΑΙΟ 5^{ο} : ΠΕΡΙΓΡΑΦΗ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Σε αυτό το κεφάλαιο θα περιγράψω και θα αναλύσω βήμα βήμα τον σκοπό της εργασίας μου, πως αντιμετωπίστηκαν τα διάφορα προβλήματα και στο τελικό στάδιο την υλοποίηση και την σύγκριση των αποτελεσμάτων με άλλον παρεμφερή movie recommendation αλγόριθμο.

5.1) ΠΕΡΙΓΡΑΦΗ ΠΡΟΒΛΗΜΑΤΟΣ

Όπως ανέφερα σε προηγούμενα κεφάλαια, η αλματώδης ανάπτυξη των Big Data και το κέρδος που η ανάλυση τους μπορεί να επιφέρει, έχει οδηγήσει τις ερευνητικές κοινότητες να δαπανούν τεράστια ποσά για την βελτίωση τεχνικών ανάλυσης αλλά και τεχνολογιών, όπως επίσης και την δημιουργία νέων. Οι εταιρίες που ασχολούνται με την εκμετάλλευση των Big Data έχουν ήδη σημαντικό κέρδος στην αποτελεσματική λειτουργία τους. Η εργασία μου θα ασχοληθεί με το κομμάτι του recommendation system ή αλλιώς το σύστημα επιλογής προϊόντων που μπορεί να αρέσουν σε κάθε χρήστη, βάση της επεξεργασίας των πληροφοριών του ιστορικού του, δηλαδή τις προτιμήσεις του, αλλά ακόμη και από θετικά ή αρνητικά σχόλια στις κριτικές του. Αυτό ήδη εφαρμόζεται με πολύ μεγάλη επιτυχία σε ιστοσελίδες με ηλεκτρονικές αγορές προϊόντων όπως η Amazon, το EBay κλπ.

Κατι παρόμοιο είναι και το αντικείμενο της εργασίας μου. Αφορά την επεξεργασία δεδομένων έτσι ώστε ο υπολογιστής να μάθει να προτείνει σε κάθε χρήστη ταινίες ανάλογα με τις προτιμήσεις του, και τα αποτελέσματα των προτιμήσεων από άλλους χρήστες. Έτσι λειτουργούν σελίδες όπως το IMDB, ή το Netflix όπου ανάλογα με την βαθμολογία που βάζει ο κάθε χρήστης ανάλογα με την ταινία, του προτείνει ταινίες βάση των προτιμήσεων άλλων ανθρώπων που είχαν βαθμολογήσει με το ίδιο σκορ την ταινία. Επίσης πέρα από την βαθμολογία γίνεται ανάλυση και στο είδος των ταινιών που έχει βαθμολογήσει κάθε χρήστης πριν του γίνει πρόταση για το ποιες ταινίες μπορεί να του αρέσουν. Αυτό είναι πάρα πολύ σημαντικό και αρκετά βοηθητικό προς τους χρήστες καθώς τους δίνεται η δυνατότητα να ανακαλύψουν ταινίες ή μουσική ή κάποιο προϊόν που θα τους ενδιέφερε αλλά δεν το είχαν ακούσει στο παρελθόν.

Για παράδειγμα ας υποθέσουμε ότι έχω βαθμολογήσει στο IMDB 50 ταινίες με υψηλή βαθμολογία εκ των οποίων οι 40 αφορούν ταινίες δράσης και επιστημονικής φαντασίας, οι 7 αφορούν κομωδίες και οι 3 τελευταίες ρομαντικές ταινίες. Βάση των προτιμήσεων μου ο αλγόριθμος δημιουργεί τον συσχετισμό μεταξύ των υψηλών μου βαθμολογιών αλλά και το είδος της πλειοψηφίας των ταινιών που βαθμολόγησα με μεγάλο σκορ και προσαρμόζεται ανάλογα στις προτιμήσεις μου. Οπότε 4/5 ταινίες που θα μου προτείνει θα αφορούν ταινίες δράσεις και επιστημονικής φαντασίας καθώς κατάλαβε ότι αυτές είναι οι αγαπημένες μου. Αν ο αλγόριθμος έβλεπε μόνο την βαθμολογία, τότε θα μου πρότεινε έναν ίδιο αριθμό από ρομαντικές ταινίες και ταινίες επιστημονικής φαντασίας, που όμως δεν θα ικανοποιούσε τα ενδιαφέροντα μου και ουσιαστικά δεν θα με βοηθούσε να βρω παρόμοιες ταινίες βάση των προτιμήσεών μου. Υπάρχουν δύο κύριες τεχνικές που χρησιμοποιούνται: **User-Based Collaborative Filtering** και **Item-Based Collaborative filtering**.

5.1.1) User-Based Collaborative Filtering

Η μια τεχνική, όπως φαίνεται και στον τίτλο, ονομάζεται User-Based Collaborative Filtering και είναι ουσιαστικά ένας ‘περίεργος’ τρόπος να πεις ότι είναι σύστημα σύστασης προϊόντων ή ταινιών βάση του συνδυασμού των προτιμήσεων του χρήστη, και όλων των υπόλοιπων χρηστών που είχαν παρόμοιες προτιμήσεις. Δηλαδή ελέγχει την δική σου διαδικτυακή συμπεριφορά σε σύγκριση με όλων των άλλων και καταλήγει να προτείνει βάση του συγκεκριμένου συνδυασμού. Ποιο συγκεκριμένα ο τρόπος με τον οποίον λειτουργεί η συγκεκριμένη τεχνική είναι ο εξής:

1. **Φτιάχνει έναν πίνακα προϊόντων, ταινιών κλπ. που κάθε χρήστης αγόρασε, είδε ή βαθμολόγησε.**

Ποιο συγκεκριμένα δημιουργείται ένας πίνακας όπου σε κάθε χρήστη που έχω στο database αντιστοιχώ τις προτιμήσεις του, δηλαδή τι ταινία έχει δει, ποια ιστοσελίδα επισκέφτηκε, τι αγόρασε στο Internet, ανάλογα με ποιες θέλουμε να είναι οι πληροφορίες τις οποίες θέλουμε να αναλύσει ο αλγόριθμος μας. Ένα μικρό παράδειγμα είναι ο παρακάτω πίνακας.

Users	**Movies**	**Products**	**Websites**
Ζαννής Λεμός	Star Wars	Witcher 3	Gazzetta.gr

	Die Hard	FIFA 2017	
	Lord of the Rings		
Γιώργος Δώνης	American History X	NBA 2k17	In.gr
			Google.gr
			Nba.com

Η κάθε εταιρία συλλέγει και αναλύει τα στοιχεία που την ενδιαφέρουν δηλαδή το IMDB τις προτιμήσεις των χρηστών σε ταινίες ή Amazon σε προϊόντα κλπ.

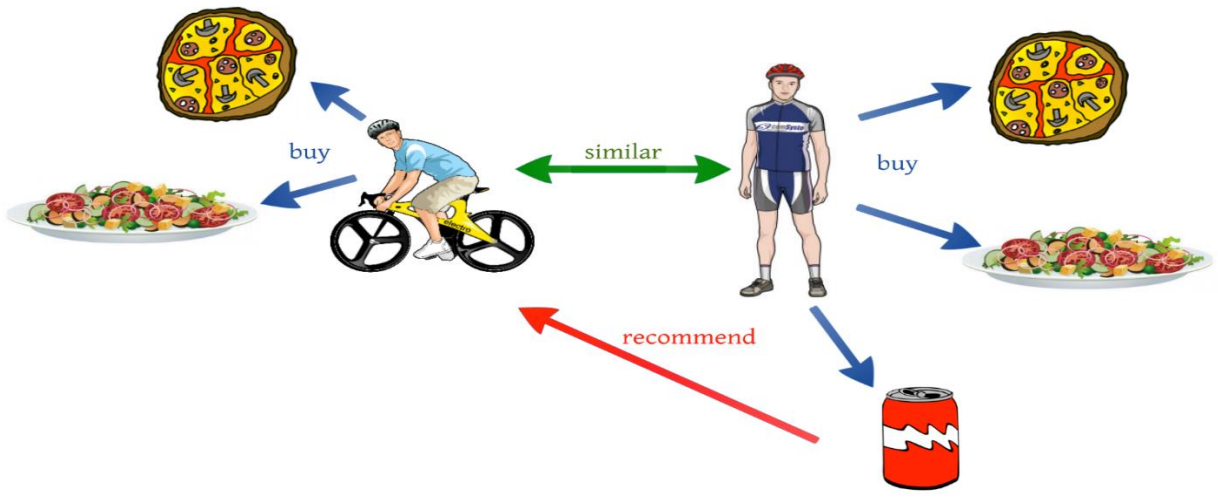
2. **Ελέγχει τα κοινά στοιχεία και τις προτιμήσεις μεταξύ των χρηστών**

Δηλαδή αυτό που κάνει είναι π.χ. αν στον χρήστη Ζαννή άρεσε η ταινία Star Wars βρίσκει και άλλους χρήστες που τους άρεσε η ταινία. Μετά ελέγχει τις άλλες ταινίες που άρεσαν σε αυτούς τους δύο χρήστες και υπολογίζει ένα σκορ ομοιότητας των ενδιαφερόντων μεταξύ των χρηστών.

3. **Βάση της ομοιότητας μεταξύ δύο χρηστών προτείνει πράγματα που δεν έχουν δει ακόμη.**

Από το προηγούμενο βήμα, έχει γίνει ο υπολογισμός της ομοιότητας μεταξύ όλων των χρηστών και του χρήστη Ζαννή, και κατατάσσει τους χρήστες με το μεγαλύτερο σκορ ομοιότητας. Από αυτούς βλέπει ποιες ταινίες (στο συγκεκριμένο παράδειγμα) δεν έχει δει ο Ζαννής και τις έχουν βαθμολογήσει υψηλά χρήστες με μεγάλο σκορ ομοιότητας και τις προτείνει στον Ζαννή.

Εικόνα 25) User-Based Collaborative Filtering



Η τεχνική αυτή έχει και κάποια αρνητικά στοιχεία. Η τεχνική αυτή βασίζεται σε ομοιότητες και ίδια ενδιαφέροντα μεταξύ των ανθρώπων και ακριβώς εκεί έγκειται ο περιορισμός της. Σκεφτείτε ότι ας πούμε στο παράδειγμα των ταινιών, ο χρήστης Ζαννής που είναι λάτρης των ταινιών επιστημονικής φαντασίας και περιπέτειας, είχε ίδιο σκορ ομοιότητας με έναν χρήστη Χ ο οποίος μπορεί για πολλούς λόγους εκείνη της περίοδο της ζωής του να είχε μια περίοδο όπου μπορεί να έβλεπε και αυτός παρόμοια είδη ταινιών, αλλά να μην είναι αυτές οι ταινίες που πραγματικά τον ενδιαφέρουν και μετά να άλλαζε προτιμήσεις και να βαθμολογούσε υψηλά ταινίες με διαφορετικό περιεχόμενο από αυτές που αρέσουν στον χρήστη Ζαννή. Η τεχνική αυτή όμως είχε αναλύσει ότι οι δύο χρήστες μεταξύ τους είχαν υψηλό σκορ οπότε θα προτείνει στον Ζαννή τα νέα είδη ταινιών που βαθμολόγησε υψηλά ο χρήστης Χ, οι οποίες δεν θα έχουν κανένα ενδιαφέρον στον χρήστη Ζαννή.

Οπότε όπως καταλαβαίνουμε από αυτό το παράδειγμα, το να υπολογίζουμε σκορ μεταξύ ανθρώπων δεν είναι πολύ ασφαλές για τον πολύ απλό λόγο ότι οι άνθρωποι και οι προτιμήσεις τους αλλάζουν. Ένα άλλο πρόβλημα είναι ότι υπάρχουν 7 δισεκατομμύρια άνθρωποι στον πλανήτη, αλλά δεν υπάρχει αντίστοιχα τόσο μεγάλος αριθμός ταινιών. Οπότε υπολογίζοντας βάση τις ομοιότητες των χρηστών απαιτεί πολύ περισσότερο χρόνο και υπολογιστική ισχύ από ότι να βρεις ομοιότητες μεταξύ των αντικειμένων και όχι των χρηστών (όπως θα δούμε στην επόμενη τεχνική).

Τέλος άλλο ένα πρόβλημα με αυτήν την τεχνική είναι ότι πολλές φορές τα αποτελέσματα δεν είναι αντικειμενικά. Αυτό συμβαίνει γιατί μεγάλες εταιρίες παραγωγής δίνουν πολλά λεφτά για να προωθήσουν την ταινία τους με υψηλές ψεύτικες βαθμολογίες από ψεύτικους χρήστες και αυτό επηρεάζει το αποτέλεσμα υπολογισμού και σύστασης ταινιών (ή οτιδήποτε άλλων θέλουμε να προτείνουμε).

Γι’ όλα αυτά τα προβλήματα υπάρχουν τεχνικές που αντιμετωπίζουν, αλλά δεν μπορούν να εξαλείψουν τελείως τις παραπλανητικές βαθμολογίες και σίγουρα δεν μπορούν να υπολογίσουν με ακρίβεια την αλλαγή προτιμήσεων των χρηστών. Γι’ αυτό και η πλειοψηφία των Data Scientists χρησιμοποιεί την επόμενη τεχνική που ονομάζεται Item-Based Collaborative Filtering.

5.1.2) Item-Based Collaborative Filtering

Αυτή η τεχνική χρησιμοποιείται από σχεδόν όλες τις μεγάλες εταιρίες σύστασης προϊόντων ή ταινιών ή μουσικής κλπ. όπως η Amazon και έρχεται να αντιμετωπίσει όλα τα προβλήματα εκείνα που έκαναν την προηγούμενη τεχνική να μην είναι τόσο αξιόπιστη.

Στο προηγούμενο κεφάλαιο μιλήσαμε για την User-Based Collaborative Filtering τεχνική όπου βρίσκει ανθρώπους παρόμοιους με εσένα και προτείνει πράγματα που δεν έχεις δει. Η διαφορά με την Item-Based Collaborative Filtering είναι ότι συγκρίνει παρόμοιες προτιμήσεις και όχι ανθρώπους. Για ποιον λόγο όμως αυτό είναι καλύτερο από την προηγούμενη τεχνική? Ας το αναλύσουμε:

- **Μια ταινία θα είναι πάντα η ίδια ταινία, δεν αλλάζει.** Αυτή είναι και η βασικότερη διαφορά με την προηγούμενη τεχνική. Όταν αναλύεις συσχετισμούς βάση πραγμάτων, αυτά πάντα παραμένουν ίδια, η ταινία Star Wars θα είναι πάντα η ίδια (εκτός αν αποφασίσει ο George Lucas να την κάνει κομωδία), σε αντίθεση με την ανάλυση των ανθρωπίνων συμπεριφορών όπου οι προτιμήσεις μπορούν να αλλάξουν με την πάροδο του χρόνου.

- **Λιγότερα πράγματα για σύγκριση από ανθρώπους.** Λογικά ούτε η Amazon δεν διαθέτει 7 δισεκατομμύρια προϊόντα (όσο ο παγκόσμιος πληθυσμός) οπότε το να ελέγξει ο αλγόριθμος ομοιότητες μεταξύ προϊόντων χρειάζεται πολύ λιγότερη υπολογιστική δύναμη από το να έλεγχε ομοιότητες μεταξύ ανθρώπων.

- **Δυσκολότερο να εκμεταλλευτείς το σύστημα.** Όπως αναφέραμε πολλές εταιρίες με την προηγούμενη τεχνική δημιουργούσαν ψεύτικους λογαριασμούς με ψεύτικες βαθμολογίες για να ανεβάσουν την βαθμολογία των ταινιών τους και να δημιουργήσουν μη αντικειμενικά κριτήρια πρόβλεψης για το σύστημα μας. Αυτό δεν μπορεί να συμβεί με αυτήν την τεχνική γιατί δεν μπορείς να δημιουργήσεις ένα ψεύτικο προϊόν συσχετισμένο με πολλά άλλα ψεύτικα προϊόντα οπότε είναι αρκετά ποιο ασφαλής η πρόβλεψη μας.

Ο τρόπος με τον οποίο λειτουργεί η συγκεκριμένη τεχνική μοιάζει πολύ με τον τρόπο που λειτουργεί η User-Based με την πολύ βασική διαφορά ότι ελέγχει ομοιότητες μεταξύ προϊόντων και όχι μεταξύ ανθρώπων. Ποιο αναλυτικά με παράδειγμα όπως λειτουργεί και ο αλγόριθμος της εργασίας μου :

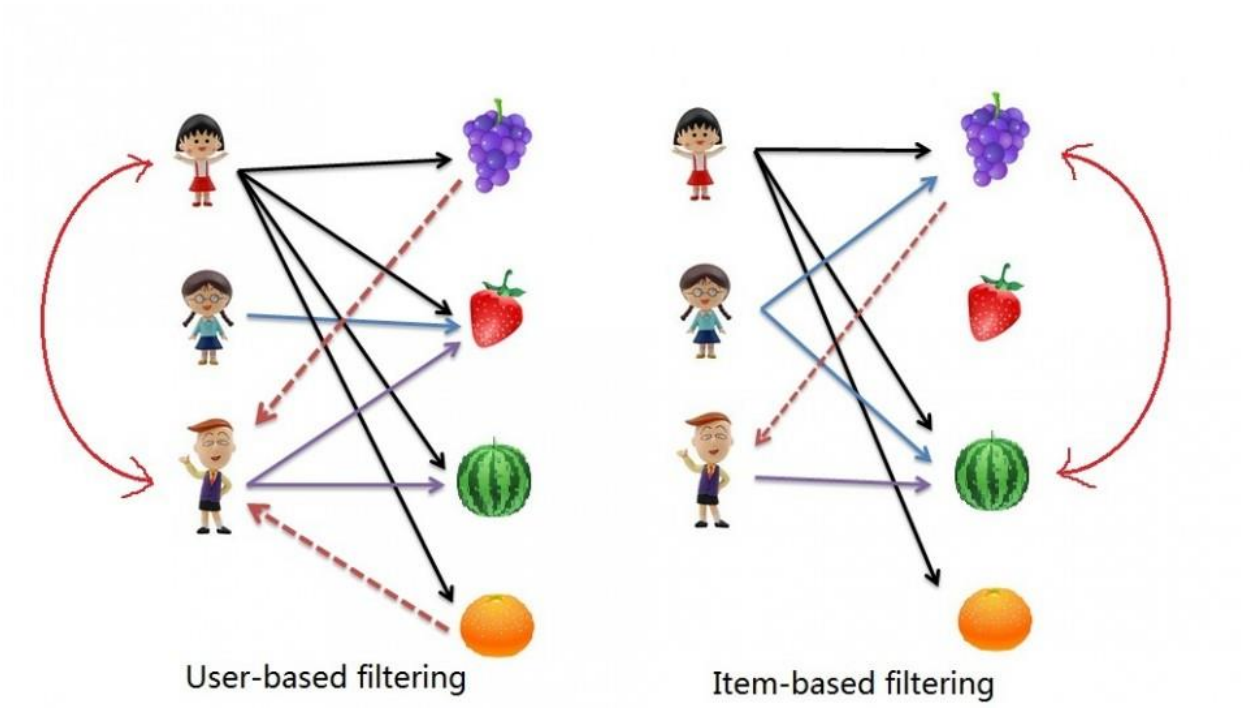
1. **Βρίσκει κάθε ζευγάρι από ταινίες που είδαν δύο άνθρωποι.** Ελέγχει λοιπόν τις παρόμοιες ταινίες μεταξύ δύο χρηστών και τις κάνει ζευγάρι, μετά συνεχίζει και ελέγχει τις υπόλοιπες μεταξύ των δύο χρηστών και τις κάνει και αυτές ζευγάρι και ούτω κάθε εξής.

2. **Ελέγχει τις βαθμολογίες των ταινιών από όλους τους χρήστες που είδαν αυτές τις δύο ταινίες.** Οπότε έχουμε ένα ζευγάρι ταινιών, ας πούμε (Lord of the Rings 1, Harry Potter 1), φτιάχνει μια λίστα με όλους τους χρήστες που έχουν δει αυτές τις ταινίες, συγκρίνει τις βαθμολογίες μεταξύ αυτών των χρηστών στις συγκεκριμένες ταινίες, και αν οι βαθμολογίες είναι παρόμοιες, τότε ο αλγόριθμος καταλαβαίνει ότι αυτές οι δύο ταινίες είναι παρόμοιες επειδή η βαθμολογία τους ήταν ίδια ή παραπλήσια από ανθρώπους που είδαν και τις δύο.

3. **Ταξινομεί τα αποτελέσματα βάση των ταινιών από το σκορ ομοιότητας.** Μετά το πρώτο ζευγάρι ταινιών ελέγχει μεταξύ των δύο χρηστών αν υπάρχουν και άλλα ζευγάρια με παρόμοιες ταινίες και παρόμοιες βαθμολογίες.

Και αυτός είναι ένας από τους τρόπους (υπάρχουν και άλλοι τρόποι με τους οποίους λειτουργεί η συγκεκριμένη τεχνική) με τον οποίον όταν έχετε βαθμολογήσει μια ταινία με υψηλή βαθμολογία στο IMDB από κάτω βλέπεται ‘people who liked this also liked’ ή ‘people who rated this highly also rated’ κλπ.

ΕΙΚΟΝΑ 26) User-Based Collaborative Filtering vs Item-Based Collaborative Filtering



Αυτό που καταφέρνουμε δηλαδή με την συγκεκριμένη τεχνική είναι αντί να εστιάζουμε σε σχέσεις μεταξύ ανθρώπων, να εστιάζουμε σε σχέσεις μεταξύ πραγμάτων (είτε ταινιών είτε προϊόντων κλπ.)

5.2) ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Τώρα που έχω αναλύσει όλα τα βασικά συστατικά που χρησιμοποίησα στον αλγόριθμό μου, ήρθε η ώρα της υλοποίησης! Θα χρησιμοποιήσω σαν γλώσσα προγραμματισμού την Python (μπορούσε να γίνει και σε R, Scala, Java, απλά είναι προσωπική συμπάθεια!) πάνω στην Apache Spark (κεφ. 4.2.3) που εν πολύς η συγκεκριμένη τεχνολογία μας επιτρέπει την γρήγορη και αποτελεσματική ανάλυση μεγάλων datasets και είναι σαφώς γρηγορότερη (όπως έδειξα) από το Hadoop άρα και προτιμότερη. Το άλλο πλεονέκτημα για την χρήση του συνδυασμού Python-Spark είναι ότι ο κώδικας που χρειάζεται μεταξύ των δύο για την υλοποίηση ενός αξιόπιστου movie recommendation system είναι εντυπωσιακά μικρότερος από το αν ας πούμε χρησιμοποιούσα Map Reduce μέσω Hadoop με γλώσσα την Java (για την ακρίβεια σχεδόν το 1/3 του κώδικα) και πολύ γρηγορότερος και ευκολότερα κατανοητός όπως θα δείτε.

Τα στοιχεία που συλλέξαμε είναι αληθινά από το MovieLens project όπου έχουν datasets με βαθμολογίες χρηστών που κυμαίνονται από μερικούς χιλιάδες user votes έως 10 εκατομμύρια user votes το οποίο και χρησιμοποίησα. (MovieLens, 19). Ποιο συγκεκριμένα στην εργασία έχουμε δύο αρχεία όπου το πρώτο είναι της μορφής UserID, MovieID, Ratings, Timestamp, όπου UserID είναι ένας μοναδικός αριθμός για κάθε χρήστη, MovieID ο μοναδικός αριθμός που αντιστοιχεί σε κάθε ταινία στο άλλο αρχείο, Ratings η βαθμολογία κάθε ταινίας από τον κάθε χρήστη σε κλίμακα από 0-5 με δυνατότητα ενδιάμεσης βαθμολόγησης (0.5, 1.5 κλπ.) και το Timestamp αντιπροσωπεύει την ώρα που βαθμολόγησε ο κάθε χρήστης την ταινία σε δευτερόλεπτα ξεκινώντας από την 1 Ιανουαρίου του 1970. Το άλλο αρχείο περιέχει τις ταινίες και είναι της μορφής MovieID, Title, Genres, όπου MovieID ο μοναδικός αριθμός κάθε ταινίας, Title ο τίτλος της ταινίας και Genre το είδος της ταινίας, όπου τα είδη χωρίζονται στις εξής κατηγορίες: Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Dram, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-fi, Thriller, War και Western.

Ποιο συγκεκριμένα για την υλοποίηση του αλγορίθμου χρησιμοποίησα 2 interfaces, το Canopy και το Jupyter Notebook πάνω στα οποία έγραψα κώδικα σε Spark-Python. Για την επεξεργασία και προσπέλαση των αρχείων παρατήρησα ότι η δημιουργία cluster στην Spark είχε μικρή διαφορά στην ταχύτητα των αποτελεσμάτων (μεγάλη διαφορά θα παρουσιαζόταν σε αρχεία πολλών GB).

Ο τρόπος με τον οποίον υλοποιήθηκε ο αλγόριθμος είναι μέσω της Item-Based Collaborative Filtering τεχνικής, ακολουθώντας τα βήματα που περιέγραψα στο κεφ 5.1.2 έχοντας ως σκοπό του τελικού αποτελέσματος την επιλογή μιας ταινίας από τον χρήστη, και την παρουσίαση από τον αλγόριθμο αποτελεσμάτων παραπλήσιων ταινιών, λίγο πολύ όπως το 'people who rated this highly also rated' του IMDB.

Θα έχουμε δηλαδή την δυνατότητα να βάζουμε ως input το MovieID που αντιστοιχεί σε μια ταινία στο αρχείο μας και μέσω της Item-Based Collaborative Filtering τεχνικής αυτό που θα κάνει ο αλγόριθμος μας είναι να βρίσκει ταινίες και να τις προτείνει στον user.

Ένας από τους πολλούς τρόπους με τους οποίους μπορούμε να υλοποιήσουμε την συγκεκριμένη τεχνική είναι **α)** βρίσκουμε κάθε ζευγάρι ταινιών που παρακολούθησε ένα άτομο **b)** βρίσκουμε όλους τους χρήστες που παρακολούθησαν αυτό το ζευγάρι ταινιών και ελέγχουμε

τις βαθμολογίες τους σε σχέση με εμάς **c)** ταξινομούμε τις ταινίες βάση του σκορ ομοιότητας που προέκυψε από το προηγούμενο βήμα.

5.2.1) Υλοποίηση σε Spark

Ωραία ως εδώ, αλλά για την μετατροπή του από την θεωρία σε Spark πρόγραμμα είναι αρκετά περίπλοκο. Υπάρχουν αρκετοί διαφορετικοί τρόποι, αλλά αυτός που μου κέντρισε το ενδιαφέρον και θεώρησα ότι είναι ο ποιο εύκολος και κατανοητός είναι ο εξής:

- **Δημιουργούμε key-value ζευγάρια της μορφής (UserID, (MovieID, Rating))**

```
# key-value ζευγάρια                                    UserID,     MovieID,  Ratings
ratings = data.map(lambda l: l.split("::")).map(lambda l: (int(l[0]), (int(l[1]), float(l[2]))))
```

Αυτό δηλαδή που κάνουμε είναι να αντιστοιχούμε σε κάθε user μια τιμή που προκύπτει από την ταινία και την βαθμολογία του user σε αυτήν

- **Βρίσκουμε κάθε ζευγάρι ταινιών που βαθμολογήθηκε από τον ίδιο User**

```
# Self-join για να βρούμε όλα τα πιθανά ζευγλαρια ταινιών.
vathmologies_Partitioned = vathmologies.partitionBy(1000)
self_join = vathmologies_Partitioned.join(vathmologies_Partitioned)
```

Αυτό μπορεί να γίνει εύκολα με μια ‘self-join’ εντολή όπου ουσιαστικά ενώνει την database με τον ευατό της και έτσι μπορούμε να έχουμε κάθε πιθανό ζευγάρι ταινιών που βαθμολογήθηκε από τον ίδιο χρήστη και καταλήγουμε να έχουμε μια τεράστια λίστα από UserIDs ακολουθούμενο από ζευγάρια ταινιών και την βαθμολογία τους (αυτό το κομμάτι του αλγορίθμου είναι το ποιο χρονοβόρο στην επεξεργασία). Αυτην την στιγμή δηλαδή βρισκόμαστε στην συγκεκριμένη μορφή (UserID, ((MovieID1, rating1), (MovieID2, rating2))). Όπου ένας χρήστης είδε την ταινία 1 και την βαθμολόγησε και είδε την ταινία 2 και την βαθμολόγησε και αυτό θα συμβεί για όλους τους συνδυασμούς ταινιών του κάθε User.

- **Φιλτράρουμε και διώχνουμε τα ζευγάρια που δημιουργήθηκαν δύο φορές κατά την self-join εντολή**

```
def filtrarisma_tainiwn( (userID, vathmologies) ):
    (tainia1, rating1) = vathmologies[0]
    (tainia2, rating2) = vathmologies[1]
    return tainia1 < tainia2
```

```
# Φιλτράρουμε τα διπλά ζευγάρια
monadika_zeugaria = self_join.filter(filtrarisma_tainiwn)
```

- **Μετά κάνουμε ως κλειδιά τα ζευγάρια των ταινιών που προέκυψαν**

```
def zeugaria_tainiwn_vathmologiwn((user, vathmologies)):
    (tainia1, vathmologies1) = vathmologies[0]
    (tainia2, vathmologies2) = vathmologies[1]
    return ((tainia1, tainia2), (vathmologies1, vathmologies2))
```

```
# Τώρα το key είναι (movie1, movie2) τα ζευγάρια των ταινιών.
zeugaria_tainiwn = monadika_zeugaria.map(zeugaria_tainiwn_vathmologiwn).partitionBy(1000)

# Τώρα έχουμε (ταινία1, ταινία2) => (βαθμολογία1, βαθμολογία2)
# μαζεύουμε όλες τις βαθμολογίες για κάθε ζευγάρι ταινιών
sullogi_vathmologiwn = zeugaria_tainiwn.groupByKey()
```

Αυτό το κάνουμε γιατί αυτό που θέλουμε είναι το σκορ ομοιότητας μεταξύ των ζευγαριών των ταινιών το οποίο θα είναι της μορφής ((MovieID1, MovieID2), (rating1, rating2)) δηλαδή η ταινία 1 και 2 που είδε ο ίδιος άνθρωπος έχουν τις βαθμολογίες 1 και 2. Αυτήν την στιγμή το UserID δεν μας ενδιαφέρει ούτε μας χρειάζεται καθώς το μόνο που μας ενδιαφέρει είναι ότι κάποιος άνθρωπος είδε αυτές τις δύο ταινίες και τις βαθμολόγησε με αυτά τα νούμερα (εδώ είναι η βασική διαφορά με την User-Based Collaborative Filtering τεχνική).

- **Στη συνέχεια θα συγκρίνουμε, θα υπολογίσουμε την ομοιότητα στις βαθμολογίες για κάθε ζευγάρι ταινιών και θα τις ταξινομήσουμε έχοντας προσπελάσει πλέον όλο το dataset.**

```
def ypologismos_omoiotitas(ratingPairs):
    numPairs = 0
    sum_xx = sum_yy = sum_xy = 0
    for ratingX, ratingY in ratingPairs:
        sum_xx += ratingX * ratingX
        sum_yy += ratingY * ratingY
        sum_xy += ratingX * ratingY
        numPairs += 1

    numerator = sum_xy
    denominator = sqrt(sum_xx) * sqrt(sum_yy)

    score = 0
    if (denominator):
        score = (numerator / (float(denominator)))

    return (score, numPairs)
```

```
# Τώρα έχουμε (ταινία1, ταινία2) => (βαθμολογία1, βαθμολογία2), (βαθμολογία1, βαθμολογία2)...
# υπολογισμος ομοιότητας
skor_omoiotitas = sullogi_vathmologiwn.mapValues(ypologismos_omoiotitas).persist()
```

- **Και στο τέλος θα έχουμε το αποτέλεσμα!** (αναλυτικά όλοι οι αλγόριθμοι στο τέλος της εργασίας)

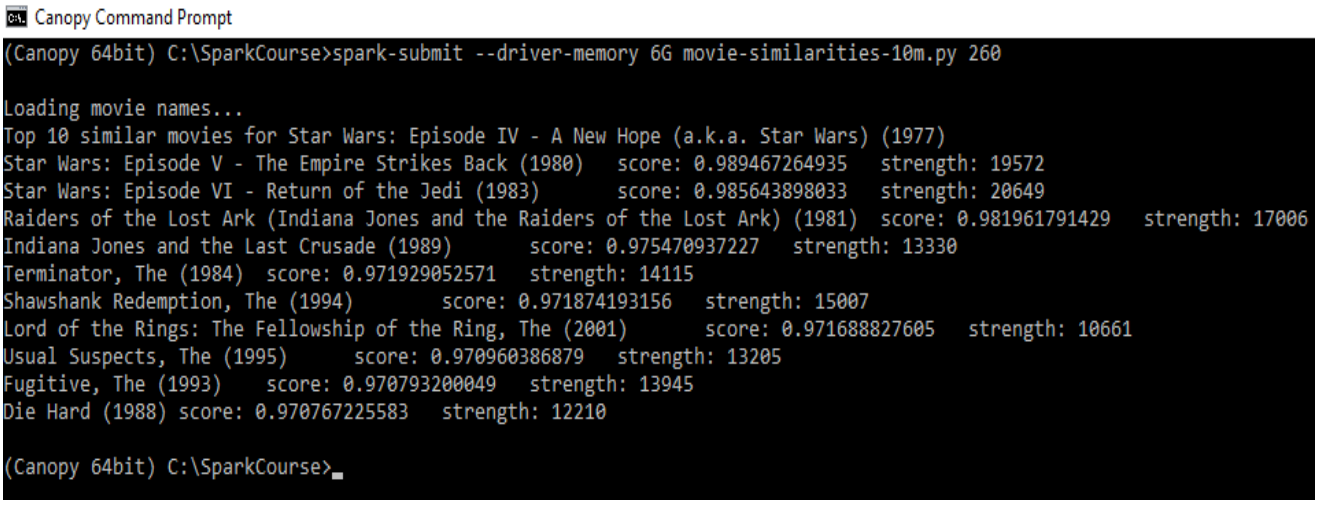
5.3) ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΓΚΡΙΣΕΙΣ ΑΛΓΟΡΙΘΜΩΝ

Έφτασε η στιγμή να δούμε αν ο αλγόριθμος μας δουλεύει σωστά. Όπως είχαμε αναφέρει ο user θα βάζει ως input το ID μιας ταινίας και ο αλγόριθμος θα προτείνει 10 παρόμοιες ταινίες που πιθανώς να του αρέσουν.

Για να αποφύγουμε την περίπτωση κάποιος χρήστης να έχει βαθμολογήσει μόνο μια ή δύο ταινίες, πράγμα που θα επηρέαζε την αντικειμενικότητα των αποτελεσμάτων, ή η ταινία μας να έχει βαθμολογηθεί από λίγους ανθρώπους, εισάγουμε κάποιες παραμέτρους.

```
    scoreThreshold = 0.97
    coOccurenceThreshold = 10000
```

Ουσιαστικά λέμε στον αλγόριθμο ότι το σκορ ομοιότητας μεταξύ δύο ταινιών θα πρέπει να είναι μεγαλύτερο του 97% και επίσης την ταινία να την έχουν παρακολουθήσει (και βαθμολογήσει) τουλάχιστον 10000 άνθρωποι. Να επισημάνω ότι δεν υπάρχει σωστό ή λάθος στις παραμέτρους που θέτουμε στον αλγόριθμο. Μέσω του trial and error ρυθμίζουμε εμείς τις παραμέτρους σε κάθε αλγόριθμο ανάλογα το dataset και το project, μέχρι να βρούμε αξιοπρεπή αποτελέσματα. Στο πείραμα μας έβαλα την ταινία **Star Wars: Episode IV – A New Hope** ως input και οι πρώτες δέκα συστάσεις ήταν οι εξής:



Όπως βλέπουμε στα αποτελέσματα στις πρώτες επιλογές επέστρεψε άλλες δύο ταινίες star wars με σκορ ομοιότητας σχεδόν 99% και ταινίες που αφορούν το είδος των ταινιών star wars δηλαδή επιστημονικής φαντασίας και περιπέτειας όπως το Indiana Jones ή το Terminator. Βλέποντας λοιπόν τα αποτελέσματα συμπεραίνουμε ότι ο αλγόριθμος μας είναι προς την σωστή κατεύθυνση (ευτυχώς δεν επέστρεψε ταινίες τύπου Toy Story ή 50 αποχρώσεις του γκρι, εκεί κάτι δεν θα πήγαινε καλά!) αλλά πρέπει να εξασφαλίσουμε μέσω συγκρίσεων αν τα αποτελέσματα μας είναι όντως σωστά. Αυτό θα γίνει με δύο τρόπους: **α)** θα συγκρίνω τα αποτελέσματα με άλλον αλγόριθμο που υπάρχει για movie recommendations και **β)** θα αφαιρέσω από το dataset 9 εκατομμύρια users και θα τον ξανατρέξω για να συγκρίνουμε και κατά πόσο η ποσότητα των πληροφοριών σε ένα dataset επηρεάζει θετικά ή αρνητικά τα αποτελέσματα μας.

Ξεκινώντας το α) ερώτημα βρήκα έναν αλγόριθμο για movie recommendations γραμμένο εξολοκλήρου σε python χρησιμοποιώντας τις δικές της βιβλιοθήκες και έτρεξα το αρχείο των 10 εκατομμυρίων users που μπορείτε να βρείτε στο MovieLens.org (αναλυτικά και

αυτός ο αλγόριθμος στο παράρτημα στο τέλος). Τα αποτελέσματα είναι τα εξής:

	(rating, size)	(rating, mean)	similarity
title			
Star Wars: Episode IV - A New Hope (1977)	2991	4.453694	1.000000
Star Wars: Episode V - The Empire Strikes Back (1980)	2990	4.292977	0.661552
Star Wars: Episode VI - Return of the Jedi (1983)	2883	4.022893	0.574808
Raiders of the Lost Ark (1981)	2514	4.477725	0.421425
Star Wars: Episode I - The Phantom Menace (1999)	2250	3.409778	0.363200
Back to the Future (1985)	2583	3.990321	0.259374
E.T. the Extra-Terrestrial (1982)	2269	3.965183	0.253247
Ghostbusters (1984)	2181	3.905548	0.251105
Jurassic Park (1993)	2672	3.763847	0.240746
Matrix, The (1999)	2590	4.315830	0.234341

Η διαφορά με τον δικό μου αλγόριθμο είναι ότι αλλάζουν κάποιες παράμετροι όπως ο αριθμός του δείγματος που έχουμε ως threshold αλλά τα αποτελέσματα είναι παρεμφερή, βλέπουμε δηλαδή ότι στις πρώτες προτεινόμενες ταινίες τα αποτελέσματα είναι σχεδόν ίδια όπως και επίσης όλες οι ταινίες ανήκουν στην ίδια κατηγορία δηλαδή ένας συνδυασμός περιπέτειας με επιστημονικής φαντασίας.

Βλέποντας λοιπόν ότι τα αποτελέσματα είναι παρόμοια ανάμεσα σε δύο διαφορετικά υλοποιήσιμα movie recommendation συστήματα, θα ελέγξω στο δικό μου αλγόριθμο τα αποτελέσματα του αρχικού dataset μειώνοντας τις βαθμολογίες των χρηστών από 10 εκατομμύρια σε 1 εκατομμύριο. Τα αποτελέσματα είναι τα εξής:

```
Loading movie names...
Top 10 similar movies for Star Wars: Episode IV - A New Hope (1977)
Star Wars: Episode V - The Empire Strikes Back (1980)   score: 0.989791710657   strength: 2355
Raiders of the Lost Ark (1981)  score: 0.985554827857   strength: 1972
Star Wars: Episode VI - Return of the Jedi (1983)       score: 0.984124835993   strength: 2113
Indiana Jones and the Last Crusade (1989)       score: 0.977444002865   strength: 1397
Shawshank Redemption, The (1994)        score: 0.976833270875   strength: 1412
Usual Suspects, The (1995)      score: 0.976687513683   strength: 1194
Godfather, The (1972)   score: 0.975928450362   strength: 1583
Sixth Sense, The (1999) score: 0.974688767431   strength: 1480
Schindler's List (1993) score: 0.974682012195   strength: 1422
Terminator, The (1984)  score: 0.974582199182   strength: 1746
```

Παραμετροποίησα τον αλγόριθμο ώστε να ανταποκρίνεται καλύτερα πλέον στο νέο dataset μειώνοντας το threshold ώστε να είναι ποιο ακριβή τα αποτελέσματα σε αρχείο με 1 εκατομμύριο ψήφους. Και πάλι όπως προηγουμένως βλέπουμε τις ίδιες επιλογές στις πρώτες ταινίες σύστασης οπότε μέσω συγκρίσεων καταλήγουμε στο συμπέρασμα ότι ο αλγόριθμος είναι αρκετά κοντά στο επιθυμητό και τα αποτελέσματα είναι αρκετά ενθαρρυντικά.

5.4) ΣΥΜΠΕΡΑΣΜΑΤΑ-ΣΥΝΟΨΗ

Η είσοδος των Big Data στην καθημερινότητα των επιχειρήσεων είναι πλέον γεγονός και η διαχείρηση μεγάλων όγκων δεδομένων είναι από τα σημαντικότερα ζητήματα που καλούνται να επιλύσουν οι Data Scientists και οι Data Analysts στις μέρες μας. Τα δεδομένα που παράγονται αυξάνονται εκθετικά χρονό με τον χρόνο γι'αυτό και είναι επιτακτική η επεξεργασία τους και η ανάλυσή τους καθώς μέσα στον τεράστιο όγκο των μη επεξεργασμένων δεδομένων κρύβονται πληροφορίες που μπορούν να αποφέρουν τεράστια κέρδη σε επιχειρήσεις και οργανισμούς. Για τον λόγο αυτό οι τεχνικές που χρησιμοποιόντουσαν στο παρελθόν δεν επαρκούν πλέον για να ανταπεξέλθουν στα νέα δεδομένα των ημερών μας και τις νέες προκλήσεις που παρουσιάζονται στο γενικότερο κλάδο των Big Data. Η χρησιμοποίηση νέων τεχνολογιών ανάλυσης μεγάλων δεδομένων όπως το Hadoop και η Spark αποτελούν όπλα στην φαρέτρα των Data Scientists που τους επιτρέπουν να μειώσουν των χρόνο της ανάλυσης και της εξαγωγής συμπεράσματων καθώς επίσης τους δίνει την δυνατότητα να προτείνουν νέες στρατηγικές στις εταιρίες που εργάζονται.

Πραγματοποιήσαμε στην συγκεκριμένη εργασία μια θεωρητική ανάλυση των κορυφαίων τεχνολογιών και τεχνικών που χρησιμοποιούνται από τους Data Scientists-Analysts στις μέρες μας (MongoDB, Apache Spark, Hadoop, Machine Learning, Data Mining κλπ.) και είδαμε ότι με την χρήση της Spark με την Python το μέγεθος των προγραμμάτων που χρειάζονται πλέον για την επεξεργασία αρχείων οποιασδήποτε μορφής είναι αρκετά μικρός σε σχέση με παλιότερα και γίνεται εύκολα κατανοητός στους ανθρώπους που παρουσιάζεται. Παρατηρήσαμε λοιπόν ότι με έναν τέτοιον αλγόριθμο φτιάξαμε ένα σύστημα το οποίο μαθαίνοντας από την συμπεριφορά των χρηστών μπορεί να προτείνει με ακρίβεια ταινίες που θα αρέσουν στους χρήστες βάση των προτιμήσεων τους.

Πλέον, έχοντας την δυνατότητα να διαχειριστούμε αρχεία πολλών Gigabytes και Terabytes σε μικρό χρονικό διάστημα, ανοίγονται νέοι ορίζοντας στους τομείς του Machine Learning, Deep Learning, Data Mining-Analysis καθώς και στην τεχνητή νοημοσύνη.

ΑΝΑΦΟΡΕΣ ΠΗΓΩΝ-ΕΙΚΟΝΩΝ

Bibliography

Adamson, Doug. 2016. *Healthcatalyst.* [Online] 2016. https://www.healthcatalyst.com/big-data-in-healthcare-made-simple.

brownlee, jason. 2016. *machinelearningmastery.* [Online] August 16, 2016. http://machinelearningmastery.com/what-is-deep-learning/.

Chakraborty, Apran. 2016. *Udacity.* [Online] April 21, 2016. http://blog.udacity.com/2016/04/languages-and-libraries-for-machine-learning.html.

desjardins, jeff. 2017. *visualcapitalist.* [Online] 2017. http://www.visualcapitalist.com/order-from-chaos-how-big-data-will-change-the-world/.

Galit Shmueli, Peter Bruce, Mia Stephens, Nitin Patel. 2017. *Data mining for business analytics.* 2017.

Gantz, John F. 2007. *The expanding digital universe.* 2007.

Griffith, Eric. 2016. pcmag. [Online] May 3, 2016. http://www.pcmag.com/article2/0,2817,2372163,00.asp.

Inc., SAS Institute. 2016. [Online] 2016. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html#dmhistory.

investopedia. 2016. investopedia. [Online] 2016. http://www.investopedia.com/terms/r/regression.asp.

ISSAC, Luke P. 2014. *The Geek Stuff.* [Online] January 14, 2014. http://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/?utm_source=tuicool.

MongoDB. 2017. [Online] March 20, 2017. https://www.mongodb.com/.

MovieLens. 19. *GroupLens.* [Online] April 2017, 19. https://grouplens.org/datasets/movielens/.

O'Dowd, Sean. 2016. mapr. *converge blog.* [Online] 2016. https://www.mapr.com/blog/top-10-big-data-trends-2016-financial-services.

O'Neill, Eleanor. 2016. *icas.* [Online] Σεπτεμβρίου 23, 2016. https://www.icas.com/ca-today-news/10-companies-using-big-data.

Peter Lyman, Hal R. Varian. 2003. *How much information?* California : s.n., 2003.

Rijmenam, Mark van. 2016. *datafloq.* [Online] January 7, 2016. https://datafloq.com/read/big-data-history/239.

rouse, Margaret. 2016. *techtarget.* [Online] 2016. http://whatis.techtarget.com/definition/machine-learning.

sas. 2017. sas. [Online] March 22, 2017. https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

spark, apache. 2017. [Online] January 4, 2017. http://spark.apache.org/.

technologies, softqube. 2014. [Online] May 1, 2014. http://www.softqubes.com/blog/role-and-uses-of-hadoop-in-your-business.

Theresa Payton, Theodore Claypoole. 2014. *Privacy in the age of big data.* 2014.

Wikepedia. 12. [Online] Μαρτίου 2017, 12. https://en.wikipedia.org/wiki/Artificial_neural_network.

Wikipedia. 2017. [Online] February 21, 2017. https://en.wikipedia.org/wiki/Natural_language_processing.

ΕΙΚΟΝΕΣ

1) http://www.visualcapitalist.com/order-from-chaos-how-big-data-will-change-the-world/

2) https://www.google.gr/search?q=big+data&rlz=1C1GIWA_enGR670GR670&espv=2&biw=1745&bih=885&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiEx6X1yNjRAhWrZpoKHdNeBpAQ_AUIBigB#imgrc=9_bW5Nnpt5xOZM%3A

3) https://blog.varonis.com/big-data-analytics/

4) http://www.tibco.com/blog/2011/06/28/critical-shortage-of-data-geek-talent-predicted-by-2018/

5) https://www.flickr.com/photos/adactio/9276962702

6) https://www.flickr.com/photos/adactio/9276962702

7) http://itmsoftdev.com/data-mining/

9) http://haragroup.ir/blog/

10) http://blog.pangeanic.com/2014/06/19/3-types-of-machine-translation/#.WLWBrzuGOUk

11) https://www.slideshare.net/comeur/ibm-cec-big-data-2011-0611-final

12) http://machinelearningmastery.com/what-is-deep-learning/

13) https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

14) http://jblomo.github.io/datamining290/slides/2013-03-01-Neural-Network.html

15) https://kvaes.wordpress.com/2015/01/21/database-variants-explained-sql-or-nosql-is-that-really-the-question/

16) https://www.mongodb.com/big-data-explained

17) https://en.wikipedia.org/wiki/Cloud_computing#/media/File:Cloud_computing.svg

18) http://www.softqubes.com/blog/role-and-uses-of-hadoop-in-your-business

19) https://hortonworks.com/apache/yarn/

20) https://jaceklaskowski.gitbooks.io/mastering-apache-spark/content/spark-rdd.html

21) http://news.techgenie.com/latest/how-to-use-a-tag-cloud-for-your-websites/

22) http://stackoverflow.com/questions/3033594/how-to-create-a-clustergram-plot-in-r

23) https://www.slideshare.net/maigva/bimcv-the-perfect-big-data-storm

24) http://blog.udacity.com/2016/04/languages-and-libraries-for-machine-learning.html

25) https://blog.guillaumeagis.eu/recommendation-algorithms-with-apache-mahout/

26) http://www.salemmarafi.com/code/collaborative-filtering-r/

ΠΑΡΑΡΤΗΜΑ

```
import sys
from pyspark import SparkConf, SparkContext
from math import sqrt


def loadMovieNames():
    movieNames = {}
    with open(r"C:\data_project") as f:
       for line in f:
```

```
            fields = line.split("::")
             movieNames[int(fields[0])] = fields[1].decode('ascii', 'ignore')
        return movieNames

def zeugaria_tainiwn_vathmologiwn((user, vathmologies)):
    (tainia1, vathmologies1) = vathmologies[0]
    (tainia2, vathmologies2) = vathmologies[1]
     return ((tainia1, tainia2), (vathmologies1, vathmologies2))

def filtrarisma_tainiwn( (userID, vathmologies) ):
        (tainia1, rating1) = vathmologies[0]
        (tainia2, rating2) = vathmologies[1]
        return tainia1 < tainia2

def ypologismos_omoiotitas(ratingPairs):
        numPairs = 0
        sum_xx = sum_yy = sum_xy = 0
                for ratingX, ratingY in ratingPairs:
                        sum_xx += ratingX * ratingX
                        sum_yy += ratingY * ratingY
                        sum_xy += ratingX * ratingY
                        numPairs += 1

                numerator = sum_xy
                denominator = sqrt(sum_xx) * sqrt(sum_yy)

                score = 0
```

```
        if (denominator):
            score = (numerator / (float(denominator)))

        return (score, numPairs)


conf = SparkConf().setMaster("local[*]").setAppName("movies")
sc = SparkContext(conf = conf)

print "\nLoading movie names..."
nameDict = loadMovieNames()

dataset = sc.textFile("file:///data_project/10m/ratings.dat")

# Key-value ζευγαρια UserID MovieID Rating

vathmologies = dataset.map(lambda x: x.split("::")).map(lambda x: (int(x[0]),
(int(x[1]), float(x[2]))))


# Self-join για να βρούμε όλα τα πιθανά ζευγάρια ταινιών.

vathmologies_Partitioned = vathmologies.partitionBy(1000)
self_join = vathmologies_Partitioned.join(vathmologies_Partitioned)

# Ως εδω το RDD μας είναι userID => ((movieID, rating), (movieID, rating))
# Φιλτράρουμε τα διπλά ζευγάρια
```

```
monadika_zeugaria = self_join.filter(filtrarisma_tainiwn)

# Τώρα το key είναι (tainia1, tainia2) τα ζευγάρια των ταινιών.

zeugaria_tainiwn
=monadika_zeugaria.map(zeugaria_tainiwn_vathmologiwn).partitionBy(1000)

# Τώρα έχουμε (ταινία1, ταινία2) => (βαθμολογία1, βαθμολογία2)
# μαζεύουμε όλες τις βαθμολογίες για κάθε ζευγάρι ταινιών

sullogi_vathmologiwn = zeugaria_tainiwn.groupByKey()

# Τώρα έχουμε (ταινία1, ταινία2) => (βαθμολογία1, βαθμολογία2), (βαθμολογία1,
βαθμολογία2)...
# υπολογισμος ομοιότητας

skor_omoiotitas =
sullogi_vathmologiwn.mapValues(ypologismos_omoiotitas).persist()

# ταξινόμηση

skor_omoiotitas.sortByKey()
skor_omoiotitas.saveAsTextFile("movie-sim")

# επιστρέφει αποτελέσματα με υψηλο σκορ ομοιότητας.
```

```
if (len(sys.argv) > 1):

        scoreThreshold = 0.97
        coOccurenceThreshold = 10000

        movieID = int(sys.argv[1])

        # φιλτράρουμε τις ταινίες που θα μας επιστρέψει ανάλογα με το threshold
        που του έχουμε δώσει

        filteredResults = skor_omoiotitas.filter(lambda((pair,sim)): \
                (pair[0] == movieID or pair[1] == movieID) \
                and sim[0] > scoreThreshold and sim[1] > coOccurenceThreshold)

        # Ταξινόμηση βάση αποτελέσματος

        results = filteredResults.map(lambda((pair,sim)): (sim,
        pair)).sortByKey(ascending = False).take(10)

        print "Top 10 similar movies for " + nameDict[movieID]
        for result in results:
                (sim, pair) = result
                similarMovieID = pair[0]
                if (similarMovieID == movieID):
                        similarMovieID = pair[1]
                print  nameDict[similarMovieID]  +  "\tscore:  "  +  str(sim[0])  +
                "\tstrength: " + str(sim[1])
```

ΑΛΓΟΡΙΘΜΟΣ ΣΥΓΚΡΙΣΗΣ

```
import pandas as pd


r_cols = ['user_id', 'movie_id', 'rating']

ratings = pd.read_csv(file:///data_project/10m/ratings.dat, sep='::', names=r_cols, usecols=range(3))


m_cols = ['movie_id', 'title']

movies = pd.read_csv(file:///data_project/10m/movies.dat, sep='::', names=m_cols, usecols=range(2))


ratings = pd.merge(movies, ratings)

movieRatings=ratings.pivot_table(index=['user_id'],columns=['title'],values='rating')

starWarsRatings = movieRatings['Star Wars: Episode IV - A New Hope (1977)']

similarMovies = movieRatings.corrwith(starWarsRatings)

similarMovies = similarMovies.dropna()

df = pd.DataFrame(similarMovies)

similarMovies.sort_values(ascending=False)

import numpy as np

movieStats = ratings.groupby('title').agg({'rating': [np.size, np.mean]})
```

```
popularMovies = movieStats['rating']['size'] >= 2000

movieStats[popularMovies].sort_values([('rating', 'mean')], ascending=False)[:10]

df=movieStats[popularMovies].join(pd.DataFrame(similarMovies,
columns=['similarity']))

df.sort_values(['similarity'], ascending=False)[:10]
```