ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή εργασία

BIG DATA SECURITY

Γολικίδης Μάριος – Σουλεϊμάνης Μανώλης

Σύμβουλος καθηγητής

Ηλιούδης Χρήστος

Θεσσαλονίκη 2015

Θα θέλαμε να ευχαριστήσουμε ιδιαίτερα τον Καθηγητή Κωνσταντίνο Διαμαντάρα για τις πολύτιμες συμβουλές και την στήριξη του σε κρίσιμα σημεία της πτυχιακής εργασίας μας και προπαντός τον Καθηγητή Χρήστο Ηλιούδη για την πολύτιμη βοήθειά του, τις πληροφορίες και την καθοδήγησή του σε όλη την διάρκεια του εγχειρήματός μας.

Πίνακας περιεχομένων

1 ΕΙΣΑΓΩΓΗ

❖ Περιοχή Έρευνας – Το πρόβλημα που μελετάμε

Ένα από τα μεγαλύτερα προβλήματα της σημερινής εποχής στον τομέα της Πληροφορικής αφορά τα δεδομένα και σχετίζεται με την Ασφάλεια. Πολλές περιπτώσεις μεγάλου πλήθους δεδομένων με μεγάλες διαστάσεις εμφανίζονται όλο και πιο συχνά λόγω της διάδοσης του Internet. Η ανάλυση και η επεξεργασία τέτοιων δεδομένων (BIG DATA) είναι εξαιρετικά δημοφιλές αντικείμενο έρευνας και ανάπτυξης τα τελευταία χρόνια σε διάφορους τομείς όπως και στην Ασφάλεια.

Οι συμβατικές μέθοδοι πλέον εμφανίζονται ανεπαρκείς να καλύψουν τις ανάγκες που προκύπτουν. Το πρόβλημα εμφανίζεται κατά την παρατήρηση του ραγδαίου ρυθμού αύξησης της πληροφορίας τόσο ως προς την διαθεσιμότητα στα δεδομένα, λόγω της χρήσης του διαδικτύου, όσο και από τον όγκο και την ποικιλομορφία των παραγόμενων καθημερινά δεδομένων αλλά και από την εμφάνιση νέων απειλών στο χώρο του διαδικτύου, μερικές από τις οποίες δείχνουν ιδιαίτερα αποτελεσματικές. Το πρόβλημα που προκύπτει αφενός αφορά την ίδια την Ασφάλεια των δεδομένων, είτε άμεσα είτε έμμεσα, που διακινούνται στο διαδίκτυο και αφορά τόσο επιχειρήσεις και οργανισμούς όσο και απλούς χρήστες και αφετέρου την επιτακτική ανάγκη για εκμετάλλευση της πληροφορίας αυτής για εξαγωγή συμπερασμάτων.

Η παρούσα πτυχιακή εργασία συνδυάζει αρκετούς τομείς της Πληροφορικής όπως Τεχνητή Νοημοσύνη, Παράλληλα Συστήματα, Ανακάλυψη Γνώσης, τα BIG DATA και κυρίως τον τομέα της Ασφάλειας. Η δική μας μελέτη στοχεύει να αναδείξει τη δύναμη που κρύβουν τα BIG DATA, και μέσω της ανάλυσής τους να εξαχθεί σημαντική πληροφορία η οποία αφορά την Ασφάλεια.

Τα BIG DATA Analytics είναι σε ενεργή χρήση σε διάφορους τομείς και, τα τελευταία χρόνια, έχουν προσελκύσει το ενδιαφέρον της κοινότητας της Ασφάλειας

λόγο της προδιαγραφόμενης ικανότητάς τους να αναλύουν και να συσχετίζουν τα δεδομένα που αφορούν την Ασφάλεια με αποτελεσματικό τρόπο και σε πρωτοφανή κλίμακα. Η παραδοχή μας είναι ότι μέσω της ανάλυσης των BIG DATA μπορούμε να εξάγουμε συμπεράσματα με απώτερο σκοπό να παραχθεί χρήσιμη πληροφορία για την Ασφάλεια. Πιο συγκεκριμένα μέσω της ανάλυσης ενός συνόλου BIG DATA να ανακτήσουμε πληροφορία η οποία θα μας οδηγήσει σε συμπεράσματα τα οποία να καταδεικνύουν κάποιο είδος επίθεσης στην Ασφάλεια και συγκεκριμένα στην διαθεσιμότητα των δεδομένων.

- **Σημαντικότητα του προβλήματος**

Οι τρέχουσες προσεγγίσεις για την Ασφάλεια στον κυβερνοχώρο καταπολεμούν μόνο γνωστές απειλές, κι αυτό γιατί δεν είναι τόσο καλές στην εύρεση νέων συσχετίσεων ή στην αποκάλυψη μοτίβων. Ως αποτέλεσμα, οι οργανισμοί είναι ευάλωτοι σε Advanced Persistent Threats (APTs), όπως Spear Phishing, Ddos Attacks, Hacktivism, και άλλου είδους επιθέσεις. Οι οργανισμοί και οι επιχειρήσεις χρειάζονται εξειδικευμένες αναλύσεις σε πραγματικό χρόνο για να αντιμετωπίσουν έναν σχετικά «αθόρυβο» κίνδυνο. Χωρίς βαθιά γνώση, οι περισσότερες απειλές δεν μπορούν να ανιχνευθούν. Οι επιτιθέμενοι έχουν εξελιχθεί, έχουν κίνητρα, υπομονή, επιμονή ακόμη και επιχορήγηση. Η πρόκληση που αντιμετωπίζουν οι οργανισμοί είναι πώς να επεκτείνουν τις πολιτικές Ασφάλειας για να βρουν και να εξουδετερώσουν αυτές τις απειλές σε μια περίοδο αύξησης τόσο των κινδύνων όσο και της πολυπλοκότητας.

Για να προσαρμοστούν στις σύγχρονες απειλές για την Ασφάλεια, οι οργανισμοί πρέπει να προχωρήσουν σε νέες πιο αποδοτικές διαδικασίες πέρα από το παραδοσιακό στυλ διαδικασίας που ακολουθούνταν μέχρι σήμερα "συλλογή δεδομένων, έπειτα ανάλυση και κατόπιν απαντήσεις". Το πρόβλημα με αυτήν την προσέγγιση είναι ότι θα πρέπει να ξεκινήσουν με τις υποθέσεις και τους κανόνες που δημιουργήθηκαν από κάποια προηγούμενα ή γνωστά σενάρια επίθεσης. Στην εποχή των BIG DATA, η συλλογή και η επεξεργασία όλων των πληροφοριών θα πρέπει να

γίνεται με μη-επεμβατικό τρόπο και η ανάλυση να γίνεται σε πραγματικό χρόνο δίνοντας στον αναλυτή Ασφάλειας τη δυνατότητα να βγάλει συμπεράσματα. Είναι εμφανές το ότι θα πρέπει να μεταβούν οι διαδικασίες ανάλυσης από στατική πλήρη ανάλυση μετά τη συλλογή των δεδομένων σε συνεχή, δυναμική και διερευνητική ανάλυση σε πραγματικό χρόνο.

Τα προϊόντα Security Information & Event Management (SIEM) παρέχουν μια καλή βάση για την παρακολούθηση της Ασφάλειας με ικανότητα ανίχνευσης υπογραφών επιθέσεων σχεδόν σε πραγματικό χρόνο. Ωστόσο, οι τεχνολογίες SIEM δεν είχαν σχεδιαστεί για BIG DATA Analytics, δεν κλιμακώνονται για τον εντοπισμό των άγνωστων απειλών σε όλα τα διαθέσιμα στοιχεία με αποτέλεσμα να μην μπορούν να ανταποκριθούν στις ταχέως εξελισσόμενες ανάγκες που απαιτούνται σήμερα κάτι που πρέπει κάνουν τα προηγμένα Analytics ασφαλείας.

Τα Security Analytics προσαρμόζονται στα BIG DATA σύμφωνα με τις επιταγές της αγοράς προκειμένου να αντιμετωπίσουν αποτελεσματικά το νέο τοπίο των απειλών στον κυβερνοχώρο. Στο μέλλον, τα BIG DATA αναμένονται να επιφέρουν αλλαγές και στα συμβατικά εργαλεία προστασίας, όπως τα Firewalls, οι εφαρμογές για την αποτροπή της απώλειας δεδομένων (Data Loss Prevention Software) και τα προγράμματα anti-Malware.

Ο στόχος των BIG DATA Analytics για την Ασφάλεια είναι να αποκτήσουν νοημοσύνη σε πραγματικό χρόνο. Παρόλο που τα BIG DATA Analytics είναι πολλά υποσχόμενα, υπάρχουν διάφορες προκλήσεις που πρέπει να ξεπεραστούν για να ωφεληθούμε από τις δυνατότητές τους όπως το ζήτημα της προέλευσης των δεδομένων και κυρίως όσον αφορά την αυθεντικότητα και την ακεραιότητα των δεδομένων που χρησιμοποιούνται για τη διαδικασία της ανάλυσης, αλλά και το μείζον θέμα της Προστασίας Προσωπικών Δεδομένων. Κρίνεται επιτακτική η ανάγκη για ρυθμιστικά κίνητρα και τεχνικούς μηχανισμούς για την ελαχιστοποίηση της ποσότητας των συμπερασμάτων που παράγονται από τα BIG DATA Analytics και μπορούν να προσβάλουν την Ασφάλεια και την προστασία της ιδιωτικής ζωής.

❖ Στόχοι της πτυχιακής εργασίας

Οι στόχοι που ετέθησαν με την έναρξη της πτυχιακής εργασίας είναι:

- Η μελέτη των προβλημάτων Ασφάλειας που αντιμετωπίζουν τα δεδομένα μεγάλου όγκου – BIG DATA.

Τα BIG DATA έχουν αναδειχθεί σε πολύ σημαντικό αντικείμενο έρευνας για πολλούς τομείς και όχι άδικα. Η τεράστια δύναμη που κρύβουν κατόπιν ανάλυσης τους είναι επιθυμητή τόσο από ερευνητές, οργανισμούς, επιχειρήσεις, κυβερνήσεις, όσο και από κακόβουλους. Οι προκλήσεις που σχετίζονται με την Ασφάλεια γύρω από τα BIG DATA είναι αρκετές και πολύ σημαντικές.

- Η δυνατότητα αξιοποίησης των μεγάλων όγκου δεδομένων για περισσότερη Ασφάλεια.

Ως συνέχεια της μελέτης μας θέσαμε ως στόχο τον πειραματισμό πάνω σε BIG DATA και μέσω των κατάλληλων διαδικασιών ανάλυσης τους να εξαγάγουμε πολύτιμη πληροφορία η οποία μέσω συμπερασμάτων θα μπορούσε να αξιοποιηθεί για την αύξηση της Ασφάλειας.

❖ Επιτεύγματα της πτυχιακής εργασίας

Μελετήσαμε τις τεχνολογίες BIG DATA, KDD και αναπτύξαμε ένα σύστημα που αξιώνει να κάνει BIG DATA Analysis. Η παραγόμενη γνώση μέσω της διαδικασίας της ανάλυσης θα μεταφραστεί σε χρήσιμη πληροφορία ως προς την αναγνώριση επίθεσης κατηγορίας Ddos Attacks. Αυτού του είδους οι επιθέσεις γίνονται εναντίον ενός υπολογιστή, ή μιας υπηρεσίας που παρέχεται, και έχουν ως σκοπό να καταστήσουν τον υπολογιστή ή την υπηρεσία ανίκανη να δεχτεί άλλες συνδέσεις με αποτέλεσμα να μην είναι διαθέσιμη η παροχή υπηρεσίας.

Για την εκτέλεση των πειραμάτων ακολουθήσαμε την διαδικασία της KDD για την ανάλυση των BIG DATA. Τα δεδομένα που αναλύονται συλλέχθηκαν από

εξυπηρετητές του τμήματος πληροφορικής του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης και αφορούν μη προσωποποιημένα IP δεδομένα. Τα δεδομένα αν και πειραματικά, φέρουν τις ιδιότητες της ποικιλομορφίας και του μεγάλου όγκου οπότε ονομάζονται και αναλύονται ως BIG DATA.

Οι βασικές παραδοχές που έγιναν για την υλοποίηση του πειράματος:

1. Το σύστημα τρέχει σε χρόνο στατικό είναι όμως σχεδιασμένο για την προσαρμογή του σε Realtime εφαρμογές.
2. Τα αποτελέσματα είναι υποθετικά, βασισμένα φυσικά σε επιστημονική γνώση και η μελέτη τους έχει ως βασικό στόχο την απόδειξη της πρότασής μας και την εφαρμοσιμότητά της. Ως επέκταση της συγκεκριμένης έρευνας αποτελεί η επιβεβαίωση από κάποιο σύστημα IDS ή SIEM.

Τα εργαλεία που χρησιμοποιήθηκαν για την εκτέλεση του πειράματός μας είναι τα εξής:

- JAVASwing για το γραφικό περιβάλλον της εφαρμογής.
- Γλώσσα προγραμματισμού JAVA για την ανάπτυξη της εφαρμογής η οποία εκτελεί τα τρία πρώτα βήματα της διαδικασίας KDD, την Επιλογή, την Προεπεξεργασία και τον Μετασχηματισμό των συλλεγμένων δεδομένων.
- Το WEKA που είναι ένα εργαλείο για εφαρμογές DATA MINING και με αυτό επιτεύχθηκε το τέταρτο βήμα της KDD.
- Τα εργαλεία JavaNNS, SNNS και RSNNS για πειραματισμό με τον αλγόριθμο ART-2 ο οποίος στοχεύουμε να χρησιμοποιηθεί στην επέκταση της συγκεκριμένης εφαρμογής.

Η συγκεκριμένη εφαρμογή δεν αποτελεί μια έτοιμη εφαρμογή για παραγωγική διαδικασία αλλά μια βάση για μελλοντική ανάπτυξη μιας τελικής εφαρμογής.

❖ Συμπεράσματα

Κατά την εκπόνηση της πτυχιακής εργασίας εντρυφήσαμε σε κάποιους από τους σημαντικότερους τομείς της Πληροφορικής και μελετήσαμε διάφορες τεχνολογίες των οποίων ο συνδυασμός δίνει απάντηση σε πολλά σημαντικά προβλήματα. Μελετήσαμε τα στάδια της διαδικασίας KDD δίνοντας έμφαση στο κομμάτι του Data Mining έχοντας σαν απώτερο σκοπό την κατανόηση της εξαγωγής γνώσης μέσω της ανάλυσης δεδομένων. Μελετήσαμε την τεχνολογία των BIG DATA με στόχο να δούμε τις ιδιαιτερότητες που εμφανίζουν και να κατανοήσουμε τη σημασία της ανάλυσής τους. Δώσαμε έμφαση στα πλεονεκτήματα και τα μειονεκτήματά τους αναφορικά με την Ασφάλεια. Μελετήσαμε την περίπτωση μιας συγκεκριμένης επίθεσης στη διαθεσιμότητα ενός διαδικτυακού συστήματος. Στο πρακτικό τμήμα της πτυχιακής εργασίας συνδυάσαμε την γνώση που πήραμε μέσω της μελέτης μας με τεχνολογίες που διδαχτήκαμε και γνωρίζαμε επαρκώς. Ολοκληρώσαμε μια εφαρμογή η οποία από την αρχή της ανάπτυξής υλοποιήθηκε σύμφωνα με τις ανάγκες του εγχειρήματος μας οι οποίες μας οδήγησαν τόσο σε απόκτηση περισσότερης γνώσης όσο και αναπροσαρμογή στη διαδικασία της υλοποίησης και στο σχεδιαστικό αλλά και στο τεχνικό επίπεδο. Η ανάλυση των BIG DATA είναι μια επίπονη διαδικασία η οποία έχει απαιτήσεις τόσο σε Hardware όσο και σε Software. Απαιτείται προσοχή με τη χρήση των δεδομένων και οι κατάλληλες μέθοδοι για την αποδοτική διαχείριση και ανάλυσή τους κι αυτό σε κάθε στάδιο της διαδικασίας της KDD. Επιτακτική κρίνεται η χρήση των κατάλληλων αλγορίθμων και εργαλείων για την ανάλυση των BIG DATA και κυρίως για την διαδικασία του DATA MINING. Από τα αποτελέσματα αποφαινόμαστε πως η ανάλυση των BIG DATA μπορεί να αποφέρει αποτελέσματα όπως συνέβη στην δική μας περίπτωση αναφορικά με την ύπαρξη ή όχι επίθεσης Ddos.

Στον απόηχο της εκπόνησης της εργασίας μας, συμπεραίνουμε πως η μελέτη και η έρευνα για τις μεθόδους και τα εργαλεία ανάλυσης των BIG DATA με στόχο την εξαγωγή πολύτιμης γνώσης που αφορούν την Ασφάλεια έχει αξία και θα συνεχιστεί τόσο για την βελτίωση της εφαρμογής όσο και για υψηλότερους στόχους.

❖ **Διάρθρωση της πτυχιακής**

Στο δεύτερο κεφάλαιο αναφέρουμε κάποιες βασικές έννοιες οι οποίες αφορούν την παραγωγή γνώσης από δεδομένα. Αναλύουμε την διαδικασία της KDD σε Βάσεις Δεδομένων και συνοπτικά τα στάδια που την αποτελούν. Η εφαρμογή της KDD, γίνεται μέσω τεχνικών πρόβλεψης και περιγραφής σε μεγάλες Βάσεις Δεδομένων Τέλος γίνεται μια σύντομη αναφορά στις κατηγορίες των αλγορίθμων οι οποίοι βρίσκουν εφαρμογή στο Data Mining.

Στο τρίτο κεφάλαιο γίνεται αναφορά σε ένα πεδίο της Τεχνητής Νοημοσύνης την Μηχανική Μάθηση. Αναφέρονται οι κατηγορίες Μηχανικής Μάθησης και αναφέρεται συνοπτικά η διαδικασία. Αν και η μελέτη των μοντέλων της Μηχανικής Μάθησης δεν αποτελεί στόχο της παρούσης πτυχιακής εργασίας ο λόγος που αναφέρεται είναι γιατί χρησιμοποιείται ως εργαλείο για την εξόρυξη πληροφορίας από δεδομένα στην διαδικασία της ανακάλυψης γνώσης (KDD).

Στο τέταρτο κεφάλαιο γίνεται λεπτομερής αναφορά στην έννοια των BIG DATA, δίνεται ένας ορισμός και καταγράφονται τα χαρακτηριστικά τους. Στη συνέχεια περιγράφονται οι έννοιες της ανάλυσης και της διαχείρισης των BIG DATA. Στη συνέχεια του κεφαλαίου καταγράφεται η χρησιμότητα των BIG DATA και τέλος τα πεδία εφαρμογής τους. Στο τέλος του κεφαλαίου αυτού στόχος μας είναι η αποσαφήνιση της έννοιας, της ιδιαιτερότητας που έχουν, την δύναμη που κρύβουν, των διαδικασιών που ακολουθούνται για το χειρισμό τους και τις επιπτώσεις που έχουν σε διάφορους τομείς της ζωής μας σαν τεχνολογία τα BIG DATA.

Στο πέμπτο κεφάλαιο γίνεται μια προσέγγιση των BIG DATA από τη σκοπιά της Ασφάλειας. Εδώ αρχικά αποσκοπούμε να αναδείξουμε την δυσκολία που παρουσιάζεται λόγω των ιδιαιτεροτήτων των BIG DATA να παραμείνουν ασφαλή τόσο από έμμεσες όσο και από άμεσες απειλές. Εξηγούμε το πώς καθίσταται δυνατό να προσαρμοστούν οι υπάρχουσες τεχνικές ασφαλείας σύμφωνα με τις απαιτήσεις των BIG DATA. Στη συνέχεια προσπαθούμε να εξηγήσουμε το πώς τα BIG DATA

Analytics μπορούν να βοηθήσουν στο να παραχθεί Ασφάλεια το πώς αντιμετωπίζει η κοινότητα της Ασφάλειας IT την ιδέα αυτή και ποια βήματα έχουν γίνει προς την κατεύθυνση αυτή. Τέλος γίνεται αναφορά σε υπάρχοντα συστήματα και τις αδυναμίες τους να εφαρμοστούν στα BIG DATA και μια σύντομη αναφορά στα Behavioral Analytics.

Στο έκτο κεφάλαιο κάνουμε αναφορά στα παράλληλα συστήματα και τα κατανεμημένα συστήματα, τον παράλληλο προγραμματισμό και τις παράλληλες Βάσεις Δεδομένων. Ο λόγος είναι για να γίνει κατανοητό το πόσο αναγκαίο είναι να γίνει παράλληλη επεξεργασία και ανάλυση σε σύνολα δεδομένων όπως τα BIG DATA όπως και την ανάγκη για χρήση πολύ-επίπεδων αρχιτεκτονικών DBMSs τόσο για την αποθήκευσή τους όσο κα για την ανάκτησή τους με αποδοτικό τρόπο. Στο τέλος αναφέρουμε ενδεικτικά το μοντέλο MapReduce,ένα από τα επικρατέστερα, το οποίο χρησιμοποιείται στην παράλληλη επεξεργασία για BIG DATA Analysis.

Στο έβδομο κεφάλαιο περιγράφουμε το τεχνολογικό περιβάλλον που πλαισίωσε την εκτέλεση των πειραμάτων. Στη συνέχεια γίνεται μια εκτενής αναφορά στην εφαρμογή που αναπτύξαμε. Εξηγούμε με σαφή τρόπο τόσο την λειτουργία της εφαρμογής με αναλύσεις τόσο για τις τεχνολογίες όσο και για τη μεθοδολογία που ακολουθήθηκε και μας καθοδήγησε μέχρι την ολοκλήρωση των πειραμάτων. Καταγράφουμε αναλυτικά τα βήματα μέχρι την ολοκλήρωση των πειραμάτων όπως και τα αποτελέσματα των πειραμάτων.

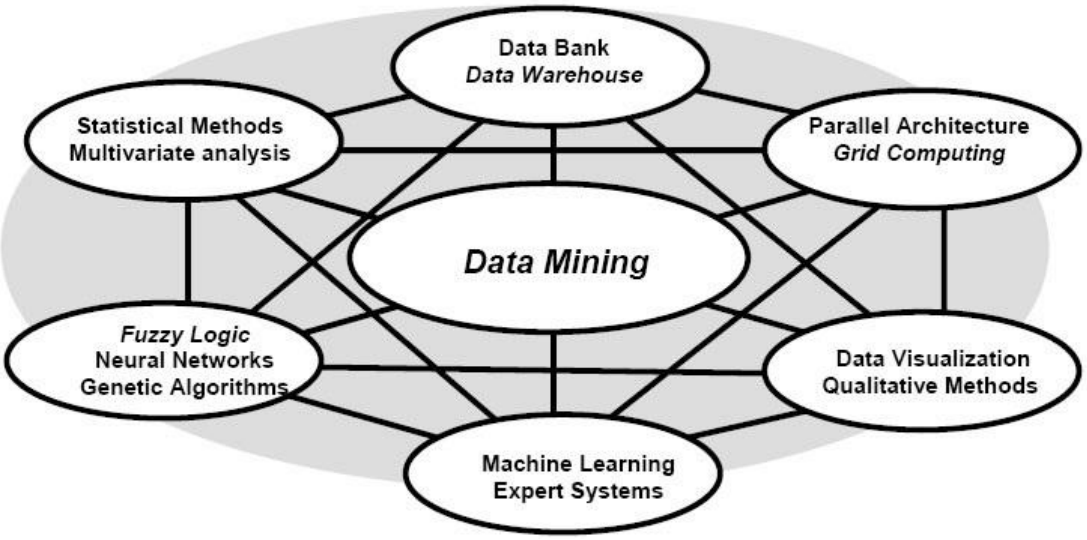
Στο όγδοο κεφάλαιο καταγράφουμε τα συμπεράσματα που βγήκαν μετά την ολοκλήρωση των πειραμάτων τα οποία εκτελέσαμε. Στη συνέχεια αναφέρουμε τις μελλοντικές επεκτάσεις που θέλουμε να έχει η εφαρμογή μας, οι οποίες θα καταστήσουν την εφαρμογή παραγωγική και αποτελεσματική τόσο σε χρόνο όσο και σε λειτουργικότητα.

2 Διεργασία Ανακάλυψης Γνώσης (Knowledge Discovery in Databases KDD)

2.1 Εισαγωγή

Τα δεδομένα αποθηκεύονται σε Βάσεις Δεδομένων. Οι Βάσεις Δεδομένων είναι πλέον παντού και αποτελούν βασικό σημείο αναφοράς για την επιστήμη της πληροφορικής και για τα πληροφοριακά συστήματα. Οι διαχείριση των Βάσεων Δεδομένων είναι κάτι πολύ σημαντικό, αφού αυτές διαθέτουν «κρυμμένη» την γνώση που οδηγεί στην επιτυχία κάθε πληροφοριακού συστήματος και κατά συνέπεια αυτών που τα χρησιμοποιούν.

Για να κατανοήσουμε την σημαντικότητα της διαχείρισης των Βάσεων Δεδομένων ας αναλογιστούμε το εξής εντυπωσιακό, ότι υπολογίζεται πως όσα δεδομένα είναι ικανός να διαβάσει ένας άνθρωπος σε ολόκληρη την ζωή του, αυτά είναι λιγότερα από τα δεδομένα που παράγει σε μία εβδομάδα ένας μεγάλος οργανισμός.



Εικόνα 2.1 Η Εξόρυξη Γνώσης [1]

Παρατηρείται πλέον ενσωμάτωση γλωσσών προγραμματισμού σε συστήματα Βάσεων Δεδομένων, με αποτέλεσμα την ενοποίηση αλγορίθμων και δεδομένων. Παρατηρείται η χρήση επεκτάσιμων αντικειμενοσχεσιακών Βάσεων Δεδομένων, τρισδιάστατες Βάσεις Δεδομένων και άλλες τεχνολογίες [53].

Επίσης τα συστήματα βάσης δεδομένων διαθέτουν πλέον ένα τεχνολογικό πλαίσιο για το Data Mining (Εικόνα 2.1), που αποτελεί μέρος της KDD. Στα συστήματα διαχείρισης Βάσεων Δεδομένων προστίθενται νέοι τύποι δεδομένων, υποστήριξη XML δομής και αιτημάτων κ.ά.

2.2 Διεργασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

2.2.1 Εισαγωγή στην KDD

Ο Όρος Data Mining δεν ταυτίζεται με τον όρο KDD σε Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD). Η KDD σε Βάσεις Δεδομένων αναφέρεται ως ένα σύνολο βημάτων, ενώ το Data Mining αποτελεί ένα από τα βήματα αυτής της διαδικασίας.

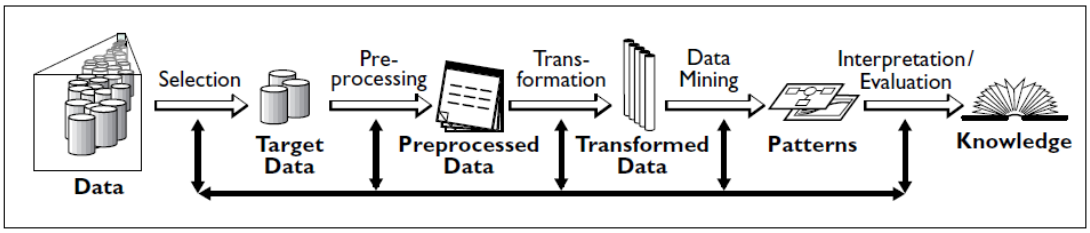
Είναι σαφές ότι η διεργασία ανακάλυψης γνώσης (KDD–Knowledge Discovery in Databases) δεν είναι κάτι εύκολο, αλλά μια μακράς διάρκειας διαδικασία που συνάδει με το τί θέλουμε να πάρουμε από τα δεδομένα μας.

Υπάρχουν σαφώς πολλά προβλήματα κατά την διαδικασία, όπως το γεγονός ότι τα δεδομένα μπορεί να μην είναι ενημερωμένα την χρονική στιγμή που λαμβάνει χώρα η διαδικασία, ενώ άλλα δεδομένα δεν μπορούν καν να χρησιμοποιηθούν στην διαδικασία διότι παρεμβαίνουν την ιδιωτικότητα και άλλα νομικά θέματα. Ο διαφορετικός τύπος των δεδομένων είναι κάτι που επίσης πρέπει να ληφθεί υπόψη, πόσο μάλλον όταν αναφερόμαστε σε BIG DATA. Ας μην ξεχνάμε επίσης τα προβλήματα σημασιολογικής διερμηνείας και νοήματος που υπάρχουν μεταξύ των ενοποιημένων Βάσεων Δεδομένων.

Αρχίζοντας μια μελέτη σχετικά με την KDD πρέπει αρχικά να επισημάνουμε κάποιες βασικές αρχές. Προφανώς το πρώτο ερώτημα που γεννιέται στον καθένα είναι τι είναι ακριβώς η KDD. Ένας ορισμός της KDD είναι ο παρακάτω:

«Η ανακάλυψη γνώσης σε Βάσεις Δεδομένων είναι μια ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα» [6].

Η KDD από Βάσεις Δεδομένων είναι μια διαλογική και επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων (Εικόνα 2.2)



Εικόνα 2.2 Διαδικασία Ανακάλυψη Γνώσης KDD [2]

Η KDD ,όπως προείπαμε, σε Βάσεις Δεδομένων είναι μία σύνθετη διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών σχέσεων-προτύπων σε δεδομένα. Αν και ως όρος είναι σχετικά πρόσφατος, λαμβάνει χώρα σε πραγματικές συνθήκες και σε μεγάλη κλίμακα των ερευνητικών αποτελεσμάτων της Μηχανικής Μάθησης και της Στατιστικής.

Η KDD είναι μία ολοκληρωμένη διαδικασία που περιλαμβάνει γενικά την επεξεργασία των δεδομένων, την εφαρμογή των **αλγορίθμων ανακάλυψης γνώσης** (κυρίως αλγόριθμοι Μηχανικής Μάθησης) και τέλος την ερμηνεία των αποτελεσμάτων. Χρησιμοποιεί τεχνικές από πολλούς τομείς, όπως Στατιστική, Μηχανική Μάθηση, Βάσεις Δεδομένων, Αναγνώριση Προτύπων, Πράκτορες κτλ. Κάτι σημαντικό, το οποίο πρέπει να ληφθεί σοβαρά υπόψη είναι ότι πολλές φορές μπορεί να χρειαστεί κάποια από τα επιμέρους βήματα να επαναληφθούν. Αυτό μπορεί να συμβεί, γιατί στην πορεία πιθανώς να εμφανιστούν προβλήματα που να έχουν

σχέση με τις αρχικές επιλογές και τα οποία δεν ήταν δυνατόν να εντοπιστούν από την αρχή της διαδικασίας.

Είναι, δηλαδή, μία διαδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πολυάριθμα βήματα με πολλές από τις αποφάσεις να λαμβάνονται από τον αναλυτή.

Φυσικά είναι πολλοί οι λόγοι που οδήγησαν στην ανάπτυξη της διαδικασίας αυτής, αναφέροντας τους βασικούς [3]

- Αύξηση της διαθεσιμότητας των δεδομένων.
- Αυτοματοποιημένη παραγωγή δεδομένων.
- Διαθέσιμη ποσότητα των δεδομένων που αυξάνεται ραγδαία

Σύμφωνα με τους Roiger και Geatz από τις Βάσεις Δεδομένων μπορούμε να λάβουμε γνώση, η οποία χωρίζεται σε τέσσερις διαφορετικές κατηγορίες [3]

- Ρηχή **(Shallow)**
- Πολυδιάστατη **(Multidimensional)**
- Κρυφή **(Hidden)**
- Σε βάθος **(Deep)**

Να σημειώσουμε εδώ ότι οι γλώσσες ερωτημάτων Βάσεων Δεδομένων , όπως η SQL, δεν είναι ανταγωνιστικές προς την διαδικασία του Data Mining, που είναι μέρος της KDD. Αυτά τα δύο εργαλεία είναι συμπληρωματικά μεταξύ τους. Οι γλώσσες αιτημάτων Βάσεων Δεδομένων ανακτούν πληροφορία από τα δεδομένα βασισμένες σε περιορισμούς και συσχετίσεις μεταξύ των δεδομένων τα οποία τα γνωρίζουν εξαρχής. Η διαδικασία του Data Mining ανακτά πληροφορία από τα

δεδομένα βασισμένη σε πρότυπα και τάσεις που υπάρχουν σε συγκεκριμένα γκρουπ δεδομένων της βάσης. Γενικά χρησιμοποιούμε SQL όταν γνωρίζουμε ακριβώς για το τί ψάχνουμε. Αντίθετα χρησιμοποιούμε την διαδικασία του Data Mining όταν είναι αόριστο αυτό που ψάχνουμε να βρούμε. Οι γλώσσες αιτημάτων Βάσεων Δεδομένων υπολογίζουν ένα σύνολο αποτελεσμάτων που είναι ένα υποσύνολο της βάσης δεδομένων ενώ η KDD, κάνοντας χρήση του Data Mining, παράγει ένα KDD Object μέσω κανόνων , ταξινομήσεων και ομαδοποιήσεων.

Όσο αναφορά την γνώση που μπορούμε να λάβουμε από τις Βάσεις Δεδομένων παρατηρούμε τα εξής:

Η πρώτη κατηγορία γνώσης **(Shallow)** αποθηκεύεται σε Βάσεις Δεδομένων και εξάγεται με απλά ερωτήματα Βάσεων Δεδομένων (SQL queries).

Η δεύτερη κατηγορία **(Multidimensional)** αποθηκεύεται σε πολλών διαστάσεων Βάσεις Δεδομένων. Χρησιμοποιείται η τεχνολογία OLAP και ανάλογα εργαλεία στα πολυδιάστατα δεδομένα.

Η τρίτη κατηγορία **(Hidden)** βασίζεται σε πρότυπα και κανόνες στα δεδομένα. Είναι δύσκολο να εντοπιστεί χρησιμοποιώντας εργαλεία αιτημάτων Βάσεων Δεδομένων αλλά πολύ εύκολο να εντοπιστεί με χρήση αλγορίθμων Data Mining.

Η τέταρτη κατηγορία **(Deep)** παρουσιάζεται και αυτή στις Βάσεις Δεδομένων. Μπορεί μόνο να βρεθεί αν μας δοθεί η κατεύθυνση αναζήτησης για το τι πρόκειται να αναζητήσουμε. Τα σημερινά εργαλεία ερωτημάτων Βάσεων Δεδομένων και οι αλγόριθμοι Data Mining που υπάρχουν δεν είναι ικανά να εντοπίσουν την γνώση αυτή.

Η πληροφορία είναι αυτή που κάνει την μεγάλη διαφορά στην αναζήτηση γνώσης και όχι τα δεδομένα. Η επεξεργασμένη ανάλυση και χρήση των δεδομένων είναι η πληροφορία και σαφώς έγκειται η ανθρώπινη παρέμβαση στην όλη διαδικασία. Υπάρχει ένα μεγάλο ερώτημα το οποίο τίθεται σήμερα στην επιστημονική κοινότητα. Θα μπορέσουμε άραγε κάποτε να σχηματοποιήσουμε και μετρήσουμε την πληροφορία;

Σαν ανακάλυψη γνώσης λοιπόν ορίζεται η διαδικασία όπου από τεράστιες ποσότητες, «συντρίμμια», δεδομένων θα καταφέρουμε να απομονώσουμε τους «σβόλους χρυσού», που είναι η παραγόμενη γνώση που θα βρεθεί. Το 80% της διαδικασίας θεωρείται η προετοιμασία των δεδομένων, όπως η απομάκρυνση διπλοεγγραφών, η επιδιόρθωση τυπογραφικών λαθών κ άλλα και έπειτα παράγεται η αναπαράσταση του μοντέλου της διαδικασίας της επεξεργασίας των δεδομένων.

Συγκεκριμένα, η KDD είναι μια διαδικασία 5 βασικών σταδίων.

- Επιλογή δεδομένων
- Προεπεξεργασία δεδομένων
- Μετασχηματισμός δεδομένων
- Εξόρυξη δεδομένων
- Ερμηνεία – Αξιολόγηση δεδομένων

2.2.2 Επιλογή δεδομένων

Στο στάδιο αυτό δημιουργείται το σύνολο δεδομένων, πάνω στο οποίο θα εφαρμοστεί η αναζήτηση γνώσης. Οι αλγόριθμοι που εκτελούν την διαδικασία, συνήθως, δεν μπορούν να χρησιμοποιήσουν τα δεδομένα με την μορφή στην οποία είναι εξ αρχής οργανωμένα. Γι’ αυτόν ακριβώς το λόγο απαιτείται η εξαγωγή των δεδομένων αυτών, από τους πολλαπλούς πίνακες και η οργάνωσή τους σε απλούστερες και πιο εύχρηστες δομές. Δημιουργούμε νέα πεδία συνδυάζοντας τα ήδη υπάρχοντα και φέρνουμε τα δεδομένα στο σχεσιακό σχήμα που θα χρησιμοποιήσουμε στην είσοδο της επεξεργασίας. Το στάδιο αυτό των δεδομένων μπορεί να περιλαμβάνει και την αποκανονικοποίηση των σχετικών πινάκων.

Συνήθως, η ανάγκη αυτή ικανοποιείται με τη χρήση συστημάτων αποθήκευσης δεδομένων (Data warehouse), τα οποία παρέχουν στη διαδικασία μία πιο εύκολη και προσβάσιμη οπτική των δεδομένων [3][4].

2.2.3 Προεπεξεργασία δεδομένων

Στο στάδιο αυτό αντιμετωπίζονται περιπτώσεις ελλιπών δεδομένων (όπως άδεια πεδία), πεδίων με τιμές που ουσιαστικά τα καθιστούν κενά, (όπως Οδός = Άγνωστο) κλπ. Το στάδιο αυτό μπορεί να ονομαστεί και στάδιο καθαρισμού των δεδομένων (Data cleaning), εξαιτίας των διαδικασιών που λαμβάνουν χώρα σε αυτό. Στο στάδιο αυτό περιλαμβάνεται ακόμα, η αφαίρεση του θορύβου από τα δεδομένα, όταν αυτό χρειάζεται, συλλέγοντας τις απαραίτητες πληροφορίες για τη διαμόρφωση ή την περιεκτικότητα του θορύβου, παίρνοντας έτσι αποφάσεις για τις στρατηγικές όσον αφορά τη διαχείριση των ελλιπών πεδίων δεδομένων [3][4].

2.2.4 Μετασχηματισμός δεδομένων

Τα δεδομένα μετασχηματίζονται έτσι ώστε να διευκολύνουν την ανακάλυψη γνώσης. Τέτοιοι μετασχηματισμοί μπορεί να περιλαμβάνουν για παράδειγμα, τη μείωση του αριθμού των υπό εξέταση χαρακτηριστικών (dimensionality reduction) με επιλογή ορισμένων εξ' αυτών (feature selection ή attribute selection), την ομοιόμορφη κωδικοποίηση της ποιοτικά ίδιας πληροφορίας, τη μετατροπή συνεχόμενων αριθμητικών τιμών σε διακριτές τιμές (διακριτοποίηση) και πολλά άλλα. Οι μετασχηματισμοί αυτοί γίνονται ανάλογα με τον στόχο της διαδικασίας [3][4].

2.2.5 Εξόρυξη δεδομένων

2.2.5.1 Μοντέλο Εξόρυξης Δεδομένων

Το Data Mining είναι μια διαδικασία με στόχο την εξαγωγή της κρυμμένης πληροφορίας από μεγάλες Βάσεις Δεδομένων. Υπάρχουν προφανώς πολλοί ορισμοί για το Data Mining, παρακάτω παραθέτουμε δυο από αυτούς:

«Είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων.» (David Hand , Heikki Mannila, and Padhraic Smyth, 2001).

Ένας πιο αυστηρός και τυπικός ορισμός του Data Mining, είναι ο εξής:

«Είναι η διαδικασία εξαγωγής υπονοούμενης και άγνωστης, αλλά ενδεχομένως χρήσιμης γνώσης, υπό την μορφή συσχετίσεων, προτύπων και τάσεων, μέσω της εξέτασης, ανάλυσης και επεξεργασίας Βάσεων Δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την Μηχανική Μάθηση, την Αναγνώριση Προτύπων, την Στατιστική, τις Βάσεις Δεδομένων και την Οπτικοποίηση». (Larose DT., 2004).

Στο στάδιο αυτό καθορίζονται οι στόχοι της KDD και γίνεται η επιλογή της στρατηγικής Data Mining που θα χρησιμοποιηθεί. Η επιλογή του αλγορίθμου έμμεσα προσδιορίζει και την κατηγορία αλγορίθμου που θα χρησιμοποιηθεί. Αν θα είναι αλγόριθμος κατηγοριοποίησης, συσταδοποίησης, και ούτως καθεξής. Η εφαρμογή του αλγορίθμου είναι ένα καθαρά υπολογιστικό στάδιο, στο οποίο γίνεται η ουσιαστική αναζήτηση της γνώσης από τα δεδομένα. Περιγράφεται και με τον όρο Data Mining, ο οποίος πολλές φορές χρησιμοποιείται καταχρηστικά για να περιγράψει ολόκληρη τη KDD [4].

Οι βασικοί στόχοι της KDD, είναι η εφαρμογή τεχνικών πρόβλεψης (prediction) και περιγραφής (description) σε μεγάλες Βάσεις Δεδομένων (Usama Fayyad, Gregory Piatetsky-ShapiroG, SmythP. , 1996).

Ποιο συγκεκριμένα:

Η **πρόβλεψη** περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της KDD(predictive), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα.

Η **περιγραφή** επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο - κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της KDD(descriptive) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

Αν και τα όρια μεταξύ της πρόβλεψης και της περιγραφής δεν είναι απολύτως ξεκάθαρα (μερικά από τα πρότυπα πρόβλεψης μπορούν να είναι περιγραφικά, στο βαθμό που είναι κατανοητά και αντίστροφα), η διάκριση είναι χρήσιμη για την κατανόηση του γενικού στόχου ανακάλυψης. Η σχετική σημασία της πρόβλεψης και της περιγραφής για συγκεκριμένες εφαρμογές Data Mining, μπορεί να ποικίλει αρκετά [4].

Για να επιτύχουμε τους παραπάνω στόχους της KDD, μπορούμε να εφαρμόσουμε διάφορες τεχνικές, Data Mining με σημαντικότερες να είναι οι εξής:

Το προβλεπτικό μοντέλο απαρτίζεται από τις ακόλουθες μεθόδους Data Mining [5]:

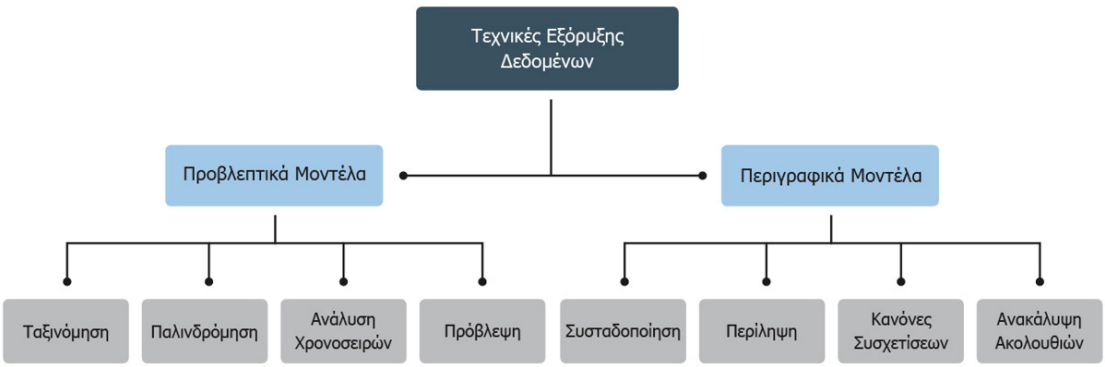
- Κατηγοριοποίηση ή Ταξινόμηση **(Classification)**
- Παλινδρόμηση **(Regression)**
- Ανάλυση Χρονοσειρών **(Time Series Analysis)**
- Πρόβλεψη **(Prediction)**

Το περιγραφικό μοντέλο απαρτίζεται από τις ακόλουθες μεθόδους Data Mining:

- Συσταδοποίηση **(Clustering)**
- Περίληψη **(Summarization)**
- Κανόνες Συσχετίσεων **(Association Rules)**
- Ανακάλυψη Ακολουθιών **(Sequential Pattern Discovery)**

2.2.5.2 Τεχνικές Εξόρυξης Δεδομένων (Data Mining)

Στην συνέχεια αναλύουμε τις αναφερόμενες κατηγορίες τεχνικών Data Mining (Εικόνα 2.3).



Εικόνα 2.3 Τεχνικές Εξόρυξης Δεδομένων [5]

- **Κατηγοριοποίηση ή Ταξινόμηση**

Η κατηγοριοποίηση **(Classification)** αποτελεί μια από τις βασικές εργασίες (tasks) της KDD, Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods).

Οι τεχνικές της κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα θα είναι συνδεδεμένα με γνωστές

κλάσεις. Ο αλγόριθμος ταξινόμησης «μαθαίνει» από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου. Το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα (testing set) στις κατάλληλες κλάσεις [7].

Άρα μπορούμε να πούμε ότι η κατηγοριοποίηση μαθαίνει σε μία λειτουργία να χαρτογραφεί (ταξινομεί) ένα στοιχείο δεδομένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες. Παραδείγματα τέτοιων μεθόδων, οι οποίες χρησιμοποιούνται ως τμήμα των εφαρμογών της KDD, περιλαμβάνουν την ταξινόμηση των τάσεων στις χρηματοοικονομικές αγορές και τον αυτοματοποιημένο προσδιορισμό των αντικειμένων ενδιαφέροντος για τις μεγάλες Βάσεις Δεδομένων.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ-κατηγοριοποιημένα σύνολα δεδομένων. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί (να ανατεθεί σε κάποια από τις κατηγορίες). Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε και να αναφέρουμε ότι οι πιο γνωστές τεχνικές ταξινόμησης είναι [4][5]:

- Naive Bayes μοντέλα
- Δένδρα αποφάσεων (Decision Trees)
- Κοντινότεροι γείτονες (K-Nearest Neighbor)
- Νευρωνικά Δίκτυα (Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (SVM)

❖ **Παλινδρόμηση**

Η παλινδρόμηση **(Regression)** είναι η παλαιότερη και η πλέον γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων του Data Mining. Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν.

Συγκεκριμένα, η παλινδρόμηση, χρησιμοποιώντας μια βάση αριθμητικών δεδομένων, αναπτύσσει μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά. Στην συνέχεια, η μαθηματική αυτή σχέση χρησιμοποιείται για την πρόβλεψη μελλοντικής συμπεριφοράς, καθώς εφαρμόζεται σε νέα αριθμητικά δεδομένα. Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (όπως π.χ. βάρος, ταχύτητα ή ηλικία). Αντίθετα, η παλινδρόμηση δεν λειτουργεί καλά με κατηγορικά δεδομένα [5][7].

Γνωστές μέθοδοι παλινδρόμησης είναι:

- Γραμμική παλινδρόμηση (Linear)
- Λογιστική παλινδρόμηση (Logistic)
- Δένδρα παλινδρόμησης (Regression Trees)
- Νευρωνικά Δίκτυα (Neural Networks)

❖ **Ανάλυση Χρονοσειρών**

Στόχος της ανάλυσης χρονοσειρών **(Time Series Analysis)** είναι η μελέτη της μεταβολής της τιμής ενός μεγέθους με το πέρασμα του χρόνου. Οι βασικές λειτουργίες αυτής της τεχνικής Data Mining είναι η εξέταση της δομής μιας χρονοσειράς, η εύρεση ομοιοτήτων μεταξύ χρονοσειρών και τέλος η χρήση διαγραμμάτων χρονοσειρών με στόχο την πρόβλεψη μελλοντικών τιμών [5].

❖ **Πρόβλεψη**

Η τεχνική της πρόβλεψης **(Prediction)**, έχοντας διαθέσιμα ιστορικά και τωρινά δεδομένα χρησιμοποιείται σε εφαρμογές που μπορούν να θεωρηθούν σαν πρόβλεψη μελλοντικών καταστάσεων. Οποιεσδήποτε από τις τεχνικές που χρησιμοποιούνται για την ταξινόμηση μπορούν να προσαρμοστούν και στην πρόβλεψη με τη χρήση των παραδειγμάτων εκπαίδευσης όπου η τιμή της μεταβλητής που προβλέπεται είναι ήδη γνωστή, μαζί με τα ιστορικά στοιχεία. Στη συνέχεια χρησιμοποιείται το ιστορικό στοιχείο για να χτίσει ένα μοντέλο που εξηγεί την παρατηρηθείσα συμπεριφορά. Τέλος όταν αυτό το μοντέλο εφαρμόζεται στις τρέχουσες εισαγωγές, το αποτέλεσμα είναι μια πρόβλεψη της μελλοντικής συμπεριφοράς.

❖ **Συσταδοποίηση**

Η συσταδοποίηση ή ομαδοποίηση **(Clustering)** είναι ένας κοινός περιγραφικός στόχος, όπου κάποιος επιδιώκει να προσδιορίσει ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων (clusters) για να περιγράψει τα δεδομένα [8]. Σύμφωνα με τους Han&Kamber, διακρίνονται τρεις βασικές κατηγορίες μεθόδων clustering:

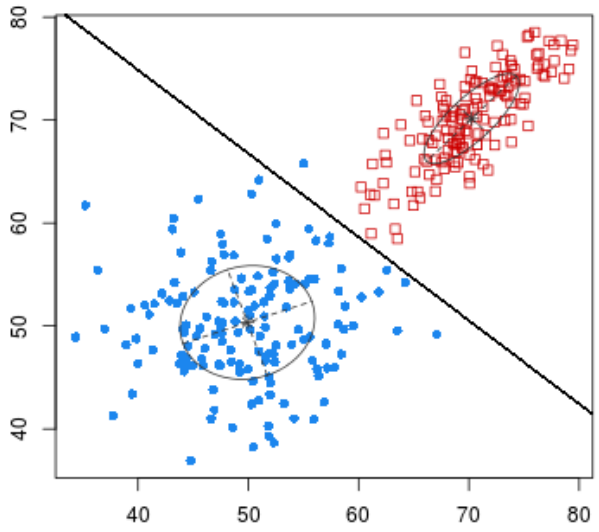
1. Μέθοδοι διαχωρισμού (partitioning methods): δημιουργούν k ομάδες από ένα δεδομένο αρχικό σύνολο n αντικειμένων με κάθε ομάδα να αντιπροσωπεύει ένα cluster και να ικανοποιούνται οι εξής δύο συνθήκες:
    a) κάθε cluster περιέχει τουλάχιστον ένα αντικείμενο και
    b) κάθε αντικείμενο ανήκει σε ένα μόνο cluster.
2. Ιεραρχικές μέθοδοι (hierarchical methods): διασπούν το αρχικό σύνολο δεδομένων δημιουργώντας μια ιεραρχική δομή από clusters και διακρίνονται σε agglomerative (bottom-up) ή divisive (top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.
3. Μέθοδοι βασισμένες σε μοντέλα (model-based methods): υποθέτουν ότι καθένα από τα clusters περιγράφεται από ένα μαθηματικό μοντέλο και

εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε cluster, ώστε να ικανοποιούν το αντίστοιχο μοντέλο.

Αυτό που διαφοροποιεί τη συσταδοποίηση από την κατηγοριοποίηση είναι ότι η συσταδοποίηση δε βασίζεται σε προκαθορισμένες κατηγορίες. Στην κατηγοριοποίηση, ο πληθυσμός διαιρείται σε κατηγορίες αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων. Όπως και στην κατηγοριοποίηση έτσι και στη συσταδοποίηση υπάρχουν πολλές εφαρμογές. Για παράδειγμα, ας θεωρήσουμε πως έχουμε διαθέσιμα τα δεδομένα πελατών μιας εταιρίας πωλήσεων. Χρησιμοποιώντας τεχνικές συσταδοποίησης, μπορούμε να βρούμε τον καταμερισμό των πελατών και της αγοράς, π.χ. μπορούμε να δούμε ποιοι πελάτες αγοράζουν για την οικογένεια τους και ποιοι για τον εαυτό τους ή ποιοι έχουν μεγάλο εισόδημα και ποιοι όχι.

Άλλο παράδειγμα των εφαρμογών συσταδοποίησης σε ένα πλαίσιο ανακάλυψης γνώσης, περιλαμβάνουν την ανακάλυψη ομοιογενών υποσυνόλων πληθυσμού για τους καταναλωτές που υπάρχουν στις Βάσεις Δεδομένων του τομέα του μάρκετινγκ.

Η Εικόνα παρουσιάζει έναν απλό διαχωρισμό στοιχείων σε δύο περιοχές **κατηγοριών (μαύρη γραμμή)**. Η Εικόνα επίσης παρουσιάζει μία πιθανή ομαδοποίηση του συνόλου των δεδομένων, σε δύο **συστάδες (μαύροι κύκλοι με κέντρο)**. Αξίζει να σημειωθεί ότι στις συστάδες υπάρχει πιθανότητα επικάλυψης αν στα κέντρα δεν αντιστοιχούν διακριτές τιμές επιτρέποντας έτσι στα σημεία των δεδομένων να ανήκουν



σε περισσότερες από μία συστάδες. Επίσης οι αρχικές ετικέτες όλων των συστάδων θα ήταν όμοιου σχήματος + , όπου αργότερα υποδεικνύονται από τα * και τα ο, για να δείξουν ότι η ιδιότητα μέλους συστάδας θεωρείται **πλέον γνωστή**[4].

- **Περίληψη**

Η περίληψη **(Summarization)** περιλαμβάνει μεθόδους με στόχο την εύρεση συμπαγούς περιγραφής υποσυνόλου των δεδομένων. Πιο αναλυτικά, χαρακτηρίζει τα δεδομένα παράγοντας αντιπροσωπευτικές πληροφορίες, γεγονός που συμβάλλει στην ανάδειξη και κατανόηση μερικών γνωρισμάτων τους [5]. Έννοιες της Στατιστικής που εξυπηρετούν αυτόν το σκοπό είναι:

- Μέσος
- Διακύμανση
- Τυπική Απόκλιση
- Ιστόγραμμα
- Διάγραμμα Διασποράς

- **Κανόνες Συσχέτισης**

Η εξαγωγή κανόνων συσχέτισης **(Association Rules)** θεωρείται μια από τις σημαντικότερες διεργασίες Data Mining. Έχει προσελκύσει μεγάλο ενδιαφέρον γιατί παρέχουν έναν συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στην ακόλουθη μορφή A→B όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα.

Οι κανόνες συσχέτισης χρησιμοποιούνται για τον υπολογισμό της πιθανότητας να συμβεί το B, με δεδομένο το ότι συνέβη το A. Η επιλογή ενός κανόνα συσχέτισης

και η αποτίμησή του ως ενδιαφέρον εξαρτάται από τις τιμές των μεγεθών support (συχνότητα εμφάνισης του item set AUB στην αρχική συλλογή) και confidence (την υπό-συνθήκη προβλεψιμότητα του B με δεδομένο το A) αλλά και άλλων μεγεθών (lift). Ο πλέον δημοφιλής αλγόριθμος για την ανακάλυψη κανόνων συσχέτισης είναι ο Apriori[3][4].

- **Πρότυπα Ακολουθιών**

Η εξόρυξη πρότυπων ακολουθιών **(Sequential Patterns)** είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες με άλλα κοινά χαρακτηριστικά. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνονται στα συμβολικά πρότυπα. Ο χρήστης εδώ μπορεί να προσδιορίσει τους περιορισμούς στα είδη των προτύπων ακολουθιών που εξάγονται με την παροχή των προσχεδίων προτύπων (template patterns) υπό μορφή σειριακών επεισοδίων, παράλληλων επεισοδίων ή κανονικών εκφράσεων. Παραδείγματα προτύπων ακολουθιών έχουμε στην καθημερινή μας ζωή όπως τα κείμενα, οι νότες, τα δεδομένα του καιρού και οι ακολουθίες του DNA [4].

2.2.6 Ερμηνεία – Αξιολόγηση δεδομένων

Στο πέμπτο και τελευταίο στάδιο της KDD, γίνεται ερμηνεία και αξιολόγηση των ευρεθέντων προτύπων, πιθανώς με υποβοήθηση γραφικών απεικονίσεων των προτύπων ή/και των δεδομένων, τα οποία περιγράφονται από το πρότυπο (pattern/Data visualization). Η γνώση που παράγεται μπορεί να χρησιμοποιηθεί σε ένα σύστημα γνώσης, όμως στην περίπτωση αυτή είναι πολύ πιθανόν να **υπάρξουν κάποιες συγκρούσεις (conflicts) μεταξύ της υπάρχουσας γνώσης και της παραγόμενης [4].**

Η KDD είναι ουσιαστικά μια διαδικασία όπου εξάγουμε από κάτι ασήμαντο, κάτι πολύ σημαντικό, τη γνώση. Η γνώση είναι οι ουσιαστικές συσχετίσεις που

υπάρχουν μεταξύ των στοιχείων των δεδομένων. Την νέα γνώση, που δεν είναι αρχικά προφανής, μπορεί κανείς να την χρησιμοποιήσει και να επωφεληθεί από αυτή. Είναι σημαντικό να τονίσουμε πως η διαδικασία της KDD μπορεί να χρησιμοποιηθεί και για πιο αδόμητα δεδομένα [3].

2.3 Συμπεράσματα

Σε αυτό το κεφάλαιο αναφέρθηκαν κάποιες βασικές έννοιες οι οποίες αφορούν την γνώση από δεδομένα. Είδαμε τα στάδια της KDD σε Βάσεις Δεδομένων ενώ αναφέρθηκε ότι το Data Mining είναι ένα από αυτά τα στάδια. Επίσης είδαμε σε ποια πεδία των επιστημών μπορούν να εφαρμοστούν οι τεχνικές. Τέλος κάναμε μια σύντομη αναφορά στις κατηγορίες των αλγορίθμων οι οποίοι βρίσκουν εφαρμογή στο Data Mining.

3 Μηχανική Μάθηση (Machine Learning)

3.1 Εισαγωγή

Η Μηχανική Μάθηση αποτελεί έναν κλάδο της Πληροφορικής, ο οποίος μελετά μοντέλα και αλγορίθμους που βασίζονται σε παρατηρούμενα δεδομένα και εφαρμόζονται σε ένα ευρύ φάσμα ερευνητικών περιοχών.

Πιο συγκεκριμένα η Μηχανική Μάθηση εντάσσεται στο πεδίο της Τεχνητής Νοημοσύνης **(Artificial Intelligence)** που ασχολείται με την κατασκευή και τη μελέτη συστημάτων τα οποία μπορούν να «μαθαίνουν» από τα δεδομένα αντί να ακολουθούν ρητές προγραμματιστικές οδηγίες. Η διαδικασία της Μηχανικής Μάθησης είναι παρόμοια με αυτήν της εξόρυξης δεδομένων (Data Mining) που περιγράφηκε στο Κεφάλαιο 1. Και τα δύο συστήματα αναλύουν δεδομένα και αναζητούν μοτίβα συμπεριφοράς. Ωστόσο αντί της εξαγωγής δεδομένων για την ανθρώπινη κατανόηση, όπως συμβαίνει σε εφαρμογές Data Mining, η Μηχανική Μάθηση χρησιμοποιεί τα δεδομένα για να βελτιώσει την κατανόηση του ίδιου του προγράμματος. Έτσι τα προγράμματα Μηχανικής Μάθησης ανιχνεύουν μοτίβα στα δεδομένα και προσαρμόζουν τις δράσεις τους με ανάλογο τρόπο[11].

Κύριος στόχος της Μηχανικής Μάθησης είναι να δημιουργήσει προγράμματα υπολογιστών που μέσα από την εμπειρική απόκτηση και την ενοποίηση γνώσεων (από μια συλλογή παρατηρήσεων που καλείται σύνολο εκπαίδευσης) κατορθώνουν να βελτιώνονται συνεχώς.

Ο βασικός λόγος, για τον οποίο κρίνεται απαραίτητη η περαιτέρω ανάπτυξή της, έγκειται στο ότι παρέχει τη δυνατότητα αυτόματης επίλυσης περίπλοκων προβλημάτων, τα οποία φαίνονται απρόσιτα στον ανθρώπινο νου, κυρίως λόγω του τεράστιου όγκου δεδομένων που πρέπει να χρησιμοποιηθούν για την επίλυσή τους.

Ένας ορισμός που θα μπορούσαμε να δώσουμε για τον όρο **Μηχανική Μάθηση** είναι ο εξής:

Μηχανική Μάθηση είναι η συλλογή αλγορίθμων και μεθόδων με τις οποίες βελτιώνεται η αποδοτικότητα μιας μηχανής στην εκτέλεση «ευφυών» εργασιών.

Ένα απλό παράδειγμα θα μπορούσε να είναι η ανακάλυψη ενός συγκεκριμένου αλλά και πολύπλοκου μοτίβου μέσα σε ένα πολύ μεγάλο σύνολο από αλυσίδες τεράστιου μήκους. Ένα τέτοιο πρόβλημα, ασφαλώς, μοιάζει απίθανο να λάβει άμεσης και έγκυρης απάντησης, ακόμα και από κάποιον εμπειρογνώμονα στο είδος, χωρίς τη χρήση κάποιου υπολογιστικού συστήματος. Σε τέτοιου είδους περιπτώσεις, λοιπόν, οδηγούμαστε υποχρεωτικά σε λύσεις που μας παρέχουν οι διάφορες μέθοδοι της Μηχανικής Μάθησης, τις οποίες προσαρμόζουμε στο εκάστοτε πρόβλημα που μας απασχολεί [10].

3.2 Κατηγορίες Μηχανικής Μάθησης

Τα προβλήματα Μηχανικής Μάθησης κατατάσσονται σε τρεις κύριες κατηγορίες ανάλογα με την είσοδο για τον αλγόριθμο, αλλά και την έξοδο που επιθυμούμε να δώσει ο αλγόριθμος που χρησιμοποιείται. Έτσι, διακρίνουμε τις κατηγορίες [10][11]:

- της μάθησης με επίβλεψη **(Supervised Learning**)

  Ο αλγόριθμος χρησιμοποιεί τα πρότυπα εισόδου x_i μαζί με την επιθυμητή τιμή εξόδου (στόχος) α.
- της μάθησης χωρίς επίβλεψη **(Unsupervised Learning)**

  Ο αλγόριθμος χρησιμοποιεί μόνο τα πρότυπα εισόδου x_i για να εκπαιδευτεί χωρίς να χρησιμοποιεί στόχους.
- της ενισχυτικής μάθησης **(Reinforcement Learning)**

  Ο αλγόριθμος χρησιμοποιεί μόνο τα πρότυπα εισόδου x_i χωρίς να γνωρίζει αν πάει καλά ή όχι, παρά μόνο στο τέλος. Τότε μόνο γνωρίζει αν πέτυχε το στόχο

του και χρησιμοποιεί αυτή την πληροφορία για να βελτιωθεί την επόμενη φορά.

Στην πρώτη κατηγορία, ανήκουν οι περιπτώσεις αλγορίθμων που δημιουργούν μια συνάρτηση τέτοια, ώστε να αντιστοιχίζει μια συγκεκριμένη είσοδο σε κάποια αυστηρά καθορισμένη έξοδο. Κάτι τέτοιο επιτυγχάνεται μέσω ενός συνόλου εκπαίδευσης, αποτελούμενο από παραδείγματα ζευγών εισόδου και επιθυμητής εξόδου [10].

Πιο συγκεκριμένα, στην επιβλεπόμενη μάθηση γνωρίζουμε ποια είναι η σωστή μορφή της εξόδου για το σύνολο των δεδομένων που εξετάζουμε, με την έννοια ότι υπάρχει μια σχέση μεταξύ της εισόδου και της εξόδου του συστήματος. Τα προβλήματα επιβλεπόμενης μάθησης διακρίνονται κυρίως σε προβλήματα παλινδρόμησης και ταξινόμησης. Στην πρώτη περίπτωση γίνεται προσπάθεια πρόβλεψης των αποτελεσμάτων για μια έξοδο συνεχούς τιμής, γεγονός που σημαίνει ότι προσπαθούμε να αντιστοιχήσουμε τις μεταβλητές εισόδου σε μια συνεχή συνάρτηση. Στη δεύτερη περίπτωση προσπαθούμε να προβλέψουμε τα αποτελέσματα για μια έξοδο διακριτής τιμής. Σκοπός επομένως είναι η αντιστοίχηση των μεταβλητών εισόδου σε διακριτές κατηγορίες [11].

Η Μάθηση χωρίς επίβλεψη είναι μια μέθοδος που κυρίως ενδιαφέρεται να εκτιμήσει μια συνάρτηση κατανομής για το σύνολο εκπαίδευσης, βάσει συμπερασμάτων που εξάγει από αυτό, ενώ ασχολείται και με προβλήματα ομαδοποίησης. Εδώ, δεν γνωρίζουμε εξαρχής τη σωστή κατηγορία για κάποιο σύνολο παραδειγμάτων, αλλά προσπαθούμε να εκτιμήσουμε έναν κατάλληλο διαμερισμό των δεδομένων σε ομάδες. Για αυτό και τα προβλήματα αυτού του είδους είναι πιο δύσκολα [10].

Πιο συγκεκριμένα, στην μη επιβλεπόμενη μάθηση γίνεται προσέγγιση του προβλήματος έχοντας λίγη σχετικά γνώση για τη σωστή μορφή των αποτελεσμάτων. Είναι δυνατό να αντλήσουμε πληροφορίες από τα δεδομένα, χωρίς να γνωρίζουμε ωστόσο την ακριβή επίδραση των μεταβλητών σε αυτά. Μπορούμε να εξάγουμε τη

δομή ομαδοποιώντας τα δεδομένα βάσει των σχέσεων. Στη μη επιβλεπόμενη μάθηση δεν υπάρχει ανατροφοδότηση (feedback) από τις τιμές πρόβλεψης [11].

Στην περίπτωση, τέλος, της ενισχυτικής μάθησης, ο αλγόριθμος αποδίδει μια επιβράβευση (θετική ή και αρνητική), μετά από μια σειρά αποφάσεων. Σκοπός της είναι να μεγιστοποιήσει αυτή την επιβράβευση, η οποία εξαρτάται από το σύνολο αυτών των αποφάσεων, καθώς κάτι τέτοιο φανερώνει ότι αυτές μας οδήγησαν πολύ κοντά στον στόχο που μας ενδιαφέρει.

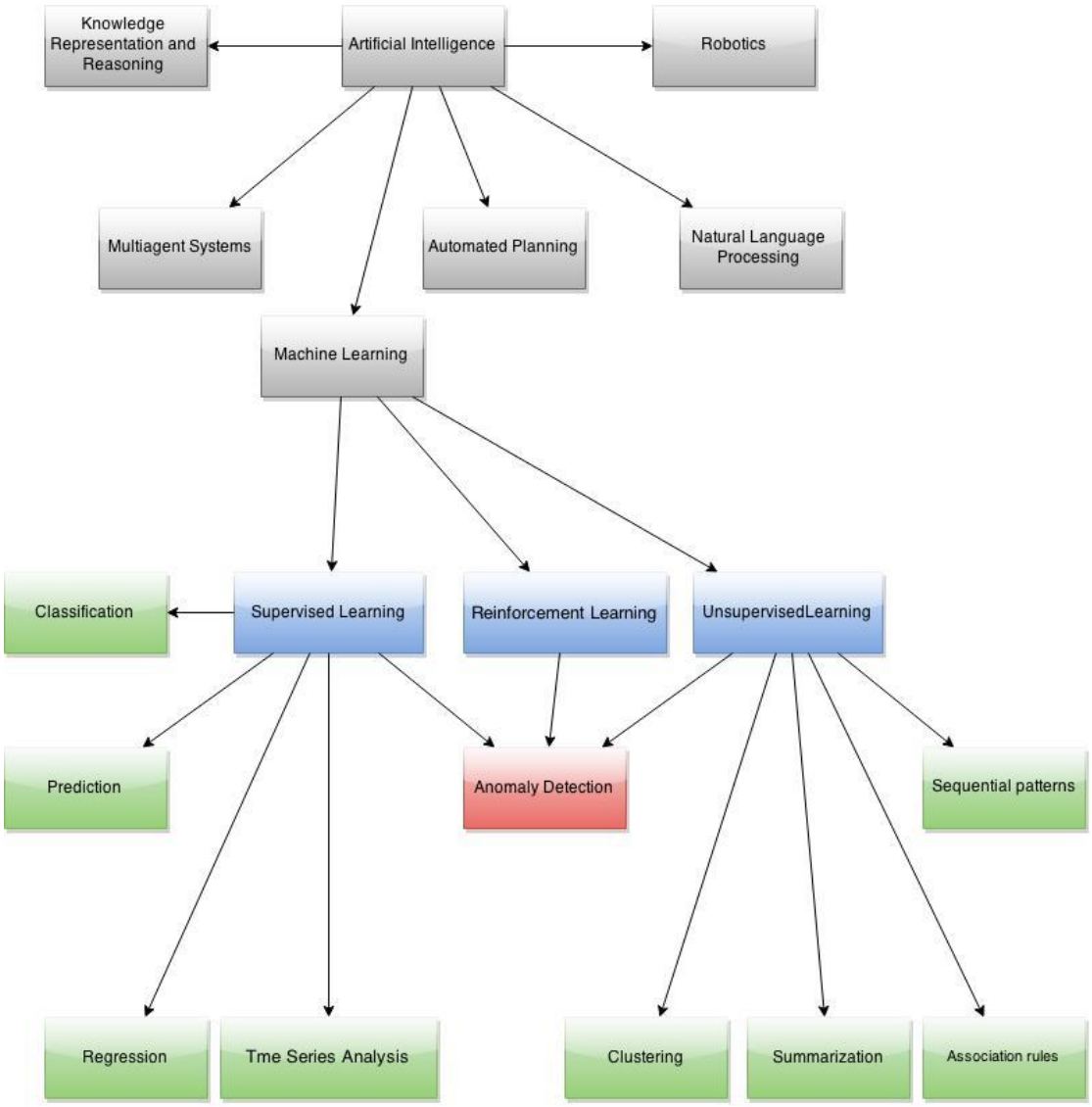
3.3 Διαδικασία Μηχανικής Μάθησης

Η διαδικασία εφαρμογής της Μηχανικής Μάθησης μπορεί να χωριστεί σε δύο βασικές φάσεις την εκπαίδευση και την ανάκληση

Κατά την εκπαίδευση πραγματώνεται η παρουσίαση πολλών προτύπων στο σύστημα (με ή χωρίς στόχους, ανάλογα με το τύπο Μάθησης) με σκοπό την ρύθμιση των παραμέτρων του ώστε αυτό να βελτιώνεται στην λειτουργία αναγνώρισης ή σε όποια άλλη λειτουργία τάχθηκε.

Κατά την ανάκληση πραγματώνεται η εισαγωγή ενός ή περισσότερων προτύπων με σκοπό την εξαγωγή της απόκρισης του συστήματος [11].

Παρακάτω παρουσιάζουμε ένα χάρτη (Εικόνα 3.1) που απεικονίζει στο πρώτο επίπεδο τις βασικές κατηγορίες της Τεχνητής Νοημοσύνης, όπου μία εξ αυτών είναι η Μηχανική Μάθηση. Στο δεύτερο επίπεδο μπορούμε να παρατηρήσουμε τις τρεις κύριες κατηγορίες τις Μηχανικής Μάθησης που αναλύσαμε προηγουμένως. Στο τρίτο επίπεδο κάνουμε την αντιστοίχηση των μεθόδων του περιγραφικού και του προβλεπτικού μοντέλου που αναλύσαμε στο Κεφάλαιο 2 με τις κατηγορίες τις Μηχανικής Μάθησης παρουσιάζοντας από μια άλλη σκοπιά την κατηγορία που ανήκει η κάθε μέθοδος.



Εικόνα 3.1 Χάρτης Βασικών Κατηγοριών Τεχνητής Νοημοσύνης.

Όπως φαίνεται στον χάρτη αναφέρουμε και μια ιδιαίτερη μέθοδο που ανήκει σε όλες τις κατηγορίες της Μηχανικής Μάθησης, παρότι δεν αποτελεί μέθοδο εξόρυξης δεδομένων.

Η έννοια του Anomaly Detection αναφέρεται στο πρόβλημα της εύρεσης μοτίβων σε δεδομένα τα οποία δεν συμμορφώνονται με την αναμενόμενη συμπεριφορά. Αυτά τα μη συμμορφούμενα πρότυπα αναφέρονται συχνά ως ανωμαλίες, ακραίες τιμές, ασύμφωνες παρατηρήσεις, εξαιρέσεις, παρεκκλίσεις, εκπλήξεις, ιδιαιτερότητες ή προσμείξεις σε διαφορετικά πεδία εφαρμογής. Ανωμαλίες και ακραίες τιμές είναι δύο όροι που χρησιμοποιούνται συχνότερα στο πλαίσιο της διαπίστωσης ανωμαλίας, μερικές φορές εναλλακτικά. Η διαδικασία Anomaly

Detection βρίσκει εκτεταμένη χρήση σε μια ευρεία ποικιλία εφαρμογών, όπως η ανίχνευση της απάτης για τις πιστωτικές κάρτες, στην ασφάλιση, στην υγειονομική περίθαλψη, στην ανίχνευση εισβολής για την Ασφάλεια στον κυβερνοχώρο, την ανίχνευση σφαλμάτων σε κρίσιμα συστήματα Ασφάλειας, καθώς και σε στρατιωτικές επιχειρήσεις επιτήρησης της δραστηριότητας του εχθρού. Η σημασία της διαδικασίας Anomaly Detection οφείλεται στο γεγονός ότι η αναγνώριση των ανωμαλιών στα δεδομένα παράγουν σε σημαντική (και συχνά κρίσιμη) πληροφοριακή άξια σε μια ευρεία ποικιλία τομέων εφαρμογής [9].

Ο λόγος που γίνεται η αναφορά της μεθόδου Anomaly Detection είναι ότι η συγκεκριμένη μέθοδος χρησιμοποιείται ευρύτατα στην περιοχή της Ασφάλειας των υπολογιστών. Η τακτική που ακολουθείται στην συγκεκριμένη πτυχιακή εργασία είναι μια άλλη προσέγγιση σε σχέση με την μέθοδο Anomaly Detection, αφού βασίζεται στο μοντέλο όχι των ακραίων – «επικίνδυνων» τιμών για να κάνει την ομαδοποίηση και κατ' επέκταση ανίχνευση των απειλών, αλλά όπως θα δούμε παρακάτω στο μοντέλο των μη ακραίων, ομαλών – «ακίνδυνων» τιμών.

3.4 Διαχωρισμός Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης

Είναι καλό να επισημάνουμε το διαχωρισμό μεταξύ των όρων Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης, ασχέτως αν η Μηχανική Μάθηση εντάσσεται στο πεδίο της Τεχνητής Νοημοσύνης και αποτελεί ένα εξελιγμένο υποσύνολό της. Η παραδοσιακή Τεχνητή Νοημοσύνη ασχολείται με άκαμπτους κανόνες (rules) σε αντίθεση με την Μηχανική Μάθηση. Για παράδειγμα θα μπορούσαμε να πούμε πως ένα σύστημα μπορεί να μάθει να παίζει σκάκι χρησιμοποιώντας κανόνες όπως

«μην θυσιάζεις τη Βασίλισσα» ή «φύλαγε το βασιλιά»

αλλά αυτοί οι κανόνες είναι τελείως άχρηστοι σε οποιοδήποτε άλλο παιχνίδι. Το σύστημα πρέπει να αποθηκεύσει νέους κανόνες για το τάβλι, την τρίλιζα, τη ντάμα, ή

οποιοδήποτε άλλο παιχνίδι. Αντίθετα η Μηχανική Μάθηση με χρήση πιο ασαφών και ευέλικτων κανόνων θα μάθει στο σύστημα να παίζει και σε περισσότερα παιχνίδια.

Ένα παράδειγμα ενός προβλήματος που απαιτεί χρήση της Μηχανικής Μάθησης είναι αναγνώριση ανεπιθύμητης αλληλογραφίας (Spam). Μια απολύτως απαραίτητη δυνατότητα σε κάθε σοβαρή υπηρεσία email.

Φυσικά οι κανόνες δεν μπορεί να είναι ίδιοι για όλους τους χρήστες! Απαιτείται προσαρμογή με βάση τη συμπεριφορά του χρήστη, δηλαδή ευελιξία, ενώ είναι εύκολα κατανοητό ότι είναι πρακτικά αδύνατη η χρήση εμπειρογνωμόνων.

Συχνά τα δεδομένα μας έρχονται συνήθως είτε σε μεγάλη ποικιλομορφία (πχ. εικόνες, βίντεο, κλπ) είτε είναι πολλά, είτε και τα δύο, και με μεγάλη ταχύτητα.

Πολλές περιπτώσεις μεγάλου πλήθους δεδομένων με μεγάλες διαστάσεις εμφανίζονται όλο και πιο συχνά λόγω της διάδοσης του Internet. Η ανάλυση και η επεξεργασία τέτοιων δεδομένων (BIG DATA) είναι εξαιρετικά δημοφιλές αντικείμενο έρευνας και ανάπτυξης τα τελευταία χρόνια.

Ας σκεφτούμε τους δημοφιλείς τόπους κοινωνικής δικτύωσης όπως το Facebook το Twitter ή το Youtube που παράγουν τεράστιους όγκους πληροφορίας κάθε μέρα και η συμπίεση της πληροφορίας είναι απαραίτητη. Δεν πρέπει όμως να θυσιάζεται η ικανότητα ορθής ομαδοποίησης, ή ορθής ταξινόμησης. Πρέπει να συμπιέσουμε τα δεδομένα, κρατώντας μόνο τα σημαντικότερα χαρακτηριστικά, δηλαδή αυτά που κάνουν τη διαφορά [10].

Στόχος της παρούσης πτυχιακής εργασίας δεν είναι η μελέτη των μοντέλων της Μηχανικής Μάθησης, αλλά ούτε οι αλγόριθμοι αυτών. Τα μοντέλα και οι αλγόριθμοι της Μηχανικής Μάθησης χρησιμοποιούνται στην διαδικασία της ανακάλυψης γνώσης (KDD) και συγκεκριμένα είναι βασικό και ουσιαστικό εργαλείο για την εξόρυξη πληροφορίας από δεδομένα (Data Mining) που μελετήσαμε στο προηγούμενο κεφάλαιο(Κεφάλαιο 1) αναλυτικά.

4 BIG DATA

4.1 Εισαγωγή

Ζούμε σε ένα κόσμο που θα μπορούσαμε να τον χαρακτηρίσουμε ψηφιακό. Κάθε μέρα στέλνονται e-mail, δημοσιεύονται κείμενα, φωτογραφίες και βίντεο στα κοινωνικά δίκτυα, υπάρχει επικοινωνία μέσω SMS και κάθε μέρα, τελικά, γίνεται αντιληπτό πόσο ψηφιακά συνδεδεμένος είναι ο κόσμος που ζούμε. Όμως δεν είμαστε οι μόνοι που δημιουργούμε πληροφορία ή στέλνουμε δεδομένα από τη μια άκρη του κόσμου στην άλλη [12].

Εκατομμύρια συσκευές σε όλο τον κόσμο κάνουν ακριβώς κάτι παρόμοιο. Σταθμοί μετεωρολογικών παρατηρήσεων, συστήματα γεωγραφικού εντοπισμού, αισθητήρες κάθε μορφής και ο κατάλογος είναι ατελείωτος. Άνθρωποι και συσκευές συμμετέχουμε με ξέφρενο ρυθμό στη δημιουργία ενός τεράστιου όγκου δεδομένων, που οι εκτιμήσεις των ειδικών τον ανεβάζουν σε 2,5 πεντάκις εκατομμύρια bytes, μόνο σε 24 ώρες [12].

Πιο συγκεκριμένα σε έρευνα του USC Annenberg School of Communication and Journalism ξεδιπλώνεται αυτή η έκρηξη πληροφοριών με αποκαλυπτικούς πραγματικά αριθμούς. Το 2007 υπήρχαν στον κόσμο αποθηκευμένα 300 Exabytes, όπου ένα Exabyte αντιστοιχεί σε 10^9 bytes, δεδομένων. Από αυτά τα δεδομένα μόνο το 7% ήταν σε αναλογική μορφή τα υπόλοιπα ήταν σε ψηφιακή ενώ το 2000 μόνο το 25% των δεδομένων ήταν σε ψηφιακή μορφή, το υπόλοιπο 75% ήταν σε χαρτί, φιλμ, δίσκους βινυλίου κασέτες κλπ. Ήδη υπολογίζεται ότι ο όγκος της αποθηκευμένης πληροφορίας το 2013 ήταν τα 1.200 Exabytes από τα οποία μόνο το 2% θα ήταν σε αναλογική μορφή. Μια παραστατική εικόνα του ασύλληπτου αυτού μεγέθους θα ήταν να φανταστούμε ότι μπορούμε να καλύψουμε όλη την έκταση των Η.Π.Α σε 52 στρώματα βιβλίων [30].

4.2 Η έννοια των BIG DATA

Η χρησιμοποίηση δεδομένων στη λήψη σωστών, έγκυρων και έγκαιρων αποφάσεων έχει αναχθεί σε ουσιαστικό παράγοντα επιτυχίας σε πολλές περιπτώσεις. Ταυτόχρονα, τα τελευταία χρόνια, με την ανάπτυξη νέων τεχνολογιών και εφαρμογών, όπως η εξάπλωση των κοινωνικών δικτύων, η εκτεταμένη χρήση Smartphones, η εγκατάσταση αισθητήρων, ο όγκος και η μορφή των δεδομένων έχει αλλάξει δραματικά, ενώ οι δυνατότητες ανάλυσης και επεξεργασίας αυτών είναι εντυπωσιακές. Αυτό επέφερε μεγάλες αλλαγές στο πως αντιλαμβάνονται τη λήψη αποφάσεων οι εταιρίες και οι οργανισμοί.

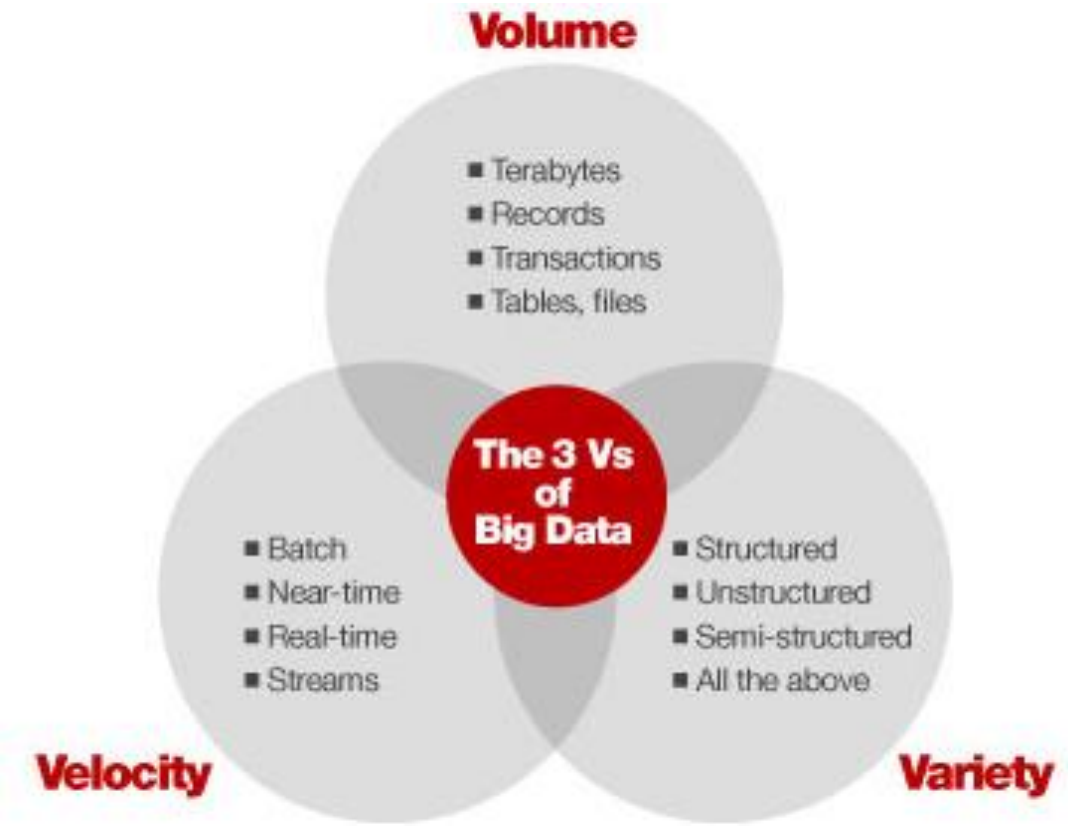
Πλέον, τις περισσότερες φορές, αντί να αναπτύσσουν μοντέλα, ξεκινούν από την ανάλυση των δεδομένων που κατέχουν, τα συνδυάζουν με αλλά που συλλέγουν από διάφορες πηγές και τα μοντέλα προκύπτουν αυτόματα με τη μορφή προτύπων.

Τα BIG DATA είναι σύνολα δεδομένων που είναι πολύ μεγάλα, πολύ πολύπλοκα, δομημένα, ημιδομημένα και αδόμητα, τα οποία χαρακτηρίζονται κυρίως από τον τεράστιο όγκο τους που είναι συνεχώς αυξανόμενος, την τεράστια ποικιλία στην αξία και την ποιοτική μορφή των δεδομένων αλλά και από την τεράστια ταχύτητα με την οποία αναπτύσσονται τα δεδομένα αυτά. Τα BIG DATA είναι σύνολα δεδομένων τα οποία υπερβαίνουν την ικανότητα και την χωρητικότητα των υπαρχόντων συμβατικών μεθόδων και συστημάτων [13].

Τα BIG DATA λοιπόν αποτελούν δεδομένα με τα εξής βασικά χαρακτηριστικά (3Vs):

- (Volume) Τεράστιος όγκος δεδομένων
- (Variety) Τεράστια ποικιλία δεδομένων
- (Velocity) Τεράστια ταχύτητα ανάπτυξης δεδομένων

Ο μεγάλος όγκος (Volume) δεδομένων χαρακτηρίζει τις μοντέρνες εφαρμογές, ενώ τα δεδομένα έρχονται σε διάφορες μορφές, όπως σχεσιακά δεδομένα, εικόνα, ήχος, video (Variety). Τέλος, η παραγωγή των δεδομένων μπορεί να γίνεται σε πολύ μεγάλους ρυθμούς και η εξαγωγή συμπερασμάτων να πρέπει να γίνει σε πραγματικό χρόνο (Velocity)[13].



Εικόνα 4.1 The 3Vs, Volume, Variety, Velocity [37] [38]

Τα BIG DATA επιτρέπουν σε όσους τα χρησιμοποιούν να επωφεληθούν από το σύνολο των πληροφοριών τους, τόσο των εσωτερικών όσο και των εξωτερικών, σε πραγματικό χρόνο. Η ανάλυσή τους είναι μια εξαιρετική και αποδοτική διαδικασία λήψης αποφάσεων και ως αποτέλεσμα αποτελεί έναν μοναδικό και καινοτόμο τρόπο για την προσφορά υπηρεσιών στην κοινωνία. Με την χρήση τους, από το 100% των πληροφοριών, συμπεριλαμβανομένων των δομημένων, των ημι-δομημένων αλλά

φυσικά και των μη δομημένων δεδομένων, μπορούμε να αντλήσουμε μεγάλη αξία που είναι χρήσιμη [21].

4.3 Αναλυτές Δεδομένων και BIG DATA

Τα τελευταία δύο χρόνια έχει αναδειχθεί ένας νέος ρόλος στις εταιρίες και τους οργανισμούς, με την ονομασία Data Scientist.

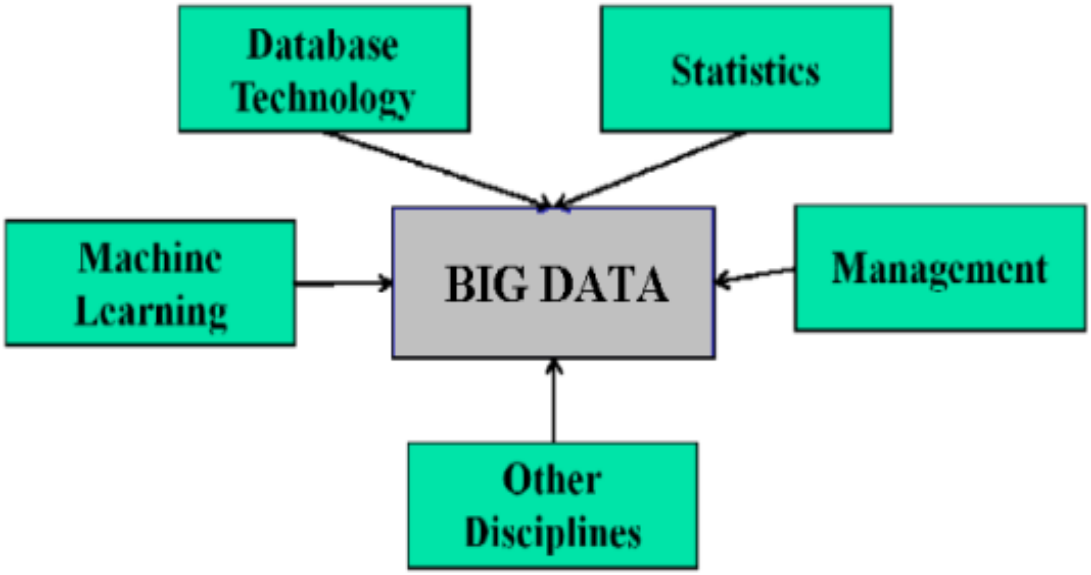
Οι Data Scientists έχουν εξελιχθεί σε μία ανώτερη θέση στην εκτελεστική ιεραρχία μίας εταιρίας ή ενός οργανισμού και μία νέα ιδιότητα στελέχους έχει αναδειχθεί, ο επιστημονικός υπεύθυνος δεδομένων (the Data Scientist), ο οποίος συνδυάζει τα προσόντα ενός ικανού προγραμματιστή, στατιστικού και αφηγητή με σκοπό να ανακαλύψει τους «σβόλους χρυσού» που υπάρχουν κάτω από τεράστιους όγκους δεδομένων [15].

Προβλέπεται ότι το επάγγελμα του Data Scientist θα γίνει σύντομα ένα από τα πλέον ελκυστικά. Τα δεδομένα πλέον είναι ευρέως διαθέσιμα, αυτό που είναι σπάνιο είναι η ικανότητα εξαγωγής γνώσης από αυτά [16].

Υπάρχει μεγάλη αναγκαιότητα τέτοιων επαγγελματιών, ενώ υπάρχει παράλληλα μεγάλη έλλειψη ανθρώπων σε αυτόν τον κλάδο [17]. Η συμβουλευτική εταιρία McKinsey προβλέπει ότι θα χρειαστούν 170.000 Data Scientists και ότι περίπου 1,5 εκατ. Managers θα πρέπει να εκπαιδευτούν στη χρήση και ανάλυση δεδομένων [18].

Υπάρχει μεγάλη συζήτηση για το ποιες είναι οι ικανότητες και γνώσεις που συνθέτουν το υπόβαθρο του Data Scientist. Καταρχάς θα πρέπει να έχει πολύ καλή, βαθιά γνώση των θεμάτων που αφορούν τη διαχείριση δεδομένων: μοντελοποίηση δεδομένων, γλώσσες ερωτημάτων και ανάκτησης δεδομένων, επεξεργασία, αποθήκευση, ενοποίηση. Θα πρέπει επίσης να έχει πολύ καλό μαθηματικό υπόβαθρο ώστε να μπορεί να κατανοήσει και να χειριστεί τις στατιστικές μεθόδους και τις τεχνικές Machine Learning που απαιτούνται στην εξαγωγή προτύπων. Τέλος, θα πρέπει να είναι σχετικά ανήσυχο πνεύμα και να γνωρίζει κάποιες αρχές σύγχρονου

Management, ώστε να ακολουθεί τις εξελίξεις του σημερινού τεχνολογικού κόσμου και να έχει μία αντίληψη πως μπορούν να χρησιμοποιηθούν τα δεδομένα σε μια εταιρία ή σε έναν οργανισμό.



Εικόνα 4.2 Συνοψίζει τους βασικούς τομείς που συνεισφέρουν στις γνώσεις ενός Data Scientist [37]

Αυτή τη στιγμή, τα πανεπιστήμια διεθνώς που προσφέρουν εξειδικευμένα προγράμματα στον τομέα των BIG DATA Analytics είναι ελάχιστα. Τα πρώτα εξειδικευμένα προγράμματα δημιουργήθηκαν στα πανεπιστήμια Northwestern και Carnegie Mellon το Σεπτέμβριο του 2012. Πολύ σοβαρές προσπάθειες είναι η δημιουργία του Institute of Data Sciences and Engineering στο Columbia University και του Data Science Center στο New York University [19][20][37].

4.4 Ανάλυση των BIG DATA

Η ανάλυση των BIG DATA, δεν αφορά μόνο το μέγεθος, τον τύπο ή το είδος των δεδομένων, αλλά ένα σύνολο διαδικασιών που απαιτούν γνώσεις Πληροφορικής,

Στατιστικής, Μηχανικής Μάθησης και Διοίκησης Επιχειρήσεων, που εμπλέκουν και χρησιμοποιούν δεδομένα.

Η ανάγκη ανάλυσης και κατόπιν διαχείρισης των BIG DATA οδήγησε στην ανάπτυξη μίας νέας γενιάς συστημάτων, μοντέλων και προγραμματιστικών εργαλείων – που ακόμα βρίσκονται σε εμβρυακό στάδιο όπως: MapReduce, Hadoop, NoSQL, κ.α. τεχνολογίες που επιτρέπουν την παράλληλη επεξεργασία δεδομένων σε μεγάλη κλίμακα και με fault-tolerant τρόπο. Ταυτόχρονα, η αξιοποίηση αυτών των δεδομένων για την παραγωγή Analytics απαιτούν ικανότητες και γνώσεις σε ένα ευρύ πεδίο αντικειμένων, όπως έχει ειπωθεί. (Ανακάλυψη και η Εξόρυξη Πληροφορίας, Στατιστική, Τεχνητή Νοημοσύνη - Μηχανική Μάθησης, Επιχειρησιακή Έρευνα, Διοίκηση Επιχειρήσεων, κ.α.) [37].

4.5 Διαχείριση των BIG DATA

Με την εξαγωγή ειδικών γνώσεων που είναι «κρυμμένες» μέσα στα BIG DATA οι επιχειρήσεις και οι οργανισμοί μπορούν να εξορθολογήσουν τις βασικές αποφάσεις που θα ληφθούν και τις οργανωτικές διαδικασίες όπως είναι, οι προσφορές, οι προμήθειες, η εφοδιαστική αλυσίδα και οι λειτουργίες απογραφής, υποστηρίζουν εταιρείες όπως η Hewlett-Packard, η IBM και άλλες εταιρείες που παρέχουν Hardware, Software και υπηρεσίες.

Σε πρόσφατη έρευνά της για το «Ψηφιακό Σύμπαν» η IDC προβλέπει ότι το 2020 οι διαθέσιμες ποσότητες δεδομένων παγκοσμίως θα ανέρχονται στα 40000 Exabytes. Ο όγκος των δεδομένων που συσσωρεύετε σε κάθε επιχείρηση είναι τεράστιος και αυξάνεται μέρα με τη μέρα, ωστόσο πόσα από αυτά, άραγε, χρησιμοποιούνται στην πραγματικότητα; Πώς μπορούν οι Βάσεις Δεδομένων να αποφορτιστούν, ώστε να μπορέσουν να επεξεργαστούν την ποικιλία των νέων πληροφοριών;

Μια βάση αρχειοθέτησης δεν εκπληρώνει μόνο τους κανόνες συμμόρφωσης, αλλά φροντίζει να κάνει χώρο και για τις τεράστιες ποσότητες δεδομένων. Οι

επιχειρήσεις και οι οργανισμοί που επικεντρώνονται στα νέα δεδομένα και φροντίζουν να τα χρησιμοποιούν στοχευόμενα, αποκτούν ένα σαφές ανταγωνιστικό πλεονέκτημα - εφόσον απομακρύνουν από τα παραγωγικά συστήματα τα ανενεργά δεδομένα και τα αρχειοθετήσουν ξεχωριστά. Με αυτόν τον τρόπο αυξάνονται οι επιδόσεις των παραγωγικών συστημάτων και ταυτόχρονα μειώνονται τα κόστη για τα μέσα αποθήκευσης [24].

Εικόνα 4.3 BIG DATA Pyramid Volume [34]

Ωστόσο, ποια στρατηγική είναι κατάλληλη για την αρχειοθέτηση των υπαρχόντων Βάσεων Δεδομένων; Σε αυτήν την περίπτωση σημαντικό ρόλο παίζουν οι κανονισμοί ελέγχου και συμμόρφωσης, που πρέπει να ληφθούν υπόψη.

Έχει δημιουργηθεί η εντύπωση ότι οι επιχειρησιακά σχετιζόμενες πληροφορίες πρέπει να φυλαχθούν κατά τέτοιο τρόπο, ώστε να μην μπορούν να αλλάξουν αργότερα, αλλά και να είναι διαθέσιμες ανά πάσα στιγμή σε ένα εύλογο χρονικό διάστημα. Σε αυτό ακριβώς το σημείο οι λύσεις δημιουργίας αντιγράφων ασφαλείας για τις Βάσεις Δεδομένων φθάνουν τα όριά τους [24].

Επίσης, παρά το γεγονός ότι ανώνυμα τμήματα των δεδομένων είναι αρκετά εύκολα και απλά στη διαχείρισή τους, ένας μεγαλύτερος κίνδυνος προκύπτει όταν διαφορετικά τμήματα δεδομένων μπορούν να συσχετιστούν με τρόπο που μπορούν να μειώσουν σημαντικά το σύνολο των ανθρώπων από όπου προέρχονται φτάνοντας ακόμη και σε επίπεδο ατόμου [23].

Το μεγαλύτερο εμπόδιο λοιπόν, είναι η ανάγκη για συμμόρφωση με ολοένα και πιο αυστηρούς νόμους που σχετίζονται με τη διαχείριση δεδομένων, την προστασία των ιδιωτικών πληροφοριών και την εμπιστευτικότητα όσον αφορά στα προσωπικά δεδομένα, ειδικά στην Ευρωπαϊκή ένωση, όπου το καθεστώς προστασίας των προσωπικών δεδομένων είναι σαφώς πιο αυστηρό σε σχέση με αυτά που ισχύουν στις Η.Π.Α [23].

4.6 Χρησιμότητα των BIG DATA

Ένας στους τρεις οργανισμούς που επιχειρούν να εξάγουν ειδικές γνώσεις από την ανάλυση όλων των δεδομένων τους αποτυγχάνουν . Η ανάλυση των επονομαζόμενων BIG DATA αποδεικνύεται δύσκολη υπόθεση, εντούτοις έξι στους δέκα οργανισμούς δηλώνουν πρόθυμοι να δεσμεύσουν το 10% του προϋπολογισμού τους για την καινοτομία και για να «εξορύξουν» πολύτιμη πληροφορία από δομημένα, ημιδομημένα και μη-δομημένα δεδομένα [14].

Ο αυξανόμενος όγκος, η ποικιλομορφία αλλά και η «ευπάθεια» των δεδομένων που ρέουν από το εσωτερικό αλλά και το εξωτερικό περιβάλλον κάθε επιχείρησης και οργανισμού αποτελούν την αιτία της αποτυχίας της χρήσης των BIG DATA. Περισσότεροι από έναν στους τρεις οργανισμούς ή επιχειρήσεις που έχει επιχειρήσει την ανάλυση, έχει αποτύχει. Περισσότερα από ένα στα δύο στελέχη ανέφεραν ότι οι επιχειρήσεις και οι οργανισμοί τους δεν είναι εξοπλισμένοι με τις σωστές λύσεις, έτσι ώστε να είναι σε θέση να αντλήσουν ειδικές γνώσεις από τα BIG DATA. Επιπλέον, δεν διαθέτουν την τεχνογνωσία καθώς και τη συνεκτική στρατηγική για να συγκεντρώσουν όλα τα στοιχεία και στη συνέχεια να ενσωματώσουν νέα αλλά και παλιά δεδομένα. Παρά τις συνεχείς αποτυχίες όμως, όπως προαναφέρθηκε το 60% των εταιρειών και των οργανισμών που συμμετείχαν στην νεότερη έρευνα δηλώνουν πρόθυμες να δεσμεύσουν για τα BIG DATA το 10% του προϋπολογισμού τους με απώτερο στόχο την καινοτομία [14][22].

Αυτό δείχνει ότι υπηρεσίες και οργανισμοί έχουν πειστεί ότι τα επονομαζόμενα BIG DATA κρύβουν πλούσιες γνώσεις και η ανάλυσή τους μπορεί να προσφέρει αποτελέσματα σε πραγματικό χρόνο.

Τα BIG DATA θεωρούνται ότι είναι η λύση σε μια σειρά από προβλήματα που απασχολούν τις σύγχρονες κοινωνίες. Μόνο μερικά πλεονεκτήματα που προκύπτουν από τις εφαρμογές της ανάλυσης τους είναι [31]:

- α) Η εξαγωγή της συμπεριφοράς των διαδικτυακών χρηστών και των προτιμήσεων τους για στοχευόμενη διαφήμιση.

- β) Η ανακάλυψη μοτίβων που αφορούν στην ανταπόκριση ασθενών σε ιατρικές θεραπείες.

- γ) Η αποκάλυψη τρομοκρατών δια μέσου της ανίχνευσης υπόπτων τηλεπικοινωνιακών, μετακινησιακών και αγοραστικών συμπεριφορών.

- δ) Η προάσπιση διαδικτυακών υπηρεσιών κοινωνικών δικτύων μεγάλης κλίμακας με την ανίχνευση κακόβουλων και ψεύτικων λογαριασμών χρηστών.

Περνώντας από τις υψηλού επιπέδου έννοιες των BIG DATA στην πραγματικότητα, δηλαδή στην αξιοποίησή τους και τη δράση με βάση την πληροφορία, οι ειδικές γνώσεις που προκύπτουν από την ανάλυση των BIG DATA μπορούν να βοηθήσουν αποδοτικά μια επιχείρηση ή έναν οργανισμό να βελτιώσει [14]:

- το cross-channel marketing σε πραγματικό χρόνο,
- την εμπειρία του πελάτη και την πρόβλεψη της ζήτησης.

Άλλες περιπτώσεις χρήσης περιλαμβάνουν

- την υποστήριξη για τον εντοπισμό και την πρόληψη μιας απάτης,
- την διασφάλιση της συμμόρφωσης σε πραγματικό χρόνο και
- να χρησιμοποιηθούν τα κοινωνικά δίκτυα για να διαχειριστούν τους κινδύνους και τη φήμη της μάρκας τους.

Ένα ακόμα παράδειγμα της λειτουργίας και των αποτελεσμάτων των BIG DATA είναι το Google flu trends που είναι διαδικτυακή υπηρεσία της Google. Κάθε εβδομάδα, εκατομμύρια χρήστες σε όλο τον κόσμο αναζητούν στο διαδίκτυο πληροφορίες σχετικά με την υγεία.

Η Google ανακάλυψε μια στενή σχέση ανάμεσα στον αριθμό των ατόμων που εκτελούν αναζήτηση για θέματα σχετικά με τη γρίπη στη μηχανή αναζήτησης που διαθέτει, και στον αριθμό των ατόμων που παρουσιάζουν στην πραγματικότητα συμπτώματα γρίπης. Μετρώντας τη συχνότητα εμφάνισης αυτών των ερωτημάτων αναζήτησης, μπορεί να υπολογίσει την έκταση της γρίπης σε διαφορετικές χώρες και περιοχές σε όλο τον κόσμο [31].

Και αυτό είναι μόνο ένα από τα παραδείγματα αφού ιδιαίτερα στο χώρο της υγείας η ανάλυση των δεδομένων πραγματικά μπορεί να φέρει επανάσταση στο τομέα της πρόληψης, ενώ και στην ιατρική έρευνα πιο συγκεκριμένα η αξιοποίηση των δεδομένων αυτών έχει τεράστια σημασία στην βελτίωση των μεθόδων θεραπείας και παρασκευής νέων φαρμάκων.

4.7 Πεδία Εφαρμογής BIG DATA

Κατανοούμε λοιπόν πόσο σημαντικό ρόλο παίζουν και θα συνεχίσουν να παίζουν τα BIG DATA στη σημερινή ψηφιακή εποχή και μπορούν να χρησιμοποιηθούν σε πολλά πεδία εφαρμογής.

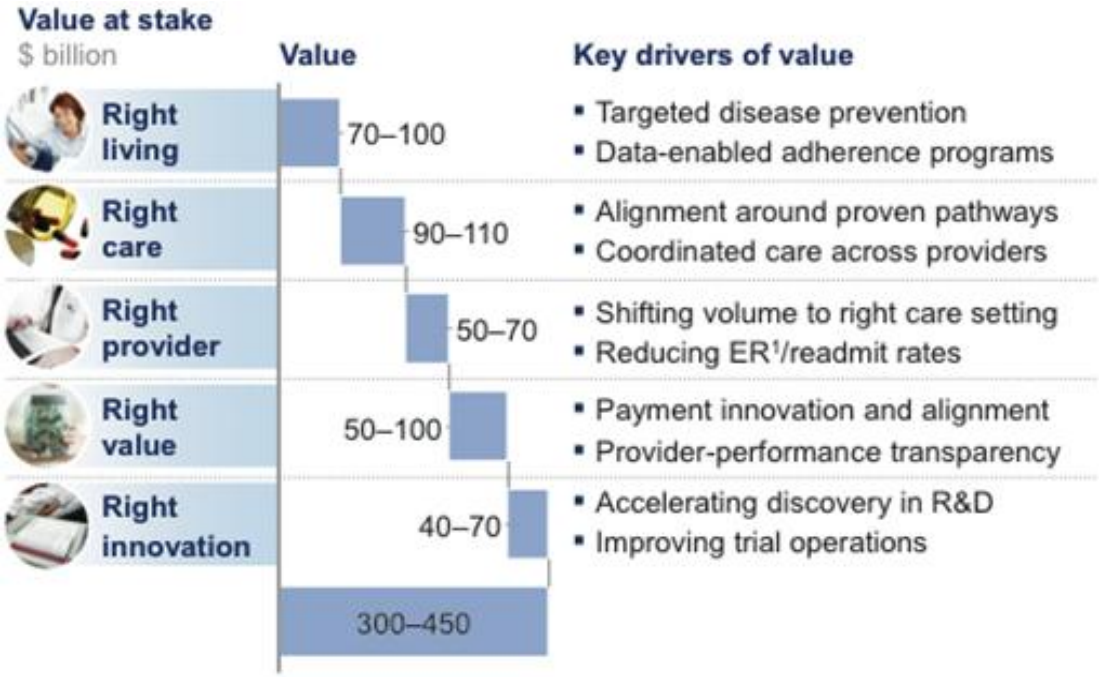
Παραθέτοντας ένα παράδειγμα, η εταιρία συμβουλευτικών υπηρεσιών McKinsey, αναλύει το πως η αξιοποίηση του μεγάλου όγκου ψηφιακών δεδομένων (BIG DATA) θα επηρεάσει σημαντικά τον τομέα της υγείας. Αναφέρουν πως υπάρχει μια αυξανόμενη ανάγκη για εφαρμογές BIG DATA κυρίως για οικονομικούς λόγους [25].

Η εν λόγω έκθεση παραθέτει πέντε τρόπους με τους οποίους τα δεδομένα αυτά θα δώσουν τη δυνατότητα στον κλάδο της υγείας να μειώσει τα κόστη και να βελτιώσει την ποιότητα των υπηρεσιών περίθαλψης [25].

- ✓ Οι ασθενείς με την αξιοποίηση των δεδομένων μπορούν να αποκτήσουν ενεργό ρόλο στη διαχείριση της υγείας τους.
- ✓ Η κατάλληλη ενσωμάτωση και εφαρμογή των εργαλείων διαχείρισης των BIG DATA θα προωθήσει την συντονισμένη και κατ επέκταση εξατομικευμένη περίθαλψη, με όλους τους παρόχους περίθαλψης να έχουν πρόσβαση στις ίδιες πληροφορίες ώστε να αποφεύγονται τα ιατρικά λάθη.
- ✓ Τα BIG DATA μπορούν να βοηθήσουν στην αντιστοίχηση των ικανοτήτων του εκάστοτε επαγγελματία υγείας με τις ανάγκες του ασθενή με απώτερο σκοπό την καλύτερη λήψη ιατρικής φροντίδας.
- ✓ Τα BIG DATA μπορούν να διασφαλίσουν την οικονομική αποτελεσματικότητα της υγειονομικής περίθαλψης μέσω της εξάλειψης ενδεχόμενης απάτης, κατάχρησης του συστήματος, κλπ.
- ✓ Τα BIG DATA θα χρησιμοποιηθούν για τη βελτίωση της καινοτομίας, στον τομέα της υγείας. Χρησιμοποιώντας παρελθοντικά στοιχεία από κλινικές

δοκιμές, καθώς και την ανάλυση των τάσεων από τα τρέχοντα δεδομένα, οι φορείς που αναζητούν καινοτόμες λύσεις θα είναι σε θέση να εντοπίσουν όλες τις πτυχές της θεραπευτικής καινοτομίας – την ανακάλυψη, την ανάπτυξη και την Ασφάλεια.

Όσον αφορά το οικονομικό όφελος, η έκθεση επισημαίνει ότι το εθνικό κόστος για την υγεία στις ΗΠΑ, με τις παρεμβάσεις των BIG DATA θα μειωθεί κατά 300-450 δισεκατομμύρια δολάρια [25].



Εικόνα 4.4 Επιρροές των BIG DATA στον τομέα της Υγείας [25]

Όπως αντιλαμβανόμαστε από τα παραπάνω τα BIG DATA μπορούν να χρησιμοποιηθούν και να δώσουν λύσεις και σε άλλα πεδία εφαρμογής με σκοπό να αυξήσουν τα οικονομικά οφέλη των οργανισμών, των εταιριών αλλά και των ανθρώπων, την αποδοτικότητα, την αποτελεσματικότητα σε όποιο πεδίο εφαρμόζεται η ανάλυσή τους, αποτελώντας κάτι το εξαιρετικό. Φυσικά όμως υπάρχουν και προβλήματα. Τίθεται το βασικότερο ζήτημα από όλα, το θέμα της Ασφάλειας των BIG DATA. Πως γίνεται να προστατευθούν;

Κάτι πολύ ενδιαφέρον και θέμα μείζονος σημασίας είναι το γεγονός ότι μπορούμε να προσπαθήσουμε να παράγουμε όσο μεγαλύτερη Ασφάλεια στα BIG DATA και όχι μόνο, από τα BIG DATA. Το πώς δηλαδή χρησιμοποιώντας τα BIG DATA μπορούμε να γίνουμε αποδοτικότεροι και πιο αποτελεσματικοί στο θέμα της προστασίας και της Ασφάλειας των "ευαίσθητων δεδομένων", συστημάτων και υπηρεσιών.

5 BIG DATA SECURITY

5.1 Εισαγωγή

Οι κυβερνοεπιθέσεις προκαλούν όλο και πιο σοβαρές συνέπειες στην υποδομή IT των σύγχρονων επιχειρήσεων και οργανισμών. Τα δίκτυα γίνονται συνεχώς πολυπλοκότερα και μεγαλύτερα, με αποτέλεσμα οι επιχειρήσεις να συναντάνε δυσκολίες, ώστε να κρατήσουν τα συστήματά τους πλήρως λειτουργικά και ασφαλή. Η ταχύτατη εξάπλωση των διαφόρων ειδών Malware δοκιμάζει σημαντικά την αποτελεσματικότητα της στατικής προστασίας, μέσω των κλασικών virus signatures. Η συνεχώς αυξανόμενη δικτυακή κυκλοφορία «φρενάρει» την ανάλυση διεισδύσεων και εκτοξεύει, παράλληλα, το κόστος διαχείρισης. Αποτελεί, δε, κοινή παραδοχή ότι τα παραδοσιακά εργαλεία δικτυακής προστασίας δεν επαρκούν

Το ζήτημα της Ασφάλειας και της προστασίας του ιδιωτικού απορρήτου είναι ζωτικής σημασίας γενικά, όπως επίσης και στο τομέα των BIG DATA. Πολλές εργασίες επιχειρήσεων και οργανισμών έχουν διεξαχθεί με την χρήση των BIG DATA, όπως είναι η εξόρυξη και η ανάλυση δεδομένων. Ωστόσο, τα θέματα της Ασφάλειας και του ιδιωτικού απορρήτου στα BIG DATA σπάνια αναφέρονται μέχρι σήμερα.

Λόγω του μεγάλου και συνεχώς αυξανόμενου μεγέθους των BIG DATA, η Ασφάλεια και η προστασία του ιδιωτικού απορρήτου αντιμετωπίζουν πολλές προκλήσεις.

Αρχικά γεννιέται η ανάγκη για νέους αποτελεσματικότερους αλγόριθμους hashing, αλγορίθμους κρυπτογράφησης, αλγορίθμους αποκρυπτογράφησης για την ανάκτηση κρυπτογραφημένων πληροφοριών, την αποκρυπτογράφηση βασισμένη σε χαρακτηριστικά κ.ά. Γενικά όλη η τεχνολογία που σχετίζεται με τα ζητήματα ασφαλείας πρέπει να βελτιστοποιηθεί. Εδώ είναι σημαντικό να τονίσουμε και μία

άλλη πρόκληση, **την Ασφάλεια και την προστασία μέσω της ανάλυσης των δεδομένων αυτών.** Το ζήτημα αυτό είναι θέμα ανάλυσης της επόμενης ενότητας.

Ωστόσο και από την πλευρά των κακόβουλων υφίστανται νέες πιο αποδοτικές επιθέσεις στη διαθεσιμότητα, την εμπιστευτικότητα, την αξιοπιστία και την ακεραιότητα των BIG DATA [27][31].

Οι συνέπειες της αδιαφορίας για την Ασφάλεια των BIG DATA, αλλά και για την Ασφάλεια στις συναφείς «third-platform» τεχνολογίες (cloud computing, κινητές συσκευές και social media) μπορεί να είναι μεγάλες, συμπεριλαμβανομένων της απώλειας χρόνου και της λειτουργικής αποτυχίας.

Επιπλέον, η αυξανόμενη νομική υποχρέωση για τα ανθρώπινα δικαιώματα, πρέπει να λαμβάνεται υπόψη. Πέρα από τη σαφή ηθική επιταγή, η υπόληψη και η οικονομική επίπτωση σε μια ολοένα και πιο φιλόδικη κοινωνία δεν θα πρέπει να αγνοηθεί [28].

Η Ασφάλεια στα BIG DATA είναι ένα θέμα πολύ δύσκολο και πολύ πιο περίπλοκο ζήτημα από όσο θα μπορούσε κανείς να φανταστεί. Άλλωστε ασφαλή συστήματα δεν υπάρχουν πραγματικά, μόνο αξιόπιστα. Ας αναλογιστούμε πόσο δύσκολο είναι να ασφαλίσουμε τα προσωπικά μας δεδομένα, πόσο μάλλον να ασφαλίσουμε τα BIG DATA, δηλαδή να ασφαλίσουμε έναν πολύ μεγαλύτερο, ποικίλο και μεταβαλλόμενο όγκο δεδομένων από τα προσωπικά μας δεδομένα. Είναι μόνο ο όγκος των δεδομένων που αποτελεί πρόβλημα Ασφάλειας; Θα μπορούσε κανείς να σκεφτεί πως τα χαρακτηριστικά των BIG DATA, αυτά καθαυτά, αποτελούν το πρόβλημα (3Vs). Σαφώς, τόσο η ποικιλία, όσο και η ταχύτητα κάνουν το εγχείρημά μας ακόμη πιο δύσκολο. Κι όμως υπάρχει κάτι το οποίο δεν είναι τόσο άμεσο, αλλά είναι τόσο καίριο. Αυτό το φοβερό πλεονέκτημα των BIG DATA που είναι η λήψη γνώσης μέσω της ανάλυσής τους, αποτελεί και το βασικό μειονέκτημα στον τομέα της Ασφάλειας που σχετίζεται με την προστασία του ιδιωτικού απόρρητου και των προσωπικών δεδομένων.

Η χρήση των δεδομένων αυτών προκαλεί προβληματισμούς αφού κυρίως συγκεντρώνεται στα χέρια των τηλεπικοινωνιακών παρόχων, των κατασκευαστών

λογισμικού, των διαφημιστικών εταιριών, αλλά και των οργανισμών και των κυβερνητικών υπηρεσιών. Οι ενθουσιώδεις υποστηρικτές για την ευεργετική επίδραση της ανάλυσης των δεδομένων αυτών σε όλους τους τομείς της κοινωνίας από την οικονομία έως την ιατρική βρίσκουν τον αντίλογο τους στους φόβους ότι η τρομακτική αυτή δύναμη που έχουν όλοι όσοι έχουν στη διάθεση τους αυτές τις πληροφορίες θα έχει δυσμενέστατες επιπτώσεις στο επίπεδο προστασίας της ιδιωτικής ζωής των πολιτών [31].

Καταλήγουμε λοιπόν στο συμπέρασμα ότι τα BIG DATA πρέπει να ασφαλιστούν ακολουθώντας δύο διαφορετικούς άξονες, αφού τα δεδομένα αυτά απειλούνται τόσο έμμεσα από τις τεχνικές ανάλυσής τους, όσο και άμεσα από οποιονδήποτε και οτιδήποτε κακόβουλο.

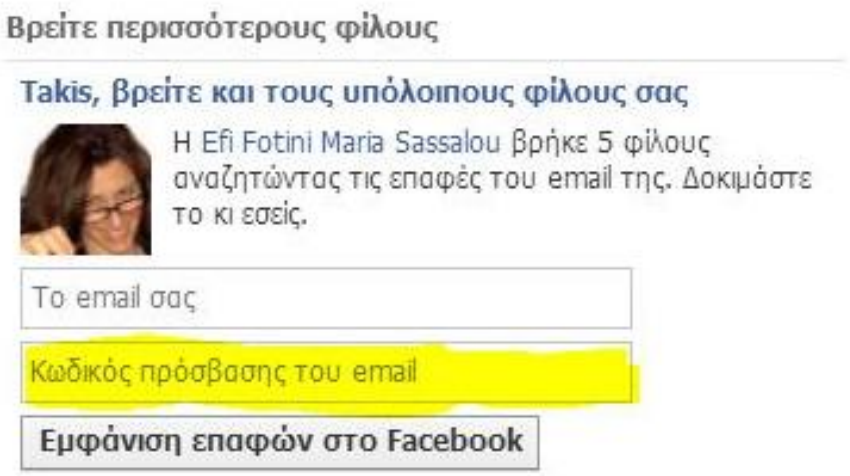
5.2 Ασφάλεια στα BIG DATA από έμμεσες απειλές

Αρχικά είναι πολύ βασικό να τονίσουμε τις πηγές άντλησης των προσωπικών δεδομένων από τα BIG DATA, οι οποίες μπορούν χωριστούν σε τρεις βασικές κατηγορίες [31]:

1. Σε προσωπικά δεδομένα που παραδίδονται εκουσίως από τους χρήστες των νέων τεχνολογιών επικοινωνίας. Στα κοινωνικά δίκτυα, ανεβάζουμε video φωτογραφίες, εκφράζουμε μηνύματα και απόψεις για την πολιτική τη θρησκεία ακόμα και για ευαίσθητα προσωπικά δεδομένα, όπως η υγεία μας.
2. Προσωπικά δεδομένα που αποθηκεύονται και γίνονται αντικείμενο επεξεργασίας χωρίς τη θέληση ή τη γνώση των χρηστών. Η χρήση του διαδικτύου και των κινητών συσκευών επικοινωνίας παράγει μια σειρά από δεδομένα, όπως η γεωγραφική θέση, οι οικονομικές συναλλαγές (τραπεζικοί λογαριασμοί , αγορές μέσω πιστωτικών καρτών κτλ), τα οποία χωρίς να γνωρίζουμε γίνονται αντικείμενο επεξεργασίας είτε από τους οργανισμούς με

τους οποίους συναλλασσόμαστε είτε από τρίτους που αγοράζουν τα στοιχεία αυτά (Εικόνα 5.1).

3. Προσωπικά δεδομένα που εξάγονται από το συνδυασμό των δύο προηγούμενων κατηγοριών και που βρίσκονται στην κατοχή μεγάλων επιχειρήσεων ή οργανισμών.



Εικόνα 5.1 Εδώ, το Facebook ζητάει, με πρόσχημα την αναζήτηση και τον εντοπισμό γνωστών μας, όχι μόνο το e-mail μας αλλά και τον κωδικό του χρήστη, για να μπορέσει να «ανιχνεύσει» και να διασυνδέσει τις επαφές του mailbox μας με την βάση δεδομένων του. Αυτό όμως είναι επικίνδυνο! [40]

Η κινητήρια δύναμη των BIG DATA, είναι ακριβώς μια συλλογή ολοένα περισσότερων bytes πληροφοριών για το σύνολο των πιο βαθιά προσωπικών, ενδόμυχων πληροφοριών για το μυαλό, το σώμα, το γεωγραφικό προσδιορισμό, την άσκηση, τις διαιτητικές συνήθειες κ.α. του καθενός, και θα χρησιμοποιηθεί για έρευνα ή ανάλυση δεδομένων [32]. Το πρόβλημα για την ιδιωτικότητα στην περίπτωση των BIG DATA δεν είναι απλό. Όπως αναφέρουν οι Schonberg και Kuner «το σημαντικό ερώτημα δεν είναι μόνο εάν τα BIG DATA αυξάνουν τον κίνδυνο για την ιδιωτικότητα, αλλά αν αλλάζουν αυτόν τον κίνδυνο».

Το πρόβλημα είναι ότι ο κίνδυνος μεταλλάσσεται γιατί ο κίνδυνος δεν βρίσκεται στον σκοπό για τον οποίο συλλέχθηκαν, ούτε στο ζήτημα της παράνομης

επεξεργασίας από αυτόν που τα έχει συλλέξει αλλά από την αυξανόμενη «δευτερεύουσα χρήση» τους.

Οι υποστηρικτές των BIG DATA πιστεύουν ότι ο θόρυβος γύρω από την ιδιωτικότητα είναι υπερβολικός, αφού τα δεδομένα ανωνυμοποιούνται πριν χρησιμοποιηθούν σύμφωνα με τις επιταγές της νομοθεσίας. Όμως οι ειδικοί αναλυτές υποστηρίζουν ότι η ανωνυμοποίηση είτε δεν γίνεται όπως πρέπει ή σε κάθε περίπτωση είναι εύκολος τελικά ο επαναπροσδιορισμός της ταυτότητας των υποκειμένων της επεξεργασίας. Έτσι, για παράδειγμα οι πληροφορίες που σχετίζονται με την υγεία είναι εύκολο να χρησιμοποιηθούν για τον εντοπισμό και τη δημιουργία ενός συγκεκριμένου προφίλ. Αυτό θολώνει τη λεπτή γραμμή ανάμεσα στα προσωπικά και μη προσωπικά δεδομένα, καθώς επιτρέπει τη σύνδεση τμημάτων δεδομένων στην πραγματική ταυτότητα ενός προσώπου. Υποστηρίζεται ότι ερευνώντας μια βάση δεδομένων συχνά και θέτοντας διαφορετικές ερωτήσεις σε αυτήν είναι δυνατόν να ανακαλύψεις πληροφορίες όπως που κατοικεί κάποιος, ή να βρεις τον ιατρικό φάκελο ενός συναδέλφου ή φίλου κ.α. [33]

Η έννοια της εμπιστοσύνης στον τρόπο με τον οποίο οι πληροφορίες χρησιμοποιούνται, διαμοιράζονται, αρχειοθετούνται και διαχειρίζονται είναι κρίσιμη σε αυτό το πολύπλοκο και εξαιρετικά ρευστό περιβάλλον .

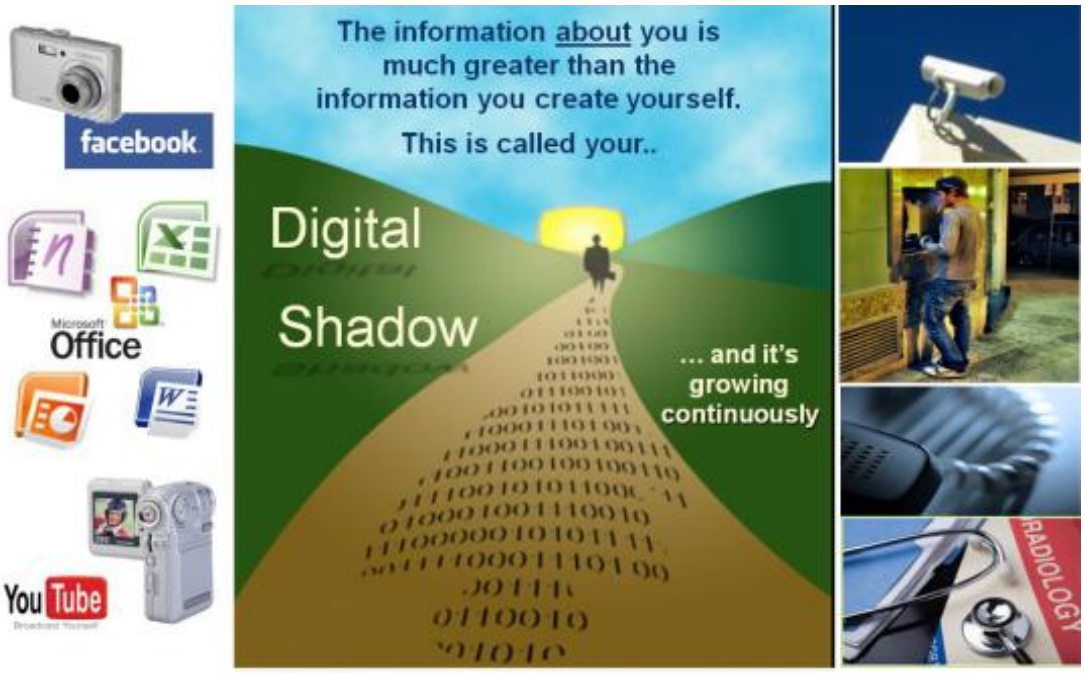
Η εμπιστοσύνη σχετίζεται αρχικά με την προέλευση των πληροφοριών, δηλαδή την ακεραιότητα των διαδικασιών και των συστημάτων πληροφορικής που παράγουν τις πληροφορίες.

Επίσης, η εμπιστοσύνη σχετίζεται με τη σύλληψη και τη διαχείριση των πληροφοριών, δηλαδή τα διαπιστευτήρια και τις ταυτότητες των ατόμων και των οργανισμών που αγγίζουν ή έχουν πρόσβαση στις πληροφορίες [35].

Βλέπουμε αυτή τη συζήτηση γύρω από την εμπιστοσύνη να αποτελεί κάτι σημαντικό. Η on-line συλλογή δεδομένων γίνεται όλο και πιο επεμβατική, η ανακάλυψη και η εξόρυξη δεδομένων στα BIG DATA, καθιστούν δυνατό για τις επιχειρήσεις και τους οργανισμούς να δημιουργήσουν προφίλ για καταναλωτές και για άτομα που επεκτείνουν την ψηφιακή τους «σκιά» μέσα από τη χρήση εφαρμογών

κινητών συσκευών και τη συμμετοχή τους σε δικτυακούς τόπους κοινωνικής δικτύωσης .

Ως αποτέλεσμα , υπάρχουν αυξανόμενες εκκλήσεις από δικηγόρους, ακαδημαϊκούς, και ρυθμιστικές αρχές να τροποποιήσουν τα τρέχοντα καθεστώτα προστασίας της ιδιωτικής ζωής και προστασίας των δεδομένων [35]. Δυστυχώς αυτό αποτελεί το μοναδικό τρόπο ασφαλείας από έμμεσες απειλές.



Εικόνα 5.2 Ψηφιακή Σκιά [44]

Ένα χαρακτηριστικό παράδειγμα της μεγάλης δύναμης και των επιπλοκών των BIG DATA μας δίνει ένα αξιοσημείωτο γεγονός που αποκάλυψαν οι NewYorkTimes [34].

Η εταιρία Target συνδυάζοντας τις αγοραστικές συνήθειες και τις αγορές συγκεκριμένων προϊόντων υγιεινής αλλά και ρουχισμού με άλλες συνήθειες κατέληξαν στην ταυτοποίηση εγκύων γυναικών μέσα από μια τεράστια βάση δεδομένων [34].

Το πρόβλημα σε αυτή τη περίπτωση είναι σύνθετο. Είναι προσβολή του δικαιώματος στην προσωπική ζωή αλλά και το κυριότερο είναι ότι αυτές οι ικανότητες πρόβλεψης μπορούν να προκαλέσουν πραγματικά κοινωνικά προβλήματα.

Συνεπώς βρισκόμαστε πια σε μια νέα πραγματικότητα όπου η έννοια της ανωνυμοποίηση όπως και η έννοιας της προηγούμενης συγκατάθεσης ή έγκρισης μιας επεξεργασίας χάνει την έννοια της. Το πως θα μπορέσουμε να δημιουργήσουμε ένα νομικό πλαίσιο που θα προστατεύει όλες τις χρήσεις δεδομένων και θα δημιουργεί υποχρεώσεις σε όλα τα μέρη που εμπλέκονται σε αυτή τη διαδικασία δεν είναι αντικείμενο της παρούσης πτυχιακής εργασίας [31].

5.3 Ασφάλεια στα BIG DATA από άμεσες απειλές

Όταν για πρώτη φορά, υπολογίσαμε την ποσότητα των πληροφοριών στο ψηφιακό σύμπαν που απαιτεί κάποιο επίπεδο Ασφάλειας, κάναμε την τρομακτική διαπίστωση ότι η ποσότητα των πληροφοριών που πρέπει να ασφαλίζεται αυξάνεται γρηγορότερα από την ικανότητά μας για να την ασφαλίσουμε, καθώς οι εργαζόμενοι χρησιμοποιούν όλο και περισσότερες φορητές συσκευές πλέον, οι καταναλωτές εν γνώσει τους και εν αγνοία τους διαμοιράζουν περισσότερα προσωπικά δεδομένα, αλλά και οι επιχειρήσεις και οι οργανισμοί, όπως προείπαμε μπορούν να παράγουν τεράστια ποσότητα δεδομένων σε πολύ μικρό χρονικό διάστημα.

Στον αντίποδα υπάρχουν, οι κακόβουλοι που «εκμεταλλεύονται» τα BIG DATA και σύμμαχός τους είναι η σχετικά λιγότερο δομημένη φύση τους, με τεράστιο όγκο, ταχύτητα ανάπτυξης και μεγάλη ποικιλία [35].

Αντιμετωπίζουμε λοιπόν ουσιώδη προβλήματα, σχετικά με το αν τα «εμπλεκόμενα» δεδομένα είναι ευαίσθητα σχετικά με [36]

- την προστασία της ιδιωτικής ζωής των πολιτών
- την Ασφάλεια των οργανισμών και των επιχειρήσεων

Οι διευθυντές IT αναφέρουν ότι υπάρχουν πολλά εμπόδια για την υιοθέτηση λύσεων «BIG DATA» με την Ασφάλεια να βρίσκεται στην κορυφή της λίστας, ακολουθούμενη από τον προϋπολογισμό και το προσωπικό για τις λύσεις αυτές [39].

- Περισσότεροι από ένας στους τέσσερις ερωτηθέντες σε παγκόσμιο επίπεδο (27%) δήλωσαν ότι η Ασφάλεια των δεδομένων και η διαχείριση των κινδύνων αποτελούν σημαντική πηγή προβληματισμού. Σύμφωνα με τα όσα δήλωσαν, ο τεράστιος όγκος δεδομένων, ο αριθμός των μεθόδων πρόσβασης στα δεδομένα και η έλλειψη οικονομικών πόρων για την Ασφάλεια συγκαταλέγονται ανάμεσα στους κύριους λόγους για τους οποίους η Ασφάλεια των δεδομένων σε έργα «BIG DATA» αποτελεί τόσο μεγάλη πρόκληση.
- Ο προβληματισμός σχετικά με την Ασφάλεια ήταν εντονότερος στην Κίνα (45%), την Ινδία (41%), τις ΗΠΑ (36%) και τη Βραζιλία (33%).
- Η έλλειψη οικονομικών πόρων (16%) σε συνδυασμό με την έλλειψη χρόνου για τη μελέτη των BIG DATA (14%) αναφέρονται ως τα κύρια εμπόδια από το ένα τρίτο των ερωτηθέντων.
- Σχεδόν ένας στους τέσσερις (23%) δήλωσε έλλειψη επαρκούς προσωπικού IT (13%) ή τεχνογνωσίας του προσωπικού σχετικά με τα BIG DATA (10%) ως το κύριο ζήτημα, ειδικά στην Ιαπωνία (31%) και τη Βραζιλία (30%).

Ας αναλογιστούμε ότι τα συστήματα διαχείρισης Βάσεων Δεδομένων υποστηρίζουν πολιτικές ασφαλείας που είναι αρκετά ευέλικτες και προστατεύουν τα δεδομένα , τόσο σε ψηλό όσο και σε χαμηλό επίπεδο από την ακατάλληλη πρόσβαση.

Αντίθετα στο ψηφιακό σύμπαν γενικά, δεν μπορούμε να ακολουθήσουμε τέτοιες πολιτικές ασφαλείας. Οι επιχειρήσεις και οι οργανισμοί που διαχειρίζονται ευαίσθητα δεδομένα πρέπει να διασφαλίζουν ότι τα ίδια τα δεδομένα είναι ασφαλή και ότι οι ίδιες πολιτικές Ασφάλειας, που ισχύουν για δεδομένα που υπάρχουν σε Βάσεις Δεδομένων ή αρχεία, εκτελούνται επίσης στο πλαίσιο του ψηφιακού σύμπαντος. Η αποτυχία να το πράξουν αυτό μπορεί να έχει σοβαρές αρνητικές συνέπειες στην Ασφάλεια των δεδομένων σχετικά με:

- Την Εμπιστευτικότητα
- Την Ακεραιότητα
- Την Διαθεσιμότητα

Για λόγους κατανόησης της ασφαλείας των BIG DATA, εξετάζουμε αρχικά την Ασφάλεια στο ψηφιακό σύμπαν, όπου έχουμε διαβαθμίσει τις πληροφορίες που απαιτούν την Ασφάλεια σε πέντε κατηγορίες, που καθεμία απαιτεί διαδοχικά υψηλότερο επίπεδο:

- ✓ **Privacy**- όπως μια διεύθυνση ηλεκτρονικού ταχυδρομείου σε μια αποστολή στο Youtube

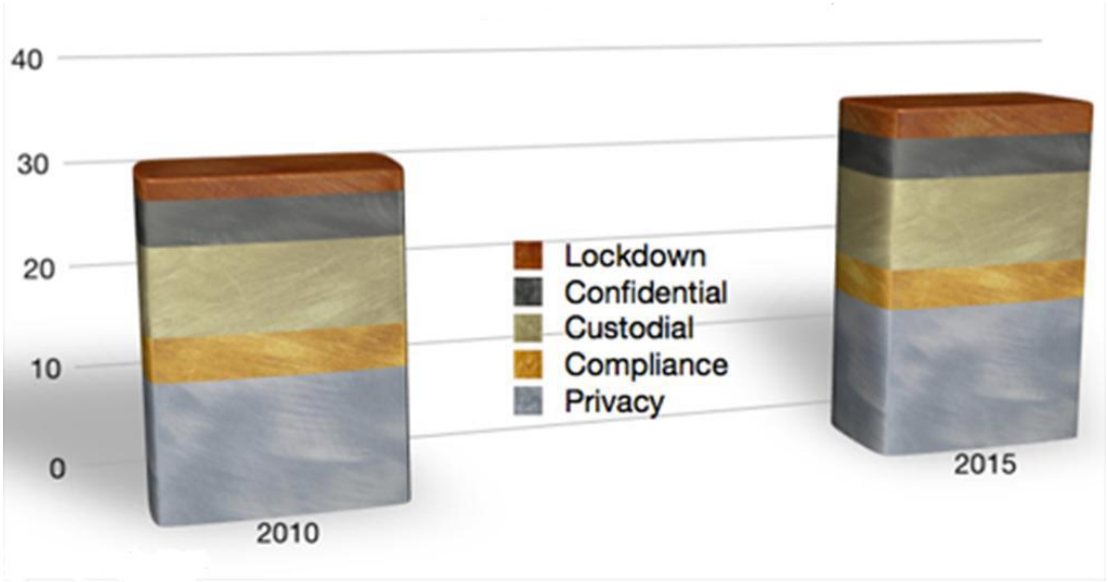
- ✓ **Compliance** - όπως ηλεκτρονικά μηνύματα που θα μπορούσαν να είναι ανιχνεύσιμα σε δικαστικές διαμάχες ή υπόκεινται σε κανόνες διατήρησης

- ✓ **Custodial** - τα στοιχεία του λογαριασμού, η παραβίαση των οποίων θα μπορούσε να οδηγήσει σε ενίσχυση ή κλοπή ταυτότητας

- ✓ **Confidential**- πληροφορίες τις οποίες ο δημιουργός θέλει να προστατεύσει, όπως εμπορικά μυστικά, πελατολόγια, εμπιστευτικά υπομνήματα κ.λπ.

- ✓ **Lockdown** - πληροφορίες που απαιτούν τη μέγιστη δυνατή Ασφάλεια, όπως οι χρηματοπιστωτικές συναλλαγές, τα αρχεία του προσωπικού μιας επιχείρησης, ιατρικά αρχεία, αρχεία στρατιωτικών υπηρεσιών κλπ.

Το 2010, το 28% του ψηφιακού σύμπαντος απαιτεί κάποιο επίπεδο Ασφάλειας (Εικόνα 5.3).



Εικόνα 5.3 Οι πέντε κατηγορίες ασφαλείας [44]

Μέσω της ανάλυσης και κατόπιν της εφαρμογής της πολιτικής Ασφάλειας που θα ορίσουμε, είναι προφανές ότι την ίδια στρατηγική θα ακολουθήσουμε και στην διαδικασία της διαχείρισης των BIG DATA, που αποτελούν μια «μικρογραφία» του ψηφιακού σύμπαντος [36].

Περισσότερες επιχειρήσεις και οργανισμοί υιοθετούν πολιτικές ασφαλείας, όπου οι χρήστες, δηλαδή οι εργαζόμενοι, έχουν τον έλεγχο πάνω σε φορητούς υπολογιστές, ταμπλέτες και Smartphones που μπορούν να χρησιμοποιήσουν για τη διεξαγωγή διαφόρων λειτουργιών [36].

Οι επιχειρήσεις και οι οργανισμοί που διαχειρίζονται τα BIG DATA, πρέπει να βρουν τρόπους να περιορίσουν την περιττή ή τυχαία έκθεση των διαβαθμισμένων πληροφοριών.

Οι χρήστες δεν μπορούν θεωρηθούν έμπιστοι ότι θα φέρονται σωστά όλη την ώρα. Αντί αυτού, πρέπει να ενσωματώσουν τις πολιτικές ασφαλείας και τις διαδικασίες για τη δημιουργία ενός κλίματος εμπιστοσύνης, που είναι αυτοματοποιημένο και πλήρες και να μην εξαρτάται από χειροκίνητη λειτουργία . Τα

ακόλουθα ειδικά μέτρα που απαιτούνται από τις επιχειρήσεις και τους οργανισμούς είναι:

- Πρέπει να κατανοήσουν τις εξαρτήσεις των ανθρώπων, των διαδικασιών, των πληροφοριών και των βασικών υπολογιστικών πόρων.
- Πρέπει να είναι σε θέση να εντοπίζουν πιθανές συγκρούσεις πολιτικών ασφαλείας, καθώς και να αναλύουν τον αντίκτυπο των νέων τεχνολογιών και διαδικασιών στο συνολικό τους ρίσκο.
- Θα πρέπει να είναι σε θέση να συσχετίσουν τις εξαρτήσεις μεταξύ γεγονότων και αλλαγών που γίνονται σχετικά με τα γεγονότα αυτά.
- Πρέπει να είναι σε θέση να αποδείξουν ότι τηρούν σταθερά τις ρυθμιστικές και νομικές τους υποχρεώσεις.

Επίσης πρέπει να χρησιμοποιηθεί το αρμόδιο τεχνολογικό περιβάλλον ασφαλείας, για το κάθε επίπεδο ευαισθησίας των δεδομένων. Ένα πλαίσιο για τη διαχείριση των κινδύνων είναι μια ουσιώδης λύση [36].

Επειδή οι πόροι είναι περιορισμένοι, το πλαίσιο για τη διαχείριση των κινδύνων θα επιτρέπει επίσης να δοθεί προτεραιότητα στην οργάνωση των πόρων της εταιρίας η του οργανισμού, ώστε να είναι σε θέση να προσδιορίσουν αν θα ανταποκριθούν πιο γρήγορα στις πιο κρίσιμες απειλές σχετικά με την Ασφάλεια και την συμμόρφωση [36].

Όπως αντιλαμβανόμαστε για την Ασφάλεια στα BIG DATA, οι τακτικές που ακολουθούνται δεν αφορούν τόσο την βελτίωση στο τεχνικό επίπεδο, που είναι βέβαια αναγκαία, όσο στο διαχειριστικό επίπεδο.

5.4 Εξόρυξη Ασφάλειας από τα BIG DATA

5.4.1 Εισαγωγή

Ο αντίκτυπος της μη εκμετάλλευσης των δεδομένων που αφορούν και την άμυνα των ίδιων των δεδομένων και την Ασφάλεια, θα μπορούσε να είναι πολύ υψηλότερος από εκείνον που εμφανίζεται στον ιδιωτικό τομέα όσον αφορά την απώλεια κέρδους ή το μερίδιο αγοράς. Το στοιχείο της νοημοσύνης απαιτείται όλο και περισσότερο για να αλιεύεις πληροφορίες μέσα στα BIG DATA με στόχο την ανακάλυψη γνώσης [28].

Τα χρονοδιαγράμματα για λόγους ασφαλείας μετριούνται σε δευτερόλεπτα, όχι ώρες ή ημέρες , απαιτώντας συνεχή και ταχεία ανάλυση. Για τους αναλυτές, η ουσία είναι η μετάβαση όλο και περισσότερο σε μια πιο άμεση ανταπόκριση στις «διαφυγούσες» υπογραφές και μακριά από τη χρονοβόρα παρακολούθηση ρουτίνας, προκειμένου να ανακαλύψουν αυτές τις υπογραφές.

Οι προσδοκίες από τους διοικούντες συνεχίζουν να αυξάνονται καθώς όλο και περισσότεροι ψηφιακοί αισθητήρες και συλλέκτες τίθενται σε κάθε υπηρεσία, και νέες υπογραφές προσδιορίζονται με αποτέλεσμα να χρειάζεται να ανανεώνονται συνεχώς οι Βάσεις Δεδομένων . Οι προσδοκίες αυτές δεν θα επιτευχθούν χωρίς να υπάρξουν οι αντίστοιχες βελτιώσεις στα εργαλεία και τις τεχνικές για την αναζήτηση και την ανάλυση των δεδομένων [28].

Τα δεδομένα, λοιπόν, μπορούν να αποτελέσουν είτε απειλή είτε ευκαιρία για τις επιχειρήσεις και τους οργανισμούς ανεξαρτήτως μεγέθους και τύπου. Ένα ιδιαίτερα ενδιαφέρον παράδειγμα δίνει η Bridget van Kralingen, senior vice-president στο τμήμα Global Business Services της IBM και αφορά στον τομέα της υγείας των Η.Π.Α. Σύμφωνα με το FBI οι απάτες στον τομέα της υγείας στις Η.Π.Α. κοστίζουν περί τα 250 δις δολάρια ανά έτος. Οι λύσεις των BIG DATA Analytics σε αυτή την περίπτωση θα μπορούσαν να βοηθήσουν τις επιχειρήσεις και τους οργανισμούς να ανακαλύψουν ταχύτερα τυχόν απάτες. Χαρακτηριστικό παράδειγμα είναι αυτό μιας

ασφαλιστικής εταιρείας, η οποία με τη χρήση του κατάλληλου λογισμικού στα BIG DATA μπόρεσε να απομονώσει τις ψεύτικες δηλώσεις αποζημίωσης μέσα σε 24 ώρες. Μέχρι πρότινος ο χρόνος που απαιτούνταν ήταν 14 μέρες [23].

Τα θέματα που αφορούν στην Ασφάλεια και στην προστασία της ιδιωτικότητας, μεγεθύνονται στην περίπτωση των BIG DATA, μιας και επηρεάζονται από την ταχύτητα δημιουργίας, της ροής, τον όγκο και την ποικιλία των δεδομένων. Για να το κατανοήσουμε αυτό, αρκεί να σκεφτούμε παραδείγματα μεγάλων υποδομών cloud, την ύπαρξη πολλαπλών πηγών ψηφιακών δεδομένων, καθώς και πολλαπλών μορφών δεδομένων σε κάποιο συγκεκριμένο οργανισμό. Ως εκ τούτου, οι παραδοσιακές δικλείδες Ασφάλειας οι οποίες είναι προσαρμοσμένες σε έλεγχο κυρίως, στατικών δεδομένων (χωρίς συνεχή ροή) με μικρό σχετικά όγκο, καθίστανται ανεπαρκείς [29].

Οι επιχειρήσεις και οι οργανισμοί οφείλουν να αναπτύσσουν και να εξελίσσουν μηχανισμούς πρόβλεψης κακόβουλων εισβολών, να επιταχύνουν τον εντοπισμό των κινδύνων ώστε να προστατεύουν τα πολύτιμα δεδομένα τους. Για να αναγνωρίζουν επιτυχώς και να αποκαθιστούν τις παρατεταμένες μη εξουσιοδοτημένες προσβάσεις στο δίκτυο τους, που είναι γνωστές και ως Advanced Persistent Threats - APTs, οι οργανισμοί πρέπει να είναι προετοιμασμένοι [29]:

- Να χειρίζονται και να επεξεργάζονται πληροφορίες υψηλής ταχύτητας, μεγάλες σε όγκο και ποικιλία.
- Να αναλύουν δομημένα, ημιδομημένα και μη δομημένα δεδομένα τόσο εντός όσο και εκτός του δικτύου τους.
- Να παρακολουθούν τα περιστατικά σε cloud, mobile και virtual περιβάλλοντα
- Να αναλαμβάνουν δράση αυτόματα μόλις εντοπιστεί μια απειλή.

Λόγω της φύσης τους, τα BIG DATA Analytics επιτρέπουν σε έναν οργανισμό να οργανώνει και να αναλύει τεράστιες ποσότητες δομημένων, ημιδομημένων και μη δομημένων πληροφοριών για να διευκολύνει τον εντοπισμό των ανέντιμων εργαζομένων και συνεργατών, καθώς και της εγκληματικής ή αθέμιτης δραστηριότητας. Ένα σημαντικό συστατικό της επιτυχίας είναι η δυνατότητα για γρήγορη και εύκολη ενσωμάτωση όλων των τύπων πληροφοριών σε όλο το εύρος πολλαπλών εσωτερικών και εξωτερικών πηγών πληροφόρησης [29].

Τα BIG DATA Analytics μπορούν να προσφέρουν ένα σημαντικό οπλοστάσιο στο πλαίσιο των μέτρων που παίρνει μια εταιρεία για την προστασία του δικτύου της, είναι γνωστό στους περισσότερους επαγγελματίες του IT. Παρόλο αυτά, οι τεχνολογίες αυτού του είδους δεν έχουν βρει την αποδοχή που τους αρμόζει και ενσωματώνονται με πολύ αργούς στην υποδομή IT.

Αυτό αποτελεί, άλλωστε, το συμπέρασμα μιας πρόσφατης έρευνας της εταιρείας ανάλυσης δικτυακής Ασφάλειας Ponemon. Στην έρευνα με τον τίτλο "BIG DATA Analytics in Cyber Defense", η οποία διεξήχθη από την Ponemon σε συνεργασία με την Teradata, ρωτήθηκαν περισσότεροι από 700 ειδικοί στο χώρο του IT Security. Σύμφωνα με την άποψη των ερωτηθέντων, οι κυβερνοεπιθέσεις στις επιχειρήσεις και στους κρατικούς οργανισμούς προκαλούν όλο και σοβαρότερες συνέπειες. Ωστόσο, μόλις 20% των ερωτηθέντων αναφέρει ότι η επιχείρησή τους έχει βελτιώσει το τελευταίο διάστημα τους μηχανισμούς προστασίας και αντιμετώπισης απειλών.

Τα μεγαλύτερα κενά ασφαλείας εντοπίζονται στις εφαρμογές mobile, στην αδυναμία ελέγχου και παρακολούθησής τους, στα πολυάριθμα δικτυωμένα συστήματα, γενικότερα, που χρησιμοποιούνται για διαφορετικούς σκοπούς. Μπορεί το 56% των ερωτηθέντων να γνωρίζουν τεχνολογίες όπως είναι τα BIG DATA Analytics και το 61% να παραδέχεται ότι αυτές οι τεχνολογίες μπορούν να δώσουν λύση στα επείγοντα θέματα ασφαλείας, ωστόσο μόνο 35% των ερωτηθέντων χρησιμοποιούν, ήδη, αυτές τις τεχνολογίες.

Λιγότεροι από τους μισούς οργανισμούς (42%) που συμμετείχαν στην έρευνα προσπαθούν να εμποδίζουν την ασυνήθιστη ή δυναμικά επικίνδυνη κυκλοφορία δεδομένων στα δίκτυα τους, ενώ σχεδόν οι μισοί (49%) επιχειρούν να ανακαλύψουν την ύπαρξη τέτοιου είδους κυκλοφορίας δεδομένων στο δίκτυό τους.

Η μελέτη της Ponemon κρούει τον κώδωνα του κινδύνου, καθώς οι επιχειρήσεις οφείλουν όσο γίνεται πιο γρήγορα να επανεξετάσουν τα προγράμματα που χρησιμοποιούν για την Ασφάλεια της υποδομής IT και να τα διευρύνουν με δυνατότητες, που θα λαμβάνουν υπόψη τους και τα BIG DATA. Με αυτό τον τρόπο θα μπορούσαν να συρρικνώσουν το χρόνο που μεσολαβεί μεταξύ της επίθεσης, της αναγνώρισής και αντιμετώπισής της, με απώτερο στόχο τον περιορισμό της ζημιάς. Όταν χρησιμοποιούνται πολυδομημένα δεδομένα από διαφορετικές πηγές, οι επιχειρήσεις μπορούν να προσφέρουν μια αποτελεσματική προστασία έναντι των κυβερνοεπιθέσεων.

Τα BIG DATA Analytics, όταν χρησιμοποιούνται σε τεχνολογίες Ασφάλειας, προσφέρουν μια αποτελεσματική προστασία έναντι των επιθέσεων. Το 82% των ερωτηθέντων θα επιθυμούσαν τα προγράμματα antivirus και antimalware να εξοπλιστούν με BIG DATA Analytics, ενώ το 80% είπαν ότι τα προγράμματα προστασίας έναντι των επιθέσεων “Denial of Service“ ή “Distributed Denial of Service“ θα έκανε τους οργανισμούς τους ασφαλέστερους.

Αν και η εξάπλωση και η πολυπλοκότητα των δεδομένων συνιστούν σημαντικές προκλήσεις σε σχέση με την προστασία από τις κυβερνοεπιθέσεις, οι εταιρείες χρησιμοποιούν όλο και πιο συχνά νέα εργαλεία και τεχνολογίες BIG DATA για τη διαχείριση των δεδομένων, τα οποία μπορούν να αντιπαρέλθουν στον όγκο και στην πολυπλοκότητα τους. Τα νέα αναλυτικά εργαλεία για τις Βάσεις Δεδομένων μπορούν να βελτιστοποιήσουν την ένταση και την ακρίβεια μιας στρατηγικής ασφαλείας και να βοηθήσουν τους οργανισμούς ώστε να ικανοποιήσουν τις απαιτήσεις, οι οποίες προκύπτουν από τις σύνθετες και απέραντες δομές δεδομένων.

Οι ειδικοί στην Ασφάλεια IT βλέπουν στα δεδομένα τόσο μια πρόκληση (50% κατονόμασαν την ανάπτυξη των δεδομένων, 39% την ενσωμάτωσή τους), όσο και μια

ευκαιρία (61% για την ενσωμάτωση δεδομένων, 53% για την πολυπλοκότητα των δεδομένων). Ενώ πολλές επιχειρήσεις έχουν δυσκολίες με τις υπάρχουσες in-house τεχνολογίες τους και την υφιστάμενη πραγματογνωμοσύνη τους, μόλις 35% δήλωσε ότι χρησιμοποιεί ήδη λύσεις BIG DATA, ενώ λίγοι παραπάνω από τους μισούς είπαν ότι η εταιρεία τους διαθέτει το απαραίτητο προσωπικό και εμπειρογνωμοσύνη.

Τα BIG DATA Analytics μπορούν να κλείσουν τα υπάρχοντα κενά ανάμεσα στις τεχνολογίες και στους χρήστες στο θέμα της Ασφάλειας IT, απλοποιώντας σημαντικά την προστασία των δικτύων. Μπορούν να συλλάβουν δεδομένα από τη δραστηριότητα των χρηστών, να επεξεργαστούν και να παρέχουν αλγόριθμους, με τους οποίους κάθε κόμβος του δικτύου θα μπορεί να παρακολουθηθεί σχεδόν σε πραγματικό χρόνο. Ένα πλεονέκτημα των BIG DATA Analytics στο θέμα της κυβερνο-Ασφάλειας είναι η ικανότητα αναγνώρισης προτύπων δραστηριότητας, τα οποία σηματοδοτούν κίνδυνο για το δίκτυο. Με αυτό τον τρόπο μπορεί να υπάρξει μια ταχύτερη αντίδραση σε ασυνήθιστες δραστηριότητες.

Πολλές ομάδες που ασχολούνται με το θέμα της Ασφάλειας του IT έχουν αναγνωρίσει ότι δεν είναι καθόλου εύκολο να αναγνωρίζουν τα δικτυακά δεδομένα που παραπέμπουν σε ασυνήθιστες συμπεριφορές ή αποτελούν μια πιθανή απειλή για το δίκτυό. Έτσι, η κυβερνο-Ασφάλεια και η παρακολούθηση του δικτύου έχει μετατραπεί, λίγο-πολύ, σε ένα πρόβλημα των BIG DATA. Οι επιχειρήσεις που διαχειρίζονται προσωπικά, ευαίσθητα ή σημαντικά δεδομένα, πρέπει να επεκτείνουν αντίστοιχα τα συστήματα Ασφάλειας τους, αν δεν θέλουν να εκτεθούν, τόσο οι ίδιες όσο και οι πελάτες τους, σε σημαντικούς κινδύνους.

Για τη μελέτη το ινστιτούτο Ponemon ρώτησε περισσότερους από 700 ειδικούς και στελέχη Πληροφορικής στις ΗΠΑ. Το θέμα της έρευνας ήταν κυρίως οι νέες τεχνολογίες για τη διαχείριση των δεδομένων και η ανάλυσή τους, ώστε οι οργανισμοί να προστατεύσουν προληπτικά και καλύτερα τα δίκτυα τους από τις πάσης φύσεως ψηφιακές επιθέσεις. Ανάμεσα στους ερωτηθέντες συμπεριλαμβάνονταν ειδικοί από εταιρείες παροχής χρηματοοικονομικών υπηρεσιών, την ευρύτερη βιομηχανία και από δημόσιους οργανισμούς. Η μέση τους επαγγελματική εμπειρία ανέρχονταν στα δέκα χρόνια. Όλοι οι ερωτηθέντες ήταν

οικείοι με την εφαρμογή μέτρων ασφαλείας έναντι των κυβερνοεπιθέσεων και είχαν ένα σημαντικό ποσοστό ευθύνης για αυτό τον τομέα [45].

Στα BIG DATA Analytics που διαχειρίζονται δεδομένα από τις εφαρμογές, των χρηστών, το δίκτυο και τα συστήματα, είναι επιτακτική η «εξόρυξη» Ασφάλειας με τη διαχείριση των BIG DATA [29].

5.4.2 Υπάρχοντα Συστήματα Ασφαλείας και BIG DATA

Τα Security Analytics γίνονται γρήγορα μια εφαρμογή BIG DATA για έναν απλό λόγο: μεγάλες οργανώσεις συλλέγουν, επεξεργάζονται και την αναλύουν όλο και περισσότερα δεδομένα προκειμένου να αντιμετωπίσουν αποτελεσματικά το νέο τοπίο των απειλών στον κυβερνοχώρο.

Η υπόσχεση των Security Information & Event Management (SIEM) τεχνολογιών ήταν να παράσχουν δυνατότητες εξελιγμένων Analytics. Η πραγματικότητα είναι ότι τα προϊόντα SIEM δεν είχαν σχεδιαστεί για BIG DATA Analytics και γενικά δεν μπορούν να ανταποκριθούν στις ταχέως εξελισσόμενες ανάγκες που απαιτούν τώρα οι εμπορικές οργανώσεις.

Οι τεχνολογίες SIEM παρέχουν μια καλή βάση για την παρακολούθηση της Ασφάλειας με ικανότητα ανίχνευσης υπογραφών επιθέσεων σχεδόν σε πραγματικό χρόνο ή βασίζονται σε κανόνες για να αναζητήσουν γνωστές απειλές. Οι τεχνολογίες SIEM είναι κατάλληλες για την τήρηση και την υποβολή αναφορών ασφαλείας. Ωστόσο, οι τεχνολογίες SIEM δεν κλιμακώνονται για τον εντοπισμό των άγνωστων απειλών σε όλα τα διαθέσιμα στοιχεία. Τα δεδομένα συχνά πρέπει να φιλτράρονται πριν φορτωθούν σε ένα σύστημα SIEM. Οι τεχνολογίες SIEM δεν μπορούν να κάνουν τα προηγμένα Analytics ασφαλείας που απαιτούνται σήμερα [42].

Μπορούμε να εντοπίσουμε ορισμένες από τις τάσεις εξετάζοντας τον τρόπο με τον οποίο τα εργαλεία ασφαλείας έχουν αλλάξει κατά την τελευταία δεκαετία. Όταν η αγορά IDS αισθητήρων μεγάλωσε, οι αισθητήρες παρακολούθησης του δικτύου και

τα εργαλεία καταγραφής αναπτύχθηκαν σε εταιρικά δίκτυα, ωστόσο, η διαχείριση των σημάτων από αυτές τις διαφορετικές πηγές δεδομένων έγινε ένα δύσκολο έργο.

Τώρα τα εργαλεία ανάλυσης BIG DATA βελτιώνουν τις πληροφορίες που διατίθενται στους αναλυτές ασφαλείας συσχετίζοντας, συγκεντρώνοντας και απεικονίζοντας ακόμα περισσότερο ποικίλες πηγές δεδομένων για μεγαλύτερο χρονικό διάστημα [46].

5.4.3 Οι Εφαρμογές Ασφαλείας εξελίσσονται από τα BIG DATA

Σύμφωνα με το RSA, το τμήμα Ασφάλειας της EMC Corporation, τα BIG DATA θα αποτελέσουν καταλύτη σημαντικών αλλαγών στον κλάδο προστασίας δεδομένων μέσα από τη χρήση ευφυών μοντέλων Ασφάλειας. Μέχρι το 2015 τα εργαλεία ανάλυσης των BIG DATA είναι πιθανό να έχουν καταλυτική επίδραση στις περισσότερες κατηγορίες προϊόντων του τομέα προστασίας δεδομένων, όπως η διαχείριση δεδομένων ασφαλείας (SIEM), η παρακολούθηση των δικτυακών υποδομών, η ταυτοποίηση και η εξουσιοδότηση των χρηστών, η διαχείριση της ταυτότητας του χρήστη, η ανίχνευση περιπτώσεων απάτης, η πληροφορική διακυβέρνηση, τα συστήματα διαχείρισης κινδύνων και συμμόρφωσης.

Σε βάθος χρόνου, τα BIG DATA θα επιφέρουν αλλαγές και στα συμβατικά εργαλεία προστασίας, όπως τα Firewalls, οι εφαρμογές για την αποτροπή της απώλειας δεδομένων (Data loss prevention) και τα προγράμματα anti-malware. Μέσα σε κάποια χρόνια, τα BIG DATA Analytics θα έχουν εξελιχθεί με τρόπο που θα επιτρέπει προηγμένες δυνατότητες πρόληψης κινδύνων και αυτοματοποιημένων διαδικασιών ελέγχου σε πραγματικό χρόνο.

Το σημερινό περιβάλλον IT μιας επιχείρησης, το οποίο χαρακτηρίζεται από το cloud και την εκτεταμένη κινητικότητα ενός όλο και μεγαλύτερου αριθμού χρηστών, έχει καταστήσει περιττές πολλές από τις μέχρι σήμερα δημοφιλείς πολιτικές προστασίας οι οποίες βασίζονται στην προστασία μιας εξωτερικής περιμέτρου και τους στατικούς ελέγχους εναντίον συγκεκριμένων απειλών. Γι’ αυτό, οι επικεφαλής

ασφαλείας στρέφονται προς intelligence-driven μοντέλα προστασίας, τα οποία είναι ευέλικτα και μπορούν να αναγνωρίζουν κάθε ρίσκο με βάση το περιεχόμενο κάθε εφαρμογής ή πληροφορίας (contextual), προκειμένου να υπάρχει προστασία και από άγνωστες απειλές.

«Το παιχνίδι αλλάζει. Όλο και περισσότερα δεδομένα ανεβαίνουν στο διαδίκτυο μέσα από αυτοματοποιημένες φόρμες και διαδικασίες, τάση που προβλέπεται να διατηρηθεί. Έτσι, ένα εργαλείο ασφαλείας το οποίο λειτουργούσε πολύ καλά μέχρι πριν δύο ή τρία χρόνια, δεν είναι σε θέση να ανταποκριθεί το ίδιο καλά και τώρα. Σήμερα, θα πρέπει να ψάξετε σε πολύ περισσότερα Data, προκειμένου να εντοπίσετε πολύ πιο εξελιγμένες απειλές. Τα συνηθισμένα εργαλεία που προσφέρει η αγορά αλλάζουν προκειμένου να αξιοποιήσουν τις δυνατότητες των BIG DATA που είναι διαθέσιμα online», δηλώνει ο William H. Stewart, Senior Vice President, Booz Allen Hamilton. Το Security Brief της RSA προτείνει έξι βασικές αρχές, οι οποίες μπορούν να βοηθήσουν τις εταιρείες να σχεδιάσουν τη μετάβαση των συστημάτων ασφαλείας τους στην εποχή των BIG DATA.

1. Διαμόρφωση μιας συνολικής στρατηγικής κυβερνο-προστασίας – Οι οργανισμοί θα πρέπει να αποκτήσουν μια ολοκληρωμένη στρατηγική προστασίας έναντι των κυβερνοαπειλών και να διαμορφώσουν ένα πρόγραμμα προσαρμοσμένο στους κινδύνους, τις απειλές και τις απαιτήσεις τους.
2. Καθιέρωση μηχανισμών ενημέρωσης για θέματα ασφαλείας – Επειδή τα BIG DATA Analytics βασίζονται σε πληροφορίες διαφόρων τύπων η συλλογή των οποίων γίνεται από πολλές διαφορετικές πηγές, χρειάζεται να αναπτυχθεί ένας ενιαίος μηχανισμός συλλογής, κωδικοποίησης, ανάλυσης και διαμοίρασης των πληροφοριών αυτών.
3. Μετάβαση από τις μεμονωμένες λύσεις σε μια ενιαία αρχιτεκτονική προστασίας – Οι οργανισμοί θα πρέπει να σκεφτούν στρατηγικά σε σχέση με το ποια προϊόντα ασφαλείας θα συνεχίσουν να χρησιμοποιούν και μετά από χρόνια, επειδή καθένα από αυτά χρησιμοποιεί τις δικές του δομές Data, οι

οποίες θα πρέπει να μπουν όλες κάτω από μια ενιαία ομπρέλα επεξεργασίας και ανάλυσης.

4. Αναζήτηση ανοιχτών και scalable εργαλείων BIG DATA– Οι οργανισμοί θα πρέπει να είναι βέβαιοι ότι οι σημερινές επενδύσεις σε προϊόντα ασφαλείας βασίζονται σε ευέλικτες τεχνολογίες ανάλυσης, και όχι σε στατικά εργαλεία που οριοθετούνται από τα άκρα του δικτύου τους ή από ένα συγκεκριμένο σετ γνωστών απειλών. Νέα εργαλεία, έτοιμα να αξιοποιήσουν τα BIG DATA μπορούν να εξασφαλίσουν την απαραίτητη ευελιξία στο σχεδιασμό ασφαλείας, ώστε να υπάρχει δυνατότητα προσαρμογής σε νέες επιχειρησιακές απαιτήσεις, απειλές ή υποδομές IT.
5. Ενδυνάμωση των μηχανισμών προστασίας με εξειδικευμένο προσωπικό πληροφορικής – Τη στιγμή που τα νέα εργαλεία θα είναι έτοιμα να αξιοποιήσουν την τεχνολογία BIG DATA, οι υπεύθυνοι IT ενδέχεται να μην είναι. Ο χώρος των Data Analytics χαρακτηρίζεται από την έλλειψη έμπειρων στελεχών. Οι Data Scientistsμε εξειδικευμένες γνώσεις σε θέματα Ασφάλειας και προστασίας είναι λίγοι και περιζήτητοι. Αυτό θα έχει ως πιθανό αποτέλεσμα, πολλοί οργανισμοί να στραφούν σε εξωτερικούς συνεργάτες για να καλύψουν τα κενά τους στον τομέα των Security Analytics.
6. Αξιοποίηση εξωτερικής πληροφόρησης γύρω από τις υπάρχουσες απειλές – Συμπλήρωση των εσωτερικών μηχανισμών προστασίας με εξωτερικές υπηρεσίες συλλογής και αξιολόγησης πληροφοριών σχετικά με τις εκάστοτε απειλές, από αξιόπιστες και ενημερωμένες πηγές.

Ωστόσο, οι IT Managers φαίνονται προσωρινά αρκετά διστακτικοί στην πρόταση της RSA. Η βασική ιδέα είναι η δημιουργία συστημάτων αποθήκευσης πληροφοριών, τα οποία θα βασίζονται στην πλατφόρμα Hadoop και πάνω σε αυτά θα εφαρμοστούν οι πιο εξελιγμένοι αλγόριθμοι Security Analytics. Όσο τα συστήματα εξελίσσονται ο όγκος των δεδομένων παρουσιάζει εκθετική αύξηση. Το πρόβλημα είναι ότι ένα μεγάλο ποσοστό αυτών των δεδομένων είναι σκουπίδια με αποτέλεσμα

τα συστήματα ανάλυσης να υπερφορτώνονται και να γίνονται πιθανώς αναποτελεσματικά [41].

5.4.4 Behavioral Analytics στα BIG DATA για την ανίχνευση απειλών

Τα Behavioral Analytics κατανοούν το παρελθόν της ανθρώπινης συμπεριφοράς, προβλέπουν τη μελλοντική συμπεριφορά και ταυτοποιούν τη μη ορθή συμπεριφορά. Τα Behavioral Analytics έχουν χρησιμοποιηθεί εκτενώς στην ανίχνευση απάτης και στην πρόληψη γιατί διαφορετικά άτομα, φυσικά εμφανίζουν διαφορετικές συμπεριφορές και η νόμιμη συμπεριφορά είναι πρακτικά πάντοτε διαφορετική από εκείνη που εμφανίζεται από ένα απατεώνα.

Τα Behavioral Analytics εκμεταλλεύονται αυτό το γεγονός. Αντί απλά να ψάχνουν για συγκεκριμένους δείκτες, συνδυάζουν τη γνώση με την παρακολούθηση για να προσδιοριστεί εάν η συμπεριφορά είναι αναμενόμενη και θεμιτή, ή ύποπτη. Τα BIG DATA είναι μια πρόκληση για τα Behavioral Analytics όχι μόνο λόγο του όγκου των δεδομένων που εμπλέκονται, αλλά και λόγω της ανάγκης να φέρουν μια ευρεία ποικιλία των πηγών δεδομένων και σχημάτων μαζί για να δημιουργήσουν μια πλήρη εικόνα.

Τα Cyber Security Analytics υιοθετούν όλο και περισσότερο τα Behavioral Analytics, από το πεδίο ανίχνευσης της απάτης, προκειμένου να αντιμετωπίσουν την πραγματικότητα ότι οι παραδοσιακές λύσεις ασφαλείας έχουν αποδειχθεί αναποτελεσματικές απέναντι στην απίστευτη ποικιλία και τον όγκο της ψηφιακής εγκληματικότητας, όπως για την κατασκοπεία στον κυβερνοχώρο, το έγκλημα στον κυβερνοχώρο, Hacktivism και την εσωτερική απειλή . Αυτή η σύγκλιση ανίχνευσης τόσο απειλής στον κυβερνοχώρο όσο και απάτης σημαίνει ότι τα οφέλη που κατακτώνται από την ανάλυση συμπεριφοράς στα δεδομένα, για την καταπολέμηση και των δυο, θα μας οδηγήσει τόσο μεγαλύτερη λειτουργική αποτελεσματικότητα όσο και σε καλύτερες επενδυτικές αποφάσεις. Τα Behavioral Analytics αποδεικνύονται ότι

είναι πιο ισχυρά, διαρκή και αποτελεσματικά από ό, τι τα Analytics ανίχνευσης υπογραφών και αυτά τα βασισμένα σε κανόνες [42].

Τα παραδοσιακά εργαλεία άμυνας αποτυγχάνουν να προστατεύσουν τις επιχειρήσεις από τις προηγμένες στοχευμένες επιθέσεις (ATAs) και το πρόβλημα των προηγμένων κακόβουλων προγραμμάτων (Advanced Malware). Το 2013, οι επιχειρήσεις δαπάνησαν αστρονομικά ποσά για Firewalls, συστήματα πρόληψης εισβολών (IDSs) και πλατφόρμες προστασίας. Ωστόσο, οι προηγμένες στοχευμένες επιθέσεις και προηγμένα κακόβουλα προγράμματα εξακολουθούν να πλήττουν τις επιχειρήσεις.

Σύμφωνα με έρευνα της Gartner οι υπεύθυνοι για την Ασφάλεια μπορούν να εφαρμόσουν αποδοτικές μεθόδους άμυνας απέναντι στις προηγμένες επίμονες απειλές (APTs) όπου κάθε μια εφαρμόζεται σε διαφορετικό επίπεδο από το επίπεδο δικτύου μέχρι το επίπεδο εφαρμογής.

- **Ανάλυση της κυκλοφορίας του δικτύου (Network Traffic Analysis)**

Σε αυτό το επίπεδο περιλαμβάνεται ένα ευρύ φάσμα τεχνικών για την ανάλυση της κυκλοφορίας του δικτύου. Για παράδειγμα, ανώμαλη μοτίβα κίνησης DNS είναι μια ισχυρή ένδειξη της δραστηριότητας μιας προσπάθειας επίθεσης ενός δικτύου υπολογιστών στην διαθεσιμότητα ενός κατανεμητή (botnet for Ddos Attack).Η καταγραφή του NetFlow παρέχει τη δυνατότητα να καθιερωθούν οι βασικές γραμμές των κανονικών ρευμάτων κυκλοφορίας και να αποκαλυφθεί τυχόν κίνηση ανώμαλων μοτίβων που είναι επικίνδυνα. Μερικά εργαλεία συνδυάζουν ανάλυση πρωτοκόλλων και ανάλυση περιεχομένου.

- **Forensics Δικτύου (Network Forensics)**

Τα Forensics Δικτύου είναι εργαλεία παρέχουν δέσμευση και αποθήκευση της κίνησης του δικτύου ολόκληρων πακέτων και παρέχουν εργαλεία για ανάλυση και αναφορά που καλύπτουν τις ανάγκες για υποστήριξη της αντιμετώπισης των

περιστατικών, της έρευνας και της προηγμένη ανάλυση των απειλών. Η ικανότητα αυτών των εργαλείων να εξάγουν και να διατηρούν τα μετά-δεδομένα τα διαφοροποιεί από τα εργαλεία σύλληψης πακέτων του δικτύου.

- **Ανάλυση Ωφέλιμου Φορτίου (Payload Analysis)**

Χρησιμοποιώντας ένα περιβάλλον sandbox, η τεχνική Ανάλυση Ωφέλιμου Φορτίου χρησιμοποιείται για την ανίχνευση κακόβουλου λογισμικού (Malware) και στοχευμένων επιθέσεων σχεδόν σε πραγματικό χρόνο. Η Ανάλυση Ωφέλιμου Φορτίου προσφέρει λύσεις οι οποίες παρέχουν λεπτομερείς εκθέσεις σχετικά με τη συμπεριφορά κακόβουλου λογισμικού , αλλά δεν προσφέρει τη δυνατότητα παρακολούθησης της συμπεριφοράς στο Endpoint περιβάλλον για μεγάλη χρονική περίοδο.

- **Endpoint Ανάλυσης Συμπεριφοράς (Endpoint Behavior Analysis)**

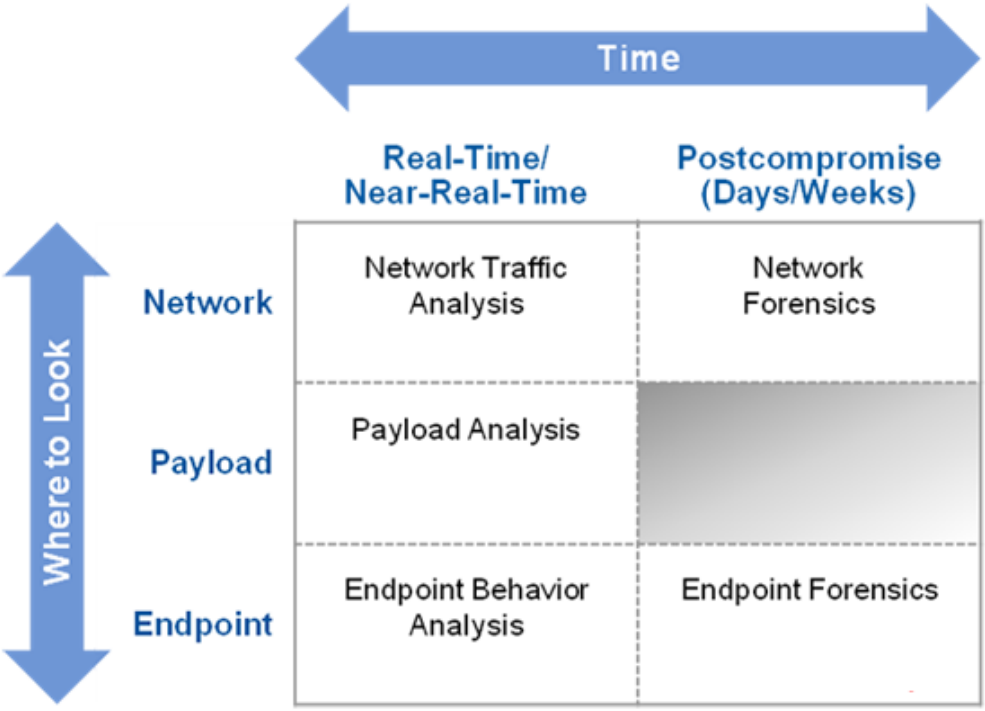
Υπάρχουν περισσότερες από μία προσεγγίσεις για την Endpoint Ανάλυση Συμπεριφοράς για να αμυνθούμε απέναντι στις στοχευμένες επιθέσεις. Μια λύση βρίσκεται στον περιορισμό των εφαρμογών η οποία επιτυγχάνεται απομονώνοντας τις εφαρμογές και τα αρχεία σε εικονικούς χώρους αποθήκευσης. Άλλες καινοτομίες περιλαμβάνουν τη διαμόρφωση του συστήματος, τη μνήμη και την παρακολούθηση της διαδικασίας για την παρεμπόδιση των επιθέσεων, καθώς και τεχνικές για ανταπόκριση σε πραγματικό χρόνο για την αντιμετώπιση των περιστατικών. Μια εντελώς διαφορετική στρατηγική για την άμυνα έναντι των APTs είναι ο περιορισμός στην εκτέλεση των εφαρμογών που είναι γνωστές "καλές" εφαρμογές, τεχνικής η οποία είναι γνωστή ως εφαρμογές «λευκής λίστας».

- **Endpoint Forensics**

Το Endpoint Forensics χρησιμεύει ως ένα εργαλείο για τις ομάδες αντιμετώπισης κινδύνων. Endpoint πράκτορες συλλέγουν δεδομένα από τους hosts

που παρακολουθούν. Οι λύσεις αυτές είναι χρήσιμες καθώς επισημαίνουν ποιοί υπολογιστές έχουν παραβιαστεί από κακόβουλο λογισμικό, και μπορούν και καταγράφουν την όποια συγκεκριμένη συμπεριφορά του κακόβουλου λογισμικού.

Λόγω των προκλήσεων όσον αφορά την καταπολέμηση των στοχευμένων επιθέσεων και του κακόβουλου λογισμικού, οργανισμούς ασφαλείας αποτελεσματική λύση για την προστασία απέναντι στις ATAs επιτυγχάνεται με χρήση συνδυασμού τουλάχιστον δυο μεθόδων και με συμπληρωματικό τρόπο. Αποτελεσματική προστασία προέρχεται από το συνδυασμό των τεχνολογιών από διαφορετικές σειρές (για παράδειγμα: δίκτυο / ωφέλιμου φορτίου, ωφέλιμου φορτίου / τελικό σημείο ή το δίκτυο / καταληκτικό σημείο). Η πιο αποτελεσματική προσέγγιση είναι να συνδυάσει μεθόδους διαγώνια μέσα από το πλαίσιο [54].



Εικόνα 5.4 Five Styles of Advanced Threat Defense [54]

Τα BIG DATA αλλάζουν το τοπίο των τεχνολογιών ασφαλείας για την παρακολούθηση του δικτύου, τα συστήματα SIEM και τις μεθόδους για Forensics. Ωστόσο, στην αιώνια κούρσα εξοπλισμών της επίθεσης και της άμυνας, τα BIG

DATA Analytics δεν είναι πανάκεια, και οι ερευνητές ασφαλείας πρέπει να συνεχίσουν τη διερεύνηση νέων τρόπων για να αντεπεξέλθουν σε εξελιγμένες επιθέσεις. Τα BIG DATA μπορούν επίσης να δημιουργήσουν έναν κόσμο όπου η διατήρηση του ελέγχου σχετικά την αποκάλυψη των προσωπικών μας πληροφοριών συνεχώς αμφισβητείται.

Ως εκ τούτου, πρέπει να εντείνουμε τις προσπάθειές μας για να εκπαιδευτεί μια νέα γενιά επιστημόνων που σέβονται την αξία της ιδιωτικής ζωής και να υπάρξει συνεργασία για την ανάπτυξη εργαλείων και για το σχεδιασμό συστημάτων BIG DATA που ακολουθούν κοινώς αποδεκτές κατευθυντήριες γραμμές της ιδιωτικής ζωής [46].

6 Παραλληλισμός και BIG DATA

6.1 Εισαγωγή

Η εξέλιξη στην τεχνολογία των υπολογιστών υπήρξε ραγδαία τις τελευταίες δεκαετίες. Το 1945 δεν υπήρχαν υπολογιστές με δυνατότητα αποθήκευσης προγραμμάτων. Σήμερα υπάρχουν σαφώς υπολογιστές που έχουν καλύτερη απόδοση, περισσότερη μνήμη, και μεγαλύτερη χωρητικότητα δίσκου από έναν παλαιότερο υπολογιστή. Στην αρχή τις δεκαετίας του 1970 έκαναν την εμφάνισή τους τα πρώτα ολοκληρωμένα κυκλώματα. Τη δεκαετία εκείνη η επίδοση των υπολογιστών βελτιωνόταν με ρυθμό περίπου 25% με 30% κάθε χρόνο. Το 1971 έκανε την εμφάνισή του ο πρώτος μικροεπεξεργαστής μέσα σε ένα chip: ο επεξεργαστής 4004 της Intel ο οποίος διέθετε 2300 τρανζίστορ ενώ είχε ταχύτητα ρολογιού 740 kHz, εύρος διαύλου bus 4 bits και επίδοση 70 χιλιάδες εντολές το δευτερόλεπτο (0,07 MIPS).

Από την αρχή της ιστορίας των υπολογιστών η βάση εξέλιξης της αρχιτεκτονικής τους στηριζόταν σε τρεις βασικές αρχές:

- Στο **pipelining**, το οποίο επέτρεπε την ταυτόχρονη εκτέλεση πολλαπλών εντολών, έστω και σε διαφορετικές φάσεις η κάθε μια, από μια μόνο κεντρική επεξεργαστική μονάδα
- Στον **παραλληλισμό**, ο οποίος με χρήση πολλών επεξεργαστικών μονάδων επιτρέπει την ταυτόχρονη εκτέλεση πολλαπλών εντολών και
- Στην **πασίγνωστη αρχή της τοπικότητας των κλήσεων** που καθιστά ανταποδοτική την χρήση της κρυφής μνήμης (cache) και της εικονικής μνήμης (virtual memory).

Είναι σημαντικό να τονίσουμε ότι οι παραπάνω αρχές στοχεύουν στην βελτίωση της απόδοσης ενός υπολογιστικού συστήματος, όπου συνεργασία πολλών (**κατανεμημένο σύστημα**) κάνουν ακόμα πιο αποδοτική και γρήγορη την εκτέλεση μιας διαδικασίας. Παρακάτω αναφέρουμε αυτές τις τεχνικές αύξησης της αποδοτικότητας του συστήματος κατά την εκτέλεση της διαδικασίας [47].

6.2 Αρχή της Τοπικότητας των Κλήσεων

Η σπουδαιότερη ιδιότητα που χαρακτηρίζει την εκτέλεση των προγραμμάτων των υπολογιστών είναι η *τοπικότητα της αναφοράς* (locality of reference). Όταν ένα πρόγραμμα εκτελείται από την ΚΜΕ, η επόμενη εντολή που θα ανακληθεί κάθε φορά από τη μνήμη και θα εκτελεστεί στην ΚΜΕ θα ανήκει πιθανότατα σε κάποια στην ίδια ή κάποια γειτονική θέση σε σχέση με αυτή που εκτελέστηκε αμέσως προηγουμένως [49].

6.3 Pipeline

Είναι μία αρχιτεκτονική που χρησιμοποιείται σε πολλούς σύγχρονους επεξεργαστές, η αρχιτεκτονική διοχέτευσης (pipeline architecture). Στους επεξεργαστές αυτούς, ο κύκλος εκτέλεσης μίας εντολής διαιρείται σε πολλές διαδοχικές φάσεις, που ονομάζονται κύκλοι διοχέτευσης(pipeline cycles). Μέσα στην ΚΜΕ υπάρχει ένα διακεκριμένο τμήμα που αντιστοιχεί σε κάθε ένα κύκλο και ονομάζεται βαθμίδα (stage). Κάθε εντολή ξεκινά από την πρώτη βαθμίδα της διοχέτευσης και περνά από όλες με τη σειρά μέχρι να φθάσει και στην τελευταία. Μόλις μία βαθμίδα τελειώσει την επεξεργασία μίας εντολής, μπορεί αμέσως να προχωρήσει στην επόμενη χωρίς καθυστέρηση [48].

Έστω ότι υπάρχουν πέντε διαφορετικές βαθμίδες σε έναν υποθετικό επεξεργαστή και αναλαμβάνουν να εκτελέσουν τις πέντε υποθετικές φάσεις

εκτέλεσης μια εντολής που αντιστοιχούν στην κάθε βαθμίδα όπως φαίνεται παρακάτω (Εικόνα 6.1) (Α – Β – Γ – Δ – Ε).

Εντολές	Χρόνος → 1	2	3	4	5	6	7	8	9
1	Α	Β	Γ	Δ	Ε				
2		Α	Β	Γ	Δ	Ε			
3			Α	Β	Γ	Δ	Ε		
4				Α	Β	Γ	Δ	Ε	
5					Α	Β	Γ	Δ	Ε

Εικόνα 6.1 Παράδειγμα Pipelineμε 5 εντολές και 5 βαθμίδες- Φάσεις [47]

Τη χρονική περίοδο 1 η βαθμίδα Α είναι απασχολημένη με την εντολή 1, ενώ οι υπόλοιπες βαθμίδες περιμένουν τη σειρά τους. Τη χρονική περίοδο 2 η βαθμίδα Α ασχολείται με την εντολή 2 αφού η εντολή 1 τώρα βρίσκεται στη φάση Β. Τη χρονική περίοδο 3 η βαθμίδα Α ασχολείται με την εντολή 3, η βαθμίδα Β με την εντολή 2 και η βαθμίδα Γ με την εντολή 1. Με τον ίδιο τρόπο συνεχίζουμε την εκτέλεση των εντολών ώστε κάθε βαθμίδα παραλαμβάνει μια εντολή από την προηγούμενη,

Έτσι η εντολή 1 είναι εκτελείται μετά την χρονική περίοδο 5, η εντολή 2 είναι έτοιμο μετά τη χρονική περίοδο 6 κλπ. Μέσα σε 9 χρονικές περιόδους έχουν ολοκληρωθεί 5 η εντολές. Είναι εύκολο να δει κανείς ότι μετά από Ν περιόδους θα έχουν ολοκληρωθεί Ν+4 εντολές (ομοίως Ρ εντολές απαιτούν Ρ+4 περιόδους για την παραγωγή τους). Ο λόγος είναι ότι τις 4 πρώτες περιόδους δεν απασχολούνται και οι 5 βαθμίδες. Οι 4 αυτές πρώτες χρονικές περίοδοι λέγονται και χρόνος γεμίσματος της pipeline. Μέχρι να «γεμίσει» η pipeline δεν εκτελείται καμία ολοκληρωμένη εντολή. Αφού όμως γεμίσει εκτελείται μια εντολή κάθε μια χρονική περίοδο. Αν δεν κάναμε χρήση της αρχιτεκτονικής διοχέτευσης τότε 5 εντολές θα τις εκτελούσαμε 25 χρονικές περιόδους αντί για 9, επομένως καταφέραμε να έχουμε επιτάχυνση!

Πολύ συχνά ωστόσο συμβαίνει οι διάφορες φάσεις να μην απαιτούν τον ίδιο χρόνο εκτέλεσης.

Από την παραπάνω συζήτηση εξάγονται τα ακόλουθα συμπεράσματα:

- σκοπός της τεχνικής pipelining είναι η επιτάχυνση της εκτέλεσης μιας διαδικασίας με τον τεμαχισμό της σε επί μέρους φάσεις οι οποίες μπορούν να εκτελούνται ταυτόχρονα για ξεχωριστές διαδικασίες
- η επιτάχυνση που επιτυγχάνεται δεν είναι μόνο συνάρτηση του αριθμού των φάσεων αλλά και της εξισορρόπησης μεταξύ των χρόνων εκτέλεσης των διαφόρων φάσεων. Με άλλα λόγια αν τεμαχίσουμε μια διαδικασία σε 10 επί μέρους φάσεις αυτό δε σημαίνει ότι θα επιτύχουμε επιτάχυνση της διαδικασίας κατά 10 φορές αν δεν έχουμε προηγουμένως φροντίσει οι φάσεις αυτές να απαιτούν τον ίδιο (ή περίπου τον ίδιο) χρόνο εκτέλεσης.

Στην τεχνολογία των επεξεργαστών η τεχνική του pipelining είναι εξαιρετικά διαδεδομένη και πλέον αποτελεί αναπόσπαστο κομμάτι της αρχιτεκτονικής [47].

Το σημαντικότερο πρόβλημα στην επίτευξη μέγιστης επιτάχυνσης από την pipeline είναι οι λεγόμενοι κίνδυνοι (hazards). Παρακάτω κάνουμε μια απλή αναφορά των κινδύνων οι οποίοι προκύπτουν από τρεις διαφορετικές αιτίες [47]:

- οι **δομικοί κίνδυνοι** προκύπτουν στις περιπτώσεις όπου το υλικό δεν αρκεί για να καλύψει τις ανάγκες όλων των δυνατών συνδυασμών εντολών που μπορούν να προκύψουν στην pipeline.
- οι **κίνδυνοι των δεδομένων** προκύπτουν όταν η εκτέλεση μιας εντολής εξαρτάται από το αποτέλεσμα μιας προηγούμενης εντολής
- οι **κίνδυνοι ελέγχου** προκύπτουν από τις εντολές άλματος οι οποίες αλλάζουν τη ροή του προγράμματος.

6.4 Παραλληλία

Η ακολουθία είναι μια από τις θεμελιώδεις πλευρές της ανθρώπινης και φυσικής δραστηριότητας. Ο κάθε άνθρωπος ακολουθεί μια καθημερινή ρουτίνα αποτελούμενη από μια σειρά βημάτων δραστηριότητας: σηκώνεται το πρωί, τρώει το

πρωινό του, πηγαίνει στη δουλειά ή το σχολείο, τρώει το μεσημεριανό, το βραδινό κ.λπ. Η ανθρώπινη ομιλία είναι σειριακή και η γνώση εκφράζεται και μεταδίδεται με μία σειρά από λέξεις είτε προφορικά είτε γραπτά. Όλη η έννοια του χρόνου, που είναι τόσο σημαντική για τον προγραμματισμό και την εκτέλεση επιτυχημένων ενεργειών, βασίζεται στην έννοια της ακολουθίας. Παντού στη φύση και στην ανθρώπινη ύπαρξη, η δραστηριότητα ξεδιπλώνεται και εκφράζεται μέσα από την αρχή της ακολουθίας. Αυτός είναι και ο λόγος που οι πρώτοι αλγόριθμοι και τα πρώτα προγράμματα υλοποιήθηκαν με βάση την έννοια της ακολουθίας. Ακόμη πριν κατασκευαστούν οι πρώτοι Η/Υ, το μαθηματικό πρότυπο ενός αλγορίθμου καθοριζόταν ως μία πεπερασμένη ακολουθία πράξεων.

Ωστόσο η επιστήμη των υπολογιστών βαθμιαία ωρίμασε και έγινε φανερό, ιδιαίτερα τα τελευταία χρόνια, ότι η ακολουθία είναι τμήμα μόνο της υπόθεσης. Οι ανθρώπινες και φυσικές δραστηριότητες δεν είναι μόνο σειριακές αλλά και παράλληλες: η δράση δεν αναπτύσσεται μόνο σειριακά αλλά η δράση συμβαίνει ταυτόχρονα σε πολλά διαφορετικά σημεία. Η παραλληλία είναι εξίσου θεμελιώδης και σημαντική όσο και η ακολουθία.

Κάθε άνθρωπος μιλάει και ενεργεί σειριακά, ωστόσο αποτελεί τμήμα κάποιου οργανισμού, ο οποίος περιέχει πολλούς ανθρώπους που όλοι λειτουργούν παράλληλα. Φανταστείτε σε πιο σημείο θα βρισκόταν οι ανθρωπότητα σήμερα αν δεν υπήρχε η συντονισμένη δραστηριότητα πολλών μεμονωμένων ανθρώπων που δουλέψανε ταυτόχρονα ως μία ομάδα. Η ίδια ακριβώς αρχή παρατηρείται και στη φύση: οι κύκλοι που αναπτύσσονται στη φύση είναι σίγουρα δράσεις σειριακές αλλά η φύση λειτουργεί παράλληλα, παντού χωρίς όρια. Η επιστήμη της Φυσικής τώρα αρχίζει να διαμορφώνει τους πιο θεμελιακούς φυσικούς νόμους με βάση τους όρους των πεδίων: ενεργειακά πεδία και πεδία ύλης. Το πεδίο είναι μια οντότητα που λειτουργεί παντού την ίδια ακριβώς χρονική στιγμή με τέλεια παραλληλία. Γι' αυτό μια ολοκληρωμένη εικόνα της ανθρώπινης αλλά και της φυσικής δραστηριότητας απαιτεί οπωσδήποτε και τις δύο βασικές έννοιες: την ακολουθία και την παραλληλία [51].

6.5 Παράλληλα Συστήματα και Κατανεμημένα Συστήματα

Η ανάγκη για την επίλυση μεγάλων προβλημάτων και η εξέλιξη της τεχνολογίας του διαδικτύου, είχε ως αποτέλεσμα την διαρκή ανάγκη για την εύρεση όλο και περισσότερων πόρων. Η ανάγκη αυτή οδήγησε στην δημιουργία δομών συνεργαζόμενων υπολογιστικών συστημάτων, με απώτερο σκοπό την επίλυση προβλημάτων που απαιτούν μεγάλη υπολογιστική ισχύ ή την αποθήκευση μεγάλου όγκου δεδομένων. Η ύπαρξη τέτοιων δομών αλλά και κεντρικών μονάδων επεξεργασίας με περισσότερους από έναν επεξεργαστές, δημιούργησε πρωτόκολλα για την δημιουργία εφαρμογών που θα εκτελούνται και θα επιλύουν ένα πρόβλημα σε περισσότερους από έναν επεξεργαστές, ώστε να επιτευχθεί η μείωση του χρόνου εκτέλεσης. Ένα παράδειγμα τέτοιου πρωτοκόλλου είναι αυτό της ανταλλαγής μηνυμάτων (MPI).

Τα τελευταία χρόνια λοιπόν παρατηρείται το γεγονός να αυξάνεται η ζήτηση που σχετίζεται με αύξηση της απόδοσης των υπολογιστικών συστημάτων. Ταυτόχρονα παρατηρείται και μια ραγδαία εξέλιξη στον τομέα των τηλεπικοινωνιών και των δικτύων που καθιστά εφικτό την πρόσβαση σε απομακρυσμένα υπολογιστικά συστήματα με στόχο την ανταλλαγή δεδομένων.

Στην περίπτωση όπου το ζητούμενο είναι **η αύξηση της απόδοσης ενός υπολογιστή** τότε το πρόβλημα είναι θέμα *υπολογιστικής αρχιτεκτονικής*.

Στην περίπτωση που το ζητούμενο είναι **η επικοινωνία μεταξύ απομακρυσμένων υπολογιστικών σταθμών για την ανταλλαγή δεδομένων και την συνεργασία για την αντιμετώπιση κάποιας εφαρμογής** τότε το πρόβλημα είναι περισσότερο θέμα *κατανεμημένης επεξεργασίας*.

Η παράλληλη επεξεργασία είναι ένα είναι ένας τομέας που κατηγοριοποιείται στην αρχιτεκτονική των υπολογιστών με στόχο τη βελτίωση της ταχύτητας επεξεργασίας χωρίς να βασίζεται στην βελτίωση της τεχνολογίας του υλικού.

Θα μπορούσαμε να πούμε σε γενικές γραμμές πως ένας υπολογιστής θεωρείται παράλληλος αν αποτελείται από πολλές επεξεργαστικές μονάδες που συνεργάζονται

στενά για την λύση ενός προβλήματος σε χρόνο μικρότερο από το χρόνο, από τον χρόνο που θα χρειαζόταν ένας επεξεργαστής μόνος του για να λύσει το πρόβλημα αυτό.

Η παράλληλη επεξεργασία άρχισε να αποδίδει καρπούς τα τελευταία έτη. Αποτέλεσμα αυτού, είναι η εμφάνιση εμπορικών παράλληλων συστημάτων με διπλούς ή τετραπλούς επεξεργαστές Η παράλληλη επεξεργασία μπήκε πλέον στο χώρο του λεγόμενου desktop computing, δηλαδή στους υπολογιστές γραφείου.

Μια άλλη καινοτομία των τελευταίων ετών βασίζεται στη ραγδαία ανάπτυξη των δικτύων. Η διάδοση και η πλέον σχετικά φθηνή δημιουργία πολύ γρήγορων τυποποιημένων δικτύων τοπικής εμβέλειας ή ακόμη και ευρείας περιοχής έδωσε τη δυνατότητα ανάπτυξης ενός νέου υπολογιστικού μοντέλου όπου πολλοί, απλοί ή όχι υπολογιστές, όντας συνδεδεμένοι μέσα στο ταχύ δίκτυο μπορούν να λειτουργούν ως μια μεγάλη, εικονική παράλληλη μηχανή. Αυτή η τεχνολογία είναι γνωστή ως grid-computing ενώ οι ομάδες τέτοιων υπολογιστών καλούνται clusters(μια ακόμα πιο νέα αναφορά είναι το cloud computing). Η τεχνολογία αυτή αποτελεί μια γέφυρα μεταξύ παράλληλης και κατανεμημένης τεχνολογίας [47][50].

6.6 Παράλληλος Προγραμματισμός

Στον σειριακό προγραμματισμό γράφουμε ένα πρόγραμμα προκειμένου να εκτελεστεί σε σειριακούς υπολογιστές. Με τον όρο αυτό εννοούμε ότι η εκτέλεση θα γίνει σε έναν επεξεργαστή, ο οποίος θα εκτελέσει την μια εντολή μετά την άλλη και θα εκτελείται μόνο μια εντολή την φόρα μέχρι να φτάσουμε στο τέλος του προγράμματος. Αντιθέτως, στον παράλληλο προγραμματισμό δεν συμβαίνει αυτό. Τα συστήματα που χρησιμοποιούμε σε αυτή την περίπτωση διαθέτουν περισσότερους επεξεργαστές, και ορισμένα τμήματα του προγράμματος εκτελούνται ταυτόχρονα, εφόσον βέβαια κάτι τέτοιο είναι εφικτό. Για αυτό, το πρόγραμμα αρχικά χωρίζεται σε τμήματα τα οποία μπορούν να εκτελεστούν ταυτόχρονα και ο κάθε επεξεργαστής αναλαμβάνει την εκτέλεση ενός τμήματος από αυτά με τον κλασικό σειριακό τρόπο. Είναι λοιπόν φανερό ότι κάτι τέτοιο δίνει την δυνατότητα να εκτελεστεί ένα

πρόγραμμα σε λιγότερο χρόνο αφού κάποια κομμάτια του μπορούν να εκτελεστούν ταυτόχρονα.

Επομένως, ως ορισμό, θα μπορούσαμε να πούμε ότι ο παράλληλος προγραμματισμός είναι ο προγραμματισμός σε μία γλώσσα που επιτρέπει διαφορετικά τμήματα της υπολογιστικής διαδικασίας να εκτελούνται ταυτόχρονα σε διαφορετικούς υπολογιστές. Γιατί όμως παράλληλος προγραμματισμός; Ένας λόγος λοιπόν, για τον οποίο ο παράλληλος προγραμματισμός είναι σημαντικός, αναφέρθηκε ήδη και αφορά την ταχύτητα εκτέλεσης ενός προγράμματος. Επίσης θα πρέπει να σημειωθεί ότι οι ταχύτητες των επεξεργαστών δεν αυξάνονται με τον ίδιο ρυθμό που αυξάνονταν τα προηγούμενα χρόνια και για αυτό, προκειμένου να πετύχουμε γρηγορότερες ταχύτητες, δεν στοχεύουμε στην αύξηση της ταχύτητας του επεξεργαστή αλλά στην δημιουργία αρχιτεκτονικών με περισσότερους επεξεργαστές, Τελικώς, με τον τρόπο αυτό, πετυχαίνουμε γρηγορότερες ταχύτητες αν χρησιμοποιούμε τους επεξεργαστές αυτούς ταυτόχρονα. Με τον τρόπο αυτό λοιπόν μπορούμε να κάνουμε χρήση επεξεργαστών ή μηχανημάτων που έχουμε ήδη και να δημιουργήσουμε παράλληλα συστήματα τα οποία να είναι αρκετά γρήγορα κερδίζοντας σε κόστος αφού δεν θα είναι αναγκαία η αγορά νέου, γρηγορότερου, επεξεργαστή.

Είναι γεγονός πως υπάρχουν αρκετά μεγάλα προβλήματα, τα οποία χρειάζονται πολλούς πόρους, ειδικά σε μνήμη και είναι αδύνατον να λυθούν σε έναν υπολογιστή και γι αυτό γίνεται αναγκαία η χρήση περισσότερων. Οι υπολογιστές που θα χρησιμοποιηθούν δεν είναι απαραίτητο να βρίσκονται στον ίδιο χώρο αλλά οπουδήποτε αλλού. Επομένως αν τοπικά διαθέτουμε ένα υπολογιστικό σύστημα με περιορισμένες δυνατότητες, μπορούμε να χρησιμοποιήσουμε άλλους που βρίσκονται κάπου αλλού και μας παρέχουν περισσότερους υπολογιστικούς πόρους [50].

Σκοπός λοιπόν της σχεδίασης και κατασκευής παράλληλων υπολογιστών είναι η εκτέλεση προγραμμάτων σε μικρότερο χρόνο από εκείνον που θα χρειαζόταν ένας υπολογιστής με ένα μόνο επεξεργαστή της ίδιας τεχνολογίας, δηλαδή ένας μονοεπεξεργαστής (Uniprocessor). Η επίτευξη αυτού του στόχου μας εισάγει στην τεχνολογία του παράλληλου προγραμματισμού [47].

6.7 Παραλληλισμός σε σύνολα Δεδομένων

Ένας αλγόριθμος εξόρυξης δεδομένων δίνει καλύτερα αποτελέσματα με περισσότερα δεδομένα και μπορεί να φτάσει στην ίδια ακρίβεια αποτελεσμάτων από έναν καλύτερο ή πιο πολύπλοκο αλγόριθμο. Ο βασικός στόχος ενός συστήματος ανάλυσης και εξόρυξης δεδομένων είναι να επεξεργαστεί περισσότερα δεδομένα με καλύτερους αλγόριθμους. Σε πολλούς τομείς η μεγαλύτερη ποσότητα δεδομένων είναι σημαντική, διότι παρέχει μια πιο ακριβή περιγραφή του αναλυθέντος φαινομένου. Με περισσότερα δεδομένα, οι αλγόριθμοι εξόρυξης δεδομένων είναι σε θέση να εξάγουν μια ευρύτερη ομάδα παραγόντων συσχετίσεων και πιο διακριτές συσχετίσεις. Σήμερα, με τα BIG DATA έχουμε ποσότητες εκατοντάδων Terabytes ή Petabytes και αυτά είναι πραγματικά σενάρια. Το πρόβλημα της αποθήκευσης αυτών των μεγάλων συνόλων δεδομένων δημιουργείται από την μια λόγω της αδυναμίας να έχουμε μια μονάδες δίσκου με το κατάλληλο μέγεθος και το πιο σημαντικό, το ότι χρειάζεται μεγάλο χρονικό διάστημα για να έχουμε πρόσβαση σε αυτήν. Η ταχύτητα πρόσβασης στα BIG DATA επηρεάζεται από την ταχύτητα του δίσκου, την εσωτερική μεταφορά δεδομένων, την εξωτερική μεταφορά δεδομένων, την κρυφή μνήμη, το χρόνο πρόσβασης, την αναμονή περιστροφής, παράγοντες οι οποίοι δημιουργούν καθυστερήσεις και σημεία συμφόρησης.

Μια απλή λύση είναι να διαβάσουμε τα δεδομένα από όλους τους δίσκους ταυτόχρονα. Αν λάβουμε υπόψη το κανάλι επικοινωνίας, τότε τα σημεία συμφόρησης που δημιουργούνται είναι από το διαθέσιμο εύρος ζώνης. Τέλος, η απόδοση μειώνεται με την ταχύτητα της πιο αργής συνιστώσας.

Άλλα χαρακτηριστικά των BIG DATA προσθέτουν συμπληρωματικά επίπεδα δυσκολίας:

- πολλές πηγές εισόδου, σε διαφορετικό οικονομικό και κοινωνικό τομέα, υπάρχουν πολλές πηγές πληροφοριών.

- πλεονασμός, καθώς τα ίδια δεδομένα μπορούν να παρέχονται από διαφορετικές πηγές.
- έλλειψη κανονικοποίησης στα δεδομένα ή έλλειψη προτύπων αναπαράστασης των δεδομένων, τα δεδομένα μπορεί να έχουν διαφορετικές μορφές, μοναδικές ταυτότητες, διαφορετικές μονάδες μέτρησης, διαφορετικούς βαθμούς ακεραιότητας και συνέπειας. Δεδομένα που περιγράφουν το ίδιο φαινόμενο μπορεί να διαφέρουν σε σχέση με τη μέτρηση των χαρακτηριστικών τους, τις μονάδες μέτρησης, το χρόνο της εγγραφής, τις μεθόδους που χρησιμοποιούνται.

Για περιορισμένα σύνολα δεδομένων η αποτελεσματική λύση διαχείρισης των δεδομένων δίνεται από τις σχεσιακές Βάσεις Δεδομένων SQL, αλλά για τα BIG DATA μερικές από τις θεμελιώδεις αρχές τους δεν αρκούν.

6.8 Παράλληλες Βάσεις Δεδομένων

Κατά την πρόσβαση στα BIG DATA, το σύστημα αποθήκευσης αρχείων μπορεί να αποτελέσει εμπόδιο. Για αυτό πολλή σκέψη τέθηκε σε επανασχεδιασμό του παραδοσιακού συστήματος αρχείων για την καλύτερη απόδοση κατά την πρόσβαση σε μεγάλα αρχεία δεδομένων. Σε μια κατανεμημένη προσέγγιση, το σύστημα αρχείων έχει ως στόχο να επιτύχει τους ακόλουθους στόχους:

- η πρόσβαση θα πρέπει να είναι επεκτάσιμη, το σύστημα αρχείων θα πρέπει να επιτρέπει να προστεθεί επιπλέον Hardware για την αύξηση της ικανότητας αποθήκευσης ή / και των επιδόσεων.
- θα πρέπει να προσφέρει υψηλή απόδοση, το σύστημα αρχείων θα πρέπει να είναι σε θέση να εντοπίσει τα δεδομένα που μας ενδιαφέρουν επί των διανεμόμενων κόμβων σε εύθετο χρόνο.

- θα πρέπει να είναι αξιόπιστο, το σύστημα αρχείων θα πρέπει να είναι σε θέση να αναδημιουργήσει από τους κατανεμημένους κόμβους τα αρχικά δεδομένα με τρόπο πλήρη και χωρίς στρεβλώσεις.
- θα πρέπει να έχει υψηλή διαθεσιμότητα, το σύστημα αρχείων θα πρέπει να χειρίζεται τις αποτυχίες και να ενσωματώσει μηχανισμούς για την παρακολούθηση και ανίχνευση σφαλμάτων, την ανοχή σφαλμάτων και την αυτόματη αποκατάσταση.

Μία κατανεμημένη βάση δεδομένων (DDB) είναι μια συλλογή πολλαπλών, λογικά διασυνδεδεμένων Βάσεων Δεδομένων που διανέμονται μέσω ενός δικτύου υπολογιστών. Ένα κατανεμημένο σύστημα διαχείρισης Βάσεων Δεδομένων, κατανεμημένο DBMS, είναι ένα σύστημα λογισμικού που επιτρέπει τη διαχείριση μιας κατανεμημένης βάση δεδομένων και καθιστά την κατανομή διαφανή στους χρήστες. Ένα παράλληλο DBMS,είναι ένα σύστημα διαχείρισης που υλοποιείται σε έναν υπολογιστή με πολλούς επεξεργαστές. Το παράλληλο DBMS υλοποιεί την έννοια του οριζόντιου κατακερματισμού διανέμοντας τα μέρη ενός μεγάλου σχεσιακού πίνακα σε πολλούς κόμβους με δυνατότητα να υποβάλλονται σε επεξεργασία με παραλληλισμό. Αυτό απαιτεί διαμερισμένη εκτέλεση SQL λειτουργιών. Κάποιες βασικές λειτουργίες, όπως μια απλή SELECT, μπορεί να εκτελεστεί ανεξάρτητα σε όλους τους κόμβους. Οι πιο σύνθετες λειτουργίες εκτελούνται μέσω ενός αγωγού πολλαπλών-λειτουργιών. Διαφορετικές αρχιτεκτονικές πολλαπλών παράλληλων συστημάτων, όπως είναι η κοινής μνήμης, με κοινούς δίσκους ή τίποτα κοινό, καθορίζουν τις πιθανές στρατηγικές για την υλοποίηση ενός παράλληλου DBMS, όπου η κάθε μια με τα δικά της πλεονεκτήματα και μειονεκτήματά. Η αρχιτεκτονική με τίποτα κοινό διανέμει δεδομένα σε ανεξάρτητους κόμβους και έχει υλοποιηθεί από πολλά εμπορικά συστήματα, καθώς παρέχει επεκτασιμότητα και διαθεσιμότητα.

Με βάση τους παραπάνω ορισμούς, μπορούμε να συμπεράνουμε ότι τα συστήματα παράλληλων Βάσεων Δεδομένων βελτιώνουν τις επιδόσεις της επεξεργασίας των δεδομένων μέσω του παραλληλισμού του φόρτου, θέτοντας κατάλληλα ευρετήρια και επερωτήσεις στα δεδομένα. Στα κατανεμημένα συστήματα

Βάσεων Δεδομένων, τα δεδομένα αποθηκεύονται σε διαφορετικά DBMSs που μπορούν να λειτουργήσουν ανεξάρτητα. Επειδή τα συστήματα παράλληλων Βάσεων Δεδομένων μπορούν να διανέμουν δεδομένα για να αυξήσουν την απόδοση της αρχιτεκτονικής, υπάρχει μια λεπτή γραμμή που χωρίζει τις δύο έννοιες σε πραγματικές εφαρμογές. Παρά τις διαφορές μεταξύ των παράλληλων και κατανεμημένων DBMSs, τα περισσότερα από τα πλεονεκτήματά τους είναι κοινά:

- Τα αποθηκευμένα δεδομένα είναι σύμφωνα με ένα καλά καθορισμένο σχήμα, αυτό επικυρώνει τα δεδομένα και παρέχει την ακεραιότητα των δεδομένων.
- Τα δεδομένα είναι δομημένα σε ένα σχεσιακό μοντέλο γραμμών και στηλών.
- Τα ερωτήματα SQL είναι γρήγορα.
- Η γλώσσα ερωτημάτων SQL είναι ευέλικτη εύκολη να τη μάθει κανείς και να διαβάσετε και επιτρέπει στους προγραμματιστές να εφαρμόσουν σύνθετες εργασίες με ευκολία.
- Χρησιμοποιούν κατακερματισμό ή B-δεντρικά ευρετήρια για να επιταχύνουν την πρόσβαση στα δεδομένα.
- Μπορούν να επεξεργαστούν αποτελεσματικά σύνολα δεδομένων μεγέθους έως και δύο Petabytes. Γνωστές εμπορικές παράλληλες Βάσεις Δεδομένων Teradata, Aster Data, Netezza , DATAllegro, Vertica, Greenplum, η IBMDB2 και OracleExaData, έχουν αποδειχθεί επιτυχής, διότι: επιτρέπουν γραμμική κλιμάκωση.
- Το σύστημα μπορεί να διατηρήσει σταθερή απόδοση καθώς το μέγεθος της βάσης δεδομένων αυξάνεται με την προσθήκη περισσότερων κόμβων στο παράλληλο σύστημα.
- Επιτρέπουν γραμμική επιτάχυνση, για μια βάση δεδομένων με ένα σταθερό μέγεθος, η απόδοση μπορεί να αυξηθεί με την προσθήκη περισσότερων συστατικών, όπως επεξεργαστές, μνήμη και τους δίσκους.
- Εφαρμόζουν εσωτερικά ερωτήματα, ενδο-ερωτήματα και ενδο-λειτουργία παραλληλισμού.

- Απαιτούν μειωμένη προσπάθεια εφαρμογής.
- Απαιτούν μειωμένη προσπάθεια διοίκησης.
- Προσφέρουν υψηλή διαθεσιμότητα.

Στην αρχιτεκτονική μαζικής παράλληλης επεξεργασίας (MPP), προσθέτοντας περισσότερο υλικό δίνεται η δυνατότητα για μεγαλύτερη χωρητικότητα αποθήκευσης και αύξηση της ταχύτητας ερωτημάτων. Η MPP αρχιτεκτονική, υλοποιείται ως συσκευή αποθήκευσης δεδομένων, μειώνει την προσπάθεια εφαρμογής λόγω του ότι το υλικό και το λογισμικό είναι προεγκατεστημένα και δοκιμασμένα για να λειτουργήσουν τη συσκευή, πριν την προμηθευτούμε. Επίσης, μειώνει την προσπάθεια διαχείρισης, καθώς έρχεται ως ανεξάρτητη συσκευή από το υπόλοιπο σύστημα. Οι συσκευές αποθήκευσης δεδομένων προσφέρουν υψηλή διαθεσιμότητα με ενσωματωμένες δυνατότητες ανακατεύθυνσης χρησιμοποιώντας πλεονασμό δεδομένων για κάθε δίσκο. Ιδανικά, κάθε μονάδα επεξεργασίας της συσκευής αποθήκης δεδομένων πρέπει να επεξεργάζεται την ίδια ποσότητα δεδομένων σε κάθε δεδομένη στιγμή. Για να επιτευχθεί αυτό, τα δεδομένα θα πρέπει να κατανέμονται ομοιόμορφα σε κάθε μονάδα επεξεργασίας. Η σκεβρώματος των δεδομένων είναι ένα μέτρο για να αξιολογήσει πώς γίνεται η κατανομή των δεδομένων σε κάθε μονάδα επεξεργασίας. Μια τιμή σκεβρώματος ίση με 0 σημαίνει ότι ο ίδιος αριθμός των εγγραφών κατανέμεται σε κάθε μονάδα επεξεργασίας και θεωρείται ιδανική. Έχοντας λοιπόν κάθε μονάδα επεξεργασίας να κάνει το ίδιο ποσό εργασίας εξασφαλίζεται ότι όλες οι μονάδες επεξεργασίας θα τελειώσουν το έργο τους περίπου την ίδια χρονική στιγμή, ελαχιστοποιώντας τυχόν χρόνους αναμονής. Μια άλλη πτυχή που έχει σημαντικό αντίκτυπο στην απόδοση του ερωτήματος είναι ότι υπάρχουν μαζί όλα τα δεδομένα που σχετίζονται με την ίδια μονάδα επεξεργασίας. Με αυτό τον τρόπο ο χρόνος που απαιτείται για τη μεταφορά δεδομένων μεταξύ των μονάδων επεξεργασίας εξαλείφεται. Ο τρόπος με τον οποίο τα δεδομένα κατανέμονται σε παράλληλους κόμβους της βάσης δεδομένων επηρεάζουν τη συνολική απόδοση. Αν και η ισχύς του παράλληλου DBMS δίδεται από τον αριθμό των κόμβων, μπορεί να

είναι επίσης ένα μειονέκτημα. Για απλές ερωτήσεις ο πραγματικός χρόνος επεξεργασίας μπορεί να είναι πολύ μικρότερος από το χρόνο που απαιτείται για να ξεκινήσει η παράλληλη λειτουργία. Επίσης, οι κόμβοι μπορούν να γίνουν hotspots ή εμπόδια καθώς καθυστερούν το όλο σύστημα.

6.9 MapReduce

Το μοντέλο MapReduce (MR) είναι ένα μοντέλο προγραμματισμού και μια σχετική εφαρμογή για την επεξεργασία και τη δημιουργία μεγάλων συνόλων δεδομένων. Το μοντέλο αναπτύχθηκε από τον Jeffrey Dean και Sanjay Ghemawat στη Google. Το μοντέλο MapReduce στηρίζεται στην λειτουργία μιας συνάρτησης Map που χρησιμοποιεί ως αρχική διαδικασία ζεύγη κλειδιού-τιμής και μια συνάρτηση Reduce που συγχωνεύει όλες τις ενδιάμεσες τιμές του ίδιου κλειδιού. Το σύνολο των δεδομένων χωρίζεται σε μικρότερα υποσύνολα τα οποία υποβάλλονται σε επεξεργασία παράλληλα με μια μεγάλη συστάδα υπολογιστικών μηχανών. Η συνάρτηση Map, παίρνει ένα δεδομένο εισόδου και παράγει ένα σύνολο ενδιάμεσων υποσυνόλων. Η βιβλιοθήκη του μοντέλου MapReduce ομαδοποιεί όλα τα ενδιάμεσα υποσύνολα που συνδέονται με το ίδιο ενδιάμεσο κλειδί και να τα προωθεί στην συνάρτηση Reduce. Η συνάρτηση Reduce, δέχεται επίσης ένα ενδιάμεσο κλειδί και τα υποσύνολα που σχετίζονται με αυτό. Η συνάρτηση αυτή συγχωνεύει μαζί αυτά τα υποσύνολα και το κλειδί για να σχηματίσουν ένα πιθανώς μικρότερο σύνολο τιμών. Κανονικά η μια συνάρτηση Reduce παράγει καμία ή μία τιμή εξόδου. Πολλές εργασίες γίνονται στον πραγματικό κόσμο με τη χρήση του μοντέλου MapReduce. Αυτό το μοντέλο χρησιμοποιείται για την υπηρεσία αναζήτησης στο διαδίκτυο, για τη διαλογή και την επεξεργασία των δεδομένων, για την εξόρυξη δεδομένων, για μηχανική μάθηση και για ένα μεγάλο αριθμό από άλλα συστήματα. Το πλαίσιο του μοντέλου MapReduce διαχειρίζεται τον τρόπο με τον οποίο τα δεδομένα χωρίζονται μεταξύ των κόμβων και κατά πόσο τα ενδιάμεσα αποτελέσματα του ερωτήματος είναι καθολικά για το σύστημα.

Τα πλεονεκτήματα του μοντέλου MapReduce είναι:

- το μοντέλο είναι εύκολο στη χρήση, ακόμη και για τους προγραμματιστές χωρίς εμπειρία σε παράλληλα και κατανεμημένα συστήματα.
- παρέχει ανεξαρτησία αποθήκευσης του συστήματος, δεδομένου ότι δεν απαιτεί ιδιόκτητα συστήματα αρχείων δεδομένων ή προκαθορισμένα μοντέλα δεδομένων.
- Τα δεδομένα αποθηκεύονται σε αρχεία απλού κειμένου και το σύστημα δεν είναι υποχρεωμένο να συμμορφώνεται στα σχεσιακά συστημάτων δεδομένων ή οποιαδήποτε άλλη δομή.

Στην πραγματικότητα η αρχιτεκτονική MapReduce :

- Μπορεί να χρησιμοποιήσει δεδομένα που έχουν μια αυθαίρετη μορφή.
- Έχει ανοχή σε σφάλματα.
- Είναι διαθέσιμη για γλώσσες προγραμματισμού υψηλού επιπέδου.
- Παρέχει γλώσσα επερωτήσεων επιτρέπει χειρισμό με μεγάλη ταχύτητα.

6.10 Συμπεράσματα

Η επεξεργασία των BIG DATA λόγω της προέλευσής τους από πολλαπλές πηγές είναι ένα δύσκολο έργο, καθώς απαιτεί τεράστια αποθηκευτική και επεξεργαστική δυνατότητα. Επίσης, η επεξεργασία και ανάλυση αυτών των τεράστιων όγκων δεδομένων καθίσταται μη εφικτή χρησιμοποιώντας μια παραδοσιακή σειριακή προσέγγιση. Η διανομή των δεδομένων σε πολλαπλές μονάδες επεξεργασίας και η παράλληλη επεξεργασία αποδίδει γραμμικά βελτιωμένες ταχύτητες επεξεργασίας. Κατά την κατανομή των δεδομένων είναι ζωτικής σημασίας ότι σε κάθε μονάδα επεξεργασίας κατανέμεται ο ίδιος αριθμός εγγραφών και φυσικά όλα τα σχετικά σύνολα δεδομένων που πρέπει να βρίσκονται στην ίδια μονάδα επεξεργασίας. Χρησιμοποιώντας μια πολύ-επίπεδη αρχιτεκτονική για την απόκτηση, τον μετασχηματισμό, την φόρτωση και την ανάλυση των δεδομένων, εξασφαλίζεται

ότι κάθε στρώμα έχει τα εφόδια ώστε να εκτελέσει τη συγκεκριμένη διαδικασία που του έχει ανατεθεί. Παρά τον αριθμό των στρωμάτων που εκτελούνται στο παρασκήνιο, ο χρήστης χειρίζεται είναι ένα φιλικό προς αυτόν περιβάλλον εργασίας που παρέχεται από το frond-end του διακομιστή εφαρμογής. Τελικά, αφού τα θέματα αποθήκευσης και επεξεργασίας έχουν λυθεί, το πραγματικό πρόβλημα είναι να ψάξουμε για τις σχέσεις μεταξύ των διαφόρων τύπων δεδομένων, όπως έχουν κάνει κάποιοι με μεγάλη επιτυχία (όπως η Google στην αναζήτηση στο Web ή η Amazon στο ηλεκτρονικό εμπόριο).

7 Εφαρμογή

7.1 Εισαγωγή

Στην παρούσα εργασία ακολουθήθηκε η διαδικασία που περιγράψαμε στο θεωρητικό υπόβαθρο για την ανάλυση των BIG DATA. Με την διαδικασία της KDD να λαμβάνει χώρα, βασικός στόχος είναι η εξόρυξη γνώσης και συμπερασμάτων μέσω μιας διαδικασίας ανάλυσης δεδομένων. Εφαρμόστηκε βήμα προς βήμα η διαδικασία της KDD και συγκεκριμένα **η Επιλογή , η Προεπεξεργασία. ο Μετασχηματισμός, η Εξόρυξη και η Ερμηνεία(αξιολόγηση)** των δεδομένων.

Αρχικά είναι ουσιαστικό να πούμε με τι σχετίζεται η συγκεκριμένη εφαρμογή. Η εφαρμογή αφορά την ανάλυση δεδομένων ακολουθώντας της αρχές της διαδικασίας KDD, σε δεδομένα που συλλέχθηκαν από εξυπηρετητές του τμήματος πληροφορικής του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης και αφορούν μη προσωποποιημένα IP δεδομένα. Στόχος είναι η εύρεση πιθανών επιθέσεων ασφαλείας στη διαθεσιμότητα της υπηρεσίας των εξυπηρετητών(Ddos Attacks). Είναι σημαντικό να τονίσουμε πως η συγκεκριμένη εφαρμογή δεν αποτελεί μια έτοιμη εφαρμογή για παραγωγική διαδικασία αλλά μια βάση για μελλοντική ανάπτυξη μια τελικής εφαρμογής, δίνοντας την ευκαιρία σε εμάς να έχουμε μια πρώτη επαφή με όλα τα διαφορετικά τεχνολογικά περιβάλλοντα που συμμετέχουν σε ένα τόσο πολύ δύσκολο εγχείρημα. Μιλώντας για το τεχνολογικό περιβάλλον τα εργαλεία που χρησιμοποιήθηκαν είναι τα εξής:

- Το BIG DATA Anal που είναι μια εφαρμογή JAVA που αναπτύχθηκε στα πλαίσια αυτής της εργασίας. Η εφαρμογή αναπτύχθηκε από τους φοιτητές: Γολικίδη Μάριο και Σουλεϊμάνη Εμμανουήλ.

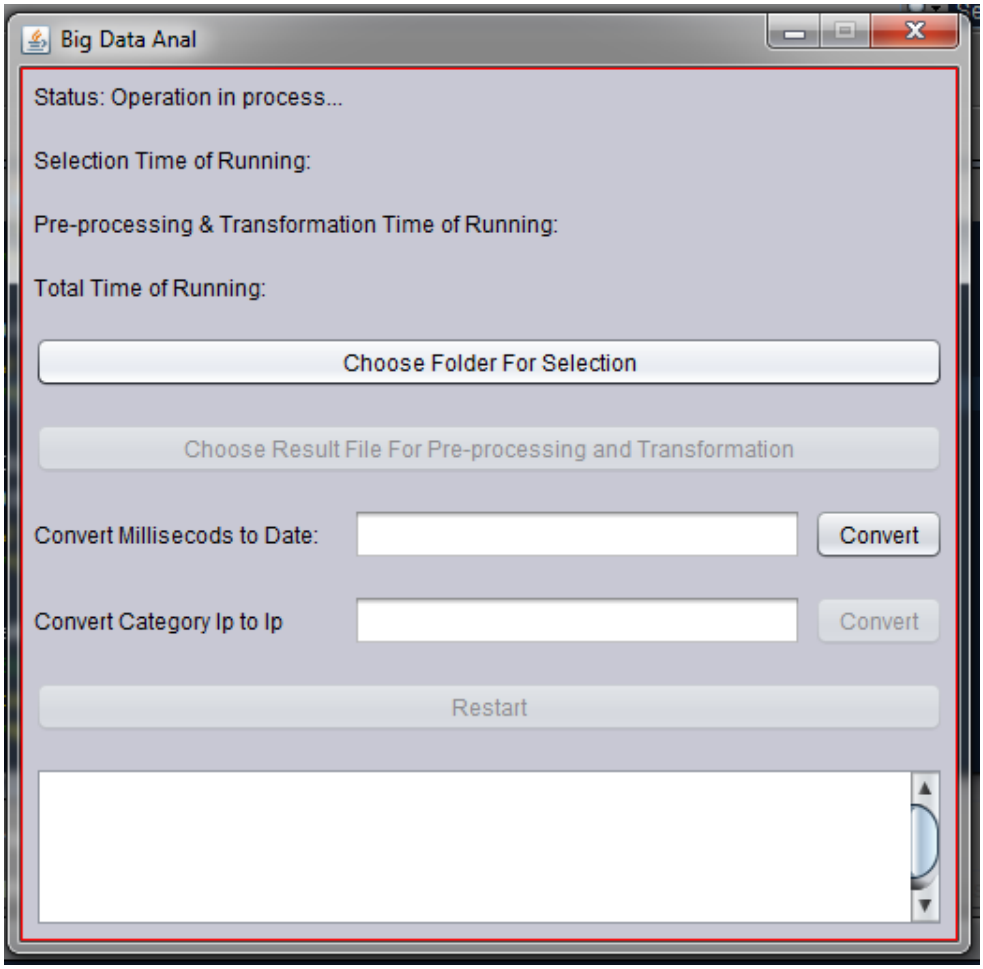
- Τα εργαλεία JavaNNS, SNNS και RSNNS για δοκιμαστικά πειράματα στην επέκταση της συγκεκριμένης εφαρμογής που αφορούν τον αλγόριθμο ART και την ενσωμάτωσή του στην εφαρμογή αυτή.
- Το WEKA που είναι ένα εργαλείο για εφαρμογές DATA MINING και συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος XMeans που είναι μια eXtended παραλλαγή του αλγορίθμου KMeans, αλλά και ο Density based clustering Algorithm.

Πριν όμως συνεχίσουμε την ανάλυση της εφαρμογής και των πειραμάτων θα πρέπει να ορίσουμε ξεκάθαρα τις παραδοχές που έγιναν για την ανάλυση αυτή.

1. Αρχικά για την πρώτη διαδικασία της KDD που αφορά την συλλογή των δεδομένων έγινε η παραδοχή πως τα δεδομένα που δεχόμαστε από τους εξυπηρετητές είναι δεδομένα από λειτουργικό σύστημα Linux. Αυτό αφορά περισσότερο τα συμπιεσμένα αρχεία.
2. Το σύστημα τρέχει σε χρόνο στατικό είναι όμως σχεδιασμένο για την προσαρμογή του σε Realtime εφαρμογές, κάτι το οποίο ισχύει για όλη την διαδικασία της KDD.
3. Τα αποτελέσματα είναι υποθετικά, βασισμένα φυσικά σε επιστημονική γνώση και η μελέτη τους έχει ως βασικό στόχο την απόδειξη της πρότασής μας και την εφαρμοσιμότητά της. Ως επέκταση της συγκεκριμένης έρευνας αποτελεί η επιβεβαίωση των αποτελεσμάτων από κάποιο σύστημα IDS ή SIEM.
4. Μπορεί και πρέπει να ενσωματωθεί η χρήση του παραλληλισμού τόσο σε επίπεδο Software όσο και σε επίπεδο Hardware για την βελτίωση της απόδοσης του συστήματος.
5. Πραγματοποιείται προσομοίωση τις διαδικασίας λόγο των δεδομένων που μας δόθηκαν (μόλις 10GB).
6. Όταν αναφερόμαστε σε IP αναφερόμαστε στην πραγματικότητα σε destination IP.

7.2 Η εφαρμογή BIG DATA Anal

Η διαδικασία εκκινείτε με την εκτέλεση της εφαρμογής. Η εφαρμογή μετά την εκκίνησή της εμφανίζει την παρακάτω φόρμα

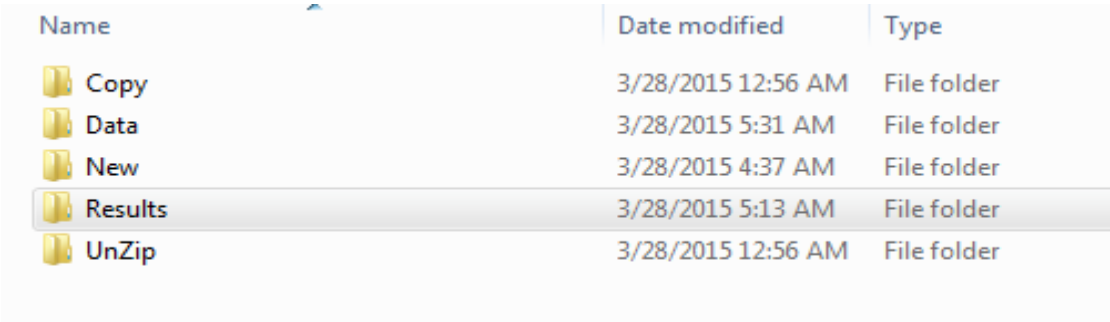


Φόρμα 7.1

Όπως μπορεί να παρατηρήσει κανείς οι επιλογές του χρήστη είναι απόλυτα συγκεκριμένες. Ο χρήστης μπορεί είτε να κλείσει την εφαρμογή, είτε να μετατρέψει Milliseconds σε Ημερομηνία, κάτι το οποίο βοηθάει και είναι ουσιαστικό στο τέλος της διαδικασίας, είτε να επιλέξει τον φάκελο ΤΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ που «περιέχει και» τα δεδομένα ως προς επεξεργασία. Το πρόγραμμα περιμένει μέσα στο φάκελο της επιλογής του χρήστη (ΕΠΕΞΕΡΓΑΣΙΑΣ) έναν φάκελο με το όνομα **New**. Αυτός ο φάκελος εμπεριέχει τα δεδομένα προς επεξεργασία. Τον ονομάσαμε **New**, διότι αναμένουμε κάθε μια χρονική μονάδα το σύστημα να «ξαναδεί» δεδομένα μέσα στο

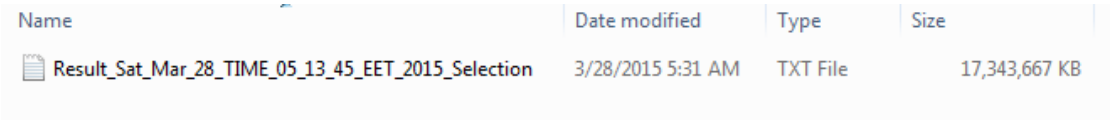
φάκελο **New** και να κάνει επανάληψη της διαδικασίας προσθέτοντας κάθε φορά νέα δεδομένα στο σύστημα για ανάλυση. Ο φάκελος επεξεργασίας που τυχαία στην δική μας εκτέλεση ονομάζεται **Program** είναι ο φάκελος στον οποίο θα δημιουργηθούν και άλλοι φάκελοι στο πρώτο στάδιο της επεξεργασίας δίπλα στον **New**.

C:\ProccessDirectory\Program



Εικόνα 7.1

C:\ProccessDirectory\Program\Results

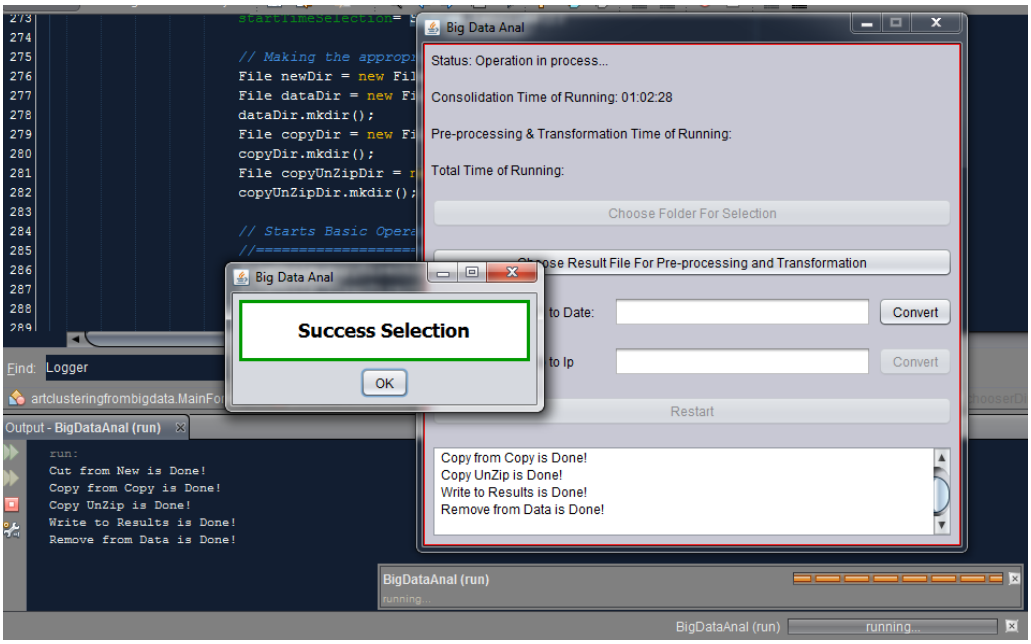


Εικόνα 7.2

Οι φάκελοι αυτοί είναι οι **Copy,UnZip,Data**και **Results**, όπου ο κάθε ένας έχει σημαντικό ρόλο στην επεξεργασία.

- **Copy:** Κρατάει αντίγραφο των αρχικών δεδομένων
- **Data:** Ο φάκελος που δέχεται τα δεδομένα από τον New, εκεί αποσυμπιέζονται τα συμπιεσμένα αρχεία και μετακινούνται στο φάκελο UnZip.
- **UnZip:** Κρατάει αντίγραφο από όλα τα αρχεία μετά την αποσυμπίεση
- **Results:** Εκεί παράγεται το πρώτο δείγμα αποτελεσμάτων έχοντας ως ονομασία το «Directory_Date_Time_Selection», ένα παράδειγμα μιας ονομασίας, η οποία σχετίζεται με την δική μας περίπτωση είναι Result_Sat_Mar_28_TIME_05_13_45_EET_2015_Selection. Το αρχείο αυτό εμπεριέχει ενοποιημένη όλη μα όλη την πληροφορία, ακόμα και τα «σκουπίδια» από τα δεδομένα που επεξεργαστήκαμε. Αν όλα κυλήσουν ομαλά

στο πρώτο κομμάτι της επεξεργασίας αυτό που περιμένουμε να δούμε είναι το εξής

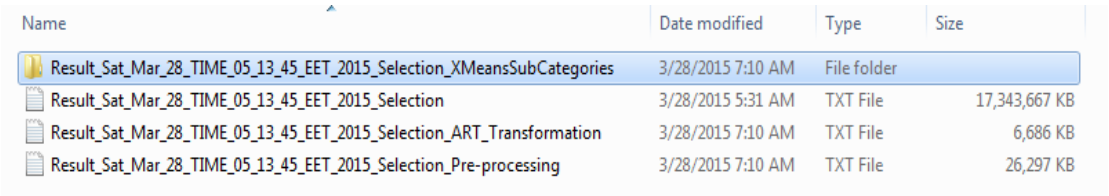


Φόρμα 7.2

Αφού τελειώσει το πρώτο μέρος που αποτελεί την διαδικασία της **Επιλογής της KDD**, εμφανίζεται στο χρήστη η παραπάνω φόρμα. Όπως διαφαίνεται το δεύτερο κουμπί έχει αποδεσμευθεί. Ο χρήστης μπορεί επίσης να διακρίνει τις πληροφορίες που εμφανίζονται σαν Logger στην κάτω περιοχή. Το δεύτερο κουμπί είναι εκείνο από όπου μπορεί ο χρήστης να επιλέξει το ενοποιημένο αρχείο, δηλαδή αυτό το οποίο έχει δημιουργηθεί στο πρώτο μέρος. Αφού τελειώσει η διαδικασία στο δεύτερο μέρος, στα αποτελέσματα που αναμένουμε είναι η αντιστοίχιση της κάθε IP και Ημερομηνία σε μια ξεχωριστή κατηγορία αντίστοιχα. Επίσης κάθε τέτοια οντότητα, στο πρόγραμμά μας δομείται σε ένα αντικείμενο τύπου ipAddressMapping και είναι μια εγγραφή στο ένα από τα παραγόμενα αρχεία. Η εγγραφή θα είναι ως έξης:

(IP>Date<>CategoryIp>CategoryDate)

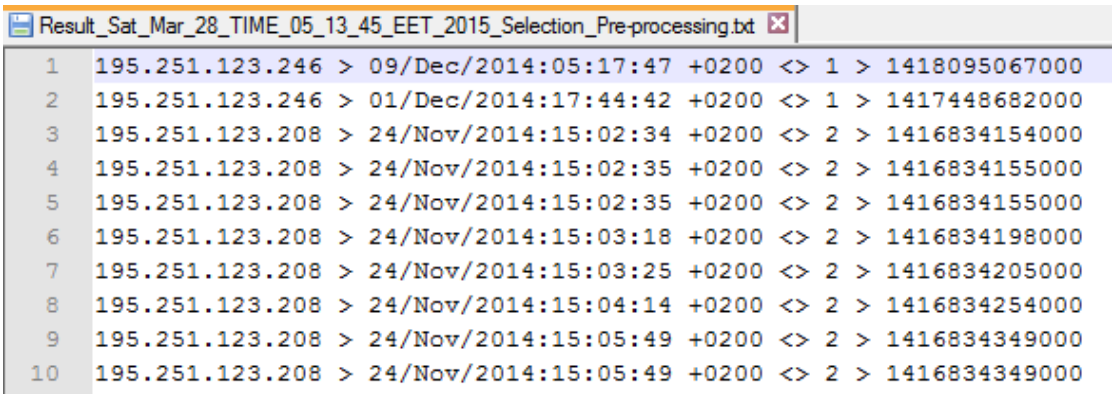
Φυσικά μια τέτοια εγγραφή εξυπηρετεί μόνο τον χρήστη και ως εκ τούτου θα πρέπει να υπάρχουν αρχεία που εξυπηρετούν τόσο τον αλγόριθμο ART, όσο και τον αλγόριθμο XMeans για την διαδικασία του Data Mining. Εν κατακλείδι παράγονται τα εξής αρχεία στο φάκελο Results:



Εικόνα 7.3

Το αρχείο με κατάληξη Pre-processing:

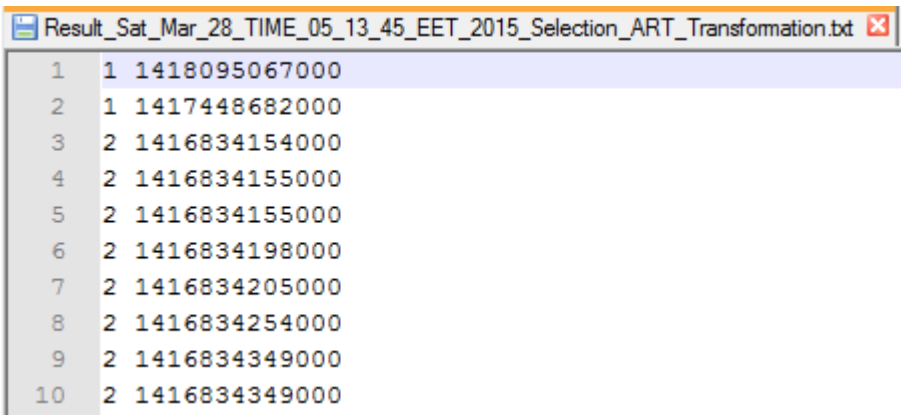
Είναι το αρχείο που περιέχει τις αντιστοιχήσεις (Result_Sat_Mar_28_TIME_05_13_45_EET_2015_Selection_Pre-processing)



Εικόνα 7.4

Το αρχείο με κατάληξη ART_Transformation:

Είναι το αρχείο που εξυπηρετεί τον αλγόριθμο ART (Result_Sat_Mar_28_TIME_05_13_45_EET_2015_Selection_ART_Transformation)



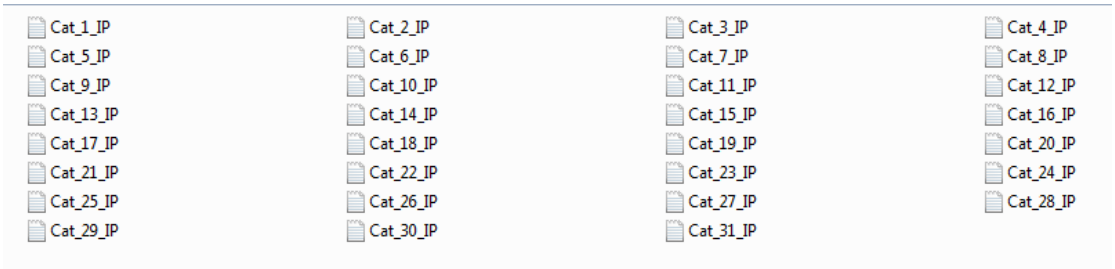
Εικόνα 7.5

Ο φάκελος με κατάληξη XMeansSubCategories:

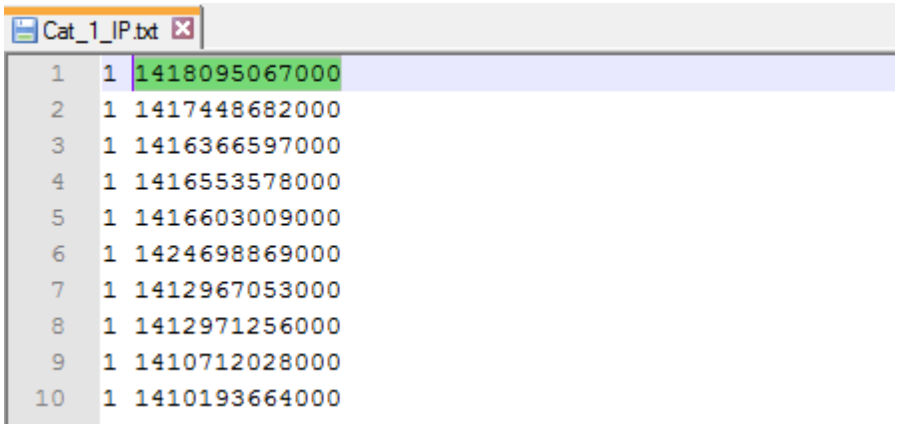
Είναι ο φάκελος με τα αρχεία που εξυπηρετεί τον αλγόριθμο XMeans, ο οποίος ενθυλακώνεται στο Density based clustering Algorithm και είναι ο γρηγορότερος από όλους [1]. Παρατηρούμε εσωτερικά του φακέλου πως τα δεδομένα έχουν χαρακτηριστικά ονόματα ανάλογα με την κατηγορία της κάθε IP.

Αυτό σημαίνει πως έχουμε ήδη ξεχωρίσει από τα ποικίλα δεδομένα την κίνηση σε κάθε εξυπηρετητή ξεχωριστά, φυσικά με βάση την κατηγορία της κάθε IP διεύθυνσης και μπορούμε να την αναλύσουμε ως προς τις επιθέσεις ασφαλείας.

(Cat_X_IP)

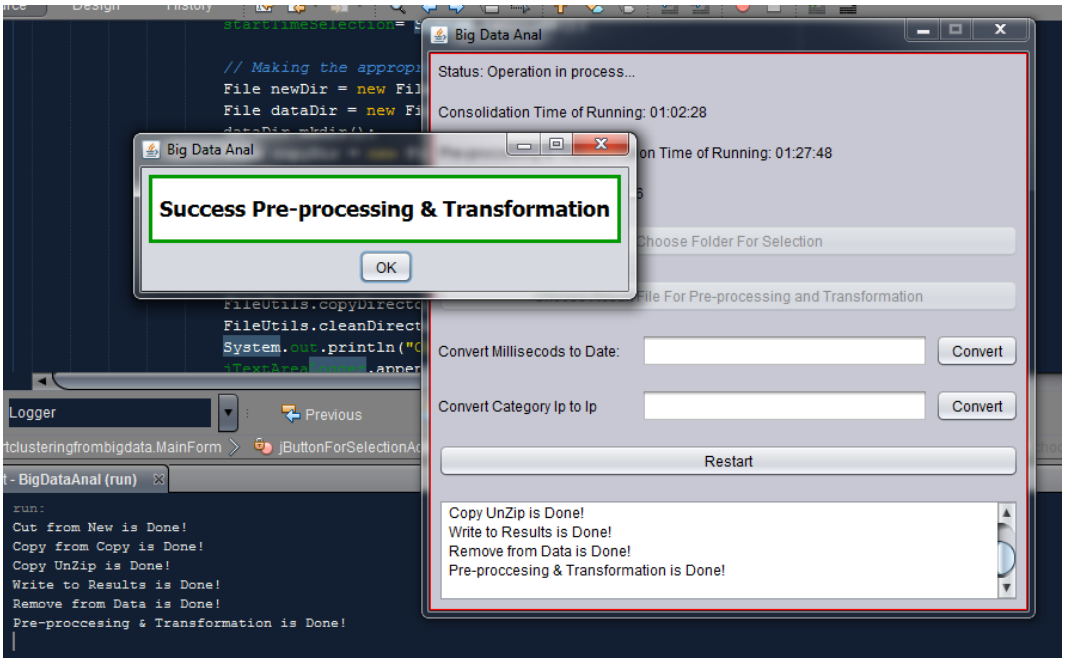


Εικόνα 7.6



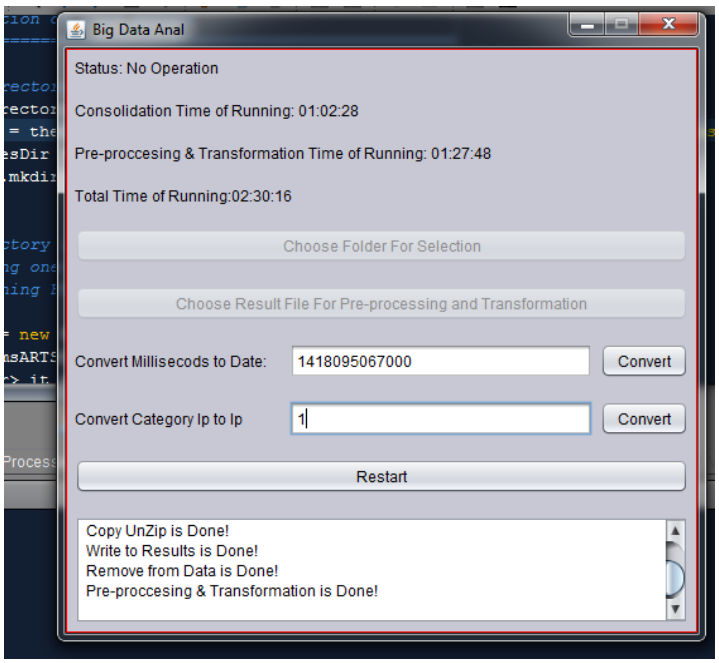
Εικόνα 7.7

Αν όλα κυλήσουν ομαλά και χωρίς σφάλματα στο δεύτερο κομμάτι της επεξεργασίας, που είναι το σημαντικότερο και το πιο ουσιαστικό σχετικά με το κομμάτι της εφαρμογής που χρησιμοποιούμε, αφού είναι το πιο βεβαρημένο, αυτό που περιμένουμε να δούμε είναι το εξής

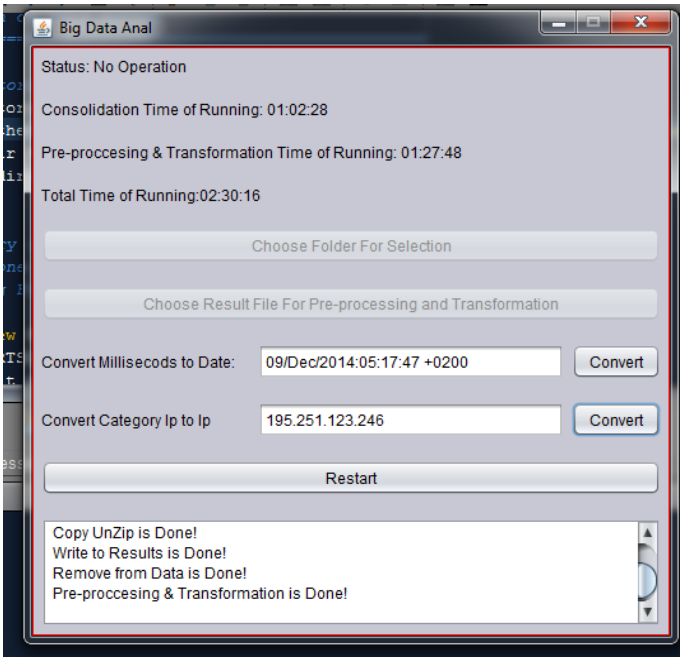


Φόρμα 7.3

Είναι σημαντικό να παρατηρήσουμε την δυνατότητα μετατροπής των κατηγοριών στις αντιστοιχήσεις τους. Αυτό μπορούμε να το εφαρμόσουμε τόσο για τις ημερομηνίες, όσο και για τις διευθύνσεις IP. Η μετατροπή ημερομηνιών παίζει σημαντικό ρόλο στο επόμενο στάδιο της διαδικασίας.



Φόρμα 7.4



Φόρμα 7.4

Αυτό αποτελεί το δεύτερο μέρος που αποτελεί την διαδικασία της **Προεπεξεργασία και του Μετασχηματισμού της KDD**

7.3 Επεξήγηση του κώδικα

Έχουμε επαναλάβει πως ένα από τα χαρακτηριστικά των BIG DATA είναι το τεράστιο μέγεθος των δεδομένων. Αυτό έχει άμεσο αντίκτυπο στον χρόνο που χρειάζεται ο κώδικας για την εκτέλεση των διαδικασιών, όπου ο χρόνος είναι η βασική παράμετρος που μας ενδιαφέρει.

Για να βελτιστοποιήσουμε λοιπόν το χρόνο μπορούμε να ακολουθήσουμε τεχνικές παραλληλισμού στο Hardware και στο Software και μάλιστα αυτό είναι επιτακτικό. Σχετικά με τους πόρους που μπορούμε να διαθέσουμε στο σύστημά μας θα λέγαμε ότι το Hadoop είναι η πλέον ενδεδειγμένη λύση. Κάτι που αποτελεί μελλοντική επέκταση της εφαρμογής αυτής. Όσο αναφορά τον παραλληλισμό στον κώδικα σε ένα τέτοιο πολύπλοκο σύστημα τα βήματα του κώδικα θα πρέπει να

αποτελούνται αποκλειστικά από έτοιμες βιβλιοθήκες αυξάνοντας έτσι στο μέγιστο την χρησιμότητα της ίδιας της γλώσσας αλλά και βελτιστοποιώντας τον κώδικα όσο το δυνατόν. Από αυτή τη σκοπιά θα πρέπει να μην υπάρχει δυνατότητα παραλληλισμού και βελτιστοποίησης του κώδικα εκτός μόνο ελάχιστων σημείων. Όμως είναι πολύ σημαντική η επιλογή της κάθε βιβλιοθήκης, των κλάσεων και των μεθόδων που θα χρησιμοποιηθούν στο κώδικα για την απόδοση του.

Χαρακτηριστικά παραδείγματα κλάσεων που έχουν χρησιμοποιηθεί βελτιώνοντας έως και δέκα φορές την χρονική απόδοση του συστήματος είναι RandomAccessFile για γρηγορότερη πρόσβαση στο αρχείο και η FileChannel που γράφει και διαβάζει τα δεδομένα χωρίς την χρήση Buffer με αποτέλεσμα η μνήμη να είναι πολύ περισσότερο αποδοτική και ο χρόνος να ελαττώνεται στο έπακρο. Υπάρχει όμως και ένα σημείο όπου ο όγκος αποτέλεσε σημαντικότερο παράγοντα από την χρονική απόδοση και αφορά την αποσυμπίεση.

Γίνεται η χρήση της βιβλιοθήκης Gzip που είναι βραδύτερη από της ταχύτατες Snappy και LZ4, αφού η Gzip παράγει το μικρότερο μέγεθος συμπίεσης κάτι που είναι μείζονος σημασίας στον τομέα των BIG DATA. Σκεφτείτε πως όλα τα αποτελέσματα μπορεί να συμπιέζονται και να αποθηκεύονται εκ νέου. Η ιδανικότερη λύση θα ήταν η υβριδική λύση χρησιμοποιώντας την Gzip για συμπίεση και την LZ4, που είναι ελαφρός ταχύτερη από την Snappy, για αποσυμπίεση.

7.4 WEKA

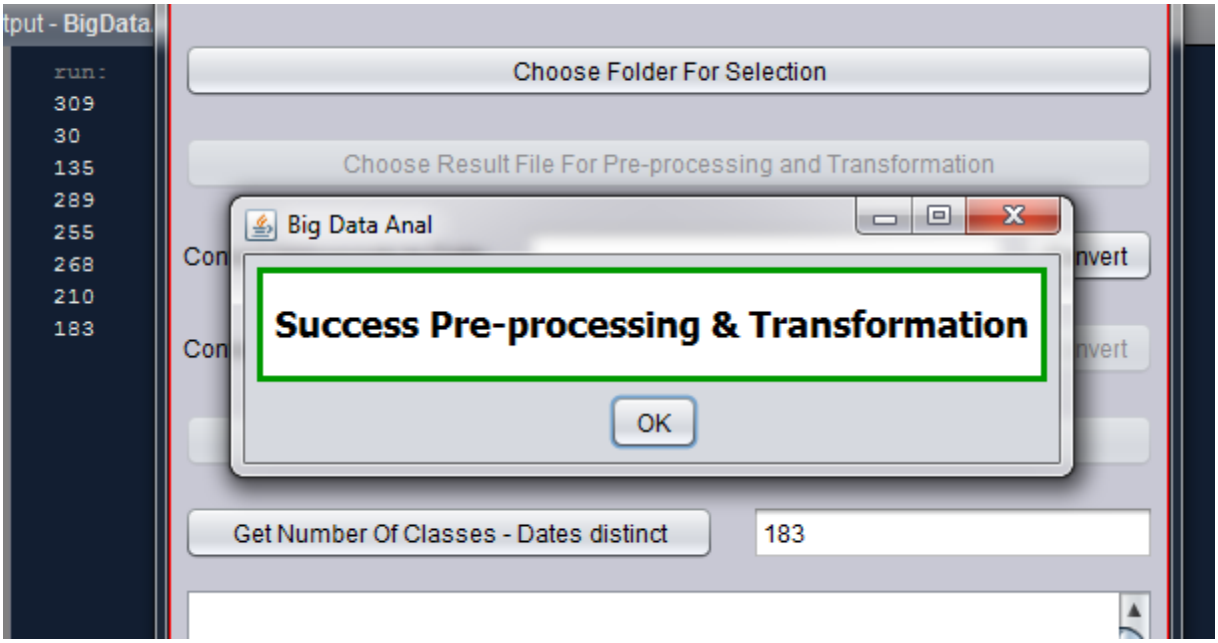
Στην συνέχεια της μελέτης μας κάνουμε την χρήση του εργαλείου WEKA για να προχωρήσουμε στην διαδικασία Data Mining με την χρήση του αλγορίθμου XMeans σε συνδυασμό με τον Density based clustering Algorithm που λαμβάνει υπόψη την πυκνότητα των δεδομένων και που είναι ο γρηγορότερος από του αλγορίθμους που μας διατίθενται [55].

Κάτι πολύ σημαντικό για την συγκεκριμένη ανάλυση είναι ότι τα δεδομένα δεν πρέπει να ομαδοποιηθούν σε μεγάλες ομάδες. Αφού η κάθε εγγραφή αντιπροσωπεύει

μια ημερομηνία και συγκεκριμένα έχει εύρος 15 λεπτών, η ομαδοποίηση πρέπει γίνεται σε ένα εύρος 2 ωρών που σχετίζεται με την κίνηση στο δίκτυο του κάθε εξυπηρετητή. Οι διαφορετικές IP που έχουν βρεθεί είναι 31, άρα είναι 31 τα ξεχωριστά αρχεία που πρέπει να μελετηθούν.

Για το κάθε αρχείο όμως παρατηρούμε αρχικά το βαθμό της σημαντικότητάς του. Μόνο τα 8 από τα 31 αρχεία έχουν σημαντικό αριθμό εγγραφών έστω άνω των χιλίων και μόνο σε αυτά το συμπέρασμα μπορεί να έχει σημαντική αξία και βάση.

Επίσης ο αλγόριθμος που χρησιμοποιούμε πρέπει να γνωρίζει τον αριθμό των κλάσεων που ομαδοποιεί και αυτό το ορίζουμε από τις ξεχωριστές κατηγορίες ημερομηνιών που βρίσκουμε στα αρχεία μας. Επανερχόμαστε λοιπόν στην εφαρμογή που αναπτύξαμε που υπολογίζει τον αριθμό τις κάθε κλάσης όπως φαίνεται παρακάτω



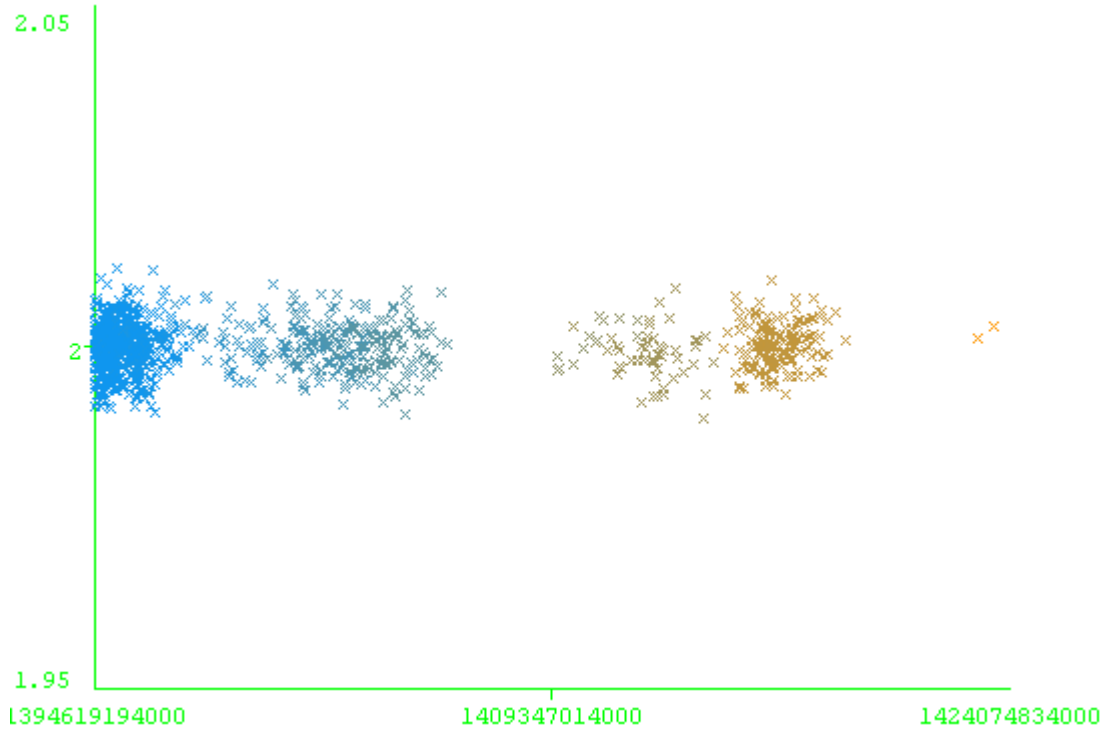
Φόρμα 7.5

Τα συγκεντρωτικά αποτελέσματα των αρχείων για ανάλυση φαίνονται στον παρακάτω πίνακα

Αρχείο	Κατηγορία IP	Αριθμός διαφορετικών Ημερομηνιών	Αριθμός κλάσεων στο WEKA
Cat_1_IP	1	309	309
Cat_2_IP	2	30	30
Cat_3_IP	3	135	135
Cat_7_IP	7	289	289
Cat_17_IP	17	255	255
Cat_18_IP	18	268	268
Cat_24_IP	24	210	210
Cat_25_IP	25	183	183

Πίνακας 7.1

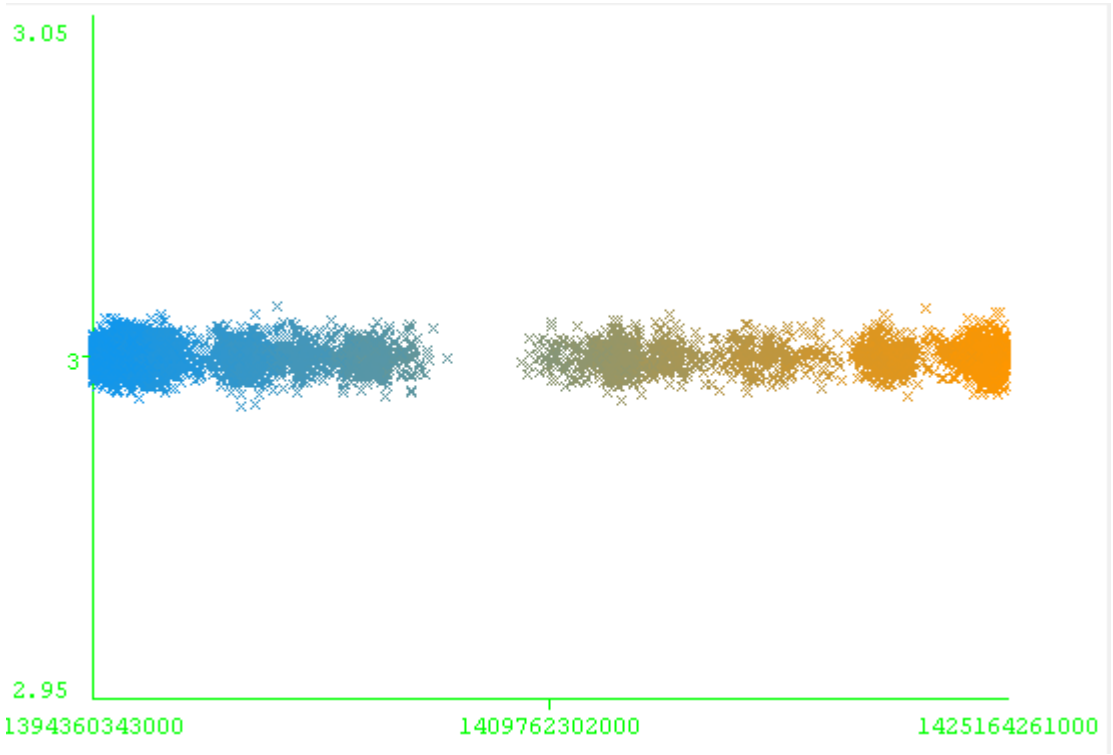
Παρακάτω παρατίθενται τα 2 από τα οκτώ πειράματα που παρατηρείται μεγάλη διακύμανση σε συγκεκριμένο εύρος τιμών και μπορεί να υπάρξει υποψία επιθέσεων στην διαθεσιμότητα (Ddos Attacks).



Εικόνα 7.8

Cluster centers : 30 centers			Clustered Instances		
Cluster 0	2.0	1.400573757375E12	0	8	(1%)
Cluster 1	2.0	1.3958295435652173E12	1	23	(2%)
Cluster 2	2.0	1.401184569E12	2	2	(0%)
Cluster 3	2.0	1.3959217990136987E12	3	73	(7%)
Cluster 4	2.0	1.3959232425714285E12	4	7	(1%)
Cluster 5	2.0	1.3959260205625E12	5	16	(2%)
Cluster 6	2.0	1.4109513113333333E12	6	12	(1%)
Cluster 7	2.0	1.3947188854836064E12	7	129	(13%)
Cluster 8	2.0	1.3947157796153845E12	8	13	(1%)
Cluster 9	2.0	1.3958298294285715E12	9	7	(1%)
Cluster 10	2.0	1.3990239017777778E12	10	18	(2%)
Cluster 11	2.0	1.3965031944E12	11	6	(1%)
Cluster 12	**2.0**	**1.4165539534705881E12**	**12**	**153**	**(15%)**
Cluster 13	2.0	1.4036691396296296E12	13	27	(3%)
Cluster 14	2.0	1.4127661642857144E12	14	50	(5%)
Cluster 15	2.0	1.4018870271481482E12	15	27	(3%)
Cluster 16	2.0	1.40169525975E12	16	28	(3%)
Cluster 17	2.0	1.4023901556444443E12	17	45	(4%)
Cluster 18	2.0	1.3998957547916667E12	18	24	(2%)
Cluster 19	2.0	1.404285280171875E12	19	64	(6%)
Cluster 20	2.0	1.395990198E12	21	1	(0%)
Cluster 21	2.0	1.404293748E12	22	51	(5%)
Cluster 22	2.0	1.3946210072941177E12	24	81	(8%)
Cluster 23	0.0	0.0	25	2	(0%)
Cluster 24	2.0	1.3952241938024692E12	26	22	(2%)
Cluster 25	2.0	1.424074832E12	27	27	(3%)
Cluster 26	2.0	1.4026572865E12	29	89	(9%)
Cluster 27	2.0	1.394708155E12			
Cluster 28	2.0	1.412771231E12			
Cluster 29	2.0	1.39471775428125E12			

Πίνακας 7.2



Εικόνα 7.9

Cluster centers	: 135 centers		Clustered Instances		
Cluster 0	3.0	1.4173159877972974E12	0	76	(2%)
Cluster 1	3.0	1.425164257E12	1	5	(0%)
Cluster 2	3.0	1.3955981903409092E12	2	88	(2%)
Cluster 3	3.0	1.4011218939090908E12	3	11	(0%)
Cluster 4	3.0	1.403689167E12	4	3	(0%)
Cluster 5	3.0	1.394400962081081E12	5	37	(1%)
Cluster 6	3.0	1.4110365520967742E12	6	62	(1%)
Cluster 7	3.0	1.4029995182727273E12	7	11	(0%)
Cluster 8	3.0	1.3993562615333333E12	8	90	(2%)
Cluster 9	3.0	1.4247780735945945E12	9	37	(1%)
Cluster 10	3.0	1.424987999E12	10	10	(0%)
Cluster 11	3.0	1.3993549286E12	11	10	(0%)
Cluster 12	3.0	1.401112947E12	12	8	(0%)
Cluster 13	3.0	1.3953801035625E12	13	96	(2%)

Cluster 14	3.0	1.4214633913076924E12	14	65	(1%)
Cluster 15	3.0	1.41179117078E12	15	57	(1%)
Cluster 16	3.0	1.3951608743913044E12	16	23	(0%)
Cluster 17	3.0	1.395645980755102E12	17	49	(1%)
Cluster 18	3.0	1.3986818296E12	18	5	(0%)
Cluster 19	3.0	1.405028072568182E12	19	44	(1%)
Cluster 20	3.0	1.396491609E12	20	13	(0%)
Cluster 21	3.0	1.4037495611428572E12	21	7	(0%)
Cluster 22	3.0	1.3971160711590908E12	22	44	(1%)
Cluster 23	3.0	1.4243893694444443E12	23	24	(0%)
Cluster 24	3.0	1.3948143165333333E12	24	45	(1%)
Cluster 25	3.0	1.404077291875E12	25	8	(0%)
Cluster 26	3.0	1.4210045884910715E12	26	112	(2%)
Cluster 27	3.0	1.4169145511147542E12	27	59	(1%)
Cluster 28	3.0	1.3951631139166667E12	28	23	(0%)
Cluster 29	3.0	1.4238697999615386E12	29	26	(1%)
Cluster 30	3.0	1.4029950765833333E12	30	12	(0%)
Cluster 31	3.0	1.424440871888889E12	31	9	(0%)
Cluster 32	3.0	1.4156934111923076E12	32	26	(1%)
Cluster 33	3.0	1.3993499775384614E12	33	13	(0%)
Cluster 34	3.0	1.39944911E12	34	1	(0%)
Cluster 35	3.0	1.400628359809524E12	35	42	(1%)
Cluster 36	3.0	1.401111924E12	37	14	(0%)
Cluster 37	3.0	1.3958535088E12	38	28	(1%)
Cluster 38	3.0	1.4230134148928572E12	39	71	(1%)
Cluster 39	3.0	1.424253884802817E12	40	40	(1%)
Cluster 40	**3.0**	**1.3964349239512195E12**	**41**	**158**	**(3%)**
Cluster 41	3.0	1.401318438700637E12	42	1	(0%)
Cluster 42	3.0	1.399753408E12	43	57	(1%)
Cluster 43	3.0	1.3987493193684211E12	44	35	(1%)
Cluster 44	3.0	1.4121657180571428E12	45	23	(0%)
Cluster 45	3.0	1.4035358462608696E12	46	10	(0%)
Cluster 46	3.0	1.3952516161E12	47	8	(0%)
Cluster 47	3.0	1.424381028375E12	49	6	(0%)
Cluster 48	0.0	0.0	50	31	(1%)
Cluster 49	3.0	1.4036505198333333E12	51	24	(0%)

Cluster 50	3.0	1.3993574451290322E12	52	35	(1%)
Cluster 51	3.0	1.4019994482083333E12	53	71	(1%)
Cluster 52	3.0	1.3968524003428572E12	54	37	(1%)
Cluster 53	3.0	1.3969365900140845E12	55	14	(0%)
Cluster 54	3.0	1.3946241679210527E12	56	26	(1%)
Cluster 55	3.0	1.3968702245714285E12	57	34	(1%)
Cluster 56	3.0	1.3982626986538462E12	58	11	(0%)
Cluster 57	3.0	1.4243726226764707E12	59	12	(0%)
Cluster 58	3.0	1.3996189432727273E12	60	27	(1%)
Cluster 59	3.0	1.3996187686666667E12	61	29	(1%)
Cluster 60	3.0	1.412253551925926E12	62	23	(3%)
Cluster 61	3.0	1.3996181768275862E12	63	18	(0%)
Cluster 62	3.0	1.3955951647258064E12	64	48	(1%)
Cluster 63	3.0	1.3996191970555557E12	65	20	(0%)
Cluster 64	3.0	1.395162445125E12	66	4	(0%)
Cluster 65	3.0	1.421667862E12	67	30	(1%)
Cluster 66	3.0	1.40371314975E12	68	16	(0%)
Cluster 67	3.0	1.4118832962432432E12	69	11	(0%)
Cluster 68	3.0	1.399618027125E12	70	20	(0%)
Cluster 69	3.0	1.396950559E12	71	17	(0%)
Cluster 70	3.0	1.41458354705E12	72	58	(1%)
Cluster 71	3.0	1.3954370118235293E12	73	11	(0%)
Cluster 72	3.0	1.4023915746896553E12	74	12	(0%)
Cluster 73	3.0	1.420895842181818E12	75	12	(0%)
Cluster 74	3.0	1.3993931754166667E12	76	98	(2%)
Cluster 75	3.0	1.4127963985833333E12	77	16	(0%)
Cluster 76	3.0	1.4244166715714285E12	78	74	(2%)
Cluster 77	3.0	1.3993499408125E12	79	20	(0%)
Cluster 78	3.0	1.3999580971756758E12	80	22	(0%)
Cluster 79	3.0	1.4212271359E12	81	45	(1%)
Cluster 80	3.0	1.3996186343913044E12	82	5	(0%)
Cluster 81	3.0	1.396853008488889E12	83	21	(0%)
Cluster 82	3.0	1.3988501284E12	84	14	(0%)
Cluster 83	**3.0**	**1.39584335915E12**	**85**	**160**	**(3%)**
Cluster 84	3.0	1.3996190049285715E12	86	17	(0%)
Cluster 85	3.0	1.39632008866875E12	87	5	(0%)

Cluster 86	3.0	1.4122901940588235E12	88	35	(1%)
Cluster 87	3.0	1.424210299E12	89	18	(0%)
Cluster 88	3.0	1.3959685208571428E12	90	2	(0%)
Cluster 89	3.0	1.3968526597E12	91	55	(1%)
Cluster 90	3.0	1.3964171535E12	92	32	(1%)
Cluster 91	3.0	1.4133066921636365E12	93	15	(0%)
Cluster 92	3.0	1.404458137E12	94	27	(1%)
Cluster 93	3.0	1.3946326014666667E12	95	56	(1%)
Cluster 94	3.0	1.3946319308518518E12	96	6	(0%)
Cluster 95	3.0	1.3947212880714285E12	97	94	(2%)
Cluster 96	3.0	1.3957043251666667E12	98	5	(0%)
Cluster 97	3.0	1.4120853879574468E12	99	9	(0%)
Cluster 98	3.0	1.403858631E12	100	100	(2%)
Cluster 99	3.0	1.3996188796666667E12	101	28	(1%)
Cluster 100	3.0	1.4138700697E12	102	26	(1%)
Cluster 101	3.0	1.3950301603928572E12	103	41	(1%)
Cluster 10 2	3.0	1.39961850032E12	104	20	(0%)
Cluster 103	3.0	1.3968525407435898E12	105	3	(0%)
Cluster 104	3.0	1.3949092947E12	106	21	(0%)
Cluster 105	3.0	1.3946925953333333E12	107	85	(2%)
Cluster 106	3.0	1.3996190874761904E12	108	51	(1%)
Cluster 107	3.0	1.4036828981294119E12	109	35	(1%)
Cluster 108	3.0	1.4186092973333333E12	110	19	(0%)
Cluster 109	3.0	1.3951281429428572E12	111	53	(1%)
Cluster 110	3.0	1.4243920715789473E12	112	19	(0%)
Cluster 111	3.0	1.4251337846037737E12	113	66	(1%)
Cluster 112	3.0	1.3995329342105264E12	114	136	(3%)
Cluster 113	3.0	1.4243907294285715E12	115	58	(1%)
Cluster 114	3.0	1.3955936532E12	116	1	(0%)
Cluster 115	3.0	1.4100100381724138E12	117	72	(1%)
Cluster 116	3.0	1.395525461E12	118	68	(1%)
Cluster 117	3.0	1.4243878074166667E12	119	13	(0%)
Cluster 118	3.0	1.403689748030303E12	120	45	(1%)
Cluster 119	3.0	1.4231037505384614E12	121	40	(1%)
Cluster 120	3.0	1.4164291451555557E12	122	44	(1%)
Cluster 121	3.0	1.4206242787E12	123	22	(0%)

Cluster 122	3.0	1.403688668068182E12	124	41	(1%)
Cluster 123	3.0	1.411475282818182E12	125	15	(0%)
Cluster 124	3.0	1.395164507975E12	126	11	(0%)
Cluster 125	3.0	1.3969393076666667E12	128	10	(0%)
Cluster 126	**3.0**	**1.401174045E12**	**129**	**181**	**(4%)**
Cluster 127	0.0	0.0	130	31	(1%)
Cluster 128	3.0	1.4028154253E12	131	61	(1%)
Cluster 129	3.0	1.425151424917127E12	132	78	(2%)
Cluster 130	3.0	1.3967984248064517E12	133	25	(1%)
Cluster 131	3.0	1.4210559580819673E12	134	9	(0%)
Cluster 132	3.0	1.3945499133246753E12			
Cluster 133	3.0	1.39512529248E12			
Cluster 134	3.0	1.3966300473333333E12			

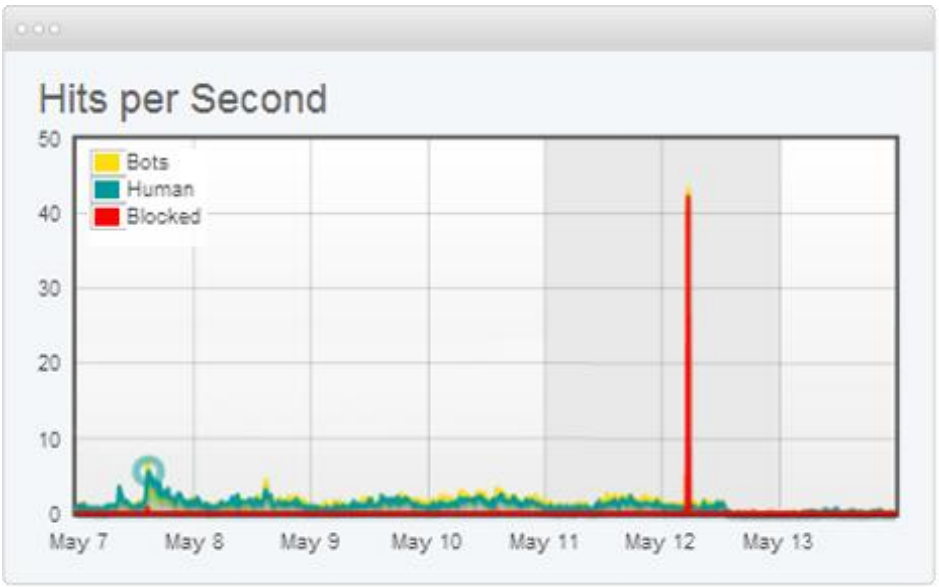
Πίνακας 7.3

Παρατηρείται στις άνω σκιασμένες περιπτώσεις ότι έχουμε πάνω από 150 αλληλεπιδράσεις συνεχόμενες με το εξυπηρετητή σε διάστημα 2 ωρών που σημαίνει άνω των 2,5 αλληλεπιδράσεων το λεπτό και συνεχόμενα. Κάνοντας τον έλεγχο των ημερομηνιών στο παρακάτω πίνακα παρατηρούμε ότι οι ημερομηνίες δεν αναφέρονται σε περιόδους μεγάλης αναμενόμενης κίνησης.

ΕΛΕΓΧΟΣ ΗΜΕΡΟΜΗΝΙΩΝ	
1.4165539534705881E12	Παρασκευή/21/Νοεμβρίου2014 - **2:12**
1.3964349239512195E12	Τετάρτη/2/Απριλίου2014 - **6:35**
1.39584335915E12	Τετάρτη/26/Μαρτίου2014 - **10:15**
1.401174045E12	Τρίτη/27/Μαΐου2014 - **03:00**

Πίνακας 7.4

Αυτό όμως δεν συνιστά απαραίτητα επίθεση, αλλά συνιστά περαιτέρω μελέτη. Είναι μια πρώτη ένδειξη για να δούμε των αριθμό των πακέτων αλλά και το μέγεθος τους για να έχουμε την δυνατότητα να βγάλουμε σαφέστερα συμπεράσματα.



Εικόνα 7.10 Μια πραγματική Επίθεση [56]

8 Συμπεράσματα και Μελλοντικές Επεκτάσεις

8.1 Συμπεράσματα

Η συγκεκριμένη πτυχιακή εργασία είχε ως στόχο την μελέτη διαφόρων τεχνολογιών, που η ένωσή τους αποτελεί τον βασικό πυλώνα του τομέα της Πληροφορικής σήμερα.

Αναφέρθηκαν κάποιες βασικές έννοιες οι οποίες αφορούν την γνώση που μπορεί να παραχθεί από τα δεδομένα. Είδαμε τα στάδια της KDD, ενώ αναφέρθηκε ότι το Data Mining είναι ένα από αυτά τα στάδια. Ο στόχος δεν ήταν ούτε η μελέτη των Βάσεων Δεδομένων, ούτε των μοντέλων της Μηχανικής Μάθησης και των αλγορίθμων, ούτε του Παραλληλισμού στα δεδομένα, αλλά πώς αυτές οι τρεις τεχνολογίες μπορούν να συνδυαστούν και να δώσουν λύσεις σε καίρια ζητήματα της σημερινής εποχής, όπως είναι η Ασφάλεια.

Οι προαναφερθείσες τεχνολογίες μπορούν να χρησιμοποιηθούν και να δώσουν λύσεις και σε πολλούς άλλους τομείς εκτός από την Ασφάλεια, με επιθυμητό σκοπό να αυξήσουν τα οικονομικά οφέλη των οργανισμών, των εταιριών αλλά και των ανθρώπων, την αποδοτικότητα, την αποτελεσματικότητα σε όποιο πεδίο ενσωματωθεί η εφαρμογή τους.

Στον τομέα της Ασφάλεια οι τακτικές που προαναφέρθηκαν αποτελούν το βασικό εργαλείο, και μπορούν να γίνουν πλήρως αποδοτικές. Τόσο σε επίπεδο θεωρητικό όσο και σε πρακτικό ο συνδυασμός των διαδικασιών δίνει αποτελέσματα και γνώση.

Στον αντίποδα υπάρχουν και τα μειονεκτήματα του συνδυασμού αυτών των τεχνολογιών που μπορούν να δημιουργήσουν ένα περιβάλλον που θα τίθεται σοβαρότατο ζήτημα σχετικά με την αποκάλυψη, των προσωπικών μας πληροφοριών παρά τη θέλησή μας.

Σε αυτήν την εργασία έγινε μια προσπάθεια να περιγραφούν οι επιθέσεις τύπου Ddos με σκοπό να στοχοποιηθούν για περαιτέρω παρακολούθηση.

Δημιουργήθηκε μια εφαρμογή JAVA η οποία διαχειρίζεται τα BIG DATA και προχωρεί σε όλη την διαδικασία εξόρυξης γνώσης βήμα προς βήμα ακολουθώντας την διαδικασία της KDD, ακριβώς έτσι όπως μελετήθηκε. Χρησιμοποιήθηκαν οι βέλτιστες βιβλιοθήκες κώδικα για τέτοιες διαδικασίες σχετικές με τα BIG DATA, όπως η FileChannel και η RandomAccessFile. Τα μη προσωποποιημένα δεδομένα που μας δόθηκαν τέθηκαν υπό επεξεργασία από την εφαρμογή που αναπτύξαμε. Σαν αποτέλεσμα μπορέσαμε να ανακαλύψουμε τα δεδομένα των οποίων η επεξεργασία έχει αξία με πλήρη αρχειοθέτηση για κάθε συγκεκριμένο εξυπηρετητή(destination IP). Παράχθηκαν δηλαδή δεδομένα όπου κάθε μια destination IP αντιστοιχεί σε ξεχωριστό αρχείο που αποθηκεύει τις ημερομηνίες της κίνησης του δικτύου.

Σε ένα δεύτερο στάδιο καταφέρνουμε να τα χρησιμοποιήσουμε στο τεχνολογικό περιβάλλον WEKA, να τα *ομαδοποιήσουμε*(ημερομηνίες εύρους 2 ωρών) και να βγάλουμε ουσιαστικά συμπεράσματα. Χρησιμοποιήθηκαν οι αλγόριθμοι Xmeans και Density based clustering Algorithm που είναι πιο αποδοτικοί από άποψη χρόνου, και ταιριάζουν σε καταστάσεις απαιτήσεων γρήγορης εκτέλεσης.

Ερευνήθηκε διεξοδικά ο καταλληλότερος αλγόριθμος για την συγκεκριμένη διαδικασία, ο αλγόριθμος ART, και συγκεκριμένα οι ART-2 και ART-3, που όχι μόνο είναι ακόμα πιο γρήγοροι αλλά και αποδοτικότεροι και κυρίως κλιμακωτοί και ταιριάζουν εξαιρετικά στο τελικό προϊόν που θα παραχθεί.

Πραγματοποιήθηκε το 100% των πειραμάτων αλλά παρατέθηκε το 25% λόγο της μεγάλης έκτασης του κάθε αποτελέσματος, Μας δόθηκαν 10 GB που **προσομοιώνουν** την διαδικασία διότι δεν υπήρχε η δυνατότητα στο πρακτικό μέρος της χρήσης του εργαλείου Hadoop που αποτελεί αναπόσπαστο κομμάτι της διαδικασίας και θα προστεθεί στην συνέχεια. Το Hadoop θα έδινε την δυνατότητα τις μείωσης του χρόνου του κάθε πειράματος κατά ένα πολύ μεγάλο ποσοστό.

Βρέθηκαν 4 πιθανές περιπτώσεις με μη ομαλή συμπεριφορά που χρήζουν περαιτέρω εξερεύνηση για το αν συντελούν επίθεση!

8.2 Μελλοντικές Επεκτάσεις

Για να μπορεί να είναι χρήσιμο ένα τέτοιο εργαλείο θα πρέπει να απαιτεί από τον χρήση την λιγότερο δυνατή συμμετοχή σε αυτές τις τεχνικές ανάλυσης, αλλά να συμμετέχει μόνο στην ερμηνεία. Για να γίνει κάτι τέτοιο θα πρέπει να υπάρξει ένας ενιαίος σχεδιασμός και μια υλοποίηση ενός λογισμικού το όποιο αυτόματα θα είναι σε θέση από τα αρχικά δεδομένα να δημιουργεί τα *περιγραφικά μοντέλα* πρόγνωσης ώστε να ερμηνευτούν τα αποτελέσματά τους.

Όσο αφορά τα χαρακτηριστικά του εφαρμογής που υλοποιήσαμε θα πρέπει να επεκταθεί τόσο από άποψη λογισμικού, όσο και από άποψη υλικού και πόρων ώστε να αντιμετωπιστούν και να μηδενιστούν κάποιες παραδοχές της εφαρμογή μας:

1. Αρχικά για την πρώτη διαδικασία της KDD που αφορά την συλλογή των δεδομένων έγινε η παραδοχή πως τα δεδομένα που δεχόμαστε από τους εξυπηρετητές είναι δεδομένα από λειτουργικό σύστημα Linux. Αυτό αφορά περισσότερο τα συμπιεσμένα αρχεία: **Υλοποίηση λογισμικού για αρχεία κάθε τύπου.(Το υπάρχων λογισμικό – κώδικας παραθέτετε στο ΠΑΡΑΡΤΗΜΑ Ι)**

2. Το σύστημα τρέχει σε χρόνο στατικό είναι όμως σχεδιασμένο για την προσαρμογή του σε Realtime εφαρμογές, κάτι το οποίο ισχύει για όλη την διαδικασία της KDD: **Αλγόριθμος ART, είναι ο καταλληλότερος για τον σύστημα λόγο χρόνου και δυνατότητας κλιμάκωσης – πλήρης και ολοκληρωμένη επεξήγηση του αλγορίθμου στο τέλος του κεφαλαίου)**

3. Τα αποτελέσματα είναι υποθετικά, βασισμένα φυσικά σε επιστημονική γνώση και η μελέτη τους έχει ως βασικό στόχο την απόδειξη της πρότασής μας και

την εφαρμοσιμότητά της. **Ως επέκταση της συγκεκριμένης έρευνας αποτελεί η επιβεβαίωση από κάποιο σύστημα IDS ή SIEM.**

4. Μπορεί και πρέπει να ενσωματωθεί χρήση του παραλληλισμού τόσο σε επίπεδο Software όσο και σε επίπεδο Hardware για την βελτίωση της απόδοσης του συστήματος: **Βελτιστοποίηση κώδικα και Hadoop.**

8.3 ART

Το μοντέλο ART βασίζεται στην λεγόμενη Θεωρία του Αυτό-προσανατολιζόμενου Συντονισμού (Adaptive Resonance Theory) που αναπτύχθηκε από τον Grossberg. Η θεωρία αυτή ασχολείται με το ερώτημα πώς τα πραγματικά νευρωνικά συστήματα μπορούν να διαθέτουν ταυτόχρονα πλαστικότητα, δηλαδή να μαθαίνουν προσαρμοζόμενα στα νέα και σημαντικά ερεθίσματα από το περιβάλλον, αλλά και σταθερότητα, δηλαδή να τροποποιούνται διαρκώς πέφτοντας σε κατάσταση αστάθειας όταν δέχονται άσχετα ή συχνά επαναλαμβανόμενα ερεθίσματα. Ένα επαρκές νευρωνικό σύστημα πρέπει να είναι ικανό να δημιουργεί νέες καταστάσεις ισορροπίας ή νέα σημεία έλξης ώστε να απομνημονεύει καινούργια πρότυπα (πλαστικότητα) αποφεύγοντας ωστόσο την ακατάπαυστη επανακωδικοποίηση των καταστάσεων ισορροπίας (σταθερότητα).

Ο Grossberg είδε τον συμβιβασμό μεταξύ πλαστικότητας και σταθερότητας σαν ένα δίλημμα. Στα πιο διαδεδομένα τεχνητά νευρωνικά δίκτυα με επίβλεψη , όπως το MLP, δεν υπάρχει ξεκάθαρη πολιτική σταθερότητας, αφού όλα τα συναπτικά βάρη είναι υποψήφια για τροποποίηση κάθε χρονική στιγμή.

Οι Carpenter και Grossberg ανέπτυξαν τα δίκτυα ART-1, ART-2 και ART-3 για να αντιμετωπίσουν το πρόβλημα. Θεωρούν ότι η κάθε κλάση αντιπροσωπεύεται από το διάνυσμα των συναπτικών βαρών ενός νευρώνα. Το διάνυσμα αυτό καλείται αρχέτυπο για αυτήν την κλάση. Κάθε διάνυσμα εισόδου γίνεται αποδεκτό σε μια κλάση εφόσον μοιάζει αρκετά με το αρχέτυπο της κλάσης. Στην περίπτωση αυτή λέμε ότι το διάνυσμα εισόδου συντονίζεται με την κλάση. Το αρχέτυπο τροποποιείται

κατάλληλα κάθε φορά που ένα νέο πρότυπο εισέρχεται στην κλάση έτσι ώστε να παραμένει πάντα αντιπροσωπευτικό της κλάσης αυτής.

Αν το διάνυσμα εισόδου δεν συντονίζεται με καμία κλάση τότε δημιουργεί μια καινούργια κλάση από μόνο του και ταυτόχρονα αποτελεί το αρχέτυπο της. Η κλάση αυτή στο μέλλον μπορεί να προσελκύσει και άλλα πρότυπα. Έτσι υλοποιείται η έννοια της σταθερότητας , αφού πρότυπα που είναι άσχετα με κάποια κλάση δεν θα επηρεάσουν το αρχέτυπο της κλάσης αυτής, αλλά και η έννοια της πλαστικότητας αφού το αρχέτυπο κάθε κλάσης προσανατολίζεται ανάλογα με τα πρότυπα που ταιριάζουν στην κλάση.

Το βασικό μοντέλο ART είναι ένα σύστημα μάθησης χωρίς επίβλεψη. Τυπικά αποτελείται δύο επίπεδα, το επίπεδο σύγκρισης και το επίπεδο αναγνώρισης τα οποία αποτελούνται από ένα πλήθος Ν ενεργών νευρώνων, που αντιπροσωπεύουν τους υποψήφιους νευρώνες για το ρόλο του αρχέτυπου κάποιας κλάσης. Στην αρχή οι νευρώνες δεν είναι δεσμευμένοι από κάποια κλάση. Υπάρχει επίσης μία μονάδα επαναφοράς και μία μεταβλητή που ονομάζεται μεταβλητή επαγρύπνησης που είναι ουσιαστικά το κατώφλι αναγνώρισης που χρησιμοποιείται από το επίπεδο αναγνώρισης.

Το επίπεδο σύγκρισης παίρνει σαν είσοδο ένα διάνυσμα και το αποθέτει στο καλύτερο ταίριασμα του επιπέδου αναγνώρισης. Το καλύτερο ταίριασμα είναι εκείνος ο νευρώνας του οποίου το σύνολο των βαρών είναι το πιο κοντινό στις τιμές του διανύσματος εισόδου.

Κάθε νευρώνας του επιπέδου αναγνώρισης παράγει μία τιμή, ανάλογη της ποιότητας του ταιριάσματος με το διάνυσμα εισόδου. Αυτό γίνεται με τον κάθε νευρώνα του επιπέδου αναγνώρισης. Κατά αυτή την άποψη το επίπεδο αναγνώρισης επιτρέπει κάθε νευρώνα να αναπαριστά μια κατηγορία, κλάση, στην οποία το κάθε διάνυσμα εισόδου μπορεί να ομαδοποιείται. Το διάνυσμα εισόδου που ταίριαξε με κάθε νευρώνα έπειτα συγκρίνεται με την παράμετρο επαγρύπνησης.

Αν η παράμετρος επαγρύπνησης έχει μικρότερη τιμή από την τιμή ταιριάσματος των διανυσμάτων τότε το διάνυσμα εισόδου συντονίζεται με την κλάση

στην οποία πλέον αντιστοιχεί ο νευρώνας και αρχίζει να εφαρμόζεται η προσαρμογή στα βάρη του νευρώνα αναγνώρισης με τον οποίο έγινε το κατάλληλο ταίριασμα. Αυτά τα βάρη του νευρώνα είναι που θα προσαρμοστούν ως προς τα χαρακτηριστικά του διανύσματος εισόδου(masking).

Αντίθετα, αν η τιμή εξόδου είναι κατώτερη της παραμέτρου επαγρύπνησης τότε ο νευρώνας αναστέλλεται και μια διαδικασία αναζήτησης λαμβάνει χώρα.

Σε αυτή την διαδικασία αναζήτησης οι νευρώνες αναγνώρισης απενεργοποιούνται ένας προς ένας από την διαδικασία εάν δεν κάνουν masking έως ότου η τιμή της παραμέτρου επαγρύπνησης γίνει μικρότερη από μια τιμή ταιριάσματος των διανυσμάτων εισόδου και νευρώνα.

Πιο συγκεκριμένα σε κάθε κύκλο της διαδικασίας αναζήτησης ο κάθε νευρώνας του επιπέδου αναγνώρισης επιλέγεται και έπειτα απενεργοποιείται ή όχι. Αν δεν υπάρξει ταίριασμα με κάποιο νευρώνα αναγνώρισης που να υπερτερεί σχετικά με την παράμετρο επαγρύπνησης, τότε ένας νέος νευρώνας αναγνώρισης δημιουργείται στο επίπεδο αναγνώρισης του οποίου τα βάρη προσαρμόζονται ως προς το διάνυσμα εισόδου.

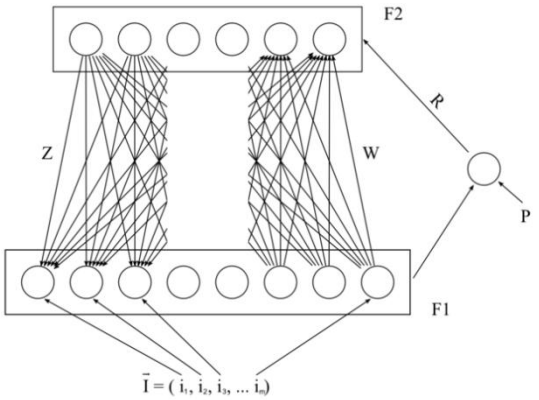
Δύο ερωτήματα μένει να απαντηθούν:

- Πώς αποφασίζουμε αν ένα διάνυσμα εισόδου x συντονίζεται με μια κλάση C ή όχι

- Πως τροποποιείται το αρχέτυπο διάνυσμα w της κλάσης C όταν αποδεχόμαστε το νέο διάνυσμα x στην κλάση αυτή

Για την απάντηση του πρώτου ερωτήματος χρησιμοποιείται μια παράμετρος επαγρύπνησης ρ η οποία αποτελεί το κατώφλι αποδοχής ενός προτύπου σε οποιαδήποτε κλάση. Ο χρήστης έχει την δυνατότητα να επιλέγει την τιμή της παραμέτρου αυτής μεταξύ 0 και 1 [57].

Η παράμετρος επαγρύπνησης έχει σημαντική επιρροή στο σύστημα. Η μεγαλύτερη παράμετρος επαγρύπνησης παράγει μεγαλύτερη λεπτομέρεια ενώ η μικρότερη παράμετρος ενεργοποίησης δίνει πιο γενικά αποτελέσματα [58][57]



Υπάρχουν δύο βασικές μέθοδοι ART: αργές και γρήγορες. Στην μέθοδο αργής μάθησης, ο βαθμός εκπαίδευσης των βαρών του νευρώνα αναγνώρισης προς το διάνυσμα εισόδου υπολογίζεται σε συνεχείς τιμές με διαφορικές εξισώσεις και επομένως εξαρτάται από το χρονικό διάστημα που το διάνυσμα εισόδου παρουσιάζεται.

Με τη γρήγορη μάθηση, αλγεβρικές εξισώσεις χρησιμοποιούνται για τον υπολογισμό του βαθμού το προσαρμογών των βαρών που πρέπει να γίνουν.

Ενώ η γρήγορη μάθηση είναι αποτελεσματική και αποδοτική για μια ποικιλία εργασιών, η μέθοδος αργή μάθηση είναι πιο εύλογη και μπορεί να χρησιμοποιηθεί σε δίκτυα συνεχούς χρόνου (δηλαδή όταν το διάνυσμα εισόδου μπορεί να ποικίλει συνεχώς)[60].

Για την απάντηση του δεύτερου ερωτήματος θα επικεντρωθούμε στον συγκεκριμένο αλγόριθμο ART που πρόκειται να χρησιμοποιήσουμε, από την οικογένεια αλγορίθμων ART που υπάρχουν, ανάλογα με το ποια είναι η χρήση του καθενός.

Η οικογένεια ART παρουσιάζεται με μία μεγάλη ποικιλία. Υπάρχουν αλγόριθμοι που αναφέρονται είτε στην μάθηση με επίβλεψη, είτε στην μάθηση χωρίς επίβλεψη

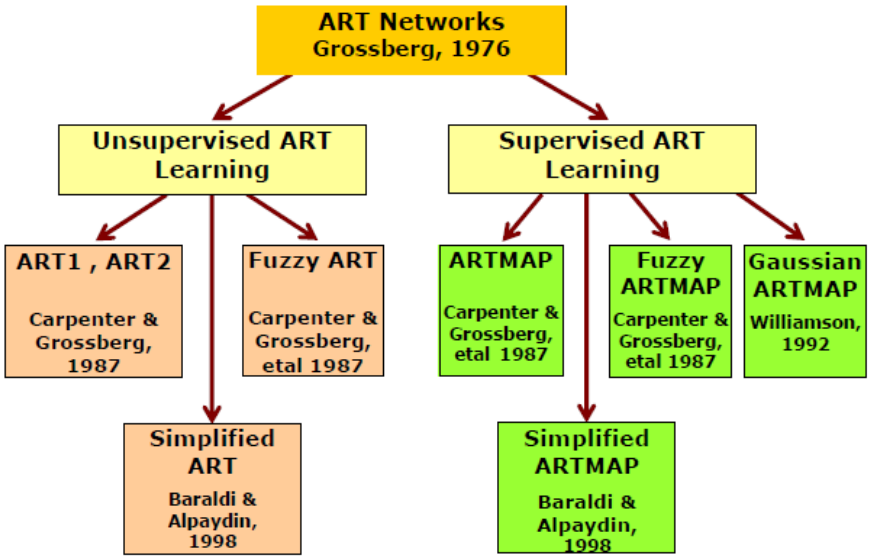


Fig. Important ART Networks

Στην περίπτωση της μάθησης χωρίς επίβλεψη διακρίνουμε τους παρακάτω μοντέλα-αλγορίθμους που ομοιάζουν αρκετά και είναι παρόμοιοι με επαναληπτικούς αλγορίθμους ομαδοποίησης:

ART-1: Σχεδιάστηκε το 1987 για να ομαδοποιεί δυαδικά δεδομένα-πρότυπα

ART-2: Σχεδιάστηκε το 1987 για να ομαδοποιεί συνεχή ή πραγματικά δεδομένα-πρότυπα. Η ικανότητα στο να αναγνωρίζει συνεχή ή πραγματικά πρότυπα ενισχύει σημαντικά το μοντέλο αυτό. Οι διαφορές μεταξύ των δύο μοντέλων ART-1 και ART-2 είναι:

- Ο ART-2 έχει τις κατάλληλες τροποποιήσεις για να μπορέσει να επεξεργαστεί συνεχή διανύσματα
- Το πρώτο επίπεδο του ART-2 είναι πολύ πιο περίπλοκο γιατί τα συνεχή διανύσματα εισόδου δεν είναι διακριτά μεταξύ τους. Το πρώτο επίπεδο του ART-2 χωρίζεται σε περισσότερα υπό επίπεδα τα οποία μπορούν με την σειρά τους να επεξεργαστούν συνεχή δεδομένα.

- Το πρώτο επίπεδο του ART-2 υποστηρίζει κοινωνικοποίηση και καταστολή του θορύβου, επιπρόσθετα στη σύγκριση των άνω και κάτω σημάτων, που χρειάζονται για κάθε επανάληψη τις διαδικασίας

Ο αλγόριθμος ART-2α είναι μια βελτιωμένη μορφή του ART-2 με θεαματική επιτάχυνση εκτέλεσης, και με ποιο ποιοτικά αποτελέσματα που είναι σπανίως κατώτερα από αυτά της μορφής ART-2. βασίζεται στην μορφή ART-2 προσομοιώνοντας στοιχειωδώς την ρύθμιση των νευροδιαβιβαστών της συναπτικής δραστηριότητας με την ενσωματώνοντας τις υπαρκτές συγκεντρώσεις ιόντων νατρίου (Na+) και ασβεστίου(Ca2+), στην προσομοίωση μέσα στις εξισώσεις του συστήματος, το οποίο οδηγεί σε ένα πιο αποδοτικό αποτέλεσμα

ART-3: Σχεδιάστηκε το 1990 και αποτελεί μια βελτιστοποίηση των δύο αρχικών μοντέλων. Ενσωματώνει ένα μοντέλο των χημικών συνάψεων σε μία (ART) αρχιτεκτονική νευρωνικό δικτύου που ονομάζεται ART-3.Το δυναμικό σύστημα ART-3 προτυποποιεί έναν απλό, ολοκληρωμένο μηχανισμό για την παράλληλη αναζήτηση ενός κώδικα αναγνώρισης γνωστών από προηγούμενη μάθηση προτύπων. Αυτός ο μηχανισμός αναζήτησης σχεδιάστηκε για την εφαρμογή των υπολογιστικών αναγκών των συστημάτων ART ενσωματωμένων σε ιεραρχίες δικτύου, όπου μπορεί, σε γενικές γραμμές, να υπάρχουν είτε γρήγορης ή αργής μάθησης και κατανεμημένες ή συμπιεσμένες αναπαραστάσεις κώδικα. Ο μηχανισμός αναζήτησης ενσωματώνει μία ιδιότητα επαναφοράς του κώδικα που εξυπηρετεί τουλάχιστον τρεις διαφορετικές λειτουργίες: τη διόρθωση λανθασμένων επιλογών κατηγοριοποίησης, τη μάθηση από τα ανατροφοδοτούμενα σχόλια και την ανταπόκριση στα μεταβαλλόμενα πρότυπα εισόδου. Οι τρεις τύποι επαναφοράς απεικονίζονται, με προσομοίωση ηλεκτρονικού υπολογιστή, τόσο για μέγιστα συμπιεσμένους και μερικώς συμπιεσμένους κώδικες αναγνώρισης προτύπων[61].

Fuzzy ART: Ενσωματώνει την ασαφή λογική στο πρότυπο του αλγορίθμου ART ενισχύοντας την γενίκευση. Ο Fuzzy ART είναι ένας αλγόριθμος μη επιβλεπόμενης μάθησης (un-supervised learning) που δέχεται ως είσοδο μία ροή αναλογικών προτύπων (input patterns) ή διανυσμάτων και δημιουργεί αυτόφυτα κατηγορίες αναγνώρισης ή τάξεις. Η μάθηση έχει αποδειχθεί ότι είναι ευσταθής, καταλήγοντας

πάντα σε μία σταθερή κατανομή κατηγοριών. Κάθε κατηγορία ξεκινά ως σημείο στο χώρο εισόδου και επεκτείνεται ώστε να περιλάβει νέα διανύσματα που παρουσιάζονται, δημιουργώντας ένα ορθογώνιο υπέρ-παραλληλεπίπεδο (hyperbox).

Αυτή η διαδικασία γίνεται μέχρι να καλυφθεί ο χώρος εισόδου, δηλαδή όλα τα διανύσματα εισόδου να ανήκουν σε κατηγορίες. Το μέγιστο επιτρεπόμενο μέγεθος κάθε υπέρ-παραλληλεπιπέδου καθορίζεται από μία παράμετρο που ονομάζεται εγρήγορση, ρυθμίζοντας έτσι εμμέσως και τον αριθμό των κατηγοριών[59].

Simplied Adaptive Resonance Theory (SART): κατηγορίες δικτύων προτείνονται για χειρισμό των προβλημάτων που προκύπτουν κατά την ανίχνευση δυαδικών και αναλογικών πρότυπων με αλγόριθμους κατηγορίας ART-1. Η βασική ιδέα του SART είναι να υποκαταστήσει τις βασισμένες στον ART-1 λειτουργίες μονής κατεύθυνσης (ασύμμετρες) ενεργοποίησης και ταυτοποίησης με αμφίδρομα \" (συμμετρική) ζεύγη λειτουργιών. Αυτή η υποκατάσταση καθιστά την κατηγορία των SART αλγόριθμων δυνητικά πιο εύρωστη και λιγότερο χρονοβόρα από ό, τι τα συστήματα της κατηγορίας ART-1. Οι αλγόριθμοι SART αποτελούνται από ένα ζεύγος υποσυστημάτων προσοχής και προσανατολισμού και είναι ικανοί να επεξεργάζονται πολυδιάστατα μοντέλα πραγματικών τιμών[62].

Στην περίπτωση της μάθησης με επίβλεψη τα μοντέλα χαρακτηρίζονται από την κατάληξη MAP και ονομάζονται ARTMAP και συνδυάζουν με μία εύμορφη τροποποίηση από τα μοντέλα ART-1 ART-2 για να υλοποιήσουν ένα μοντέλο μάθησης με επίβλεψη που αποτελείται ουσιαστικά από δύο ενότητες. Στη πρώτη λειτουργεί πάνω στα δεδομένα εισόδου και τα ομαδοποιεί ενώ η δεύτερη ενότητα λειτουργεί στα δεδομένα στόχου και τα ομαδοποιεί επίσης. Έπειτα από τις ομαδοποιήσεις των δύο ενοτήτων το μοντέλο συσχετίζει κατάλληλα τα δεδομένα εισόδου και τα δεδομένα εξόδου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. http://www.scielo.br/scielo.php?pid=S1679-87592008000100001&script=sci_arttext
2. http://www.ceine.cl/the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-Data/
3. ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ, Μεταπτυχιακή Εργασία Δονάτος Παπανικολάου, Σεπτέμβριος 2010
4. http://eureka.lib.teithe.gr:8080/bitstream/handle/10184/3770/Dermentzis_Oikonomou.pdf?sequence=2
5. Βραχυπρόθεσμη πρόβλεψη ενεργειακής ζήτησης. Προσεγγίσεις βασισμένες στη Μηχανική Μάθηση ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ Παναγιώτης Γ. Λαδάς
6. Usama Fayyad, Gregory Piatetsky-Shapiro G, Smyth P. , 1996
7. Kotsiantis SB, Kanellopoulos D, Pintelas P., 2006
8. Jain AK, MurtyNM, Flynn JP., 1999
9. http://www.cs.umn.edu/tech_reports_upload/tr2007/07-017.pdf
10. Κωνσταντίνος Διαμαντάρας-Πρόγραμμα Μεταπτυχιακών Σπουδών-Ευφυείς Τεχνολογίες Διαδικτύου-Τμήμα Πληροφορικής, Α.Τ.Ε.Ι. Θεσσαλονίκης
11. Βραχυπρόθεσμη πρόβλεψη ενεργειακής ζήτησης - Προσεγγίσεις βασισμένες στη Μηχανική Μάθηση - ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ - Παναγιώτης Γ. Λαδάς
12. http://blog.tedxacademy.com/2013/06/big-Data.html
13. http://www.technologyreview.com/view/519851/the-big-Data-conundrum-how-to-define-it/ [**National Institute of Standards and Technology (**NIST)]
14. http://tech.in.gr/Analysis/article/?aid=1231253504(*Δημοσίευση: 18 Ιουν. 2013, 15:54*Ερευνα της HP)
15. Περιοδικό Economist («A special report on managing information Data, Data everywhere», Φεβρουάριος 2010)
16. Hal Varian, Chief Economist της Google
17. Tom Davenport σε άρθρο του «Data Scientist: The Sexiest Job of the 21st Century» στο Harvard Business Review (Οκτώβριος 2012)
18. Μελέτη «BIG DATA: The next frontier for innovation, competition and productivity» (Ιούνιος 2011)
19. http://idse.columbia.edu
20. http://Datascience.nyu.edu
21. George Kadifa, επικεφαλής του τμήματος Software στην HP κατά την ανακοίνωση του HAVEn. Το HAVEn αποτελεί την νεότερη πρόταση της HP για την ανάλυση BIG DATA, μια πλατφόρμα BIG DATA Analytics, η οποία αξιοποιεί το λογισμικό, το Hardware και τις υπηρεσίες Analytics της HP (συνδυάζει τις δοκιμασμένες τεχνολογίες των HP Autonomy, HP Vertica, HP ArcSight και HP Operations Management).
22. Έρευνα της Hewlett-Packard υπό τον τίτλο "BIG DATA and Cloud" από την Coleman Parkes Research, Ltd., τον Μάιο του 2013.
23. http://biztech.gr/business-Analytics-diaxeirisi-tis-pliroforias - Δημοσιεύτηκε στις 10 Απρ, 2013
24. http://www.netweek.gr/default.asp?pid=9&la=1&arId=25635 - 12 Σεπτεμβρίου 2013 | 11:20
25. http://www.pmjournal.gr/mckinsey-big-Data-boost-innovation-health(August 10, 2013)
26. http://www.autonews.com/apps/pbcs.dll/article?AID=/20130805/OEM06/130809928
27. http://www.nsp.org.au/confs/bigsecurity2014
28. https://www.rusi.org/downloads/assets/RUSI_BIG DATA_Report_2013.pdf
29. http://www.greekinformatics.gr/pliroforiki/Software/4455-hparcsight.html
30. Cukier K., Schonberger, Vm, M., "BIG DATA: A Revolution That Will Transform How We Live, Work, and Think" , John Murray, 2013 σελ8-9
31. «Η ευρωπαϊκή νομοθεσία αντιμέτωπη με τα BIG DATA. Οι επιπτώσεις από την «έκρηξη» πληροφοριών στην προστασία της ιδιωτικής ζωής» Παναγιώτης Κίτσος Δικηγόρος Ειδικός Επιστήμονας, Διεθνές Πανεπιστήμιο Ελλάδας Ομάδα Δικαίου Πληροφορικής, Τμ. Εφαρμοσμένης Πληροφορικής Πανεπιστήμιο Μακεδονίας - Παρασκευή Παππά Δικηγόρος Καθηγήτρια Εφαρμογών ΤΕΙ Ηπείρου, Σχολή και Διοίκησης και Οικονομίας
32. Elizabeth S. Roop, "BIG DATA Creates Big Privacy Concerns," For the Record 10 Sept. 2012, 11 Φεβρ. 2013 http://www.fortherecordmag.com/archives/091012p10.shtml
33. Essers, L. "Medical privacy threatened by loophole in draft EU Data protection law, professor warns". *Computerworld*. Retrieved February 12th, 2013 from http://news.idg.no/cw/art.cfm?id=905E03DE-D151-A1BC-F4FBA421C54FE418
34. http://tim.webAnalyticsdemystified.com/?p=334
35. Extracting Value from Chaos http://www.emc.com/digital_universe
36. IDC - BIG DATA: What It Is and Why You Should Care - Sponsored by: AMD Richard L. Villars, Carl W. Olofson, Matthew Eastwood, June 2011

37. http://Analytics.dmst.aueb.gr/?q=node/1
38. Πηγή: TDWI, 4th Quarter 2011
39. http://www.cisco.com/web/GR/news/13/news_020413.html
40. www.i-dialogue.org/Ασφάλεια-προσωπικών-δεδομένων-social-media/
41. http://bitdaily.gr/501/international-news/special-reports-international-news/big-Data-proklisi-ke-efkeria-gia-tous-cios
42. http://www.informationweek.com/whitepaper/download/showPDF?articleID=191741034
43. American Diabetes Association; American Hospital Association; HealthPartners Research Foundation; McKinsey Global Institute; National Bureau of Economic Research ; US Census Bureau
44. IDC's Digital Universe Study, sponsored by EMC, June 2011
45. http://www.netweek.gr/default.asp?pid=9&la=1&arId=25421
46. BIG DATA Analytics for Security, Alvaro A. Cárdenas | University of Texas at Dallas, Pratyusa K. Manadhata | HP Labs, Sreeranga P. Rajan | Fujitsu Laboratories of America
47. ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΑΡΑΛΛΗΛΑ ΣΥΣΤΗΜΑΤΑ - Κων/νος Διαμαντάρας - Τμήμα Πληροφορικής, ΑΤΕΙ Θεσσαλονίκης
48. http://digitalschool.minedu.gov.gr/modules/ebook/show.php/DSGL-C127/146/1048,3898/
49. http://digitalschool.minedu.gov.gr/modules/ebook/show.php/DSGL-C127/146/1049,3900/
50. «ΡΕΥΣΤΟΜΗΧΑΝΙΚΗ ΚΑΙ GRID» Του φοιτητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών - Κωνσταντινίδης Νικόλαος του Γεωργίου
51. http://www.it.uom.gr/project/parallel/
52. Distributed Parallel Architecture for "BIG DATA" - Catalin BOJA, Adrian POCOVNICU, Lorena BĂTĂGAN Department of Economic Informatics and Cybernetics Academy of Economic Studies, Bucharest, Romania catalin.boja@ie.ase.ro, pocovnicu@gmail.com, lorena.batagan@ie.ase.ro - http://www.revistaie.ase.ro/content/62/12%20-%20Boja.pdf
53. http://dbtech.uom.gr/mod/resource/view.php?inpopup=true&id=918
54. http://www.gartner.com/newsroom/id/2595015
55. A Comparative Study of clustering algorithms Using WEKA tools Bharat Chaudhari1, Manan Parikh2
56. http://www.incapsula.com/Ddos/Ddos-attacks/
57. ΤΕΧΝΗΤΑ ΝΕΥΡΟΝΙΚΑ ΔΙΚΤΥΑ – Κωνσταντίνος Διαμαντάρας
58. http://en.wikipedia.org/wiki/Adaptive_resonance_theory
59. Διδακτορική Διατριβή, του Κ_ΝΣΤΑΝΤΙΝΟΥ Χ. ΖΗΚΙΔΗ. Μηχανικού Τηλεπικοινωνιών – Ηλεκτρονικών Σχολής Ικάρων
60. http://www.myreaders.info/05_Adaptive_Resonance_Theory.pdf
61. http://cns-web.bu.edu/Profiles/Grossberg/CarGro1990NN.pdf
62. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.3946&rep=rep1&type=pdf

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ Ι – ΚΩΔΙΚΑΣ

```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
package artclusteringfrombigdata;

import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.RandomAccessFile;
import java.math.BigInteger;
import java.nio.channels.FileChannel;
import java.sql.Timestamp;
import java.text.ParseException;
import java.text.SimpleDateFormat;
import java.util.ArrayList;
```

```
import java.util.Date;
import java.util.HashSet;
import java.util.Iterator;
import java.util.Set;
import java.util.concurrent.TimeUnit;
import java.util.logging.Level;
import java.util.logging.Logger;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
import java.util.zip.GZIPInputStream;
import javax.swing.JFileChooser;
import org.apache.commons.io.FileUtils;

/**
 *
 *
 * @author mg
 */
public class MainForm extends javax.swing.JFrame {

    /**
     * Creates new form HomePage
     */
    public MainForm() {
```

```
        // Basic initialization for program Components
        initComponents();
        chooserDir.setFileSelectionMode(JFileChooser.DIRECTORIES_ONLY);
        chooserDir.setAcceptAllFileFilterUsed(false);
        /*
         * When program is started the second & the third button is not
         * enabled. The user is obliged to choose first directory for the Selection Operation.
         * Then user is obliged to choose the Selection File,which have been created during Selection Operation.
         * Also after Pre-proccesing & Transformation (Fill ipAdressList) Convert Ip button can be used
         * At the end of the operation the user can restart the program.
        */
        jButtonForPreProcessingTansformation.setEnabled(false);
        jButtonConvertIp.setEnabled(false);
        jButtonRestart.setEnabled(false);
    }

    /**
     * This method is called from within the constructor to initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is always
     * regenerated by the Form Editor.
     */
```

```
@SuppressWarnings("unchecked")
// <editor-fold defaultstate="collapsed" desc="Generated Code">
private void initComponents() {

    jPanel1 = new javax.swing.JPanel();
    jButtonForSelection = new javax.swing.JButton();
    jLabelSelectionRunningTime = new javax.swing.JLabel();
    jLabelOperation = new javax.swing.JLabel();
    jButtonForPreProcessingTansformation = new javax.swing.JButton();
    jLabelPreProcAndTranRunningTime = new javax.swing.JLabel();
    jLabelTotalRunningTime = new javax.swing.JLabel();
    jScrollPane1 = new javax.swing.JScrollPane();
    jTextAreaLogger = new javax.swing.JTextArea();
    jButtonRestart = new javax.swing.JButton();
    jLabelConvertDate = new javax.swing.JLabel();
    jTextFieldConvertDate = new javax.swing.JTextField();
    jLabelConvertIp = new javax.swing.JLabel();
    jTextFieldConvertIp = new javax.swing.JTextField();
    jButtonConvertDate = new javax.swing.JButton();
    jButtonConvertIp = new javax.swing.JButton();
    jButtonGetClasses = new javax.swing.JButton();
    jTextFieldGetClasses = new javax.swing.JTextField();

    setDefaultCloseOperation(javax.swing.WindowConstants.EXIT_ON_CLOSE);
```

```
        setTitle("Big Data Anal");
        setResizable(false);

        jPanel1.setBackground(new java.awt.Color(200, 200, 212));
        jPanel1.setBorder(javax.swing.BorderFactory.createLineBorder(new
java.awt.Color(255, 0, 0)));

        jButtonForSelection.setText("Choose Folder For Selection");
        jButtonForSelection.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonForSelectionActionPerformed(evt);
            }
        });

        jLabelSelectionRunningTime.setText("Selection Time of Running:");

        jLabelOperation.setText("Status: No Operation");

        jButtonForPreProcessingTansformation.setText("Choose Result File For Pre-
processing and Transformation");
        jButtonForPreProcessingTansformation.addActionListener(new
java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonForPreProcessingTansformationActionPerformed(evt);
            }
```

```
        });

        jLabelPreProcAndTranRunningTime.setText("Pre-processing & Transformation Time of Running:");

        jLabelTotalRunningTime.setText("Total Time of Running:");

        jTextAreaLogger.setEditable(false);
        jTextAreaLogger.setColumns(20);
        jTextAreaLogger.setRows(5);
        jScrollPane1.setViewportView(jTextAreaLogger);

        jButtonRestart.setText("Restart");
        jButtonRestart.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonRestartActionPerformed(evt);
            }
        });

        jLabelConvertDate.setText("Convert Millisecods to Date:");

        jLabelConvertIp.setText("Convert Category Ip to Ip");

        jButtonConvertDate.setText("Convert");
```

```
        jButtonConvertDate.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonConvertDateActionPerformed(evt);
            }
        });

        jButtonConvertIp.setText("Convert");
        jButtonConvertIp.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonConvertIpActionPerformed(evt);
            }
        });

        jButtonGetClasses.setText("Get Number Of Classes - Dates distinct");
        jButtonGetClasses.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(java.awt.event.ActionEvent evt) {
                jButtonGetClassesActionPerformed(evt);
            }
        });

        jTextFieldGetClasses.setEditable(false);

        javax.swing.GroupLayout jPanel1Layout = new javax.swing.GroupLayout(jPanel1);
```

```
        jPanel1.setLayout(jPanel1Layout);
        jPanel1Layout.setHorizontalGroup(
jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addGroup(jPanel1Layout.createSequentialGroup()
                .addContainerGap()
.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
LEADING)
                    .addGroup(jPanel1Layout.createSequentialGroup()
.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
LEADING)
                            .addComponent(jLabelConvertIp,
javax.swing.GroupLayout.PREFERRED_SIZE,                    152,
javax.swing.GroupLayout.PREFERRED_SIZE)
                            .addComponent(jLabelConvertDate))
                        .addGap(18, 18, 18)
.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
TRAILING)
                            .addComponent(jTextFieldConvertDate)
                            .addComponent(jTextFieldConvertIp))
.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED)
```

```
.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
LEADING)
                .addComponent(jButtonConvertDate)
                .addGroup(javax.swing.GroupLayout.Alignment.TRAILING,
jPanel1Layout.createSequentialGroup()
                    .addComponent(jButtonConvertIp)
                    .addContainerGap())))
            .addGroup(javax.swing.GroupLayout.Alignment.TRAILING,
jPanel1Layout.createSequentialGroup()

.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
TRAILING)
                .addComponent(jLabelSelectionRunningTime,
javax.swing.GroupLayout.Alignment.LEADING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                .addComponent(jLabelOperation,
javax.swing.GroupLayout.Alignment.LEADING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                .addComponent(jLabelPreProcAndTranRunningTime,
javax.swing.GroupLayout.Alignment.LEADING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                .addComponent(jButtonForSelection,
javax.swing.GroupLayout.Alignment.LEADING,
```

```
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                    .addComponent(jButtonForPreProcessingTansformation,
javax.swing.GroupLayout.DEFAULT_SIZE, 486, Short.MAX_VALUE)
                    .addComponent(jLabelTotalRunningTime,
javax.swing.GroupLayout.Alignment.LEADING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                    .addComponent(jScrollPane1,
javax.swing.GroupLayout.Alignment.LEADING)
                    .addComponent(jButtonRestart,
javax.swing.GroupLayout.Alignment.LEADING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                    .addGroup(javax.swing.GroupLayout.Alignment.LEADING,
jPanel1Layout.createSequentialGroup()
                        .addComponent(jButtonGetClasses,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
                        .addGap(18, 18, 18)
                        .addComponent(jTextFieldGetClasses,
javax.swing.GroupLayout.PREFERRED_SIZE,                    202,
javax.swing.GroupLayout.PREFERRED_SIZE)))
                .addContainerGap())))
    );
    jPanel1Layout.setVerticalGroup(
```

```
jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
        .addGroup(javax.swing.GroupLayout.Alignment.TRAILING,
jPanel1Layout.createSequentialGroup()
            .addContainerGap()
            .addComponent(jLabelOperation)
            .addGap(18, 18, 18)
            .addComponent(jLabelSelectionRunningTime)
            .addGap(18, 18, 18)
            .addComponent(jLabelPreProcAndTranRunningTime)
            .addGap(18, 18, 18)
            .addComponent(jLabelTotalRunningTime)
            .addGap(18, 18, 18)
            .addComponent(jButtonForSelection)
            .addGap(18, 18, 18)
            .addComponent(jButtonForPreProcessingTansformation)
            .addGap(18, 18, 18)

.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
BASELINE)
                .addComponent(jLabelConvertDate)
                .addComponent(jTextFieldConvertDate,
javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)
```

```
                .addComponent(jButtonConvertDate))
            .addGap(18, 18, 18)

.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
BASELINE)
                .addComponent(jLabelConvertIp)
                .addComponent(jTextFieldConvertIp,
javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE)
                .addComponent(jButtonConvertIp))
            .addGap(18, 18, 18)
            .addComponent(jButtonRestart)
            .addGap(18, 18, 18)

.addGroup(jPanel1Layout.createParallelGroup(javax.swing.GroupLayout.Alignment.
BASELINE)
                .addComponent(jButtonGetClasses)
                .addComponent(jTextFieldGetClasses,
javax.swing.GroupLayout.PREFERRED_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.PREFERRED_SIZE))

.addPreferredGap(javax.swing.LayoutStyle.ComponentPlacement.RELATED,      13,
Short.MAX_VALUE)
```

```
            .addComponent(jScrollPane1,
javax.swing.GroupLayout.PREFERRED_SIZE,                    86,
javax.swing.GroupLayout.PREFERRED_SIZE)
            .addContainerGap())
        );

        javax.swing.GroupLayout layout = new
javax.swing.GroupLayout(getContentPane());
        getContentPane().setLayout(layout);
        layout.setHorizontalGroup(
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addComponent(jPanel1, javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
        );
        layout.setVerticalGroup(
            layout.createParallelGroup(javax.swing.GroupLayout.Alignment.LEADING)
            .addComponent(jPanel1, javax.swing.GroupLayout.Alignment.TRAILING,
javax.swing.GroupLayout.DEFAULT_SIZE,
javax.swing.GroupLayout.DEFAULT_SIZE, Short.MAX_VALUE)
        );

        pack();
    }// </editor-fold>

    /*
```

```
    * Here is the initialization of the variables. Spicifically
    * listFiles is a List, which keeps all the file paths for Selection File
    * listFilesZips is a List, which keeps all the zip file paths for Selection File
    * The startTimeSelection and endTimeSelection are used for calculating of time Selection
    * Also startPreProccesingAndTransformation and endPreProccesingAndTransformation
    * are used for calculating of time PreProccesing And Transformation
    */
    ArrayList<String> listFiles = new ArrayList<>();
    ArrayList<String> listFilesZips = new ArrayList<>();
    ArrayList<ipAddressMapping> ipList = new ArrayList<>();
    ArrayList<String> ipListXMeansClasses = new ArrayList<>();
    Set<Integer> ipListSet = new HashSet();

    static long startTimeSelection = 0;
    static long endTimeSelection = 0;
    static long startTimePreProccesingAndTransformation = 0;
    static long endTimePreProccesingAndTransformation= 0;


    /*
     * This code is relative with the first Button(Selection)
     * This code is relative with the button Event(Selection)
```

```
    */
    @SuppressWarnings("WaitWhileNotSynced")
    private void jButtonForSelectionActionPerformed(java.awt.event.ActionEvent evt)
{

        /*
         * Set Text to jLabelOperation. Helps user to understand that the program is
running
        */
        jLabelOperation.setText("Status: Operation in process...");
        try {

            /*
             * Open the directory chooser and user will choose the appropriate
             * directory for the beginning. These directory must have the subdirectory
             * New. Inside New is our new data for analyze
             */
            int chooserDirListener = chooserDir.showSaveDialog(this);
            if(chooserDirListener == JFileChooser.APPROVE_OPTION){

                // Zip file tail-->Our pattern regex
                String pattern = ".*gz$";

                // If choosen File is correct
```

```
            if(chooserDir.getSelectedFile().isDirectory()){

                // Get the time when program starts
                startTimeSelection= System.nanoTime();

                // Making the appropriate Folders
                File newDir = new File(chooserDir.getSelectedFile().toString()+"\\New");
                File dataDir = new File(chooserDir.getSelectedFile().toString()+"\\Data");
                dataDir.mkdir();
                File copyDir = new File(chooserDir.getSelectedFile().toString()+"\\Copy");
                copyDir.mkdir();
                File copyUnZipDir = new File(chooserDir.getSelectedFile().toString()+"\\UnZip");
                copyUnZipDir.mkdir();

                // Starts Basic Operation of this Button

//=========================================================
                FileUtils.copyDirectory(newDir, copyDir);
                FileUtils.cleanDirectory(newDir);
                System.out.println("Cut from New is Done!");
                jTextAreaLogger.append("Cut from New is Done!\n");
```

```
            FileUtils.copyDirectory(copyDir,dataDir);
            System.out.println("Copy from Copy is Done!");
            jTextAreaLogger.append("Copy from Copy is Done!\n");

            listFilesForFolder(dataDir,pattern);
            FileUtils.copyDirectory(dataDir, copyUnZipDir);
            System.out.println("Copy UnZip is Done!");
            jTextAreaLogger.append("Copy UnZip is Done!\n");

            String outputFile =
FileWRITER(listFiles,chooserDir.getSelectedFile().toString());
            FileWRITERzip(listFilesZips,outputFile);
            System.out.println("Write to Results is Done!");
            jTextAreaLogger.append("Write to Results is Done!\n");

            FileUtils.cleanDirectory(dataDir);
            System.out.println("Remove from Data is Done!");
            jTextAreaLogger.append("Remove from Data is Done!\n");
            // ends Basic Operation of this Button

//=========================================================

            // Get the time when program ends and then it dispayed to user
```

```
                endTimeSelection = System.nanoTime();
                jLabelSelectionRunningTime.setText("Consolidation Time of Running: "+transformJTime(endTimeSelection - startTimeSelection));

                // Success Message to user
                new SuccessSelection().setVisible(true);

                jButtonForSelection.setEnabled(false);
                jButtonForPreProcessingTansformation.setEnabled(true);

            // If choosen File is not correct
            }else{
                jLabelOperation.setText("Status: No Operation");
                new FalseOperation().setVisible(true);
            }

        // Handles when Exception occurs
        }else{
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }

    } catch (Exception ex) {
        Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
```

```
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }
    }

    // Regex for Ip Version4
    private static final String IPADDRESS_PATTERN_V4 =
            "[1][9][5].[2][5][1].[1][2][3].(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)";

    // Regex for Date Time Style :
    private static final String DATETIME_PATTERN =
            "[0-3][0-9]\\/[a-zA-Z][a-zA-Z][a-zA-Z]\\/[1-9][0-9][0-9][0-9]:[0-2][0-9]:[0-
5][0-9]:[0-5][0-9] \\+[0-9][0-9][0-9][0-9]";


    /*
     * This code is relative with the second Button(PreProccesing And Transformation)
     * This code running after the first Button Code(Selection Button) Selection button
     * select and consolidate all data to One File into Results with tail Selection.txt
     * Here the second Button(PreProccesing And Transformation) takes this File and
     * making two Files the Pre-processing and Transformation Files..Pre-processing
has
     * the matces of ip and DateTimes with categories and Transformation has only the
     * categories(numbers), which is availiable for Art Algorithm(Data Mining)
```

```
    * This code is relative with the button Event(PreProccesing And Transformation).
    */
    @SuppressWarnings({"ConvertToTryWithResources", "UseSpecificCatch"})
    private void jButtonForPreProcessingTansformationActionPerformed(java.awt.event.ActionEvent evt) {
        /*
         * Set Text to jLabelOperation. Helps user to understand that the program is running again
        */
        jLabelOperation.setText("Status: Operation in process...");
        try {

            /*
             * Open the File chooser and user will choose the appropriate
             * file for the PreProccesing And Transformation.
             */
            int chooserFileListener = chooserFile.showSaveDialog(this);
            if(chooserFileListener == JFileChooser.APPROVE_OPTION){

                // Get the time when program starts
                startTimePreProccesingAndTransformation = System.nanoTime();

                // Making the appropriate Files
                File inFile = new File(chooserFile.getSelectedFile().toString());
```

```
String outputFile = chooserFile.getSelectedFile().toString();
outputFile = outputFile.replace(".txt", "");
outputFile = outputFile.concat("_Pre-processing.txt");
File outFile = new File(outputFile);
String outputFileART = chooserFile.getSelectedFile().toString();
outputFileART = outputFileART.replace(".txt", "");
outputFileART = outputFileART.concat("_ART_Transformation.txt");
File outFileART = new File(outputFileART);

// Initialize Classes for Files(Writers and Readers)
BufferedReader bf = new BufferedReader(new FileReader(inFile));
BufferedWriter bw = new BufferedWriter(new FileWriter(outFile.getAbsoluteFile()));
BufferedWriter bwART = new BufferedWriter(new FileWriter(outFileART.getAbsoluteFile()));

String data;//Of any Line in txt
// Get the patterns of regex
Pattern ipV4 = Pattern.compile(IPADDRESS_PATTERN_V4);
Pattern dateTime = Pattern.compile(DATETIME_PATTERN);

// Matcher is going to mathes the regex patterns
Matcher m;
```

```
/*
 * Every Line is matched, is mapped to our specific Classn ipAddressMapping
 * ArrayList<ipAdressMapping> ipList = new ArrayList<>();
 */

// Help to set categories for ips and DateTimes
int categoryOfIp = 1;
long categoryOfDateTime = 0;

while((data = bf.readLine()) != null){

   Matcher matcherIpV4 = ipV4.matcher(data);
   Matcher matcherDateTime = dateTime.matcher(data);

   // If matches ip regex and DateTime regex
   if((matcherIpV4.find()) && matcherDateTime.find()){

      String getIpV4 = matcherIpV4.group(0); // Get Line sub String Ip which is matched
      String getDateTime = matcherDateTime.group(0); // Get Line sub String  DateTime which is matched
      String result = getIpV4.concat(" > ").concat(getDateTime);

      // Convert DateTime to Millisecond:Art Category
```

```
categoryOfDateTime = new SimpleDateFormat("dd/MMM/yyyy:HH:mm:ss Z").parse(getDateTime).getTime();

String resultART = "";

//Check if ip already exists
int positionIp = -1;
for(int i=0;i<ipList.size();i++){
   if(ipList.get(i).getIp().equals(getIpV4)){
      positionIp = i;
      break;
   }
}
//Ip not exists
if(positionIp==-1){

      //Check if DateTime already exists
      int positionDateTime = -1;
      for(int j=0;j<ipList.size();j++){
         if(ipList.get(j).getDateTime().equals(getDateTime)){
            positionDateTime = j;
            break;
         }
      }
```

```
                    //DateTime not exists
                    if(positionDateTime==-1){

                        ipAddressMapping itemNew = new
ipAddressMapping(getIpV4,getDateTime,categoryOfIp,categoryOfDateTime);
                        ipList.add(itemNew);
                        ipListSet.add(itemNew.getCategoryIp());
                        result = result.concat("<>
").concat(String.valueOf(itemNew.getCategoryIp())).concat(" >
").concat(String.valueOf(itemNew.getCategoryDateTime()));
                        resultART =
resultART.concat(String.valueOf(itemNew.getCategoryIp())).concat("
").concat(String.valueOf(itemNew.getCategoryDateTime()));
                        categoryOfIp++;
                    //DateTime exists
                    }else{
                        ipAddressMapping itemNew = new
ipAddressMapping(getIpV4,getDateTime,categoryOfIp,ipList.get(positionDateTime).
getCategoryDateTime());
                        ipList.add(itemNew);
                        ipListSet.add(itemNew.getCategoryIp());
                        result = result.concat("<>
").concat(String.valueOf(itemNew.getCategoryIp())).concat(" >
").concat(String.valueOf(itemNew.getCategoryDateTime()));
```

```
                    resultART =
resultART.concat(String.valueOf(itemNew.getCategoryIp())).concat("
").concat(String.valueOf(itemNew.getCategoryDateTime()));
                    categoryOfIp++;
                }
            //Ip exists
            }else{
                //Check if DateTime already exists
                int positionDateTime = -1;
                for(int j=0;j<ipList.size();j++){
                    if(ipList.get(j).getDateTime().equals(getDateTime)){
                        positionDateTime = j;
                        break;
                    }
                }
                //DateTime not exists
                if(positionDateTime==-1){
                    ipAddressMapping itemNew = new
ipAddressMapping(getIpV4,getDateTime,ipList.get(positionIp).getCategoryIp(),categ
oryOfDateTime);
                    ipList.add(itemNew);
                    result = result.concat(" <>
").concat(String.valueOf(itemNew.getCategoryIp())).concat(" >
").concat(String.valueOf(itemNew.getCategoryDateTime()));
```

```
                    resultART =
resultART.concat(String.valueOf(itemNew.getCategoryIp())).concat("
").concat(String.valueOf(itemNew.getCategoryDateTime()));
                    //DateTime exists
                    }else{
                    result = result.concat(" <>
").concat(String.valueOf(ipList.get(positionIp).getCategoryIp())).concat(" >
").concat(String.valueOf(ipList.get(positionDateTime).getCategoryDateTime()));
                    resultART =
resultART.concat(String.valueOf(ipList.get(positionIp).getCategoryIp())).concat("
").concat(String.valueOf(ipList.get(positionDateTime).getCategoryDateTime()));
                    }
                }

                // Write results to appropriate Files
                bw.write(result);
                bw.newLine();
                bwART.write(resultART);
                bwART.newLine();

            }else{
                // TODO add your handling code here:
            }
        }
        //close the Writers and Readers(Classes for Files)
```

```
            bw.close();
            bwART.close();
            bf.close();

            System.out.println("Pre-proccesing & Transformation is Done!");
            jTextAreaLogger.append("Pre-proccesing & Transformation is Done!\n");
            // ends Basic Operation of this Button

//=============================================================
==

            // Making the SubDirectory into Results
            String theCurrentDirectory = chooserFile.getSelectedFile().toString();
            theCurrentDirectory = theCurrentDirectory.replace(".txt",
"").concat("_XMeansSubCategories");
            File artSubCategoriesDir = new File(theCurrentDirectory);
            artSubCategoriesDir.mkdir();

            /*
             * Fill the SubDirectory into Results
             * Here we are making one File for Each specific Ip category
             * These is Data Mining Files Seperated
             */
            File tempFileART[] = new File[ipListSet.size()];
            BufferedWriter bwtemsARTS[] = new BufferedWriter[ipListSet.size()];
```

```
            for(Iterator<Integer> it = ipListSet.iterator(); it.hasNext();){
                Integer i = it.next();
                String thisIP = String.valueOf(i);
                theCurrentDirectory = artSubCategoriesDir.getPath().concat("\\Cat_"+thisIP+"_IP.txt");
                tempFileART[i-1] = new File(theCurrentDirectory);
                bwtemsARTS[i-1] = new BufferedWriter(new FileWriter(tempFileART[i-1].getAbsoluteFile()));
                for(int j=0;j<ipList.size();j++){
                   if(ipList.get(j).getCategoryIp()==i){
                      String thisResult = String.valueOf(ipList.get(j).getCategoryIp()).concat(",").concat(String.valueOf(ipList.get(j).getCategoryDateTime()));
                      bwtemsARTS[i-1].write(thisResult);
                      bwtemsARTS[i-1].newLine();
                   }
                }
                bwtemsARTS[i-1].close();
            }

             // Get the times when program ends and then it dispayed to user
            endTimePreProccesingAndTransformation = System.nanoTime();
            jLabelPreProcAndTranRunningTime.setText("Pre-proccesing & Transformation Time of Running:
```

```
"+transformJTime(endTimePreProccesingAndTransformation -
startTimePreProccesingAndTransformation));
            jLabelTotalRunningTime.setText("Total Time of
Running:"+transformJTime((endTimePreProccesingAndTransformation+endTimeSel
ection) - (startTimePreProccesingAndTransformation+startTimeSelection)));

            // Success Message to user
            new SuccessPreProccesingTransformation().setVisible(true);
            jButtonForPreProcessingTansformation.setEnabled(false);
            jButtonConvertIp.setEnabled(true);
            jButtonRestart.setEnabled(true);

            // If choosen File is not correct
        }else{
            // Handles when Exception occurs
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }

    } catch (Exception ex) {
        Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
        jLabelOperation.setText("Status: No Operation");
        new FalseOperation().setVisible(true);
    }
}
```

```
    // This button Restarts the MainForm..it looks like basic method run()
    private void jButtonRestartActionPerformed(java.awt.event.ActionEvent evt) {
        dispose();
        MainForm homePage = new MainForm();
        //homePage.setExtendedState(JFrame.MAXIMIZED_BOTH);
        //homePage.setLocationRelativeTo(null);
        homePage.setVisible(true);
    }

    @SuppressWarnings("UseSpecificCatch")
    private void jButtonConvertDateActionPerformed(java.awt.event.ActionEvent evt)
{
        try{
            long getMil = Long.valueOf(jTextFieldConvertDate.getText(),10);
            jTextFieldConvertDate.setText(transformMilliSecDateFormat(getMil));
        } catch (Exception ex) {
            Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }

    }
```

```
    private void jButtonConvertIpActionPerformed(java.awt.event.ActionEvent evt) {
        try{
            int getIp = Integer.valueOf(jTextFieldConvertIp.getText());
            for(int i=0;i<ipList.size();i++){
                if(ipList.get(i).getCategoryIp()==getIp){
                    jTextFieldConvertIp.setText(String.valueOf(ipList.get(i).getIp()));
                }
            }
        } catch (Exception ex) {
            Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }
    }

    // Regex for Time Millisecond Style :
    private static final String TIME_MIL_PATTERN =
            "(([0-9]){5})";
    /*
     * Return the number of different Milliseconds(DateTime) in File
     * This helps for DataMining when we want to know the number of differnt
classes(ex: K-Means / XMeans)
     */
```

```
private void jButtonGetClassesActionPerformed(java.awt.event.ActionEvent evt) {
    /*
     * Set Text to jLabelOperation. Helps user to understand that the program is running again
     */
    jLabelOperation.setText("Status: Operation in process...");
    jTextFieldGetClasses.setText("");
    try {

      /*
       * Open the File chooser and user will choose the appropriate
       * file for getting XMeans Classes
       */
      int chooserFileListener = chooserFile.showSaveDialog(this);
      if(chooserFileListener == JFileChooser.APPROVE_OPTION){

        // Making the appropriate Files
        File inFile = new File(chooserFile.getSelectedFile().toString());


        // Initialize Classes for Files(Writers and Readers)
        BufferedReader bf = new BufferedReader(new FileReader(inFile));

        String data;
```

```
while((data = bf.readLine()) != null){


    Pattern time = Pattern.compile(TIME_MIL_PATTERN);
    Matcher matcherTime = time.matcher(data);
    //If matches time regex
    if(matcherTime.find()){
        String getDateTime = matcherTime.group(0); // Get Line sub String Time Mil First 5 which is matched
        if(!ipListXMeansClasses.contains(getDateTime)){
            ipListXMeansClasses.add(getDateTime);
        }


    }
}
// like Logger
System.out.println(ipListXMeansClasses.size());
jTextFieldGetClasses.setText(String.valueOf(ipListXMeansClasses.size()));
ipListXMeansClasses.clear();
// Success Message to user
new SuccessPreProccesingTransformation().setVisible(true);
jButtonForPreProcessingTansformation.setEnabled(false);


// If choosen File is not correct
}else{
```

```
            // Handles when Exception occurs
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }

    } catch (Exception ex) {
        Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
        jLabelOperation.setText("Status: No Operation");
        new FalseOperation().setVisible(true);
    }
}

// This method add to List the paths with all files in specific directoty
public void listFilesForFolder(File folder,String pattern) throws Exception {
    for (File fileEntry : folder.listFiles()) {
        if (fileEntry.isDirectory()) {
            listFilesForFolder(fileEntry,pattern);
        } else {
                if(fileEntry.getPath().matches(pattern)){
                    String OUTPUT_FILE = fileEntry.getPath().concat(".txt");
                    gunzipIt(fileEntry.getPath(),OUTPUT_FILE);
                    listFilesZips.add(OUTPUT_FILE);
                }else{
                    listFiles.add(fileEntry.getPath());
```

```
            }
        }
    }
}
```

```
/**
 * GunZip it
 * @param INPUT_GZIP_FILE
 * @param OUTPUT_FILE
 * @throws java.io.IOException
 */
//This method unZip a specific file
@SuppressWarnings("ConvertToTryWithResources")
public void gunzipIt(String INPUT_GZIP_FILE,String OUTPUT_FILE) throws IOException{
    File file = new File(INPUT_GZIP_FILE);
    byte[] buffer = new byte[(int)file.length()];
    FileOutputStream out;
    GZIPInputStream gzis = new GZIPInputStream(new FileInputStream(INPUT_GZIP_FILE));
    out = new FileOutputStream(OUTPUT_FILE);
    int len;
```

```
        while((len = gzis.read(buffer,0,buffer.length)) > 0) {
            out.write(buffer, 0, len);
        }
        out.close();
        gzis.close();
        file.delete();
    }
    // This method add text(TIME) to DateFormat
    public static String transformJavaDateFormat(){
        long timeForResults = System.currentTimeMillis();
        Timestamp stamp = new Timestamp(timeForResults);
        Date date = new Date(stamp.getTime());
        String forDate = date.toString().replace(' ', '_');
        forDate = forDate.replace(':', '_');
        forDate = forDate.substring(0,10)+"_TIME_"+forDate.substring(11,28);
        return forDate;
    }
    // This method Milliseconds to DateFormat
    public static String transformMilliSecDateFormat(long milliseconds){
        String theDate = "Problem with this Date";
        try{
            SimpleDateFormat date = new SimpleDateFormat("dd/MMM/yyyy:HH:mm:ss
Z");
            theDate = date.format(milliseconds);
```

```
        } catch (Exception ex) {
            Logger.getLogger(MainForm.class.getName()).log(Level.SEVERE, null, ex);
            jLabelOperation.setText("Status: No Operation");
            new FalseOperation().setVisible(true);
        }
        return theDate;
    }
    //This method transform Milliseconds to TimeFormat
    public static String transformJTime(long start){
        long help = TimeUnit.NANOSECONDS.toSeconds(start);
        long second = (help) % 60;
        help = TimeUnit.NANOSECONDS.toMinutes(start);
        long minute = (help) % 60;
        help = TimeUnit.NANOSECONDS.toHours(start);
        long hour = (help) % 24;
        String estimateTime = String.format("%02d:%02d:%02d", hour, minute, second);
        return estimateTime;
    }
    /*
     * It is very Impotand that Write Methods used FileChannels and RandomAccess Files
     * This is very importand for Analyzing BIG DATA,because is the only way for fast
     * process and the reason is No Buffer(Java) is used during this operation
```

```
     *
    */
    // This method Write simple Files to txt
    @SuppressWarnings("ConvertToTryWithResources")
    public static String FileWRITER(ArrayList<String> inputFiles,String selectedFile)
throws Exception{
        String resultsDir = selectedFile.concat("\\Results");
        File forResults = new File(resultsDir);
        forResults.mkdir();
        String                                  outputFile                                  =
forResults.toString().concat("\\Result_"+transformJavaDateFormat()+"_Selection.txt"
);
        RandomAccessFile outFile = new RandomAccessFile(outputFile, "rw" );
        for (String inputFile : inputFiles) {
            RandomAccessFile inFile = new RandomAccessFile(inputFile, "r");
            FileChannel inputChannel = inFile.getChannel();
            inputChannel.transferTo(0, inputChannel.size(), outFile.getChannel());
            inFile.close();
        }
        outFile.getChannel().close();
        outFile.close();
        return outputFile;
    }
    // This method Write zip Files to txt
    @SuppressWarnings("ConvertToTryWithResources")
```

```
    public static void FileWRITERzip(ArrayList<String> inputFiles,String outputFile) throws Exception{
        RandomAccessFile outFile = new RandomAccessFile(outputFile, "rw" );
        for (String inputFile : inputFiles) {
            RandomAccessFile inFile = new RandomAccessFile(inputFile, "r");
            FileChannel inputChannel = inFile.getChannel();
            inputChannel.transferTo(0, inputChannel.size(), outFile.getChannel());
            inFile.close();
        }
        outFile.getChannel().close();
        outFile.close();
    }

    /**
     * @param args the command line arguments
     */
    public static void main(String args[]) {
        /* Set the Nimbus look and feel */
        //<editor-fold defaultstate="collapsed" desc=" Look and feel setting code (optional) ">
        /* If Nimbus (introduced in Java SE 6) is not available, stay with the default look and feel.
         * For details see http://download.oracle.com/javase/tutorial/uiswing/lookandfeel/plaf.html
         */
```

```
        try {
            for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
                if ("Nimbus".equals(info.getName())) {
                    javax.swing.UIManager.setLookAndFeel(info.getClassName());
                    break;
                }
            }
        } catch (ClassNotFoundException | InstantiationException |
IllegalAccessException | javax.swing.UnsupportedLookAndFeelException ex) {

java.util.logging.Logger.getLogger(MainForm.class.getName()).log(java.util.logging.
Level.SEVERE, null, ex);
        }
        //</editor-fold>

        //</editor-fold>

        /* Create and display the form */
        java.awt.EventQueue.invokeLater(new Runnable() {
            @Override
            public void run() {
                MainForm homePage = new MainForm();
                //homePage.setExtendedState(JFrame.MAXIMIZED_BOTH);
                //homePage.setLocationRelativeTo(null);
```

```
            homePage.setVisible(true);
        }
    });
}

//***** Here we Make File Chooser ************************************
private final javax.swing.JFileChooser chooserDir= new javax.swing.JFileChooser("C:\\");
private final javax.swing.JFileChooser chooserFile= new javax.swing.JFileChooser("C:\\");
//
*****************************************************************************

// Variables declaration - do not modify
private javax.swing.JButton jButtonConvertDate;
private javax.swing.JButton jButtonConvertIp;
private javax.swing.JButton jButtonForPreProcessingTansformation;
private javax.swing.JButton jButtonForSelection;
private javax.swing.JButton jButtonGetClasses;
private javax.swing.JButton jButtonRestart;
private javax.swing.JLabel jLabelConvertDate;
private javax.swing.JLabel jLabelConvertIp;
public static javax.swing.JLabel jLabelOperation;
private javax.swing.JLabel jLabelPreProcAndTranRunningTime;
```

```
    private javax.swing.JLabel jLabelSelectionRunningTime;
    private javax.swing.JLabel jLabelTotalRunningTime;
    private javax.swing.JPanel jPanel1;
    private javax.swing.JScrollPane jScrollPane1;
    private javax.swing.JTextArea jTextAreaLogger;
    private javax.swing.JTextField jTextFieldConvertDate;
    private javax.swing.JTextField jTextFieldConvertIp;
    private javax.swing.JTextField jTextFieldGetClasses;
    // End of variables declaration
}
```