ΑΝΑΛΥΣΗ ΚΙΝΗΣΗΣ ΔΙΚΤΥΩΝ ΥΠΟΔΟΜΩΝ ΖΩΤΙΚΗΣ ΣΗΜΑΣΙΑΣ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ APACHE SPOT

του

ΝΙΚΟΛΑΟΥ ΣΙΓΑΝΟΥ

Ηλεκτρονικού Μηχανικού, ΤΕΙ Ηράκλειου, 1994

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων για το πτυχίο

ΠΜΣ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΕΛΛΗΝΙΚΟ ΜΕΣΟΓΕΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

2020

Εγκρίθηκε από:

Δρ. Ευάγγελο Μαρκάκη

Περίληψη

Το Apache Spot είναι ένα υπό εξέλιξη έργο του Ιδρύματος για την ανίχνευση ανωμαλιών σε δεδομένα τηλεμετρίας δικτύου μεγάλου όγκου. Παρουσιάζει ιδιαίτερο ενδιαφέρον αλλά και πολυπλοκότητα καθώς συνδυάζει τους τομείς Κυβερνοασφάλειας, Μεγάλων Δεδομένων και Μηχανικής Μάθησης. Σε αυτή την εργασία γίνεται η παρουσίαση της λειτουργίας του, η εγκατάσταση και η παραμετροποίησή του και η χρήση του στο δικτυακό περιβάλλον μιας Υποδομής Ζωτικής Σημασίας (ΥΖΣ), ενταγμένο σε ένα πλαίσιο για την αξιολόγιση των αποτελεσμάτων. Πιο συγκεκριμένα, αναλύονται δεδομένα netflow εξωτερικής, εσωτερικής και κίνησης ενδοδικτύου που συλλέχθηκαν για διάρκεια δύο μηνών από υποδομή μεσαίου μεγέθους από τον χώρο της Υγείας που χρησιμοποιεί καθιερωμένα συστήματα ασφαλείας. Στόχος είναι ο εντοπισμός ευρημάτων σε ένα πραγματικό δίκτυο αλλά και ενδεχόμενων περιορισμών στις δυνατότητές του. Παρουσιάζεται η μεθοδολογία της ανάλυσης που περιλαμβάνει τη χρήση επιπλέον υπηρεσιών πληροφοριών φήμης, τις πληροφορίες από κατάλογο των κόμβων του δικτύου και τη χρήση εργαλείου εκτέλεσης αξιολόγησης ευπάθειας των εμπλεκόμενων κόμβων.

Εξετάζοντας τη συμπεριφορά του Apache Spot κατά την ανίχνευση πολυήμερων σαρώσεων με μεταβλητά χαρακτηριστικά, διαπιστώνεται ότι δεν υπάρχει συνοχή στα αποτελέσματα καθώς οι εν λόγω σαρώσεις δεν ανιχνεύονται κάθε ημέρα. Επίσης, από την ιχνηλάτηση ύποπτων ροών προκύπτουν επιθέσεις που έγιναν άλλες μέρες και δεν έχουν εντοπιστεί στα αποτελέσματα, γεγονός, όμως, που δείχνει ότι αποτελεί άριστο εργαλείο ιχνηλάτησης. Στην ανίχνευση ανωμαλιών όπου ύποπτες ροές είναι αυτές που ο αλγόριθμός υπολογίζει ότι έχουν την χαμηλότερη πιθανότητα, οι πιθανότητες καθορίζονται από τα χαρακτηριστικά των ροών αλλά και το σύνολο της κίνησης γεγονός που εξηγεί την έλλειψη συνοχής που παρατηρήθηκε. Πάραυτα, από τα ευρήματα προκύπτει ότι μπορεί να εντοπίσει ενέργειες που έχουν διαφύγει των άλλων συστημάτων προστασίας. Για παράδειγμα, εντοπίστηκε αμφίδρομη επικοινωνία με κόμβο που φιλοξενούσε κακόβουλο λογισμικό, επικοινωνία που ολοκληρώθηκε σε συνολικά μόλις τέσσερις ροές και μεταφέρθηκαν ελάχιστα MB.

Αν η ανάλυση είναι σε βάθος χρόνου, το Spot μπορεί να αποκαλύψει κακόβουλες ενέργειες μη ορατές σε άλλα συστήματα και να μειώσει τον μέσο χρόνο ανίχνευσης και αντίδρασης σε ένα συμβάν κυβερνοασφάλειας. Οι ΥΖΣ έχουν σαν προτεραιότητά τους την ελαχιστοποίηση των επιπτώσεων των κυβερνοεπιθέσεων και το Apache Spot είναι ένα εργαλείο που μπορεί να προστεθεί στην πολύ-επίπεδη προστασία τους έχοντας έναν ξεχωριστό ρόλο.

Abstract

Apache Spot is an incubating project of the Foundation, used for anomaly detection in large volumes of network telemetry data. It is of particular interest and as it combines the fields of Cybersecurity, Big Data, and Machine Learning. This thesis contains the presentation of Spot's functionality, installation and configuration process, and its usage in the networking environment of critical infrastructure, integrated into a framework for the evaluation of the results. Specifically, it analyses internal, external, and intranet netflow data collected over a two-month period from a medium-sized healthcare facility, which uses established security systems. The purpose is to identify findings in a real network, along with possible limitations to its capabilities. The analysis methodology including the use of additional reputation information services, information from network inventory, and the use of a vulnerability assessment tool for the nodes involved is presented.

Having inspected Apache Spot's behavior when detecting several-day scans with variable characteristics, it is deduced that the results are quite inconsistent, as these scans are not detected every day. In addition, attacks that occurred on different days and cannot be found in the results are traced during the investigation of suspicious connections, proving that Spot is an excellent investigating tool. In the anomaly detection where suspicious flows are the ones that the algorithm calculates to have the lowest probability, the probabilities are determined by the characteristics of the flows and the total traffic, which explains the lack of coherence observed. However, the findings show that it can identify actions that have eluded other protection systems For instance, bidirectional communication was detected with a node hosting malware, communication completed in a total of just four flows and minimal MB transferred.

For a long-term analysis, Apache Spot can reveal malicious actions invisible to other security systems and reduce the average of the detection and reaction time to a cybersecurity incident. The priority of critical infrastructure is the minimization of the effects of cyberattacks and Apache Spot is a tool that can be added to their multi-level protection having a special role.

Περιεχόμενα

Κατάλογος Εικόνων iii

Κατάλογος Πινάκων iv

Ευχαριστίες 1

Πρόλογος 2

Κεφάλαιο 1 – Κυβερνοασφάλεια 4

1.1 Εισαγωγή 4

1.2 Σύγχρονες Απειλές 6

1.3 Κυβερνοασφάλεια υποδομών ζωτικής σημασίας 7

1.4 Επιθέσεις σε ΥΖΣ 10

1.5 Δυσκολίες των σύγχρονων μέτρων προστασίας δικτύων. 11

1.6 Security Analytics 12

1.7 Το Apache Spot σε δίκτυα ΥΖΣ 14

1.8 Σύγχρονες ερευνητικές προτάσεις για την ασφάλεια των ΥΖΣ 16

1.9 Ερευνητική δραστηριότητα για το Apache Spot. 23

Κεφάλαιο 2 - Ανάλυση κίνησης δικτύου με χρήση του Apache Spot 24

2.1 Εισαγωγή 24

2.2 Μεθοδολογία 24

2.3 Σενάριο Χρήσης 29

2.4 OpenVAS 31

Κεφάλαιο 3 - Apache Spot και Οικοσύστημα Hadoop 33

3.1 Hadoop 33

3.1.1 HDFS 33

3.1.2 MapReduce 35

3.2 Yarn 35

3.3 Apache Kafka 37
3.4 Apache Spark 38
3.5 Apache Hive 40
3.6 Apache Impala 41
3.7 Εγκατάσταση συστοιχίας Cloudera Hadoop. 41
Κεφάλαιο 4 - Apache Spot 44
4.1 Εισαγωγή 44
4.2 Αρχιτεκτονική 45
4.3 Αξιολόγηση συνδέσεων με μηχανική μάθηση 48
4.4 Οπτικοποίηση και παρουσίαση αποτελεσμάτων 50
4.4.1 Suspicious 51
4.4.2 Threat Investigation 53
4.4.3 Storyboard 54
4.5 Εγκατάσταση 54
4.6 Λειτουργία 57
Κεφάλαιο 5 - Ανάλυση κίνησης δικτύου υποδομής στο χώρο της Υγείας 60
5.1 Εισαγωγή 60
5.2 Ρύθμιση παραμέτρων αλγορίθμου LDA 61
5.3 Πλήρης ανάλυση αποτελεσμάτων μιας εβδομάδας 67
5.4 Ευρήματα 76
Συμπεράσματα 84
Βιβλιογραφία 86

Κατάλογος Εικόνων

Εικόνα 1: Υποδομές ζωτικής σημασίας στην ΕΕ 8
Εικόνα 2: IT και OT στις ΥΖΣ 9
Εικόνα 3: Ανάλυση κίνησης δικτύου με Apache Spot 25
Εικόνα 4: Ενσωματωμένες βιβλιοθήκες Apache Spark 39
Εικόνα 5: Μπλοκ διάγραμμα συστοιχίας Apache Spark 40
Εικόνα 6: Cloudera Manager - Εγκατεστημένες υπηρεσίες και λειτουργική κατάσταση 43
Εικόνα 7: Χαρακτηριστικά Apache Spot 44
Εικόνα 8: Τύποι δεδομένων και ανιχνεύσιμες απειλές 45
Εικόνα 9: Αρχιτεκτονική του Apache Spot 46
Εικόνα 10: Δομή φακέλων HDFS 56
Εικόνα 11: Περιβάλλον χρήσης Apache Spot - Suspicious Connections 59
Εικόνα 12: Δίκτυο ΥΖΣ και αξιολόγηση κίνησης 60
Εικόνα 13: Κακόβουλες IP δ/νσεις που ανιχνεύτηκαν ανά ημέρα και αριθμό θεμάτων 64
Εικόνα 14: Κακόβουλες IP δ/νσεις που ανιχνεύτηκαν συνολικά με διάφορες τιμές θεμάτων 65
Εικόνα 15: Κακόβουλες IPs δ/νσεις που αναγνωριστήκαν με διάφορες τιμές υπερπαραμέτρων 66
Εικόνα 16: Ροές ανάλογα με την συχνότητα τους 73
Εικόνα 17: Απόσπασμα πηροφοριών από το X-force σχετικά με κακόβουλη IP δ/ση 74
Εικόνα 18: Αποτελέσματα αναζήτησης ύποπτων ροών στο Hue 75
Εικόνα 19: Στοιχεία χαρακτηριστικής σάρωσης 76
Εικόνα 20: Ροές από επίθεση στην θύρα 1723 77
Εικόνα 21: Συνδέσεις με διακομιστή ανώνυμης περιήγησης 78
Εικόνα 22: Πολλαπλές ροές στην θύρα 25 του διακομιστή αλληλογραφίας 79
Εικόνα 23: Ύποπτη εξερχόμενη σύνδεση 79
Εικόνα 24: Ιστορικό χαρακτηρισμού IP δ/νσης από το X-Force 80
Εικόνα 25: Πληροφορίες για IP δ/νση από το VirusTotal 81
Εικόνα 26: Ιστορικό χαρακτηρισμού IP δ/νσης από το X-Force 82
Εικόνα 27: Πληροφορίες για IP δ/νση από το VirusTotal 82
Εικόνα 28: Επαναλαμβανόμενες συνδέσεις σε θύρες 137, 139 και 22. 83

Κατάλογος Πινάκων

Πίνακας 1: Βασικές λειτουργίες πλαισίου κυβερνοασφάλειας NIST..... 5
Πίνακας 2: Αντιστοίχιση εννοιών αλγορίθμου LDA σε στοιχεία των αρχείων καταγραφής.. 49
Πίνακας 3: Πεδία που αποθηκεύονται στο HDFS κατά την πρόσληψη των δεδομένων..... 50
Πίνακας 4: Στοιχεία που περιλαμβάνονται στις ύποπτες συνδέσεις 51
Πίνακας 5: Στοιχεία που παρουσιάζονται για την αξιολόγηση των ύποπτων συνδέσεων.... 53
Πίνακας 6: Κατανομή ρόλων στους 3 διακομιστές του οικοσυστήματος Hadoop..... 55
Πίνακας 7: εγκατεστημένο λογισμικό στοιχεία ανά κόμβο..... 57
Πίνακας 8: Παράμετροι αλγορίθμου LDA..... 62
Πίνακας 9: Ποσοτικά χαρακτηριστικά δεδομένων που συλλέχθηκαν..... 68
Πίνακας 10: Κατανομή των IP διευθύνσεων στα ημερήσια αποτελέσματα..... 68
Πίνακας 11: Αριθμός IP δ/νσεων και ροών που εμπλέκονται με επιβεβαιωμένα κακόβουλες δ/σνσεις..... 71
Πίνακας 12: Αριθμητική κατανομή κακόβουλων δραστηριότήτων..... 72

Ευχαριστίες

Ευχαριστώ τη σύζυγό μου Χρύσα για την επιμέλεια του κειμένου και την υπομονή της και τον γιο μου Διογένη. Ευχαριστώ, επίσης το προσωπικό του εργαστηρίου Pasiphae για τη συνεργασία τους και την Ελένη για τη συμβολή της στη συλλογή των δεδομένων.

Πρόλογος

Η ραγδαία επέκταση του διαδικτύου και η εισχώρησή του ακόμα και στους πιο παραδοσιακούς τομείς της ανθρώπινης δραστηριότητας έχει προκαλέσει εκθετική αύξηση των κινδύνων και των επιπτώσεων από τις απειλές που σχετίζονται με τη διασύνδεση των αυξανόμενων «έξυπνων» συσκευών και υποδομών. Η κυβερνοασφάλεια είναι ένας πολυσύνθετος τομέας με μεγάλο τεχνικό και επιστημονικό ενδιαφέρον στο οποίο δεν εργάζονται μόνο εταιρίες πληροφορικής αλλά και εξειδικευμένα ερευνητικά κέντρα, οργανισμοί και κρατικές αρχές. Τα τελευταία χρόνια, η ανάπτυξη λογισμικού ακολουθώντας στρατηγικές ασφαλείας (Secure by design), η στρατηγική «άμυνας εις βάθος» (DiD), η συνεχής παρακολούθηση των δικτύων, η διαρκής ενημέρωση σχετικά με τις αναδυόμενες απειλές (Threat Intelligence Feeds), ο έλεγχος ταυτότητας πολλαπλών συντελεστών, η εξέλιξη των αλγόριθμων κρυπτογράφησης κ.α., έχουν συμβάλει σημαντικά στη διαρκή μάχη που εξελίσσεται στον τομέα της κυβερνοασφάλειας. Οι τεχνικές των επιθέσεων έχουν όμως εξελιχθεί αντίστοιχα καθώς οι επιτιθέμενοι έχουν καλύτερη οργάνωση και μεγαλύτερη χρηματοδότηση, με αποτέλεσμα να έχουμε ποσοτική και ποιοτική αύξηση των επιτυχημένων επιθέσεων.

Σε μια τυπική υποδομή ενός Επιχειρησιακού Κέντρου Ασφάλειας (SOC), το οποίο υπάρχει σε μεγάλες επιχειρήσεις και οργανισμούς, αναλυτές ασφαλείας επεξεργάζονται δεδομένα από διάφορες πηγές. Σημάνσεις από διάφορα πρωτόκολλα (πχ. Syslog, SNMP) και ανάλυση αρχείων καταγραφής από διάφορα συστήματα (π.χ. διακομιστές διαδικτύου, εξουσιοδότησης, μεσολάβησης, DHCP, windows/linux agents και πλήθος άλλων), σε πραγματικό χρόνο, ενεργοποιούν μεγάλο αριθμό ειδοποιήσεων διαφόρων επιπέδων. Οι ειδοποιήσεις, οι οποίες προκύπτουν αυτοματοποιημένα, αξιολογούνται και συχνά η διερεύνησή τους επιβάλει τη συσχέτιση των αρχείων καταγραφής και τη διενέργεια ανά περίπτωση ελέγχων. Ο μεγάλος αυτός αριθμός ειδοποιήσεων που εμπεριέχει μεγάλο αριθμό ψευδώς θετικών, μειώνει την αποτελεσματικότητα των διαδικασιών ασφαλείας, με αποτέλεσμα σε αρκετές περιπτώσεις να δίνεται η ευκαιρία σε πραγματικές απειλές να περάσουν απαρατήρητες.

Σε αυτή την εργασία παρουσιάζεται η χρήση του Apache Spot, ένα σύστημα κυβερνοασφάλειας που βασίζεται στην ανάλυση μεγάλων δεδομένων, σε μια υποδομή που είναι ικανή να ανιχνεύσει και να αξιολογήσει απειλές με ελάχιστη παρέμβαση. Το Apache Spot (incubating) ένα έργο που έχει παραχωρηθεί στο Apache Foundation, είναι σε φάση που ακολουθείται η προβλεπόμενη διαδικασία

για την ένταξή του στα ολοκληρωμένα έργα του Ιδρύματος. Αυτό που κάνει το Apache Spot ξεχωριστό είναι η ικανότητά του να εντοπίζει ύποπτες συνδέσεις αναλύοντας μεγάλες ποσότητες δεδομένων. Αυτές οι συνδέσεις ενδέχεται να σχετίζονται με απειλές και επιθέσεις που λαμβάνουν χώρα στο δίκτυό μας. Οι ύποπτες συνδέσεις αρχικά αξιολογούνται με τη χρήση του ίδιου του Spot και με τη χρήση εξωτερικών πηγών από αναλυτές ασφαλείας και στη συνέχεια τα εμπλεκόμενα συστήματα ελέγχονται αυτοματοποιημένα ως προς την ύπαρξη ευπαθειών ασφαλείας χρησιμοποιώντας το λογισμικό OpenVAS. Μια τέτοια υποδομή δύναται να ενσωματωθεί στα συστήματα ασφαλείας υποδομών ζωτικής σημασίας.

Η παρούσα εργασία είναι δομημένη σε πέντε κεφάλαια.

Στο 1ο Κεφάλαιο γίνεται μια γενική παρουσίαση των σύγχρονων θεμάτων στον τομέα της κυβερνοασφάλειας. Σύγχρονες απειλές, επιπτώσεις, μέθοδοι αντιμετώπισής τους και σύγχρονη έρευνα για την ασφάλεια των Υποδομών Ζωτικής Σημασίας.

Στο 2ο Κεφάλαιο παρουσιάζεται το μοντέλο ανίχνευσης και αξιολόγησης απειλών που χρησιμοποιούμε στην παρούσα εργασία.

Στο 3ο Κεφάλαιο παρουσιάζεται το οικοσύστημα Apache Hadoop, ως υποστηρικτική υποδομή του Apache Spot.

Το Apache Spot ως το κεντρικό σύστημα της ανάλυσής μας, παρουσιάζεται στο 4ο κεφάλαιο. Τεχνολογίες, δομικά στοιχεία, λειτουργία, στοιχεία εγκατάστασης, παραμετροποίηση και χρήση του.

Στο Κεφάλαιο 5ο παρουσιάζεται μια περίπτωση ανάλυσης κίνησης δικτύου υποδομής ζωτικής σημασίας.

Κεφάλαιο 1 – Κυβερνοασφάλεια

1.1 Εισαγωγή

Η ασφάλεια στον κυβερνοχώρο αποτελεί μια από τις μεγαλύτερες προκλήσεις στον χώρο της Πληροφορικής. Η εικόνα, όπως διαμορφώνεται από την πληθώρα των μελετών επί του θέματος, δεν εμπνέει αισιοδοξία. Παραθέτουμε μερικά χαρακτηριστικά στοιχεία από πρόσφατες έρευνες.

Ο μέσος αριθμός διαρροών ασφαλείας είχε αύξηση 11% το έτος 2018, ανεβάζοντας το σύνολο της αύξησης κατά την τελευταία 5ετία στο 67% [1].

Σε 500 εκατομμύρια εγγραφές προσωπικών δεδομένων πελατών ανέρχεται η διαρροή δεδομένων της αλυσίδας ξενοδοχείων Marriott International [2] ενώ η διαρροή δεδομένων της Yahoo επηρέασε 3 δισεκατομμύρια λογαριασμούς χρηστών [3].

Οι επιπτώσεις της διαρροής δεδομένων είναι τόσο σημαντικές στον οικονομικό τομέα γεγονός που οδήγησε οίκους αξιολόγησης, όπως η Moody's και η S&P, να συμπεριλάβουν και ελέγχους σε θέματα κυβερνοασφάλειας στις αξιολογήσεις πιστοληπτικής ικανότητας.

Στην πρόσφατη έκθεση του Παγκόσμιου Οικονομικού Φόρουμ που αφορά τους παγκόσμιους κινδύνους, στους 10 κινδύνους με τις μεγαλύτερες επιπτώσεις περιλαμβάνονται οι κυβερνοεπιθέσεις και η κατάρρευση κρίσιμων υποδομών πληροφορικής [4]. Στην ίδια έκθεση, οι διαρροές δεδομένων και οι κυβερνοεπιθέσεις περιλαμβάνονται στους 5 πιο πιθανούς κινδύνους σε παγκόσμιο επίπεδο.

Η επίθεση του λυτρισμικό WannaCry το 2017, επηρέασε 200 εκ. υπολογιστές σε 150 χώρες προκαλώντας ζημιές εκατοντάδων εκ. δολαρίων.

Σε 2,9 εκατομμύρια θέσεις εργασίας προσδιορίζει σήμερα ο οργανισμός $(ISC)^2$ το έλλειμα σε προσωπικό, εξειδικευμένο σε θέματα κυβερνοασφάλειας [5].

Το μέσο κόστος ανά διαρροή δεδομένων ανέρχεται σε 3.9 εκ. δολάρια και περιέχει κατά μέσο όρο περίπου 25,500 εγγραφές ενώ ο μέσος χρόνος αναγνώρισης και περιορισμού των διαρροών εκτιμάται στις 279 ημέρες. Τα στοιχεία προέχονται από έρευνα που έγινε σε 500 εταιρίες ανά τον κόσμο που υπέστησαν διαρροή δεδομένων [6].

Όταν οι επιθέσεις αφορούν υποδομές ζωτικής σημασίας το κόστος δεν είναι μόνο οικονομικό. Επηρεάζεται η ασφάλεια, η καθημερινή ζωή των πολιτών και η ομαλή λειτουργία της κοινωνίας στο σύνολό της.

Το FBI ήδη από το 2014 συμπεριλαμβάνει στον κατάλογο με τους πιο καταζητούμενους ανθρώπους, εγκληματίες που έχουν διαπράξει εγκλήματα στον κυβερνοχώρο. Έκτοτε ο αριθμός τους σε αυτή τη λίστα συνεχώς και αυξάνεται. Παράλληλα, το ποσό των 15 δις δολαρίων περιλαμβανόταν στον προϋπολογισμό των ΗΠΑ του 2019 για την κυβερνοασφάλεια και τα μεγαλύτερα ποσά τα έλαβαν τα υπουργεία Αμύνης και Εσωτερικής Ασφάλειας.

Το πιο γνωστό πλαίσιο κυβερνοασφάλειας είναι το αμερικάνικο NIST Cybersecurity Framework [7] που αρχικά σχεδιάστηκε για την προστασία υποδομών ζωτικής σημασίας και έχει υιοθετηθεί από πολλά άλλα κράτη καθώς και από τον ιδιωτικό τομέα. Το πλαίσιο αυτό ορίζει σε υψηλό επίπεδο ως βασικές λειτουργίες της διαχείρισης και περιορισμού του κινδύνου κυβερνοασφάλειας, την αναγνώριση, προστασία, ανίχνευση, απόκριση και ανάκτηση. Ακολουθεί περιγραφή των βασικών λειτουργιών:

Πίνακας 1: Βασικές λειτουργίες πλαισίου κυβερνοασφάλειας NIST

Αναγνώριση	Ανάπτυξη της επιχειρησιακής κατανόησης για την διαχείριση του κινδύνου των συστημάτων, των δεδομένων, των περιουσιακών στοιχείων και των λειτουργιών.
Προστασία	Ανάπτυξη και υλοποίηση των κατάλληλων μέτρων προστασίας που διασφαλίζουν την παροχή των υπηρεσιών των υποδομών ζωτικής σημασίας.
Ανίχνευση	Ανάπτυξη και υλοποίηση των κατάλληλων ενεργειών για την ανίχνευση των περιστατικών κυβερνοασφάλειας.
Απόκριση	Ανάπτυξη και υλοποίηση των κατάλληλων ενεργειών αντιμετώπισης των περιστατικών κυβερνοασφάλειας που ανιχνεύτηκαν.
Ανάκτηση	Ανάπτυξη και υλοποίηση των κατάλληλων ενεργειών για την ανάπτυξη ανθεκτικότητας και για την αποκατάσταση των λειτουργιών που επηρεάστηκαν από τα περιστατικά κυβερνοασφάλειας.

Οι πέντε αυτές βασικές κατηγορίες αναλύονται σε υποκατηγορίες, οι οποίες εξετάζονται διεξοδικά με τη βοήθεια εκτενέστατου υποστηρικτικού υλικού, το οποίο περιλαμβάνει προδιαγραφές,

κατευθυντήριες γραμμές και πρακτικές που είναι κοινές για τους τομείς των υποδομών ζωτικής σημασίας. Το αποτέλεσμα του πλαισίου είναι η δημιουργία μιας προσαρμοσμένης θεώρησης για τη δημιουργία και βελτίωση της κυβερνοασφάλειας. Η υποδομή και η μεθοδολογία που παρουσιάζεται στην παρούσα εργασία εντάσσεται στην κατηγορία ανίχνευσης απειλών.

1.2 Σύγχρονες Απειλές

Σύμφωνα με το γενικό πλαίσιο της ασφάλειας, το ζητούμενο από τα συστήματα προστασίας είναι η διασφάλιση του απορρήτου, της ακεραιότητας των πληροφοριών και της διαθεσιμότητας των συστημάτων. Οι πρόσφατες μελέτες δείχνουν ότι οι επιθέσεις με τις μεγαλύτερες επιπτώσεις στην ασφάλεια είναι το κακόβουλο λογισμικό (malware), διαδικτυακές επιθέσεις (web-based), επιθέσεις άρνησης υπηρεσίας, κακόβουλοι εισβολείς, ηλεκτρονικό ψάρεμα (phishing) και κοινωνική μηχανική, κακόβουλος κώδικας, κλεμμένες συσκευές, λυτρισμικό (ransomware) και τα botnets.

Το παραδοσιακό μοντέλο προστασίας δικτύων περιλαμβάνει, μεταξύ άλλων, τη χρήση τοίχου προστασίας, την πρόσβαση στα συστήματα με χρήση διακομιστών εξουσιοδότησης, την ενημέρωση των προγραμμάτων και του λειτουργικού συστήματος με ενημερώσεις ασφαλείας, τη χρήση προγραμμάτων προστασίας από ιούς και την παρακολούθηση αρχείων καταγραφής μέσω ειδοποιήσεων. Δύο μεγάλες αλλαγές συντελούνται τα τελευταία χρόνια κάνοντας το έργο της ασφάλισης των συστημάτων έναντι απειλών ακόμα πιο δύσκολο.

Η άνοδος του IoT και η ραγδαία ανάπτυξη των ενσωματωμένων συστημάτων -που αναμένεται ότι θα κλιμακώνονται τα χρόνια που ακολουθούν - δημιουργεί πραγματικά ένα νέο τοπίο στον τομέα της ασφάλειας. Να σημειωθεί ότι το IoT θα έχει μεγάλη επίδραση και στη λειτουργία των υποδομών ζωτικής σημασίας. Η σύνδεση νέων και μεγάλου αριθμού συσκευών, οι οποίες διαθέτουν μικρά λειτουργικά συστήματα και δεν έχουν σχεδιαστεί με υψηλές προδιαγραφές ασφαλείας αποτελούν μια μεγάλη απειλή. Από τους απλούς δικτυακούς εκτυπωτές και τις κάμερες οικιακής χρήσης έως τους αξονικούς τομογράφους των νοσοκομείων και τα X-rays των αεροδρομίων κάθε διασυνδεμένη συσκευή μπορεί - εφόσον διαφεύγει της επίβλεψης ειδικών ασφαλείας- , να αποτελέσει πρόβλημα για την ασφάλεια ενός δικτύου. Έχει διαπιστωθεί ότι απλές δικτυακές συσκευές συνδέονται στο διαδίκτυο χωρίς να γίνει αλλαγή των προεπιλεγμένων κωδικών πρόσβασης του κατασκευαστή αποτελώντας μια σοβαρή ευπάθεια ενός δικτύου. Μια ελλιπώς θωρακισμένη δικτυακή συσκευή στο εσωτερικό ενός δικτύου μπορεί να αποτελέσει έναν πολύτιμο σύμμαχο για τους επιτιθέμενους. Χαρακτηριστική περίπτωση απειλής από συσκευές IoT και ενσωματωμένα συστήματα είναι το Mirai

Botnet [8] που δημιουργήθηκε το 2016, στο οποίο ενεπλάκησαν 600 χιλιάδες συσκευές και χρησιμοποιήθηκε κυρίως για επιθέσεις DDoS.

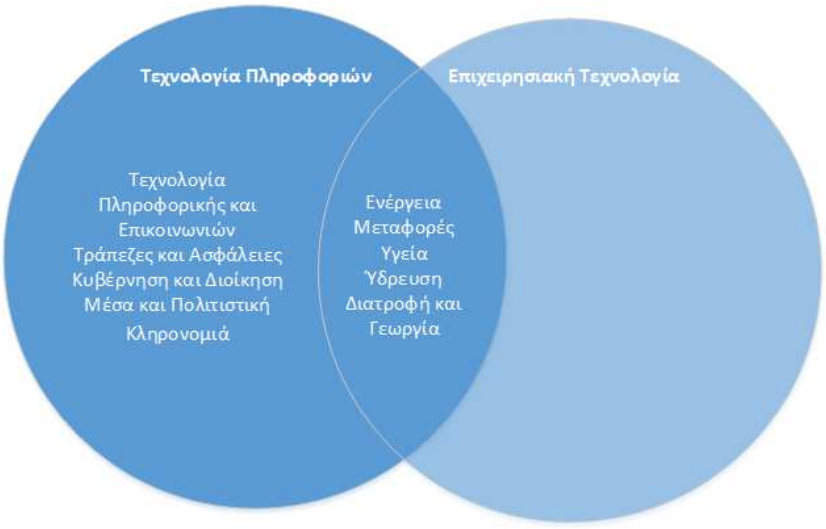
Στα σύγχρονα περιβάλλοντα εργασίας συναντούμε συχνά την πρακτική να μπορεί το προσωπικό να χρησιμοποιεί τη δική του συσκευή (φορητό υπολογιστή, κινητό τηλέφωνο, ταμπλέτα) για τη σύνδεσή του στο εταιρικό δίκτυο. Βασικό κίνητρο αυτής της σύγχρονης τάσης (BYOD) είναι η αύξηση της παραγωγικότητας και η μείωση του κόστους. Η τακτική αυτή δημιουργεί προβλήματα στον τομέα της κυβερνοασφάλειας καθώς οι συσκευές αυτές λειτουργούν και εκτός των ασφαλισμένων και υπό παρακολούθηση εταιρικών δικτύων σε οικιακά και σε δημόσια δίκτυα. Οι εργαζόμενοι μπορεί επίσης να μοιράζονται τη συσκευή αυτή με άλλα πρόσωπα (π.χ. του οικογενειακού κύκλου) τα οποία συνήθως δεν έχουν την «κουλτούρα ασφαλείας» που απαιτείται στα υπηρεσιακά και εταιρικά περιβάλλοντα εργασίας. Οι συσκευές αυτές είναι εκτεθειμένες σε πλήθος κινδύνων, με σημαντικότερο εξ αυτών την εισαγωγή κακόβουλου λογισμικού [9]. Με την σύνδεσή τους στα εταιρικά δίκτυα μπορούν να μεταφέρουν το κακόβουλο λογισμικό εντός αυτού παρακάμπτοντας τα μέτρα ασφαλείας. Για την αντιμετώπιση των θεμάτων που ανακύπτουν από την πρακτική BYOD, ιδιαίτερες πολιτικές ασφαλείας εφαρμόζονται σε συσκευές που συνδέονται και σε μη ασφαλή δίκτυα, αλλά επηρεάζουν και την γενικότερη στρατηγική ασφαλείας του δικτύου.

1.3 Κυβερνοασφάλεια υποδομών ζωτικής σημασίας

Οι υποδομές που χαρακτηρίζονται ως ζωτικής σημασίας, δεν είναι ίδιες σε όλα τα κράτη, υπάρχει σχετική διαφοροποίηση [10]. Οι υποδομές ζωτικής σημασίας (ΥΖΣ) σύμφωνα με τον πιο πρόσφατο Ευρωπαϊκό κανονισμό παρουσιάζονται στο παρακάτω σχήμα.

Εικόνα 1: Υποδομές ζωτικής σημασίας στην ΕΕ

Ορισμένοι από τους ανωτέρω τομείς που περιλαμβάνονται στις ΥΖΣ χρησιμοποιούν κυρίως κλασικές Τεχνολογίες Πληροφοριών (IT) δηλαδή αποθήκευση, μεταφορά και επεξεργασία πληροφορίας ενώ κάποιοι χρησιμοποιούν και Επιχειρησιακή Τεχνολογία - (OT), δηλαδή ανίχνευση ή πρόκληση αλλαγών σε φυσικές διαδικασίες μέσω της άμεσης παρακολούθησης ή/και ελέγχου φυσικών συσκευών. Βλέπε παρακάτω σχήμα:



Εικόνα 2: IT και OT στις ΥΖΣ

Στους τομείς που χρησιμοποιούν Τεχνολογία Πληροφοριών, π.χ. Τραπεζικός Κλάδος, Επικοινωνίες κ.τ.λ. εφαρμόζονται οι γενικές αρχές κυβερνοασφάλειας σε μια διαρκή προσπάθεια να διασφαλιστεί υψηλός βαθμός ασφάλειας, όπως συχνά ορίζεται από το σχετικό νομοθετικό και κανονιστικό πλαίσιο που ορίζουν εθνικές και κοινοτικές ή ομοσπονδιακές νομοθεσίες και κανονισμοί.

Στους τομείς που χρησιμοποιούν και Επιχειρησιακή Τεχνολογία σημαντικό και διαρκώς αυξανόμενο ρόλο έχουν τα κυβερνο-φυσικά συστήματα (Cyber Physical Systems –CPS) και το Διαδίκτυο των Αντικειμένων. Τα συστήματα αυτά θα αποτελέσουν τον πυρήνα των υποδομών ζωτικής σημασίας [11]. Είναι η βάση των ήδη αναπτυσσόμενων αλλά και των μελλοντικών έξυπνων υπηρεσιών που θα βελτιώσουν την ποιότητα ζωής σε πολλούς τομείς. Μερικοί από αυτούς είναι η προσωποποιημένη ιατρική περίθαλψη, η αντιμετώπιση καταστάσεων εκτάκτου ανάγκης και η διαχείριση ροής κυκλοφορίας. Σε τέτοιου είδους τομείς υπάρχουν σημαντικές τεχνολογικές και λειτουργικές διαφοροποιήσεις στα θέματα ασφάλειας. Για παράδειγμα, στον τομέα της Ενέργειας σημαντικότερο ρόλο έχει η διαθεσιμότητα του συστήματος έναντι της ιδιωτικότητας των πληροφοριών, στον τομέα της Αεροναυτιλίας η ακρίβεια των παρεχόμενων στα αεροσκάφη πληροφοριών έναντι της ιδιωτικότητάς τους, ενώ στον τομέα της Υγείας η ακρίβεια και η ιδιωτικότητα των πληροφοριών και η διαθεσιμότητα συστημάτων ή υπηρεσιών είναι εξίσου σημαντικές. Η ετερογένεια των συστημάτων και η τεχνολογική τους διαφοροποίηση σε σχέση με τα κλασικά πληροφοριακά συστήματα αποτελεί πρόκληση για τον τομέα της κυβερνοασφάλειας και παρουσιάζει ιδιαίτερο ερευνητικό ενδιαφέρον.

1.4 Επιθέσεις σε ΥΖΣ

Το Μάιο του 2017 έγινε η επίθεση με το λυτρισμικό WannaCry, η οποία επηρέασε επιχειρήσεις και οργανισμούς ανά τον κόσμο και είχε ιδιαίτερες επιπτώσεις στη λειτουργία του Αγγλικού συστήματος υγείας. Πιο συγκεκριμένα 80 από τις 236 μονάδες του συστήματος υγείας είτε μολύνθηκαν είτε έκλεισαν τις συσκευές ή τα συστήματά τους, ως μέτρο προστασίας, το προσωπικό κλειδώθηκε έξω από τις συσκευές του, γεγονός που εμπόδισε ή καθυστέρησε την πρόσβαση και ενημέρωση των πληροφοριών των ασθενών, την αποστολή των εξετάσεων και τη μεταφορά ασθενών σε νοσοκομεία. Παθολογικά και ακτινολογικά τμήματα διέκοψαν τη λειτουργία τους καθώς ο διαγνωστικός εξοπλισμός δεν ήταν διαθέσιμος (π.χ. τομογράφοι) [12]. Ως αποτέλεσμα, πέντε νοσοκομεία έκαναν εκτροπή των εκτάκτων περιστατικών τους σε άλλα και συνολικά ακυρώθηκαν ή δεν πραγματοποιήθηκαν περί τα 20.000 ιατρικά ραντεβού. Το λυτρισμικό WannaCry χρησιμοποίησε την - και τότε- γνωστή ευπάθεια (CVE-2017-0145) «EternalBlue», του πρωτοκόλλου διαμοιρασμού αρχείου Server Message Block (SMB) των Windows και στη συνέχεια, πέραν της εξάπλωσής του, έκανε κρυπτογράφηση των αρχείων με το WannaCrypt. Η σύνδεσή του στο κέντρο Εντολών και Ελέγχου (C&C) γινόταν με χρήση του δικτύου ανωνυμίας TOR [13].

Η επίθεση στο σύστημα διανομής ηλεκτρικής ενέργειας της Ουκρανίας τον Δεκέμβριο 2015, γνωστή ως BlackEnergy, άφησε επί οκτάωρο 225.000 κατοίκους χωρίς ρεύμα. [14] Η αρχική εισβολή έγινε με αποστολή ηλεκτρονικού ταχυδρομείου με συνημμένο μολυσμένο έγγραφο του Word με παραλήπτες προσωπικό πληροφοριακών συστημάτων και διαχειριστές. Για την επίθεση χρησιμοποιήθηκαν τα νόμιμα πιστοποιητικά των χρηστών που αποκτήθηκαν από ηλεκτρονικό ψάρεμα τύπου spear. Η επίθεση που έγινε σε τρεις τοπικές εταιρίες διανομής ήταν συγχρονισμένη και ενορχηστρωμένη ενώ είχε προηγηθεί εκτεταμένη επίθεση αναγνώρισης των δικτύων τους. Κατά την επίθεση, ο κακόβουλος χειρισμός των διακοπτών έγινε με εργαλεία απομακρυσμένης διαχείρισης του λειτουργικού συστήματος ή με λογισμικό απομακρυσμένου πελάτη βιομηχανικού συστήματος ελέγχου με πρόσβαση μέσω εικονικών ιδιωτικών δικτύων (VPN). Στο τέλος της επίθεσης και στις τρεις περιπτώσεις σε ορισμένα συστήματα διαγράφησαν κάποια αρχεία και η κύρια εγγραφή εκκίνησης (MBR), με τη χρήση του κακόβουλου λογισμικού KillDisk θέτοντάς τα εκτός λειτουργίας. Έγινε απενεργοποίηση των μετατροπέων Serial-to-Ethernet στους υποσταθμούς καταστρέφοντας το υλικολογισμικό (firmware) τους ενώ έγινε και συγχρονισμένη απενεργοποίηση

των αδιάλειπτων τροφοδοτικών ρεύματος (UPS), που τροφοδοτούσαν τους διακομιστές, με χρήση διεπαφής απομακρυσμένης διαχείρισης. Ακολούθησε επίθεση άρνησης υπηρεσίας στο τηλεφωνικό κέντρο.

Η περίπτωση του BlackEnergy δεν είναι η μοναδική. Επίθεση σε βιομηχανικά συστήματα ελέγχου με αντίστοιχη στρατηγική είχε προηγηθεί από το κακόβουλο λογισμικό Dragonfly με σκοπό την συλλογή πληροφοριών [15]. Η επιτυχία του BlackEnergy κατέδειξε την πολυπλοκότητα των σύγχρονων επιθέσεων και την εξέλιξη του εγκλήματος στον κυβερνοχώρο. Παρόλο που τα δίκτυα των συστημάτων ελέγχου ήταν καλά διαχωρισμένα από τα επιχειρησιακά δίκτυα και προστατεύονταν από στιβαρούς τοίχους προστασίας, η επίθεση δεν απετράπη. Αξιοσημείωτο είναι ότι έγινε σε υποδομές που χρησιμοποιούνται σε πολλά αντίστοιχα βιομηχανικά συστήματα ελέγχου. Ένα επίσης σημαντικό στοιχείο είναι ότι, όπως υπολογίζεται, η αρχική εισβολή στο δίκτυο έγινε έξι μήνες πριν την επίθεση και για όλο αυτό το διάστημα δεν έγινε αντιληπτή.

1.5 Δυσκολίες των σύγχρονων μέτρων προστασίας δικτύων.

Τα σύγχρονα δίκτυα εφαρμόζουν εξειδικευμένες τεχνικές για την ελαχιστοποίηση των ευπαθειών σε όλα τα επίπεδα του προτύπου OSI και προστατεύονται από πολλαπλές συσκευές. Τα συστήματα ανίχνευσης εισβολέων (IDS), βασίζουν τη λειτουργία τους κυρίως σε δύο τεχνικές. Η μια τεχνική εστιάζει στον εντοπισμό κακόβουλου λογισμικού, εξετάζοντας την εισερχόμενη κίνηση με στόχο την εύρεση κάποιας υπογραφής κακόβουλου λογισμικού. Οι υπογραφές πρέπει να έχουν ήδη ανιχνευτεί από κάποια διαδικτυακή υπηρεσία και το σύστημά μας πρέπει να έχει ενημερωθεί σχετικά, αφήνοντας έτσι ένα χρονικό περιθώριο για το οποίο δεν υπάρχει προστασία [16]. Επίσης, οι επιτιθέμενοι έχουν αναπτύξει τεχνικές πολυμορφισμού, μεταμορφισμού κ.α. και μπορούν να παρακάμπτουν την άμυνα συστημάτων που βασίζονται στην ανίχνευση υπογραφών, όπως τοίχοι προστασίας, αντιικό λογισμικό. Τα IDS χρησιμοποιούν επίσης στατιστικές μεθόδους για την ανίχνευση εισβολής στο δίκτυο, βασιζόμενες σε αλγόριθμους μηχανικής μάθησης [17]. Αντίστοιχες μέθοδοι με χρήση αλγόριθμων μηχανικής μάθησης χρησιμοποιούνται για την ανίχνευση κακόβουλου λογισμικού ή ανεπιθύμητης αλληλογραφίας και αναφέρονται εδώ γιατί συχνά αποτελούν ένα σημαντικό βήμα της εισβολής. Η τεχνική αυτή των IDS παρουσιάζει προβλήματα καθώς εμφανίζει ψευδώς θετικές ειδοποιήσεις και η απόδοσή τους σε μεγάλα δίκτυα είναι χρονοβόρα με αποτέλεσμα τη μειωμένη απόδοσή τους. Η παραμετροποίησή τους είναι σύνθετη (δεν μπορεί να είναι αυτοματοποιημένη), είναι αναγκαία η επανεκπαίδευσή τους, ενώ απαιτούνται

διαφορετικοί αλγόριθμοι για την ανίχνευση διαφορετικών απειλών [18]. Αντιστοίχως, η μέθοδος εκ βάθους ελέγχου πακέτων (DPI) που διαθέτουν οι τοίχοι προστασίας δεν μπορεί να εφαρμοστεί αποδοτικά σε δίκτυα υψηλών ταχυτήτων. Επιπλέον, ιδιαίτερη δυσκολία υπάρχει τόσο στα IDS όσο και στους τοίχους προστασίας στην επεξεργασία κρυπτογραφημένης κίνησης.

Τα τελευταία χρόνια έχει δοθεί ιδιαίτερο βάρος στην παρακολούθηση των δικτύων και της διαδικτυακής κίνησης για την ανίχνευση ανωμαλιών, συνεπικουρώντας τα πιο παραδοσιακά συστήματα ασφαλείας. Όπως περιγράψαμε ήδη η πρακτική BYOD, το IoT αλλά και οι επιχειρησιακές ανάγκες που συχνά επιβάλουν την πρόσβαση συνεργατών (άτομα ή εταιρίες) στα εταιρικά δίκτυα μέσω εικονικών ιδιωτικών δικτύων, διευρύνουν το πεδίο των απειλών. Η πολυσύνθετη δικτυακή κίνηση, η πολυπλοκότητα των δικτύων, τα πολλά και διαφορετικά μοτίβα των επιθέσεων κάνουν την αποδοτική ανίχνευση ιδιαίτερα δύσκολη. Την εποχή των Μεγάλων Δεδομένων (Big Data), που καθημερινά μεταφέρονται μεταξύ κόμβων τεράστιες ποσότητες δεδομένων, παρέχονται ευκαιρίες σε επιτιθέμενους να εισβάλουν σε δίκτυα και να αποκρύψουν την παρουσία τους. Όπως δείχνουν όλες οι μελέτες, οι εισβολείς μπορούν να παραμένουν για αρκετό καιρό εντός των δικτύων, έχοντας τον χρόνο να οργανώσουν πολύ αποδοτικές επιθέσεις. Με τον όρο χρόνος Dwell αναφερόμαστε στο χρόνο (σε ημέρες) από την στιγμή της εισβολής μέχρι την αποβολή του εισβολέα από το δίκτυο. Περιλαμβάνει τον χρόνο που χρειάστηκε για την ανίχνευσή της και τον χρόνο για την απομόνωση και αφαίρεση των στοιχείων της επίθεσης. Σε περίπτωση εισβολής - που όπως είδαμε δεν είναι απίθανη - στόχος της ομάδας ασφαλείας είναι ο περιορισμός στο ελάχιστο του Dwell.

1.6 Security Analytics

Η εφαρμογή των μεθόδων της ανάλυσης των Μεγάλων Δεδομένων (BDA) στον τομέα της κυβερνοασφάλειας είναι γνωστή ως «Security Analytics». Είναι ο συνδυασμός λογισμικού, αλγορίθμων και αναλυτικών διαδικασιών που χρησιμοποιούνται για τον εντοπισμό πιθανών απειλών για τα πληροφοριακά συστήματα.

Ο κλασικός ορισμός του όρου Μεγάλα Δεδομένα, περιλαμβάνει την ανάλυση και επεξεργασία συνόλων δεδομένων, τα οποία είναι μεγάλου μεγέθους και αρκετά πολύπλοκα, ώστε να μπορέσουμε να τα διαχειριστούμε με κλασικές μεθόδους ανάλυσης και επεξεργασίας. Η πολυπλοκότητα των δεδομένων έγκειται στον μεγάλο Όγκο τους (Volume) -που συνήθως είναι τάξης μεγαλύτερης από terrabytes-, την Ποικιλία (Variety) της δομής τους, αναφερόμενοι στην συνύπαρξη δομημένων, αδόμητων και ημιδομημένων δεδομένων και στην Ταχύτητα (Velocity) με

την οποία δημιουργούνται. Η πιο μοντέρνα προσέγγιση του όρου αναφέρεται όχι τόσο στα χαρακτηριστικά των δεδομένων αλλά στη χρήση τους, που σχετίζεται με την προγνωστική ανάλυση, την ανάλυση συμπεριφοράς ή οποιαδήποτε άλλη προχωρημένη αναλυτική μέθοδο δεδομένων με την οποία μπορούμε να εξάγουμε αξιόλογα στοιχεία μέσα από αυτά.

Η ανάλυση των Μεγάλων Δεδομένων επιτρέπει σε ερευνητές και αναλυτές διαφορετικών επιστημονικών και επιχειρηματικών κλάδων να παίρνουν ταχύτερες και ορθότερες αποφάσεις, βασιζόμενοι σε στοιχεία που προκύπτουν από δεδομένα, τα οποία με παλιότερες τεχνικές επεξεργασίας δεν ήταν χρήσιμα ή προσπελάσιμα.

Νέοι τύποι και πηγές δεδομένων έχουν προκύψει τα τελευταία χρόνια, πηγές που σχετίζονται με τον καθημερινό τρόπο ζωής, με τις κυριότερες εξ αυτών να είναι:

- τα κοινωνικά δίκτυα, όπου οι δημοσιεύσεις, οι αναρτήσεις βίντεο, η δημιουργία σχέσεων και η κοινοποίηση απόψεων και συναισθημάτων, αποτελεί μια τεράστια πηγή δεδομένων που παρέχει ανεκτίμητης αξίας στοιχεία για τα συναισθήματα και την συμπεριφορά των χρηστών. Η ανάλυση των στοιχείων αυτών προσφέρεται για εμπορικό, κοινωνιολογικό και πολιτικό τομέα.

- συνδεδεμένες συσκευές που δημιουργούν δεδομένα, όπως αισθητήρες, βιομηχανικός εξοπλισμός ή διακομιστές που παράγουν αρχεία καταγραφής κ.α. Ο όγκος αυτού του τύπου των δεδομένων, αναμένεται να αυξηθεί εκθετικά καθώς το Διαδίκτυο των Αντικειμένων θα εισέρχεται όλο και περισσότερο στην καθημερινότητά μας. Από απλούς έξυπνους μετρητές, ιατρικές συσκευές, κάμερες παρακολούθησης έως κινητές συσκευές και ρολόγια χειρός, οι συνδεδεμένες συσκευές είναι μια διαρκώς μεγεθυνόμενη πηγή μεγάλων δεδομένων.

- διακομιστές συναλλαγών δημιουργούν δεδομένα, τα οποία παράγονται από τις καθημερινές συναλλαγές είτε είναι διαδικτυακές είτε όχι. Τέτοιου είδους δεδομένα είναι αρχεία καταγραφής από αποδείξεις, εντολές αγοράς και πληρωμής, αποδείξεις παράδοσης κ.τ.λ.

Κατά την υλοποίηση ενός συστήματος Μεγάλων Δεδομένων σημαντικές παράμετροι είναι: η συλλογή, η επιμέλεια, η αποθήκευση, η αναζήτηση, η μεταφορά, ο διαμοιρασμός, η ανάλυση και παρουσίαση των δεδομένων.

Η ιδέα της αυτοματοποιημένης ασφάλειας δικτύων που βασίζεται στα ίδια τα δεδομένα εφαρμόστηκε αρχικά στα IDS με τη λειτουργία ανίχνευσης ανωμαλιών. Τα IDS όμως δεν είναι ικανά να εκτελέσουν ανάλυση μεγάλων ποσοτήτων δεδομένων σε βάθος χρόνου. Επίσης, τα συστήματα

πληροφοριών ασφάλειας και διαχείρισης συμβάντων (SIEM) δεν είναι αποδοτικά στην ανάλυση μεγάλων, μη δομημένων και συχνά με θόρυβο δεδομένων [19]. Αντιθέτως, τα εργαλεία που διαθέτει η επεξεργασία των Μεγάλων Δεδομένων είναι κατάλληλα για την ανίχνευση των Προηγμένων Επίμονων Απειλών (APT) και για εγκληματολογική ανάλυση. Οι APT είναι επιθέσεις χαμηλού προφίλ και μακροχρόνιας διάρκειας, δηλαδή οι εισβολείς παραμένουν απαρατήρητοι για μεγάλο χρονικό διάστημα καθώς είναι ιδιαίτερα δυσδιάκριτοι. Για να εντοπιστούν τέτοιες απειλές, συχνά είναι απαραίτητο να συλλεχθούν και να συσχετιστούν μεγάλες ποσότητες δεδομένων από διαφορετικές πηγές και να γίνει ανάλυση σε βάθος χρόνου του ιστορικού του δικτύου.

Οι μέθοδοι για την ανίχνευση ανωμαλιών ανήκουν σε τρεις κατηγορίες: επιβλεπόμενες, μη επιβλεπόμενες και υβριδικές μέθοδοι και μπορούν να εφαρμοστούν σε πλήθος διαφορετικών πηγών [20]. Οι επιβλεπόμενες μέθοδοι είναι κυρίως αλγόριθμοι κατάταξης όπως: Δέντρα Απόφασης (DT), Λογιστική Παλινδρόμησης (LR), Μηχανές Διανυσμάτων Υποστήριξης (SVM), Προσαρμοστική Ώθηση (AdaBoost), Τυχαίο Δάσος (RF). Οι επιβλεπόμενοι μέθοδοι εκτελούν δυαδική ταξινόμηση και πολύ-κατάταξη ταξινόμηση. Στην πρώτη περίπτωση γίνεται χαρακτηρισμός των δεδομένων ως κανονικά ή ανώμαλα ενώ στη δεύτερη σε πολλές κατηγορίες. Όλοι οι αλγόριθμοι επιβλεπόμενης μάθησης χρησιμοποιούν σημασμένα δεδομένα για την εκπαίδευσή τους. Οι μη-επιβλεπόμενες μέθοδοι διακρίνονται σε δυο κατηγορίες: τις μεθόδους ομαδοποίησης και τις μεθόδους ακραίων σημείων. Οι μέθοδοι ομαδοποίησης βασίζουν την ανίχνευση ανωμαλιών στην υπόθεση ότι δραστηριότητες που αποκλίνουν από τις κανονικές χαρακτηρίζονται ως ανωμαλίες κάτι που δεν εμπλέκει τη δημιουργία μοτίβων [21]. Συνήθεις αλγόριθμοι ομαδοποίησης είναι η μέθοδος K- μέσων, η ιεραρχική ομαδοποίηση, ο αλγόριθμος DBSCAN ενώ στις μεθόδους ακραίων σημείων, η μέθοδος GMM και η μέθοδος μιας κλάσης. Οι αλγόριθμοι επιβλεπόμενης μάθησης έχουν καλύτερη απόδοση σε ικανότητα αναγνώρισης και σε ρυθμό ψευδών ειδοποιήσεων αλλά οι μη επιβλεπόμενης μάθησης έχουν το πλεονέκτημα ότι δεν απαιτούν σημασμένα δεδομένα την εκπαίδευσή τους και αναγνωρίζουν νέες, μη γνωστές επιθέσεις. Οι υβριδικές μέθοδοι συνδυάζουν τις δυο τεχνικές για να επιτύχουν τα πλεονεκτήματα που διαθέτουν.

1.7 Το Apache Spot σε δίκτυα ΥΖΣ

Το Apache Spot είναι σχεδιασμένο να λειτουργήσει παράλληλα με τα ήδη εγκατεστημένα συστήματα ασφαλείας και μπορεί να αποτελέσει ένα τμήμα της πολυεπίπεδης προστασίας, που στοχεύει να εντοπίσει απειλές που έχουν διαφύγει των υπολοίπων συστημάτων. Λειτουργώντας

συμπληρωματικά με τα διαθέσιμα συστήματα, η χρήση αυτού του γενικού σκοπού εργαλείου έχει ιδιαίτερα πλεονεκτήματα για τις ΥΖΣ καθώς:

- Είναι ένα σύνθετο έργο που αφορά οργανισμούς που δίνουν ιδιαίτερο βάρος στην κυβερνοασφάλεια. Οι ΥΖΣ οφείλουν να παρέχουν υπηρεσίες με την μέγιστη δυνατή διαθεσιμότητα, διασφαλίζοντας τη μέγιστη ιδιωτικότητα και ακεραιότητα.
- Οι ΥΖΣ συχνά λειτουργούν υπό κεντρικό έλεγχο και διαχείριση, εφαρμόζοντας καθολικές πολιτικές ασφάλειας και συνήθως είναι οργανισμοί με μεγάλα δίκτυα που παράγουν αρχεία καταγραφής μεγάλου όγκου.
- Είναι χρήσιμο σε δίκτυα που χρησιμοποιούνται κυβερνοφυσικά συστήματα, ενσωματωμένα συστήματα και συσκευές IoT όπως οι ΥΖΣ. Το Spot λειτουργεί με δεδομένα τηλεμετρίας που παράγονται αυτόματα και δεν απαιτεί τη συνεργασία των επιμέρους συσκευών.
- Μπορεί να εντοπίσει αόρατες απειλές που σχετίζονται με στοχευμένες επιθέσεις APT που είναι ένας μεγάλος κίνδυνος για τις ΥΖΣ.
- Είναι ένα εργαλείο που βοηθά στην ταχύτατη ιχνηλάτηση συμβάντων, άρα μπορεί να βοηθήσει στον περιορισμό των επιπτώσεων μιας επιτυχημένης επίθεσης.
- Είναι κατάλληλο για εγκληματολογική ανάλυση καθώς διατηρεί και τα πρωτότυπα, μη επεξεργασμένα δεδομένα. Οι επιθέσεις σε ΥΖΣ συχνά είναι κακουργηματικές πράξεις και αποδεικτικά στοιχεία δύναται να ζητηθούν από δικαστικές αρχές.
- Μπορεί να δώσει αξία σε αρχεία καταγραφής που ήδη τηρούνται.

Εστιάζοντας στον χώρο της Υγείας, ένα παράδειγμα συνδεδεμένης συσκευής είναι ένας μαγνητικός τομογράφος, μηχάνημα υψηλού κόστους που μπορεί να λειτουργεί για πολλά χρόνια (δεκαετίες) χωρίς καμία ενημέρωση του λειτουργικού συστήματος και των εφαρμογών του. Εφόσον καλύπτει τις ιατρικές προδιαγραφές, δεν αντικαθίσταται από κάποιο νεότερο στο πλαίσιο της στρατηγικής αναβάθμισης συστημάτων που προτείνεται από πλευράς κυβερνοασφάλειας. Σε ένα τέτοιο σύστημα δεν μπορεί να εγκατασταθεί κάποιο επιπλέον πρόγραμμα προστασίας ή κάποιος πράκτορας λογισμικού που θα παράγει αρχεία καταγραφής ή μηνύματα, ώστε να ενημερωθεί κάποιο σύστημα προστασίας ότι δέχεται επίθεση. Έτσι κάποιος που έχει διαφύγει από τα συστήματα ασφαλείας και έχει εισβάλει στο εσωτερικό του δικτύου, μπορεί να κάνει ανίχνευση θυρών (η αρχή μιας σύνθετης επίθεσης) στον τομογράφο χωρίς κανείς να τον αντιληφθεί. Μια τέτοια απειλή, το Spot έχει μεγάλες πιθανότητες να την εντοπίσει.

1.8 Σύγχρονες ερευνητικές προτάσεις για την ασφάλεια των ΥΖΣ

Ο τομέας των Τηλεπικοινωνιών έχει ξεχωριστό ρόλο καθώς πολλές άλλες ΥΖΣ βασίζονται στις Τηλεπικοινωνίες για να λειτουργήσουν. Στο πλαίσιο αυτό, η εργασία της Maria Belesioti και άλλων [22] παρουσιάζουν τους στόχους του προγράμματος RESISTO της Ε.Ε., το οποίο αφορά την πολύπλευρη και ολοκληρωμένη κατανόηση της ασφάλειας και αύξησης της ανθεκτικότητας των Τηλεπικοινωνιών ζωτικής σημασίας. Η βασική ιδέα είναι η ανάπτυξη ενός πλαισίου που θα επιτρέπει διαφορετικά τμήματα του συνόλου του προσωπικού ασφαλείας των Τηλεπικοινωνιών ζωτικής σημασίας να λειτουργεί αποδοτικότερα. Ανταλλαγή δεδομένων και σημάτων, αναγνώριση επιθέσεων με πολύπλοκα μοτίβα από διαφορετικές πηγές και επίπεδα, προσομοίωση σε πραγματικό χρόνο της εξάπλωσης της απειλής εντός της ΥΖΣ και μεταξύ των συνδεδεμένων ΥΖΣ, είναι εργαλεία για την επιλογή και υλοποίηση της καταλληλότερης αντίδρασης στην απειλή. Σε περίπτωση αποτυχίας της προστασίας, στόχος είναι η καταλληλότερη επιλογή των μέτρων περιορισμού των επιπτώσεων. Επιπλέον, στόχος του RESISTO είναι η ανάπτυξη μιας δομής συστημάτων που συνθέτουν ένα οικοσύστημα τεχνολογικών καινοτομιών και λειτουργικών μοντέλων. Αυτά περιλαμβάνουν ενσωματωμένη αξιολόγηση ανθεκτικότητας, ταχύτερης ανίχνευσης απειλών/επιθέσεων, καλύτερης πληροφόρησης για λήψη αποφάσεων και ολιστική αντίληψη της κατάστασης, σε εικονικό και φυσικό επίπεδο, για τις διασυνδεδεμένες υποδομές.

Η πολυπλοκότητα των κυβερνο-φυσικών αλληλεπιδράσεων και η αλληλεξάρτηση των φυσικών δικτύων είναι σημαντικά ζητήματα που επηρεάζουν την κυβερνοασφάλεια στις ΥΖΣ, οι οποίες συνήθως δεν είναι ένα σύστημα αλλά «σύστημα συστημάτων». Ο Q. Zhu και άλλοι [23] προτείνουν στην εργασία τους μια προσέγγιση για δυναμική λήψη αποφάσεων για την προστασία των ΥΖΣ, η οποία βασίζεται στον περιορισμό του κινδύνου. Η προσέγγισή τους είναι ένα εφαρμόσιμο μοντέλο για την αξιολόγηση του κινδύνου λαμβάνοντας υπόψιν την κυβερνο-φυσική αλληλεπίδραση και αλληλεξάρτηση. Συμπεριλαμβάνει την εξάπλωση της επίθεσης σε κάθε σύστημα και τις δυσλειτουργίες που εξαπλώνονται στο φυσικό δίκτυο. Με βάση αυτή την αξιολόγηση, γίνεται υπολογισμός του καθαρού κέρδους της στρατηγικής άμυνας. Περιγράφεται ο αλγόριθμός για να φτιαχτεί η βέλτιστη ολική στρατηγική άμυνας, βασιζόμενη στην στρατηγική άμυνας κάθε

υποσυστήματος, υπερπηδώντας το πρόβλημα των ανταγωνιστικών τοπικών στρατηγικών. Η προσέγγισή τους αξιολογείται σε προσομοιωμένο περιβάλλον δικτύου παροχής νερού.

Για την ασφάλεια των υποδομών που συνδέονται σε ένα μη ασφαλές δίκτυο, όπως το Διαδίκτυο, οι Α. Π. Φούρναρης [24] και άλλοι, προτείνουν μια μονάδα υλισμικού ασφαλείας (HMS) ικανή να παρέχει ισχυρή κρυπτογράφηση και υπηρεσίες ασφαλείας στις ΥΖΣ. Πολλές από τις συσκευές που λειτουργούν σε ΥΖΣ είναι παλαιές, δεν διαθέτουν σύγχρονες προδιαγραφές ασφαλείας και δεν έχουν αρκετή υπολογιστική ισχύ, ώστε να μπορέσουν να ανταποκριθούν σε αυτές. Η μονάδα HMS, η οποία έχει αναπτυχθεί με ασφάλεια κατά τον σχεδιασμό, λειτουργεί ως έμπιστος πυλώνας και αναλαμβάνει την εμπιστευτικότητα, την ακεραιότητα, την αυθεντικότητα αλλά και την επαλήθευση και ταυτοποίηση χρηστών για λογαριασμό της φυσικά συνδεδεμένης σε αυτήν συσκευή. Η προτεινόμενη μονάδα είναι ένα σύγχρονο SoC με μικροεπεξεργαστή ARM με υποστήριξη TrustZone που συνδέεται με μια σειρά IP πυρήνων, που λειτουργούν ως επιταχυντές κρυπτογράφησης.

Η ασφάλεια του φυσικού επιπέδου των συστημάτων ΥΖΣ αφορά και την προστασία της λειτουργίας των συσκευών που αλληλεπιδρούν με το περιβάλλον, όπως οι αισθητήρες και οι ενεργοποιητές. Είναι σύνηθες οι συσκευές αυτές να συνδέονται σε κόμβους (οι οποίοι επικοινωνούν μεταξύ τους και έχουν ικανότητα εκτέλεσης υπολογισμών) και η ύπαρξη κακόβουλων κόμβων (που εσκεμμένα δεν αποστέλλουν πραγματικές μετρήσεις ή και δεν ενεργούν σύμφωνα με τις εντολές) είναι μεγάλη απειλή για την λειτουργία μιας ΥΖΣ. Χαρακτηριστική επίθεση τέτοιου τύπου είναι ο Stuxnet. Με την τεχνική ενεργούς ασφάλειας «ψηφιακής υδατοσφράγισης» που προτείνουν ο B. Satchidanandan και άλλοι [25], εισάγονται κρυφά σήματα διέγερσης στους ενεργοποιητές, τα οποία μπορούν να ανιχνευτούν κατά μήκος βρόχων κλειστών συστημάτων και να εντοπιστούν κόμβοι με κακόβουλη συμπεριφορά. Οι κόμβοι αυτοί αποκαλύπτονται καθώς στέλνουν αλλοιωμένες μετρήσεις.

Τα βιομηχανικά συστήματα ελέγχου (ICS) βασίζονται σε Σύστηματα Εποπτικού Ελέγχου και Απόκτησης Δεδομένων (SCADA), κατανεμημένα συστήματα ελέγχου (DCS) , προγραμματιζόμενους λογικούς ελεγκτές (PLC) και έχουν κεντρικό ρόλο στην λειτουργία πολλών από τις ΥΖΣ. Ο C. Chang και άλλοι [26] προτείνουν ένα σύστημα για την προστασία τους, βασισμένο στην ανίχνευση ανωμαλιών στα δεδομένα. Στην αρχιτεκτονική τους υιοθετούν δυο μεθόδους ανίχνευσης ανωμαλιών με μηχανική μάθηση οι οποίες εφαρμόζονται παράλληλα. Η μέθοδος ομαδοποίηση K-μέσων είναι υπεύθυνη για την εκτίμηση των χαρακτηριστικών κάθε ιδιότητας των δεδομένων ανά χρονική στιγμή ενώ η μέθοδος συνελικτικών αυτοκωδικοποιητών (CAE) παρακολουθεί κάθε ιδιότητα για διαδοχικές χρονικές στιγμές. Τα δύο μοντέλα είναι ημι-επιβλεπόμενης μάθησης και στη

μελέτη τους εκπαιδεύτηκαν σε σύνολο δεδομένων από συστήματα δεξαμενών αποθήκευσης νερού και συστήματα αγωγών φυσικού αερίου. Τα αποτελέσματά τους συγκρίθηκαν με άλλες αντίστοιχες μεθόδους και προέκυψαν καλύτεροι δείκτες ανάκλησης, ακρίβειας και βαθμού F1.

Αντίστοιχα ο T. Alves και άλλοι [27] προτείνουν στην εργασία τους μια λύση για την βελτίωση της ασφάλειας σε ICS, με την προσθήκη κρυπτογράφησης σε PLC και παρακολούθησή τους από IPS. Πιο συγκεκριμένα, η εναλλακτική αρχιτεκτονική τους προσθέτει ένα επίπεδο κρυπτογράφησης μεταξύ του επιπέδου δικτύου της πλατφόρμας OpenPLC (συμβατή με IEC 61131-3) και του IPS. Η κρυπτογράφηση γίνεται με AES-256 και τη χρήση προ-διαμοιρασμένου κλειδιού (PSK) και είναι συμβατή με τα πρωτόκολλα που υποστηρίζει το OpenPLC. Από την πλευρά των συστημάτων SCADA είναι απαραίτητο να εκτελείται τοπική ασφαλής πύλη για την αποκρυπτογράφηση της επικοινωνίας. Το IPS (βασισμένο σε αλγόριθμο ομαδοποίησης K-μέσων) μπορεί να ανιχνεύσει ανωμαλίες στο δίκτυο και επιτρέπει την επικοινωνία μόνο των έμπιστων κόμβων με τα PLC, λειτουργώντας ως διακομιστής μεσολάβησης TCP. Η αρχιτεκτονική είναι ανθεκτική σε επιθέσεις ένθεσης, υποκλοπής, άρνησης υπηρεσίας κ.α.

Στον τομέα της Υγείας, για την ασφάλεια των συστημάτων διαρκούς παρακολούθησης της υγείας (CHMS) προτείνεται από τον K. K. Venkatasubramanian και άλλους [28] το Physiology-based System-wide Information Security (PySIS). Τα δεδομένα των χρηστών συλλέγονται από αισθητήρες και μεταφέρονται σε αποθηκευτικό νέφος για την ενημέρωση του ιατρικού ιστορικού τους. Η μεταφορά τους χωρίς παραβίαση της ιδιωτικότητάς τους ή την παραποίησή τους είναι πολύ σημαντικό ζήτημα ακόμα και για την υγεία τους. Το προτεινόμενο σύστημα είναι μια ενοποιημένη λύση ασφάλειας κυβερνοφυσικού συστήματος για την ασφαλή συλλογή και μεταφορά τέτοιου είδους δεδομένων. Υλοποιείται (1) με διαδικασία πιστοποίησης συμφωνίας κλειδιού για απόκρυψη – αποκάλυψη βασισμένη σε σήματα φυσιολογίας και (2) σε ένα παραγωγικό φυσιολογικό μοντέλο, που παράγει συνθετικά σήματα φυσιολογίας με κλινική συσχέτιση, βασισμένα σε αναλυτικό μοντέλο που έχει εκπαιδευτεί χρησιμοποιώντας στατιστικά από φυσιολογικά σήματα που έχουν συλλεχθεί από τον χρήστη.

Η ασφάλεια των ιατρικών δεδομένων είναι από τα σοβαρότερα θέματα της κυβερνοασφάλειας στον χώρο αυτό. Ο Q. Xia και άλλοι [29] προτείνουν ένα σύστημα που διασφαλίζει τα δεδομένα κατά τη διαμοίρασή τους μεταξύ θεματοφυλάκων μεγάλων δεδομένων. Κατά την κοινή χρήση των δεδομένων που είναι αποθηκευμένα σε αποθετήρια αποθηκευτικού νέφους μεταξύ ερευνητικών, ιατρικών και άλλων ιδρυμάτων, δύναται να υπάρξει διαρροή τους, γεγονός που αποτελεί μεγάλο

κίνδυνο για τη φήμη τους και έχει σοβαρές οικονομικές επιπτώσεις. Η προτεινόμενη λύση βασίζεται στη χρήση αλυσίδας συστοιχιών (Blockchain) που παρέχει στοιχεία προέλευσης, παρακολούθησης και ελέγχου για ιατρικά δεδομένα που διακινούνται σε τέτοια αποθετήρια μεγάλων δεδομένων. Το MeDShare, όπως ονομάζεται η εφαρμογή που προτείνουν, χρησιμοποιεί έξυπνες συμβάσεις και έναν μηχανισμό ελέγχου πρόσβασης για την αποτελεσματική παρακολούθηση της συμπεριφοράς των δεδομένων και την ανάκληση της πρόσβασης σε οντότητες, όταν ανιχνευτεί παραβίαση σε κανόνες και δικαιώματα. Η εφαρμογή εκτός του επιπέδου της υπάρχουσας βάσης δεδομένων έχει τρία επίπεδα: το επίπεδο χρήστη, το επίπεδο ερωτημάτων των δεδομένων και το επίπεδο δόμησης και προέλευσης. Το τελευταίο επίπεδο επεξεργάζεται τα αιτήματα πρόσβασης στα δεδομένα και δίνει τις εξουσιοδοτήσεις. Επιπλέον, εκτελεί υπολογισμούς πάνω στα δεδομένα και προσθέτει ετικέτες με λειτουργίες που παρακολουθούν όλες τις ενέργειες που εκτελούνται πάνω σε αυτά. Αλγόριθμοι και δομές εφαρμόζονται για να αναφέρουν ενέργειες που αποθηκεύονται με ασφάλεια σε βάση δεδομένων και, αν χρειαστεί, να ενεργοποιήσουν την παρακολούθηση των δεδομένων. Τα αποτελέσματα των ενεργειών ανακοινώνονται σε ένα δίκτυο που δεν επιτρέπει την τροποποίησή τους.

Για τη θωράκιση του συνόλου των συστημάτων που χρησιμοποιούνται στον τομέα της Υγείας ο Ε. Μαρκάκης και άλλοι [30] προτείνουν τη δημιουργία ενός συστήματος Αξιολόγησης Ασφαλείας, ως Υπηρεσία, που αναγνωρίζει τις ευπάθειες των συσκευών, τις αξιολογεί προληπτικά και περιορίζει τις απειλές. Το σύστημα που δύναται να ενσωματωθεί στην Πληροφοριακή υποδομή των ιδρυμάτων, παρακολουθεί τις υπάρχουσες και νέες συσκευές, που εισέρχονται στο δίκτυο σε πραγματικό χρόνο, με τη χρήση υποδομής SDN. Οι συσκευές τοποθετούνται προσωρινά σε ένα περιορισμένης συνδεσιμότητας ουδέτερο δίκτυο, όπου γίνονται οι έλεγχοι και η αξιολόγηση. Εκεί πραγματοποιείται η βαθμολόγηση της υπό εξέτασης συσκευής με βάση ένα πρότυπο σύστημα βαθμολόγησης. Ακολουθεί –εφόσον η αξιολόγηση είναι θετική- η παροχή συνδεσιμότητας και δικαιωμάτων στο δίκτυο και η εξουσιοδότηση υπηρεσιών. Με τον τρόπο αυτό, και χωρίς ο χρήστης να έχει γνώση θεμάτων ασφαλείας, δεν επιτρέπεται η τοποθέτηση μη έμπιστων συσκευών στο επιχειρησιακό περιβάλλον των ιδρυμάτων υγείας. Επιπλέον, αξιολογήσεις ασφαλείας διενεργούνται περιοδικά για να επιβεβαιωθεί η ακεραιότητα και η διαθεσιμότητα του δικτύου.

Για την ασφάλεια του ηλεκτρονικού ιατρικού εξοπλισμού ο A.Rao και άλλοι [31] προτείνουν ένα μοντέλο για τη δυναμική αναγνώριση και αξιολόγηση του κινδύνου και την αυτοματοποιημένη λήψη μέτρων για τον περιορισμό του. Το μοντέλο αυτό, για την αναγνώριση των απειλών σε

πραγματικό χρόνο και την υλοποίηση προσαρμοζόμενων πολιτικών και σχημάτων προστασίας, αναπτύσσεται κατά τον σχεδιασμό της συσκευής. Η ασφαλής αρχιτεκτονική ανάπτυξης επιτρέπει την απομόνωση μονάδων υλικού/υλισμικού για την προσαρμογή της τοπολογίας του συστήματος, ανάλογα με τη μεταβολή του κινδύνου. Χρησιμοποιεί συναρτήσεις συσσωρευτικής κατανομής (CDFs) για την μοντελοποίηση της κανονικής συμπεριφοράς της συσκευής, με στόχο την ποσοτικοποίηση της ομοιότητας των απειλών. Αυτός ο πιθανολογικός ανιχνευτής, που αξιολογεί και διαχειρίζεται τον κίνδυνο, δίνει ακριβή ενημέρωση για τον τρέχοντα κίνδυνο του συστήματος. Μειώνεται έτσι στο ελάχιστο ο αριθμός των ψευδώς θετικών ενεργοποιήσεων των σχημάτων προστασίας, τα οποία μπορούν να προκαλέσουν περιορισμούς στις λειτουργίες της συσκευής. Η διαδικασία πρέπει να γίνεται εντός των αυστηρών χρονικών ορίων που καθορίζουν οι λειτουργίες της συσκευής. Ένα παράδειγμα, με βάση το μοντέλο αυτό, εφαρμόζεται σε έξυπνο συνδεδεμένο βηματοδότη.

Μια αρχιτεκτονική ανθεκτική σε επιθέσεις, σε ιεραρχικά συστήματα ελέγχου CPS προτείνεται από τον Y. Won και άλλους [32]. Τέτοια συστήματα είναι τα συστήματα ελέγχου τρένων, τα οποία αποτελούνται από φυσικά συστήματα εξοπλισμένα με τοπικούς ενσωματωμένους ελεγκτές, μια μονάδα υψηλής εποπτείας για τον έλεγχο της λειτουργίας του φυσικού επιπέδου και το δίκτυο που τα διασυνδέει. Για τα επιμέρους αυτά υποσυστήματα, η προτεινόμενη αρχιτεκτονική περιλαμβάνει στοιχεία που βελτιώνουν την αντοχή του συστήματος. Για τα φυσικά συστήματα έχει αναπτυχθεί ένας αλγόριθμος ανίχνευσης επιθέσεων που επιτρέπει τη λειτουργία παρά την αποτυχία αισθητήρων, αρκεί ένας μόνο να παραμένει σε λειτουργία. Για τους ενσωματωμένους ελεγκτές μετά από ανίχνευση αποτυχίας, ενεργοποιείται ένας μηχανισμός μεταγωγής από μονάδες υψηλής απόδοσης σε υψηλής διασφάλισης. Για την επίθεση στο δίκτυο προτείνεται μια μέθοδος λογισμικά καθοριζόμενου δικτύου για ανάκτηση των δικτύων σε πραγματικό χρόνο μετά από επίθεση ή σφάλμα.

Πολλές ομάδες εργασίας έχουν συσταθεί από τους εμπλεκόμενους διεθνείς οργανισμούς (ICAO, EUROCAE) για την ανάπτυξη των μελλοντικών Αεροναυτικών επικοινωνιών, που θα είναι βασισμένες στη σουίτα πρωτοκόλλων IPv6 (ATN/IP). Η μετάβαση από τις παραδοσιακές επικοινωνίες -που είναι αυτόνομα δίκτυα - σε IP εκθέτουν τις αεροπορικές επικοινωνίες στους κινδύνους κυβερνοασφάλειας. Ο M. Niraula και άλλοι [33] [14] παρουσιάζουν ένα σχήμα αμφίπλευρης, αμοιβαίας πιστοποίησης και ακεραιότητας για τις επικοινωνίες - εδάφους αέρος,

βασιζόμενο στο πρωτόκολλο ασφάλειας επιπέδου μεταφοράς δεδομενογραμμάτων (DTLS) που έχει επεκταθεί και χρησιμοποιεί κρυπτογράφηση ελλειπτικής καμπύλης (ECC). Καθώς τα αεροσκάφη(Α/Φ) έχουν διεπαφές για πολλές υπηρεσίες, στην προτεινόμενη αρχιτεκτονική κεντρικό ρόλο έχουν οι επίγειες πύλες μηνυμάτων, οι οποίες θα παρέχουν μεταξύ άλλων την απαραίτητη αδιάλειπτη κινητικότητα των κόμβων (αεροσκάφη) και τη συμβατότητα με τα παλαιότερα συστήματα σε ένα ασφαλές πλαίσιο DTLS με τη χρήση ψηφιακών πιστοποιητικών. Περιγράφεται μια αξιόπιστη και ανθεκτική υποδομή δημοσίων κλειδιών (PKI) για τη διαχείριση των πιστοποιητικών και η δημιουργία ενός πρωτοτύπου για την απόδειξη της ορθότητας της ιδέας, σε ένα σύστημα που θα αποτελέσει ένα από τα μεγαλύτερα ασύρματα δίκτυα και χρήζει της μέγιστης δυνατής ασφάλειας.

Στον ίδιο τομέα της Αεροναυτιλίας οι υπηρεσίες επιτήρησης με την χρήση ADS-B (αποστολή στίγματος Αεροσκάφους) είναι ευάλωτες καθώς το ADS-B αναπτύχθηκε χωρίς να ληφθούν υπόψιν τα σοβαρά θέματα ασφάλειας που προκύπτουν. Είναι τεχνικά πολύ απλό -ειδικά με τη χρήση λογισμικά καθοριζόμενων ασύρματων διατάξεων (SDR)- να τροποποιήσει ή να εισάγει ψεύτικα δεδομένα στο σύστημα επικοινωνίας για να αλλοιώσει τη θέση του αεροσκάφους ή να εμφανίσει ανύπαρκτα Α/Φ επηρεάζοντας την εναέρια κυκλοφορία. Πολλαπλά ψευδώς εμφανιζόμενα Α/Φ μπορούν να υπερφορτώσουν τα συστήματα επεξεργασίας εκτελώντας έτσι επιθέσεις άρνησης υπηρεσίας. Η τεχνική της ασφαλούς ταυτοποίησης με τη χρήση κρυπτογράφησης δεν αποτελεί επιλογή καθώς η δομή των πακέτων του πρωτοκόλλου ADS-B δεν επιτρέπει την προσθήκη των απαραίτητων επιπρόσθετων bits. Ο Y. Kim και άλλοι προτείνουν [34] την προσθήκη σε κάθε μήνυμα μιας μικρής χρονοσφραγίδας -κάτι το οποίο είναι εφικτό-, η οποία μπορεί να χρησιμοποιηθεί από τον δέκτη για την αναγνώριση των ψευδών μηνυμάτων. Με βάση τη διαφορά της χρονικής στιγμής που ελήφθη ένα μήνυμα και τη χρονοσφραγίδα που εμπεριέχει, μπορεί να εξακριβωθεί, αν ο αποστολέας βρίσκεται στην απόσταση στην οποία δηλώνει. Ο έλεγχος αυτός είναι αρκετά αξιόπιστος, όταν γίνεται σε μια ακολουθία μηνυμάτων.

Για τον περιορισμό της έκθεσης των δικτύων διανομής νερού σε κυβερνοφυσικές επιθέσεις, ο N. Νικολάου και άλλοι [35] προτείνουν στην εργασία τους την εκμετάλλευση των αναλυτικών εφεδρειών των φυσικών ιδιοτήτων του συστήματος. Εισάγουν μια μεθοδολογία για τον υπολογισμό του επιπέδου κυβερνοφυσικής ασφαλείας π.χ. για τον υπολογισμό των κυβερνο-στοιχείων που απαιτούνται να θιγούν για να υπάρξει επίδραση στον έλεγχο του συστήματος. Επιπλέον, εξετάζουν τη χρήση των αναλυτικών εφεδρειών με τη βοήθεια νέων πρόσθετων μονάδων λογισμικού. Με

δεδομένο τον γράφο που περιγράφει τα εικονικά και φυσικά στοιχεία του συστήματος, καταδεικνύουν ποιο γνωσιακό μοντέλο και εννοιολογική λογική μπορούν να χρησιμοποιηθούν, για να προκύψουν νέες συνδέσεις που αυξάνουν το αριθμό των υπολογισμένων καταστάσεων του συστήματος. Αυτό βοηθάει στον περιορισμό των πιθανών επιπτώσεων από την επίθεση σε κάποιο στοιχείο του συστήματος, αφού πιθανώς να μπορεί να αντικατασταθεί από μια υπολογισμένη τιμή του ίδιου ή κάποιου άλλου αλγόριθμου.

Ο A. Mathur στην εργασία του [36] προτείνει ένα πολυεπίπεδο πλαίσιο που συμβάλει στην διασφάλιση των σταθμών επεξεργασίας υδάτων. Το πλαίσιο με την ονομασία SecWater, περιλαμβάνει 7 επίπεδα ασφαλείας, κάθε ένα εκ των οποίων εμπεριέχει μονάδες υλικού/υλισμικού για την αποτροπή, ανίχνευση εικονικών ή φυσικών επιθέσεων και για να είναι εφικτός ο έλεγχος μετά την ανίχνευσή τους. Τα δυο πρώτα επίπεδα είναι τοίχος προστασίας και συστήματα ανίχνευσης εισβολής (IDS). Για την περίπτωση που ο επιτιθέμενος έχει εισβάλει στο δίκτυο, στο επόμενο επίπεδο διενεργείται εις βάθος ανάλυση πακέτων της δικτυακή κίνησης για να γίνει επεξεργασία δεδομένων ελέγχου κατάστασης (ανίχνευση ανωμαλίας διεργασιών). Στο επόμενο επίπεδο και για την άμυνα σε επιθέσεις πολλαπλών σημείων, αλληλεξαρτώμενες λειτουργίες που ισχύουν πάντα (invariants) κατανέμονται σε PLC, SCADA και HMI δίνοντας τη δυνατότητα αναγνώρισης της επίθεσης. Στο πέμπτο επίπεδο ένας ανεξάρτητος μηχανισμός ελέγχου (ορθογώνια άμυνα) προστίθεται για την αναγνώριση δυναμικών επιθέσεων πολλαπλών σημείων. Στα επόμενα επίπεδα έχουμε ανάλυση και επικύρωση σε πραγματικό χρόνο των εντολών που στέλνονται από τα PLC στους ενεργοποιητές και τη χρήση επαναπαραμετροποιήσιμου ελέγχου, για την επαναφορά του συστήματος σε λειτουργική κατάσταση σε περίπτωση επιτυχημένης επίθεσης. Το μοντέλο μπορεί να επεκταθεί για την εφαρμογή του πλήθος ΥΖΣ.

Όταν οι ΥΖΣ απλώνονται σε μεγάλη γεωγραφική περιοχή, όπως τα δίκτυα διανομής νερού και ρεύματος, σιδηροδρομικά δίκτυα κ.α., κυβερνοεπιθέσεις μπορούν να γίνουν από ευπάθειες που προϋποθέτουν φυσική παρουσία πλησίον των υποδομών π.χ. μέσω ασύρματων δικτύων. Καλά οργανωμένες και χρηματοδοτούμενες ομάδες δύναται να πραγματοποιήσουν τέτοιες επιθέσεις και μπορούν να αντιμετωπιστούν λαμβάνοντας τα κατάλληλα ανά περίπτωση μέτρα, όπως είδαμε σε κάποιες από τις πρόσφατες εργασίες που παρουσιάστηκαν παραπάνω. Ο μεγάλος κίνδυνος όμως παραμένει η σύνδεση των υποδομών στο διαδίκτυο καθώς ο επιτιθέμενος, βρισκόμενος χιλιάδες χιλιόμετρα μακριά, στην ασφάλεια του γραφείου του, έχει άπλετο χρόνο να εντοπίσει ευπάθειες, να καθορίσει την στρατηγική του και να οργανώσει τις επιθέσεις του. Όμως το ίδιο το δίκτυο, που

αποτελεί τον πιο σημαντικό κίνδυνο, μπορεί να μας δώσει στοιχεία για τις επιθέσεις που εκτελούνται.

1.9 Ερευνητική δραστηριότητα για το Apache Spot.

Το Spot είναι από τα επιλεγμένα έργα των προγράμματων SHIELD και EVOLVE του ερευνητικού προγράμματος Horizon 2020 της Ευρωπαϊκής Ένωσης, στα πλαίσια των οποίων υπάρχουν δύο ερευνητικές εργασίες:

Στην εργασία τους ο Χ. Μαθάς και άλλοι [37] κάνουν αξιολόγηση της ικανότητας του Spot να ανιχνεύει επιθέσεις σε περιβάλλοντα Δικτύωσης βάσει λογισμικού και Εικονικοποίησης δικτυακών λειτουργιών (SDN/VNF). Πιο συγκεκριμένα, εκτελούνται επιθέσεις: άρνησης υπηρεσίας Slowloris, πλημμύρας UDP και διοχέτευσης DNS. Τα αποτελέσματα είναι αρκετά ικανοποιητικά για την επίθεση Slowloris, η οποία εκτελείται με ανταλλαγή μικρού αριθμού πακέτων και λιγότερο αποδοτική στις άλλες δύο επιθέσεις που απαιτούν μεγαλύτερη κίνηση. Σύμφωνα με τη μελέτη, είναι εμφανές ότι ο αλγόριθμός έχει καλή απόδοση σε επιθέσεις χαμηλού ρυθμού και η απόδοσή του βελτιώνεται όσο αυξάνεται η φυσιολογική κίνηση σε σχέση με τον όγκο των επιθέσεων. Έτσι, προκύπτει ότι ο αλγόριθμος είναι ευάλωτος σε επιθέσεις δηλητηρίασης, που μπορούν να τον οδηγήσουν σε εσφαλμένα αποτελέσματα.

Ο Α. Πριόβολος και άλλοι [38] αντικαθιστούν τον ενσωματωμένο στο Spot αλγόριθμο μηχανικής μάθησης LDA, με την τεχνική βαθιάς μάθησης των Αυτοκωδικοποιητών, ώστε να βελιτώσουν την απόδοσή του σε ανάλυση δεδομένων netflow. Ο Αυτοκωδικοποιητής υλοποιείται σε τρία επίπεδα με το επίπεδο εισόδου να έχει 18 νευρώνες, όσες και οι παράμετροι των ροών που αποτελούν την είσοδο του αλγορίθμου. Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε είναι η tahn και η εκπαίδευση του αλγορίθμου έγινε με βελτιστοποιητή Adam για 90 εποχές. Η υλοποίηση έγινε με την επέκταση Elephas, του api Keras σε Spark. Χρησιμοποιώντας ένα δημόσιο σύνολο δεδομένων από τοπικό δίκτυο, που περιλαμβάνει κανονική κίνηση και επιθέσεις, έγινε σύγκριση των αποτελεσμάτων σε διάφορες κατηγορίες επιθέσεων. Σύμφωνα με τους συγγραφείς υπάρχει καλύτερη απόδοση σε όλες τις περιπτώσεις σε σχέση με τον αλγόριθμο LDA.

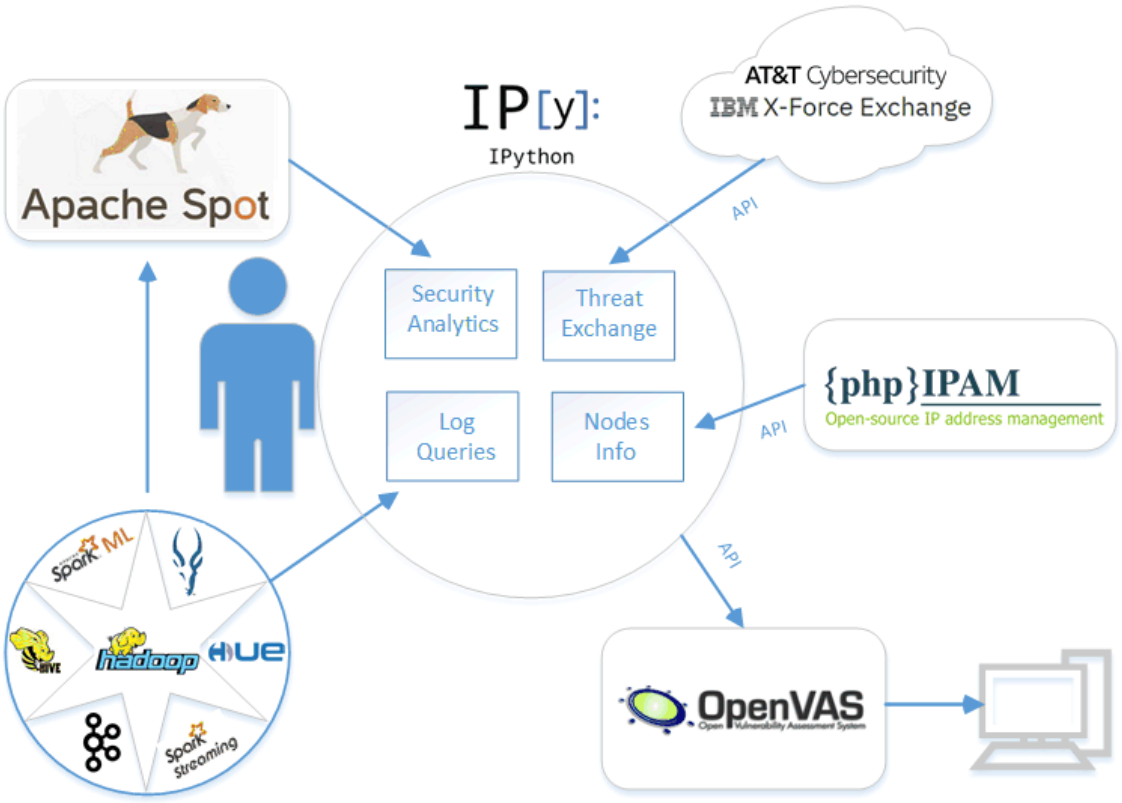
Κεφάλαιο 2 - Ανάλυση κίνησης δικτύου με χρήση του Apache Spot

2.1 Εισαγωγή

Η αξιοποίηση των αρχείων καταγραφής, που παράγονται από διαφορετικά συστήματα και εφαρμογές στον τομέα της κυβερνοασφάλειας, είναι αρκετά πολύπλοκη καθώς τα αρχεία αυτά είναι ετερογενή και περιλαμβάνουν πολλές και διαφορετικού τύπου πληροφορίες. Όταν το δίκτυο έχει μεγάλο αριθμό κόμβων –όπως είναι σύνηθες σε υποδομές ζωτικής σημασίας - η ανάλυση γίνεται ακόμα πιο δύσκολη. Επιπλέον, οι συσκευές BYOD δεν είναι εύκολο να ρυθμιστούν, ώστε να αποστέλλουν αρχεία καταγραφής σε εξωτερικούς διακομιστές και συσκευές κυβερνοφυσικές και IoT συχνά δεν παράγουν αρχεία καταγραφής ή, αν παράγουν, δεν είναι αξιοποιήσιμα από πλευράς ασφαλείας. Ένα επιπλέον πρόβλημα είναι ότι ένα κακόβουλο λογισμικό έχει τη δυνατότητα να απενεργοποιήσει την παραγωγή αρχείων καταγραφής για να αποκρύψει την δραστηριότητά του. Αντιθέτως, τα αρχεία καταγραφής που παράγονται από την δικτυακή δραστηριότητα των κόμβων, όπως αρχεία netflow, proxy κ.α. είναι ομοιογενή και δεν προϋποθέτουν τη συνεργασία ή την παραμετροποίηση του κόμβου για την παραγωγή τους. Η ανάλυση των αρχείων καταγραφής της κίνησης του δικτύου με τα κατάλληλα εργαλεία είναι μια πολύ αποδοτική μέθοδος για τον εντοπισμό απειλών καθώς μας παρέχουν ομοιογενή και πλήρη εικόνα της δικτυακής δραστηριότητας των κόμβων.

2.2 Μεθοδολογία

Στην παρούσα εργασία παρουσιάζεται η εφαρμογή ενός μοντέλου εντοπισμού και εξέτασης-αξιολόγησης συστημάτων για τα οποία υπάρχουν ενδείξεις έκθεσης σε κίνδυνο (Indication of Compromise- IoC), σε ένα δίκτυο ΥΖΣ. Το μοντέλο αυτό χρησιμοποιεί σύγχρονες τεχνολογίες Μεγάλων Δεδομένων και Τεχνητής Νοημοσύνης. Παράλληλα αξιοποιεί τις συντονισμένες προσπάθειες της κοινότητας της κυβερνοασφάλειας για τον εντοπισμό, καταγραφή και κοινοποίηση των απειλών σε παγκόσμιο επίπεδο. Στόχος είναι η ανάλυση της κίνησης του δικτύου και η ιχνηλάτηση απειλών, ώστε να προκύψει μια μικρή λίστα συστημάτων για τα οποία οι ενδείξεις έκθεσης σε κίνδυνο να είναι αρκετά ισχυρές. Η μικρή αυτή λίστα οδηγείται σε λογισμικό αξιολόγησης ευπάθειας για την εύρεση των τρωτών σημείων των συστημάτων που σχετίζονται με τις απειλές.



Εικόνα 3: Ανάλυση κίνησης δικτύου με Apache Spot

Η καρδιά του συστήματος (βλέπε παραπάνω σχήμα) είναι το Apache Spot το οποίο έχει τη δυνατότητα να αναλύει αρχεία καταγραφών -συγκεκριμένων τύπων- και να εντοπίζει τις ύποπτες συνδέσεις ανάμεσα σε δισεκατομμύρια. Για την αποθήκευση και επεξεργασία των αρχείων καταγραφής με τα οποία τροφοδοτείται, χρησιμοποιεί την ευρείας χρήσης πλατφόρμα Μεγάλων Δεδομένων, το Apache Hadoop. Στενά συνδεδεμένο με την πλατφόρμα αυτή και μετά από μια σειρά από διακριτά στάδια: πρόσληψη και εισαγωγή δεδομένων, επεξεργασία με αλγόριθμο μηχανικής μάθησης και ανάλυση ύποπτων συνδέσεων, προσθήκη εννοιολογικού πλαισίου και οπτικοποίηση, το Apache Spot παρουσιάζει στον αναλυτή ασφάλειας τις ύποπτες συνδέσεις. Η ανάλυση των ύποπτων συνδέσεων γίνεται με τον ιδιαίτερα δημοφιλή αλγόριθμό (μη επιβλεπόμενης) μηχανικής μάθησης LDA, με λίγο διαφορετικό τρόπο λειτουργίας, ώστε να λαμβάνει υπόψη του την βαθμολόγηση των συνδέσεων που κάνει ο αναλυτής, μετατρέποντας τη λειτουργία σε ημί-επιβλεπόμενης μάθησης. Το αποτέλεσμα είναι η λίστα των ύποπτων συνδέσεων, εμπλουτισμένη με συμπληρωματικά για την ανάλυση στοιχεία, όπως γεωγραφικά στοιχεία, όνομα τομέα της εξωτερικής διεύθυνσης IP και από στοιχεία που παρέχονται από υπηρεσίες φήμης σχετικά με

απειλές, όπως το McAfee Global Threat Intelligence[1] και το Facebook Threat Exchange[2]. Συμπληρωματικά με τα δεδομένα, στοιχεία από τις αλληλεπιδράσεις μεταξύ των υπόπτων συνδέσεων και αποδοτική οπτικοποίηση, δίνουν στον αναλυτή μια λειτουργική και αξιοποιήσιμη εικόνα των συνδέσεων που ο αλγόριθμος χαρακτηρίζει ύποπτες.

Για την περαιτέρω αξιολόγηση των αποτελεσμάτων, στη μέθοδο ανάλυσης που χρησιμοποιούμε, δίνεται βαρύτητα στην αξιοποίηση της διαρκούς προσπάθειας της παγκόσμιας κοινότητας και της βιομηχανίας Κυβερνοασφάλειας για εντοπισμό και άμεση κοινοποίηση των απειλών. Όπως προαναφέρθηκε, η επιχειρησιακή ανάλυση του Spot, χρησιμοποιεί τις υπηρεσίες ανταλλαγής στοιχείων απειλών McAfee GTI (παρέχεται επί μισθώματος) ή Facebook Threat Exchange (παρέχεται επιλεκτικά). Επιπρόσθετα αυτών, με τη διεπαφή Jupyter Notebook[3] και τα παρεχόμενα API, εξετάζουμε τη φήμη των ύποπτων εξωτερικών διευθύνσεων με τη χρήση των διαδικτυακών υπηρεσιών Open Threat Exchange[4] της εταιρίας AT&T (παρέχεται δωρεάν) και X-Force Exchange[5] της εταιρίας IBM (παρέχει δωρεάν για 5000 ερωτήματα/μήνα). Οι δύο αυτές ιδιαίτερα επιτυχημένες υπηρεσίες είναι συνεργατικές πλατφόρμες, στις οποίες συμμετέχουν δεκάδες χιλιάδες εταιρίες, παρέχοντας πληροφορίες από τα συστήματα ασφαλείας τους σχετικά με ιούς, κακόβουλο λογισμικό και άλλες κυβερνοεπιθέσεις. Οι πληροφορίες αυτές συγκεντρώνονται, επικυρώνονται και δημοσιοποιούνται. Έτσι IP διευθύνσεις ή ιστοχώροι από τους οποίους εκτελούνται επιθέσεις, μόλις εντοπιστούν αξιολογούνται με δείκτη επικινδυνότητας και είναι στη διάθεση όσων μετέχουν στις υπηρεσίες αυτές. Πιο συγκεκριμένα, στο μοντέλο αξιολόγησης που χρησιμοποιούμε, για κάθε εξωτερική IP διεύθυνση, την οποία ελέγχουμε ως εμπλεκόμενη σε ύποπτη δραστηριότητα του δικτύου μας, με τα παρεχόμενα APIs, μπορούμε να αντλήσουμε πληροφορία για:

- την βαθμολογία από την αξιολόγησή της. Στα κριτήρια της βαθμολόγησης συμπεριλαμβάνεται και η συνολική εικόνα του δικτύου στο οποίο ανήκει η συγκεκριμένη διεύθυνση.

[1] https://www.mcafee.com/enterprise/en-us/threat-center/global-threat-intelligence-technology.html
[2] https://developers.facebook.com/docs/threat-exchange/v2.12
[3] https://jupyter.org
[4] https://otx.alienvault.com
[5] https://exchange.xforce.ibmcloud.com

- το όνομα τομέα και από πότε αυτός είναι ενεργός. Είναι συνήθης τακτική κατά την εκτέλεση επιθέσεων να δημιουργούνται επί τούτου νέοι τομείς, ώστε να μην υπάρχει αρνητικό ιστορικό, οπότε η ημερομηνία ενεργοποίησής του είναι σημαντική πληροφορία.
- τη χώρα και την εταιρία στην οποία ανήκει. Χώρες με ελλιπή νομοθεσία σε θέματα κυβερνοασφάλειας, υψηλά επίπεδα διαφθοράς ή χαλαρή συνεργασία με διεθνείς διωκτικές αρχές είναι πρόσφορες για την εκτέλεση επιθέσεων.
- άλλα ονόματα τομέα που έχει ή ονόματα με τα οποία είχαν καταχωρηθεί παλαιότερα (passive DNS). Η συχνή μετονομασία του/των τομέων της διεύθυνσης, ενδεχομένως αποτελεί ύποπτη συμπεριφορά.

Οι ανωτέρω πληροφορίες σταθμίζονται στο σύνολό τους από τους αναλυτές καθώς το Apache Spot έχει την ιδιαίτερη ικανότητα να εντοπίζει και απειλές για τις οποίες δεν υπάρχουν πρότερες καταγραφές, άρα οι πηγές τους δεν τυγχάνουν αρνητικής αξιολόγησης.

Η εξαγωγή συμπερασμάτων είναι μια σύνθετη διαδικασία και ο αναλυτής πρέπει να έχει καλή γνώση του δικτύου, περιλαμβανομένων και λεπτομερειών σχετικά με τους εξεταζόμενους κόμβους. Καθώς το Spot είναι κατάλληλο για μεγάλα δίκτυα, ο αναλυτής αναγκάζεται συχνά να αναζητά πληροφορίες για τους κόμβους. Το λογισμικό διαχείρισης δ/νσεων IP, phpIPAM[6] είναι ένας αναλυτικός κατάλογος για τους κόμβους του δικτύου, διαθέτει πλήθος λειτουργιών και μπορεί μέσω του REST API να μας δώσει για κάθε IP δ/νση πληροφορίες, όπως το όνομα κόμβου, την περιγραφή, τον ιδιοκτήτη, σημειώσεις, θέση κ.α. Με την χρήση του πελάτη API σε γλώσσα python που διαθέτει, έχει ενσωματωθεί στο Advanced Mode και παρουσιάζει τις απαραίτητες αυτές πληροφορίες, βοηθώντας στην πιο αποδοτική και ορθή αξιολόγηση των ροών.

Έχοντας αποθηκεύσει τα αρχεία καταγραφής του δικτύου στο Apache Hadoop και χρησιμοποιώντας υπηρεσίες του οικοσυστήματός του, όπως το Apache Hive και το Apache Impala μπορούμε να έχουμε άμεση πρόσβαση σε αυτά για τη δημιουργία ερωτημάτων, που θα δώσουν επιπλέον πληροφορίες για την εξέταση της σύνδεσης. Για τον σκοπό αυτό κάνουμε χρήση του Hue[7] το οποίο είναι μια διαδικτυακή διεπαφή για εύκολη και αποδοτική υποβολή SQL ερωτημάτων σε βάσεις και αποθήκες δεδομένων. Το Hue έχει δυνατότητα αποθήκευσης των ερωτημάτων και οπτικοποίηση των αποτελεσμάτων. Να σημειωθεί ότι οι επιθέσεις ΥΖΣ είναι συνήθως κακουργηματικές πράξεις

[6] https://phpipam.net
[7] https://gethue.com

και οι φορείς που τις διαχειρίζονται πρέπει να μπορούν να στοιχειοθετήσουν συγκεκριμένες κατηγορίες ενώπιον δικαστικών αρχών. Τα αρχεία με την καταγραφή της δραστηριότητας του δικτύου, όπως αποθηκεύονται στο Hadoop, αποτελούν πολύ καλή πηγή για εγκληματολογική έρευνα παρέχοντας ιδιαίτερη ευκολία και ευελιξία για άντληση στοιχείων μέσα από τα ακατέργαστα δεδομένα.

Συνεκτιμώντας όλα τα στοιχεία, ο αναλυτής κρίνει, αν ο κόμβος που εμπλέκεται στην υπό εξέταση ύποπτη σύνδεση χρήζει αξιολόγηση ασφαλείας. Με το εργαλείο OpenVas[8] -το οποίο ακολουθεί πρότυπα του NIST και του Γερμανικό CERT-Bund - γίνονται μια σειρά ελέγχων χρησιμοποιώντας μια μεγάλη και ενημερωμένη βάση δεδομένων με τις γνωστές ευπάθειες και αδυναμίες των συστημάτων. Τα αποτελέσματα που προκύπτουν από την αξιολόγηση εξετάζονται σε συνάρτηση με την ύποπτη σύνδεση και, εφόσον υπάρχει συσχέτιση, ακολουθούνται οι διαδικασίες που προβλέπονται από τα εγχειρίδια κυβερνοασφάλειας.

Οι υπηρεσίες που χρησιμοποιούμε διατίθενται δωρεάν και τα εργαλεία είναι ανοικτού κώδικα, αλλά μπορούν να αντικατασταθούν από άλλα αντίστοιχα, εφόσον διαθέτουν τα απαιτούμενα APIs. Για παράδειγμα, το OpenVAS μπορεί να αντικατασταθεί από οποιοδήποτε λογισμικό αξιολόγησης ευπαθειών, όπως π.χ. το Nessus[9], το λογισμικό διαχείρισης IP δ/σεων phpIPAM[10] με π.χ. το NetBox[11] της DigitalOcen, το BrightCloud® Threat Intelligence[12] της Webroot για παροχή πληροφοριών σχετικά με δ/νσεις IP κα.

Αντίστοιχο έργο με το Apache Spot (ανίχνευση ανωμαλιών κυβερνοασφάλειας σε δεδομένα τηλεμετρίας μεγάλου όγκου) είναι το Apache Metron, το οποίο είναι επίσης ιδιαίτερα ενδιαφέρον και αποτελεί την εξέλιξη του έργου OpenSOC της Cisco, που παραχωρήθηκε στο Ίδρυμα Apache. Βασίζεται και αυτό στο οικοσύστημα Hadoop και χρησιμοποιεί Kafka, Apache Storm[13] και Elasticsearch[14]. Είναι μια επεκτάσιμη πλατφόρμα που υποστηρίζει πλήρη καταγραφή πακέτων, συνάθροιση δεδομένων τηλεμετρίας, επεξεργασία συνεχούς ροής κατά δεσμίδες και πραγματικού χρόνου. Αναλύει δεδομένα διαφόρων τύπων, όπως pcap, netflow, bro, snort, fireye και sourcefire και έχει δυνατότητα προσθήκης επιπλέον πηγών δεδομένων. Τα δεδομένα μπορούν να

[8] https://www.openvas.org
[9] https://www.tenable.com/products/nessus
[10] https://phpipam.net
[11] https://github.com/netbox-community/netbox
[12] https://www.brightcloud.com
[13] https://storm.apache.org/
[14] https://www.elastic.co/

εμπλουτιστούν σε πραγματικό χρόνο με πληροφορίες σχετικά με απειλές, γεωγραφικής θέσης, πληροφορίες DNS κτλ. Τα εμπλουτισμένα δεδομένα αποθηκεύονται σε λίμνες δεδομένων για μεγάλα διαστήματα με χαμηλό κόστος και παρέχει μηχανισμούς για εύκολη αναζήτηση στοιχείων και για επιχειρησιακή ανάλυση. Το Metron πραγματοποιεί επεξεργασία δεδομένων σε πραγματικό χρόνο, περιλαμβάνει συνήθεις λειτουργίες SIEM και έχει δυνατότητα προσθήκης αλγορίθμων μηχανικής μάθησης.

2.3 Σενάριο Χρήσης

Ένα σενάριο χρήσης της μεθοδολογίας που περιγράφουμε είναι η ανίχνευση beaconing σε ένα δίκτυο. Το beaconing είναι η ανά διαστήματα σύνδεση του κακόβουλου λογισμικού που έχει τοποθετηθεί σε έναν κόμβο, με κάποιο εξωτερικό διακομιστή διοίκησης και ελέγχου (C&C), ο οποίος είναι υπό τον έλεγχο του επιτιθέμενου. Σκοπός αυτής της επικοινωνίας, η οποία είναι προκαθορισμένη και ασύγχρονη είναι η λήψη εντολών για την εκτέλεση από τον κόμβο, κακόβουλων ενεργειών. Ο μολυσμένος κόμβος συνήθως, κάνει επιθέσεις σε συστήματα χωρίς να γίνεται αντιληπτός από τα συστήματα περιμετρικής ασφάλειας του δικτύου, αποστέλλει δεδομένα που έχει συλλέξει από το εσωτερικό του δίκτυο (π.χ. ανεβάζοντάς τα σε μια καθόλα νόμιμη υπηρεσία διαδικτυακής αποθήκευσης) ή συμμετέχει σε botnet για την πραγματοποίηση κατανεμημένων επιθέσεων άρνησης υπηρεσίας. Το beaconing είναι συνήθως επικοινωνία χαμηλής κίνησης, γίνεται σε αραιά χρονικά διαστήματα και χρησιμοποιεί πρωτόκολλα που χρησιμοποιούνται συχνά, οπότε είναι πολύ δύσκολο να εντοπιστεί από SIEM ή από τοίχους προστασίας. Το beaconing είναι μια από τις τεχνικές που χρησιμοποιούνται από τις APT οι οποίες είναι εκ των μεγαλύτερων απειλών κυβερνοασφάλειας των ΥΖΣ. Καθώς πρόκειται για αυτοματοποιημένη επικοινωνία, δύναται να εντοπιστεί από τον αλγόριθμο μηχανικής μάθησης του Spot αφού το μοτίβο της επικοινωνίας είναι αρκετά διαφορετικό από το μοτίβο της επικοινωνίας που παράγεται από την συνηθισμένη δικτυακή δραστηριότητα. Υπάρχουν ασφαλώς και άλλες αυτοματοποιημένες περιοδικές επικοινωνίες, οι οποίες περιλαμβάνονται στην κανονική δραστηριότητα ενός δικτύου, όπως το πρωτόκολλο NTP, περιοδικές ενημερώσεις εφαρμογών κ.τ.λ. οι οποίες, εφόσον εντοπιστούν από τον αλγόριθμο, μπορούν να χαρακτηριστούν χαμηλού ρίσκου από τον αναλυτή στο πλαίσιο της ανατροφοδότησης του αλγόριθμου.

Ας υποθέσουμε λοιπόν ότι στα αποτελέσματα του Spot υπάρχει - μεταξύ πολλών άλλων- σαν ύποπτη σύνδεση, μια καταγραφή από την επικοινωνία με έναν διακομιστή C&C (γεγονός που αρχικά δεν γνωρίζουμε). Από την πληροφορία φήμης (που έχει προστεθεί κατά την επιχειρησιακή ανάλυση των αποτελεσμάτων του αλγόριθμου) ο χαρακτηρισμός της IP δ/νσης από το McAfee GTI είναι ουδέτερος. Από την γεωτοποθεσία προκύπτει ότι η IP δ/νση είναι στη Βραζιλία, χώρα με υψηλή θέση στην εγκληματικότητα στον κυβερνοχώρο και η χρονοσφραγίδα της επικοινωνίας δείχνει ότι η επικοινωνία έγινε απογευματινή ώρα, κατά την οποία ο κόμβος δεν χρησιμοποιείται. Εξετάζοντας την εν λόγω δ/νση στις υπηρεσίες OTX και IBM X-Force προκύπτει ότι δεν υπάρχει κάποια καταγραφή ύποπτης δραστηριότητας από αυτή την IP δ/νση, αλλά το όνομα του τομέα δημιουργήθηκε πριν μερικές ημέρες. Μέσω ερωτήματος στο Apache Hue ζητούμε τις συνδέσεις του δικτύου μας προς την δ/νση αυτή από όπου προκύπτει ότι τις τελευταίες ημέρες υπάρχουν συνδέσεις ανά τακτά διαστήματα κατά τις οποίες ο κόμβος μας αποστέλλει σταθερό αριθμό bytes. Η δραστηριότητα αυτή αποτελεί ισχυρή ένδειξη ότι ο κόμβος έχει προσβληθεί από κακόβουλο λογισμικό. Ο κόμβος εξετάζεται ενδελεχώς, ως προς το λογισμικό του, και οδηγείται σε αξιολόγηση ευπάθειας από το OpenVas (βλέπε παρακάτω) για τον εντοπισμό της ευπάθειας που επέτρεψε την τοποθέτηση ή τη λειτουργία του κακόβουλου λογισμικού. Με το παραπάνω παράδειγμα βλέπουμε πως μπορούμε να εντοπίσουμε, σχετικά εύκολα, ένα πολύ σοβαρό πρόβλημα για την ασφάλεια του δικτύου μας ακόμα και αν αυτό δεν έχει καταγραφεί σε κάποια σχετική υπηρεσία.

Τη μεθοδολογία που περιγράφουμε θα χρησιμοποιήσουμε για την ανάλυση της κίνησης πραγματικού δικτύου ΥΖΣ, κίνηση που συλλέξαμε στο πλαίσιο της παρούσας εργασίας.

2.4 OpenVAS

Το OpenVAS (Open Vulnerability Assessment Scanner) είναι εργαλείο σάρωσης για την αξιολόγησης ευπαθειών δικτύων και συστημάτων. Αναπτύσσεται από την εταιρία Greenbone και είναι ανοικτού κώδικα που διατίθεται με Γενική Δημόσια Άδεια GNU. Αποτελεί μέρος του εμπορικού πακέτου διαχείρισης ευπαθειών "Greenbone Vulnerability Manager" (GVM). Είναι μια ολοκληρωμένη μηχανή σάρωσης η οποία τροφοδοτείται από διαρκώς ενημερωμένη και επεκτεινόμενη δημόσια ροή ελέγχων ευπαθειών (Network Vulnerability Tests -NVTs), η οποία περιλαμβάνει περισσότερους από 50.000 ελέγχους. Οι έλεγχοι για την αποκάλυψη ευπαθειών και κενών ασφάλειας γίνεται με την εκτέλεση επιθέσεων συχνά όμοιων με πραγματικές [39].

Το OpenVAS αξιοποιεί την NVD (National Vulnerability Database) [40], το προτυποποιημένο αποθετήριο δεδομένων της κυβέρνησης των ΗΠΑ, για την διαχείριση της ευπάθειας των συστημάτων. Τα δεδομένα παρουσιάζονται με το SCAP (Security Content Automation Protocol), τη σουίτα προδιαγραφών του NIST που επιτρέπουν την αυτοματοποίηση της διαχείρισης ευπάθειας, της μέτρησης ασφαλείας και τη συμμόρφωση. Το SCAP προτυποποιεί την ανταλλαγή περιεχομένου αυτοματισμού ασφαλείας που χρησιμοποιείται για την αξιολόγηση της συμβατότητας της παραμετροποίησης και για την ανίχνευση της ύπαρξης ευάλωτων εκδόσεων λογισμικού. Το ίδιο περιεχόμενο SCAP μπορεί να χρησιμοποιηθεί από πολλαπλά εργαλεία για την εκτέλεση μιας συγκεκριμένης αξιολόγησης που περιγράφεται από το περιεχόμενο. Περιλαμβάνει γλώσσες, συλλογές, μετρικά κ.α., εκ των οποίων το OpenVAS χρησιμοποιεί:

- τη γλώσσα OVAL (Open Vulnerability and Assessment Language) για την περιγραφή ευπαθειών, ρυθμίσεων παραμετροποίησης, κατάλογο ενημερώσεων και εφαρμογών.
- το CVE (Common Vulnerabilities and Exposure) για την κεντρική αναφορά στο μοναδικό αναγνωριστικό κάθε ευπάθειας.
- ο CPE (Common Platform Enumeration) για κοινή ονοματοδοσία και αναφορά των πληροφοριακών συστημάτων .
- το πρότυπο CVSS (Common Vulnerability Scoring System) για την κατηγοριοποίηση και βαθμολόγηση των ευπαθειών.

Με αναφορά το σημείο από όπου εκτελούνται οι έλεγχοι το OpenVAS μπορεί να χρησιμοποιηθεί για τον έλεγχο του δικτύου εξωτερικά, από την DMZ και εσωτερικά, ώστε κατά αντιστοιχία να εντοπιστούν προβλήματα στην περιμετρική ασφάλεια του δικτύου, ευπάθειες που μπορούν να

αξιοποιηθούν, αν παρακαμφθεί το τοίχος προστασίας και τέλος, αν ο κακόβουλος κώδικας ή ο εισβολέας έχει παρεισφρήσει εντός του δικτύου. Επιπλέον, υπάρχει δυνατότητα για την πραγματοποίηση και εσωτερικών ελέγχων σε συστήματα με τη χρήση εξουσιοδοτημένων λογαριασμών. Το OpenVAS χρησιμοποιεί το λογαριασμό για να συνδεθεί στο σύστημα και να εκτελέσει τοπικούς ελέγχους ασφαλείας (LSC). Για συστήματα Linux χρήση λογαριασμού με περιορισμένα δικαιώματα είναι αρκετή για την εκτέλεση των περισσοτέρων ελέγχων ενώ σε συστήματα με Windows για τον έλεγχο του μητρώου, των ενημερώσεων ασφαλείας κ.α. απαιτείται χρήστης με αυξημένα δικαιώματα. Πέραν του εντοπισμού των ευπαθειών, ιδιαίτερη βαρύτητα για τη διαχείριση και αποκατάστασή τους έχει η κατηγοριοποίηση και βαθμολόγηση. Με το Κοινό Σύστημα Βαθμολόγησης ευπαθειών (CVSS) παράγει αριθμητικά αποτελέσματα που αντικατοπτρίζουν τη σοβαρότητα κάθε ευπάθειας.

Παρόλο που έχει σχεδιαστεί για την ελάχιστη παρεμβατική επίδραση στο δικτυακό περιβάλλον, είναι απαραίτητη η αλληλεπίδραση με τα υπό εξέταση συστήματα συχνά ακολουθώντας τεχνικές που χρησιμοποιούνται σε πραγματικές επιθέσεις και συνεπώς δύναται να τα επηρεάσει (π.χ. δημιουργία αυξημένων καταγραφών και ειδοποιήσεων, καθυστέρηση στο δίκτυο και στην απόκριση συστημάτων, κλείδωμα λογαριασμού χρηστών). Ειδικά σε ενσωματωμένα συστήματα και στοιχεία ΟΤ με ελλιπή στοίβα δικτυακών πρωτοκόλλων, μπορεί να προκαλέσει κατάρρευση ή καταστροφή συσκευών. Μπορεί να γίνει περιορισμός των εκτελούμενων ελέγχων, ώστε να εξαιρεθούν έλεγχοι με παρεμβατική συμπεριφορά.

Στην παρούσα εργασία για την αξιολόγηση συστημάτων που εμπλέκονται σε ύποπτες συνδέσεις, έγινε χρήση του OpenVAS 10, με την εγκατάσταση του Greenbone Vulnerability Manager 10 (GVM-10) σε λειτουργικό σύστημα Debian 10. Οι ύποπτες συνδέσεις παρέχονται στο OpenVAS ανά ημέρα, σε αρχείο κειμένου και η εκτέλεση της αξιολόγησης σε κάθε σύστημα, επαφίεται στον διαχειριστή του δικτύου της ΥΖΣ, αφού συνυπολογίσει τυχόν επιπτώσεις στα υπό εξέταση συστήματα.

Οι δύο βασικές προσεγγίσεις για την αντιμετώπιση των ευπαθειών είναι:

- Εξάλειψη της ευπάθειας με ενημέρωση ή με αφαίρεση του εμπλεκόμενου λογισμικού ή με αλλαγή στην παραμετροποίησή του.
- Προσθήκη κάποιου κανόνα στο τοίχος προστασίας ή στο σύστημα αποτροπής εισβολής (IPS).

Κεφάλαιο 3 - Apache Spot και Οικοσύστημα Hadoop

3.1 Hadoop

Το Hadoop είναι μια πλατφόρμα ανοικτού κώδικα για κατανεμημένη επεξεργασία μεγάλων ποσοτήτων δεδομένων, από συστοιχίες υπολογιστών, οι οποίοι μπορεί να είναι ακόμα και απλοί, γενικής χρήσης. Σχεδιασμένο με γνώμονα την επεκτασιμότητα και την υψηλή διαθεσιμότητα, έχει δυνατότητα αποθήκευσης και διαχείρισης τεράστιων ποσοτήτων πληροφοριών. Μπορεί να επεκταθεί από έναν διακομιστή (κόμβο) σε αρκετές χιλιάδες και να επεξεργαστεί δεδομένα που φτάνουν σε επίπεδο petabytes [41]. Κάθε διακομιστής κάνει τοπική αποθήκευση και επεξεργασία δεδομένων. Το Hadoop δεν βασίζεται στο υλικό για την αντιμετώπιση των αποτυχιών, η ανίχνευση και διαχείριση των οποίων γίνεται σε επίπεδο εφαρμογής παρέχοντας υψηλή διαθεσιμότητα για ολόκληρη τη συστοιχία. Το Hadoop μπορεί να χειριστεί δομημένα αλλά και μη δομημένα δεδομένα παρέχοντας μεγαλύτερη ευελιξία στην συλλογή, επεξεργασία και ανάλυση δεδομένων από ότι οι σχεσιακές βάσεις και οι αποθήκες δεδομένων. Αυτή του η ευελιξία το κάνει ιδανικό για χρήση του στον τομέα των Μεγάλων Δεδομένων.

Η ανάπτυξη του Hadoop βασίστηκε στις σχετικές δημοσιεύσεις για το κατανεμημένο λειτουργικό σύστημα της Google, GFS (Google File System) και την τεχνική επεξεργασίας μεγάλων συνόλων δεδομένων MapReduce. Από το 2008 είναι έργο του ιδρύματος Apache, η πρώτη (1.0) έκδοση του είναι διαθέσιμη από το 2011, ενώ η τρέχουσα είναι η έκδοση 3.2.1.

Το Hadoop είναι μια εξελισσόμενη πλατφόρμα που συντίθεται από τέσσερα κύρια στοιχεία, το HDFS, το MapReduce, το YARN, το Hadoop Common και πλήθος άλλων συμπληρωματικών που όλα μαζί χαρακτηρίζονται ως οικοσύστημα Hadoop. Θα ακολουθήσει μια σύντομη περιγραφή των κύριων συστατικών και εκ των συμπληρωματικών μόνο αυτά που χρησιμοποιεί το Apache Spot.

3.1.1 HDFS

Το Hadoop Distributed File System [42] είναι το κύριο συστατικό και είναι υπεύθυνο για την κατανεμημένη αποθήκευση μεγάλων ποσοτήτων δεδομένων σε μεγάλο αριθμό κόμβων. Παρουσιάζει αρκετές ομοιότητες με άλλα κατανεμημένα λειτουργικά συστήματα αλλά και αρκετές διαφορές. Με την αποθήκευση να γίνεται σε μεγάλο αριθμό κόμβων (χαμηλού κόστους που μπορούν να φτάσουν τις πολλές χιλιάδες), η αποτυχία κόμβων είναι στατιστικά πολύ πιθανή και το HDFS έχει σχεδιαστεί, ώστε να εξακολουθεί να λειτουργεί αποδοτικά . Ο πυρήνας της αρχιτεκτονικής

του περιλαμβάνει την ανίχνευση και γρήγορη αυτοματοποιημένη αποκατάσταση των αποτυχιών των κόμβων. Το HDFS είναι σχεδιασμένο για την αποθήκευση μεγάλων αρχείων (μεγέθους GB ή TB) τα οποία δεν τροποποιούνται μετά την αποθήκευσή τους, ακολουθώντας το μοντέλο μια εγγραφή, πολλές αναγνώσεις. Αυτό απλοποιεί τα θέματα συνοχής των αρχείων και παρέχει πρόσβαση υψηλής ταχύτητας στα δεδομένα. Μια άλλη αρχή της αρχιτεκτονικής του είναι, ότι είναι πιο οικονομικό να επεξεργάζεσαι δεδομένα (ειδικά όταν είναι μεγάλα), από το να τα μετακινείς. Το HDFS παρέχει διεπαφές για τη μεταφορά των εφαρμογών κοντά στα δεδομένα και όχι το αντίστροφο, αποφεύγοντας τη συμφόρηση των δικτύων. Ένα άλλο χαρακτηριστικό του είναι η δυνατότητα της μεταφοράς του σε διαφορετικές πλατφόρμες λογισμικού και υλισμικού.

Τα κύρια στοιχεία του HDFS είναι ο NameNode και οι DataNodes, που ακολουθούν το μοντέλο κυρίου-υποτελούς. Κάθε συστοιχία HDFS έχει ένα NameNode, ο οποίος είναι ο κύριος διακομιστής που διαχειρίζεται την περιοχή ονομάτων (namespace) του συστήματος αρχείων και ρυθμίζει την πρόσβαση στα αρχεία. Εσωτερικά κάθε αρχείο διαιρείται σε ένα ή περισσότερα μπλοκ δεδομένων (προεπιλογή μεγέθους 128 MB) και κάθε μπλοκ αποθηκεύεται σε μια ομάδα DataNode (προεπιλογή replication factor 3). Ο NameNode εκτελεί λειτουργίες συστήματος αρχείων όπως άνοιγμα, κλείσιμο, μετονομασία αρχείων και καταλόγων. Βασικός του ρόλος είναι επίσης ο καταμερισμός των μπλοκ των αρχείων στους DataNodes. Οι DataNodes είναι υπεύθυνοι για την εγγραφή και ανάγνωση των αρχείων, διαδικασία που περιλαμβάνει τη δημιουργία, τη διαγραφή και την αναπαραγωγή των μπλοκ, υπό την καθοδήγηση του NameNode.

Η περιοχή ονομάτων του HDFS αποθηκεύεται στον NameNode, όπου κρατείται το αρχείο συναλλαγών (transaction) (το EditLog) με όλες τις αλλαγές των μεταδεδομένων του συστήματος αρχείων. Όλα τα στοιχεία της περιοχής ονομάτων μαζί με πληροφορίες για τον καταμερισμό των αρχείων και τις ιδιότητες του λειτουργικού συστήματος αποθηκεύονται στο αρχείο FsImage. Τα αρχεία EditLog και FsImage αποθηκεύονται σαν κανονικά αρχεία του τοπικού λειτουργικού συστήματος που φιλοξενεί το Hadoop. Σαν κανονικά αρχεία του τοπικού λειτουργικού συστήματος αποθηκεύονται και τα αρχεία των DataNodes. Ειδική μέριμνα έχει δοθεί για την ακεραιότητα των αρχείων και την αντιμετώπιση των προβλημάτων (π.χ. απώλεια NameNode, DataNode, προβλήματα δικτύου) με την δημιουργία σημείων ελέγχου (checkpoints) και αθροισμάτων ελέγχου (checksums) των μπλοκ, ενώ υποστηρίζεται και η ύπαρξη πολλαπλών NameNode σε διάταξη hot standby καθώς ο NameNode αποτελεί μοναδικό σημείο αποτυχίας. Επίσης, τα πολλαπλά αντίγραφα των μπλοκ των αρχείων εγγράφονται σε DataNodes που βρίσκονται σε διαφορετικά ικριώματα, ώστε να

ελαχιστοποιείται η πιθανότητα ταυτόχρονης αποτυχίας. Η πρόσβαση στα αρχεία που είναι αποθηκευμένα στο HDFS μπορεί να γίνει με πολλούς τρόπους, όπως Java API, REST API, μέσω προγράμματος περιήγησης ιστοσελίδων και διαμέσου NFS gateway από τα συστήματα αρχείων των χρηστών. Εντολές στο κέλυφος του συστήματος αρχείων (File System shell) μπορούν να κληθούν με εντολές του τύπου: *bin/hadoop fs <args>*

3.1.2 MapReduce

Είναι το δεύτερο στοιχείο, που μαζί με το HDFS αποτελούν την καρδιά του Hadoop και είναι μια προγραμματιστική προσέγγιση, που επιτρέπει την επεξεργασία μεγάλων συνόλων δεδομένων με παράλληλο, κατανεμημένο αλγόριθμο από συστοιχίες υπολογιστών. Είναι σχεδιασμένο για επεξεργασία απεριόριστου μεγέθους και οποιουδήποτε τύπου δεδομένων είναι αποθηκευμένα στο HDFS, διαιρώντας τον φόρτο εργασίας σε πολλαπλές εργασίες, οι οποίες παραλαμβάνονται από όλους τους κόμβους, που τις εκτελούν παράλληλα [43]. Ο όρος MapReduce αναφέρεται σε δύο ξεχωριστές εργασίες που εκτελεί το Hadoop. Η πρώτη είναι το Map, που παίρνει ένα σύνολο δεδομένων και το μετατρέπει σε ένα άλλο που τα στοιχεία του έχουν αναλυθεί σε πλειάδες (ζευγάρια κλειδιών/τιμών). Η εργασία Reduce παίρνει σαν είσοδό της, την έξοδο της Map και συνδυάζει τις πλειάδες που έχουν δημιουργηθεί με ένα μικρότερο σύνολο πλειάδων. Υποστηρίζει μεγάλο αριθμό γλωσσών προγραμματισμού όπως Java, C++ και Python και γλωσσών υψηλού επιπέδου μέσω Apache Hive και Apache Pig. Η επεξεργασία MapReduce είναι επεξεργασία κατά δεσμίδες (batch) και ενώ παραμένει ένας πολύ δημοφιλής τρόπος επεξεργασίας, το Apache Spot χρησιμοποιεί το μεταγενέστερο Apache Spark – θα αναφερθούμε σε αυτό παρακάτω – το οποίο εκτελεί τις ίδιες εργασίες, παρέχοντας περισσότερη ευελιξία και ταχύτητα.

3.2 Yarn

Το YARN (Yet Another Resource Negotiator) είναι μια τεχνολογία διαχείρισης πόρων και χρονοπρογραματισμού εργασιών σε συστοιχίες Hadoop και αποτελεί μετά την έκδοση 2.0 ένα από τα βασικά συστατικά του. Είναι υπεύθυνο για τη διάθεση των πόρων του συστήματος στις εφαρμογές και τον χρονοπρογραματισμό των εργασιών για την εκτέλεσή τους σε διαφορετικούς κόμβους. Πρόκειται για ένα επίπεδο που παρεμβλήθηκε μεταξύ του HDFS και του MapReduce και ανέλαβε τις ανωτέρω λειτουργίες, τις οποίες σε προηγούμενες εκδόσεις εξυπηρετούσε ο JobTracker του MapReduce. Για την ακρίβεια παρεμβάλλεται μεταξύ του Hadoop και της μηχανής επεξεργασίας, που μπορεί πλέον, να είναι άλλη από το MapReduce όπως Apache Storm, Spark ,

Hive, Pig κτλ. Ο JobTracker αντικαταστάθηκε από έναν ResourceManager για ολόκληρη τη συστοιχία και από ένα AplicationMaster ανά εφαρμογή. Ο ResourceManager εποπτεύει τη διαθεσιμότητα των κόμβων και τη χρήση των πόρων της συστοιχίας ενώ ταυτόχρονα μεριμνά για την ομοιόμορφη κατανομή των διαθέσιμων πόρων και κάνει τη διαιτησία μεταξύ των απαιτήσεων. Τα δυο βασικά στοιχεία του είναι ο Scheduler και ο Application Manager [44].

Το Yarn συνολικά αποτελείται από τέσσερα στοιχεία: τον ResourceManager, τους NodeManager τους ApplicationMaster και τους Container.

Οι NodeManager τρέχουν σε κάθε κόμβο και είναι υπεύθυνοι για την εκτέλεση των εργασιών. Παρακολουθεί τους πόρους των Container (cpu, μνήμη, δίσκο, δίκτυο) και ενημερώνει σχετικά τον ResourceManager. Με μια απλούστερη διατύπωση είναι ο διαχειριστής πόρων και εργασιών ενός κόμβου.

Ο ApplicationMaster διαπραγματεύεται τους πόρους με τον ResourceManager και συνεργάζεται με τους NodeManager για την εκτέλεση των containers και την παρακολούθηση των πόρων που καταναλώνουν. Ελέγχει την κατάσταση και την πρόοδο των containers.

Οι Containers είναι ένα σύνολο πόρων που έχουν προκύψει ως αποτέλεσμα της επιτυχημένης παραχώρησής τους από τον ResourceManager, ικανοποιώντας ένα συγκεκριμένο αίτημά ResourceRequest. Ο container δίνει τα δικαιώματα σε μια εφαρμογή να χρησιμοποιήσει καθορισμένη ποσότητα πόρων ενός συγκεκριμένου κόμβου.

Για τη διαχείριση των διαφόρων φορτίων που σχετίζονται με την υψηλή χρήση της CPU, το YARN χρησιμοποιεί την έννοια των "vcores". Είναι ένα ποσοστό χρήσης της CPU του κόμβου και διατίθεται από τον NodeManager με στόχο την αποδοτική χρήση των πόρων. Κάθε κόμβος μπορεί να ρυθμιστεί για βέλτιστη χρήση των vcores με τη παραμετροποίηση των YARN containers ως αριθμού των vcores, στο αρχείο yarn-site.xml κάθε κόμβου. Η παραμετροποίηση προκύπτει με βάση τον φόρτο εργασίας και το υλισμικό του κόμβου. Στο αρχείο αυτό καθορίζονται και παράμετροι (min, max, βήμα) για τη μνήμη που διατίθεται τους containers, με βάση τη φυσική μνήμη του κόμβου. Το YARN είναι ιδιαίτερα κρίσιμο για την εύρυθμη λειτουργία της συστοιχίας καθώς εργασίες δύναται να μην εκκινούν λόγω της περιορισμένης διάθεσης πόρων ή να τερματίζονται εξαιτίας της υπέρβασης των πόρων που τους έχουν διατεθεί. Η διεπαφή για την παρακολούθηση του ResourceManager είναι στην θύρα 8088.

Η παραμετροποίηση του YARN ειδικά για μικρές συστοιχίες με περιορισμένους πόρους, όπως αυτή που χρησιμοποιήθηκε σε αυτή την εργασία, χρήζει ιδιαίτερης προσοχής. Η ορθότητά της αποτέλεσε προϋπόθεση για τη δυνατότητα εκτέλεσης των εργασιών του Apache Spot.

3.3 Apache Kafka

Το Kafka είναι μια κατανεμημένη πλατφόρμα ροών δεδομένων. Αποτελεί ένα σύστημα ανταλλαγής ροών από εγγραφές, που ακολουθεί το μοντέλο publish/subscribe, που χρησιμοποιούν πολλά συστήματα ανταλλαγής μηνυμάτων. Εκτός από την μεταφορά των ροών, οι ροές αποθηκεύονται με αξιοπιστία ενώ παράλληλα μπορεί να υπόκεινται σε επεξεργασία. Χρησιμοποιείται για τη δημιουργία αξιόπιστων ροών δεδομένων, πραγματικού χρόνου μεταξύ συστημάτων και εφαρμογών καθώς επίσης για την μετατροπή ή την επενέργεια των ροών. Οι εγγραφές αποτελούνται από το κλειδί, την τιμή, και την χρονοσφραγίδα του και καταγράφονται σε κατηγορίες που ονομάζονται topics. Αναφερόμαστε στην αξιοπιστία καθώς το Kafka εκτελείται σε συστοιχίες κόμβων που ονομάζονται brokers και που δύναται να βρίσκονται σε διαφορετικά ανεξάρτητα κέντρα δεδομένων [45].

Το Kafka διαθέτει τέσσερα κύρια API: για την έκδοση ροών εγγραφών (Producer API), για την συνδρομή σε λήψη ροών (Consumer API), την επεξεργασία ροών (λαμβάνοντας ροή/ες εισόδου για την παραγωγή μετά την επεξεργασία άλλων ροής/ων εξόδου) (Streams API) και για την παραγωγή επαναχρησιμοποιούμενων εκτελέσιμων παραγωγών ή καταναλωτών του συνδέονται σε topic (Connector API).

Το topic είναι η κατηγορία ή το όνομα στο οποίο δημοσιεύονται οι ροές εγγραφών, οι οποίες είναι πολλαπλών συνδρομητών, δηλαδή μπορούν να υπάρχουν πολλοί συνδρομητές που λαμβάνουν τις εγγραφές του ίδιου topic. Για κάθε topic διατηρείται ένα κατατμημένο αρχείο καταγραφής που περιέχει τον διαμερισμό του topic. Κάθε διαμέρισμα (partition), διατηρεί ταξινομημένη, αριθμημένη και αμετάβλητη την ακολουθία ενός τμήματος των εγγραφών που έχουν δημοσιευτεί σε αυτό το θέμα. Ο αριθμός κάθε εγγραφής που ονομάζεται offset, προσδιορίζει μοναδικά την εγγραφή και δίνει την δυνατότητα στους συνδρομητές να ανατρέξουν σε αυτή. Ο χρόνος που φυλάσσονται οι εγγραφές είναι παραμετροποιήσιμος και καθορίζεται από την πολιτική φύλαξης του topic. Η κατάτμηση των topics δίνει την δυνατότητα, ώστε διαφορετικά τμήματά τους να αποθηκεύονται σε διαφορετικούς κόμβους, με τελικό στόχο την ευελιξία και την επεκτασιμότητα.

Τα διαμερίσματα αποθηκεύονται σε πολλαπλούς κόμβους εξασφαλίζοντας αντοχή και σε σφάλματα και υψηλή διαθεσιμότητα.

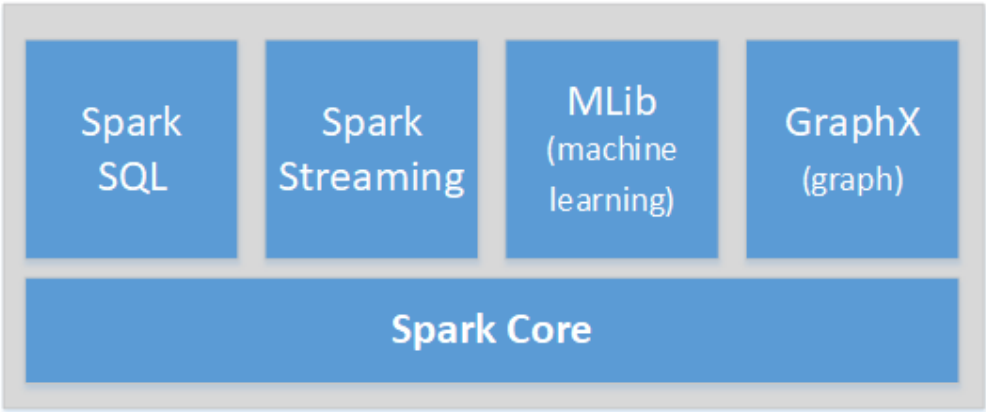
Μα βάση τα παραπάνω το Kafka μπορεί να χρησιμοποιηθεί ως σύστημα ανταλλαγής μηνυμάτων, ως σύστημα αποθήκευσης και για επεξεργασία εγγραφών σε πραγματικό χρόνο. Το Kafka επιτυγχάνει την απόζευξη μεταξύ εκδοτών και συνδρομητών σε εφαρμογές μεγάλης κλίμακας, έχοντας τη δυνατότητα προσωρινής αποθήκευσης των μη επεξεργασμένων μηνυμάτων. Έχει ευρεία χρήση σε εφαρμογές που σχετίζονται με την παρακολούθηση επιχειρησιακών δεδομένων, την συγκέντρωση αρχείων καταγραφής και εγγραφών συναλλαγών, την παρακολούθηση δραστηριοτήτων, την καταγραφή ακολουθιακών γεγονότων κ.α. Η δυνατότητα επεξεργασίας που διαθέτει το καθιστά αξιόπιστο και αποδοτικό κατά την διαδικασία εξαγωγής, μετατροπής και φόρτωσης δεδομένων σε Hadoop, HBase βάσεις δεδομένων κα.

Το Apache Spot χρησιμοποιεί το Kafka κατά την διαδικασία πρόσληψης των δεδομένων. Όπως θα δούμε εκτενέστερα κατά την ανάλυση της λειτουργίας τους, για κάθε τύπο δεδομένων που προσλαμβάνει πρέπει να δημιουργηθεί το αντίστοιχο topic και στην συνέχεια να δημιουργηθούν οι συνδρομητές οι οποίοι ονομάζονται workers.

3.4 Apache Spark

Το Apache Spark είναι μια μηχανή ανάλυσης γενικού σκοπού, για επεξεργασία μεγάλων ποσοτήτων δεδομένων. Είναι κατάλληλο για επεξεργασία τόσο ασυνεχών διεργασιών όσο και διεργασιών πραγματικού χρόνου, τις οποίες εκτελεί με υψηλή ταχύτητα. Παρέχει ευκολία στην ανάπτυξη παράλληλων εφαρμογών, υποστηρίζοντας πλήθος γλωσσών προγραμματισμού και διαθέτει κέλυφος με αλληλεπίδραση σε Java, Scala, Python, R και SQL. Μπορεί να τρέξει πάνω σε Hadoop, Amazon EC2 , Mesos και Cassandra και να δουλέψει με μεγάλο πλήθος πηγών δεδομένων, όπως για παράδειγμα HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive και πολλών άλλων [46].

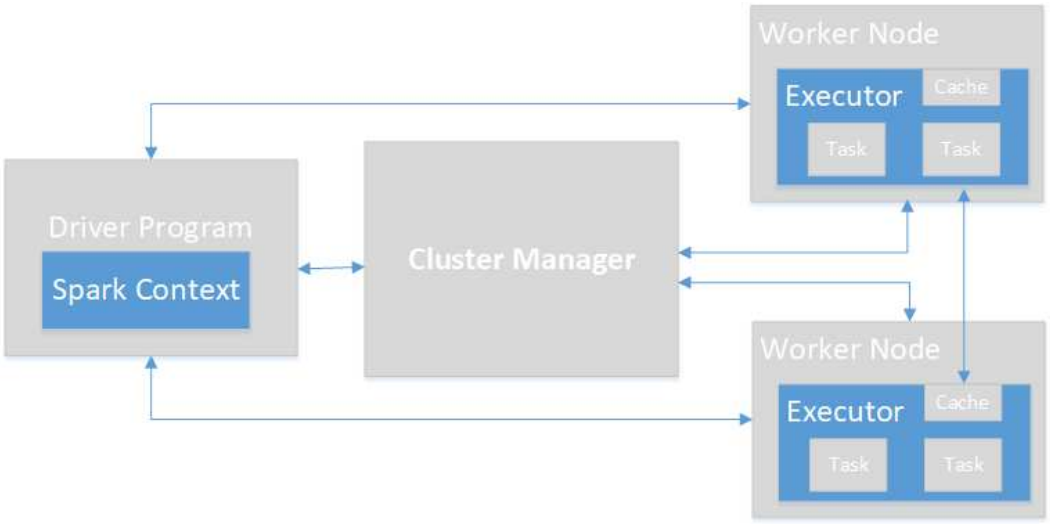
Το Spark έχει πλήθος διαφορών με το MapReduce, στο οποίο αναφερθήκαμε παραπάνω, με τις βασικότερες να είναι: υποστηρίζει ανάλυση συνεχών ροών (Spark Streaming), αποθηκεύει τα δεδομένα και στην RAM και όχι μόνο στο δίσκο επιτυγχάνοντας ταχύτητες έως 100x σε σχέση με το MapReduce και έχει ενσωματωμένες βιβλιοθήκες για Μηχανική Μάθηση και για εκτέλεση εντολών SQL (βλέπε Εικόνα 4).



Εικόνα 4: Ενσωματωμένες βιβλιοθήκες Apache Spark

Η βασική δομή δεδομένων που χρησιμοποιεί το Spark είναι η Resilient Distributed Datasets (RDD), η οποία είναι σύνολο μη τροποποιήσιμων κατανεμημένων αντικειμένων. Η δομή αυτή αποτελεί τον πυρήνα της υψηλής απόδοσης του Spark. Κάθε σύνολο δεδομένων σε RDD, κατατέμνεται σε λογικά διαμερίσματα, τα οποία μπορούν να υπολογιστούν σε διαφορετικούς κόμβους της συστοιχίας. Είναι παράλληλες δομές δεδομένων που διαθέτουν ανοχή σε σφάλματα και επιτρέπουν στον χρήστη να δημιουργεί προσωρινά αποτελέσματα που αποθηκεύονται στη μνήμη και να τα χειρίζεται με ένα μεγάλο πλήθος λειτουργιών. Ένα RDD μπορεί να περιλαμβάνει αντικείμενα Python, Java ή Scala και κλάσεις καθοριζόμενες από τον χρήστη.

Το Spark διαθέτει αρχιτεκτονική κυρίου-υποτελούς, όπου υπάρχει ένας κεντρικός συντονιστής, η διεργασία Driver, η οποία επικοινωνεί με τους κατανεμημένους κόμβους –workers, όπου τρέχουν οι executors (βλέπε σχήμα Εικόνα 5). Η Driver διεργασία είναι το πρόγραμμα που περιέχει την κύρια κλάση της εφαρμογής, δημιουργεί το SparkContext, δηλαδή τη σύνδεση με την συστοιχία Spark, καθορίζει τους μετασχηματισμούς και τις ενέργειες των RDDs και υποβάλει αιτήσεις στον master. Οι executors είναι διεργασίες που εκτελούνται στους υποτελείς κόμβους workers, εργάζονται στην ομάδα των RDD που έχουν και αποστέλλουν τα αποτελέσματά τους στον Spark Driver.



Εικόνα 5: Μπλοκ διάγραμμα συστοιχίας Apache Spark

3.5 Apache Hive

Το Hive είναι ένα πλαίσιο λογισμικού, για συστήματα αποθήκης δεδομένων που λειτουργούν με σύστημα Hadoop (ή συμβατά με αυτό όπως τα Amazon S3 και Alluxio). Βασικά πλεονεκτήματα από τη χρήση του Hive, είναι ότι μπορεί να ορίσει δομή σε μη δομημένα δεδομένα της αποθήκης και να απλοποιήσει την ανάλυση και τα ερωτήματα προς τα δεδομένα της. Διαθέτει μια γλώσσα σεναρίων την HiveQL για τη δημιουργία ερωτημάτων, η οποία είναι γλώσσα παρόμοια με την SQL. Το Hive δεν είναι σχεσιακή βάση δεδομένων, χρησιμοποιεί όμως βάση δεδομένων για να αποθηκεύσει μεταδεδομένα, ενώ τα δεδομένα είναι αποθηκευμένα στο Hadoop. Καθώς το Hadoop είναι σχεδιασμένο για ασυνεχείς διεργασίες, τα ερωτήματα Hive έχουν γενικά υψηλή λανθάνουσα καθυστέρηση και δεν είναι κατάλληλο για ερωτήματα πραγματικού χρόνου [47].

Η HiveQL είναι γλώσσα βασισμένη στην προδιαγραφή SQL-92 και δίνει τη δυνατότητα στον χρήστη, χωρίς να γνωρίζει MapReduce, να χρησιμοποιεί ερωτήματα σε SQL για να δημιουργεί τις σχετικές εργασίες του MapReduce. Οι πίνακες Hive αποτελούνται από τα δεδομένα και το σχήμα, τα οποία είναι ανεξάρτητα. Τα δεδομένα είναι αρχεία του HDFS και το σχήμα ως μεταδεδομένο αποθηκεύεται σε σχεσιακή βάση. Το σχήμα μπορεί να οριστεί σε υπάρχοντα δεδομένα, τα οποία μπορούν να προστίθενται ή να αφαιρούνται ανεξάρτητα.

3.6 Apache Impala

Το Impala είναι μια μηχανή SQL βασισμένη στη μαζική παράλληλη επεξεργασία με χαμηλή λανθάνουσα καθυστέρηση και υψηλό ταυτοχρονισμό για ερωτήματα επιχειρηματικής ευφυΐας και ανάλυσης. Είναι τεχνολογία παράλληλων βάσεων δεδομένων που είναι αποθηκευμένα σε Hadoop ή HBase και δεν απαιτεί μεταφορά ή μετασχηματισμό δεδομένων [48].

Το Impala χρησιμοποιεί τη γλώσσα ερωτημάτων και τα μεταδεδομένα του Apache Hive. Τα ερωτήματα του Impala,που μπορεί να είναι διαδραστικά, δύναται να δοθούν μέσω γραμμής εντολών ή μέσω πελάτη ODBC ή JDBC και είναι ιδιαίτερα εύκολα για χρήση από επιστήμονες δεδομένων και αναλυτές. Μπορεί κάνει ανάγνωση και εγγραφή σε πίνακες του Hive και να χρησιμοποιηθεί σε ανάλυση σε δεδομένα που έχουν παραχθεί από αυτό.

Κύρια στοιχεία του Impala είναι οι Daemon, ο Statestore και η Catalog Service. Ο Daemon είναι το βασικό στοιχείο του Impala και είναι μια διεργασία που εκτελείται σε όλους τους κόμβους της συστοιχίας. Διαβάζει και γράφει δεδομένα, δέχεται ερωτήματα και αποστέλλει τα ενδιάμεσα αποτελέσματα στον κεντρικό κόμβο συντονισμού. Κόμβος συντονισμού για κάθε ερώτημα είναι ο κόμβος που το υποβάλει. Οι Daemon της συστοιχίας επικοινωνούν διαρκώς με τον Statestore για να έχουν γνώση ποιοι κόμβοι είναι διαθέσιμοι για να τους αποδώσει έργο. Ο Statestore είναι μια διεργασία που εκτελείται σε έναν κόμβο της συστοιχίας και ελέγχει τη λειτουργική κατάσταση όλων των κόμβων και ενημερώνει όλους του υπόλοιπους. Η Catalog Service είναι επίσης μια διεργασία για όλη τη συστοιχία, που προωθεί τις αλλαγές των μεταδεδομένων που προκύπτουν από τα ερωτήματα SQL, σε όλους τους Impala Deamon.

Το Impala είναι γενικά πολύ πιο γρήγορο από το Hive καθώς οι Daemon του Impala τρέχουν διαρκώς, ώστε τα ερωτήματα να εκτελούνται άμεσα και δεν χρειάζεται να γίνει μετατροπή των SQL ερωτημάτων σε εργασίες MapReduce. Όμως, δεν υποστηρίζει σύνθετους τύπους δεδομένων, δεν διαθέτει ανοχή σε σφάλματα και έχει θέματα επεκτασιμότητας.

3.7 Εγκατάσταση συστοιχίας Cloudera Hadoop.

Όπως έχουμε προαναφέρει μια βασική επιλογή για την εγκατάσταση του Apache Spot είναι πάνω σε μια υπάρχουσα ή αφιερωμένη συστοιχία Hadoop. Για την παρούσα εργασία έγινε εγκατάσταση μιας μικρής αφιερωμένης συστοιχίας. Η εγκατάσταση έγινε με τη χρήση της διανομής Hadoop της Cloudera, αλλά μπορεί να εγκατασταθεί και σε άλλες διανομές Hadoop, αρκεί να συνεγκατασταθεί

όλο το υποστηρικτικό λογισμικό που κάνει χρήση το Apache Spot. Η συστοιχία αποτελείται από 3 κόμβους Hadoop, ο ένας εκ των οποίων είναι υπεύθυνος και για τη διαχείρισή τους. Οι κόμβοι τρέχουν λειτουργικό σύστημα Debian Linux έκδοσης 8.11 και η εγκατάσταση του Hadoop έγινε με τη χρήση της διανομής CDH 5.16.1 (με την μέθοδο των Parcels), η οποία είναι διανομή ανοικτού κώδικα και παρέχεται δωρεάν στην έκδοση που χρησιμοποιήσαμε. Η έκδοση του Hadoop που χρησιμοποιήσαμε είναι 2.6.0 και η διαχείριση της συστοιχίας γίνεται από το λογισμικό Cloudera Manager.

Σε γενικές γραμμές η δημιουργία της συστοιχίας περιλαμβάνει τα παρακάτω βήματα:

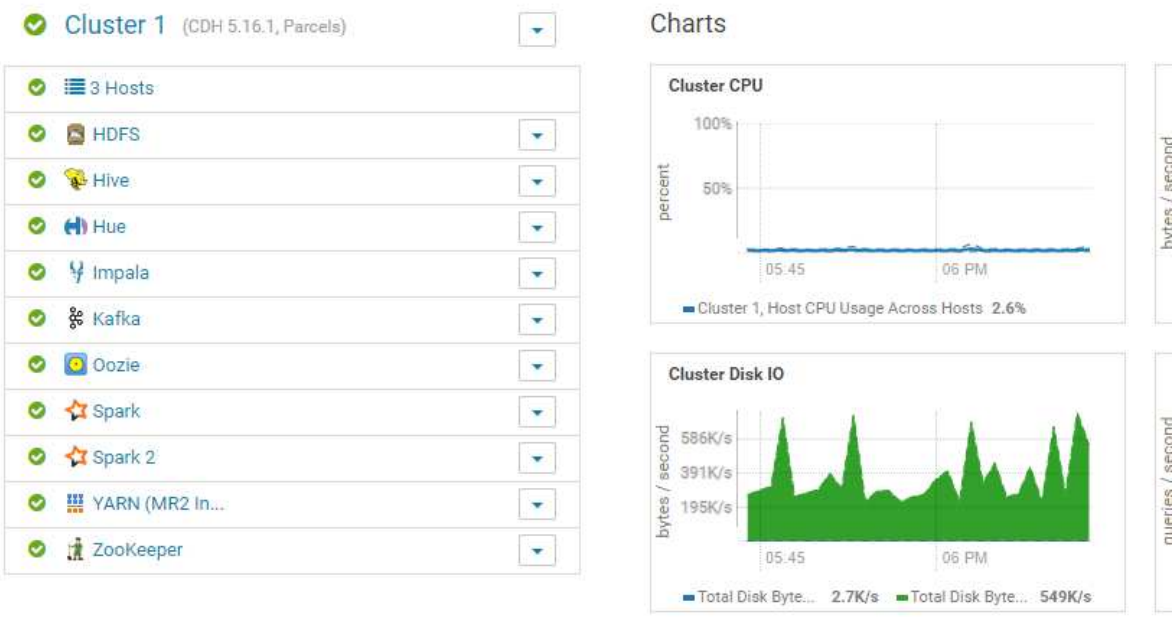
- Καθορισμός ονοματολογίας κόμβων, ρύθμιση firewall (όπου απαιτείται), δημιουργία λογαριασμού για πρόσβαση με ssh χωρίς τη χρήση κωδικών πρόσβασης σε όλους τους κόμβους.
- Ορισμός του αποθετηρίου της Cloudera. Το αποθετήριο προσδιορίζεται από τους σχετικούς πίνακες, ανάλογα από τις εκδόσεις Linux και CDH που θα χρησιμοποιήσουμε.
- Εγκατάσταση του JDK. Απαιτείται χρήση έκδοσης 64bit, η οποία πρέπει να εγκατασταθεί σε όλους τους κόμβους που θα συμμετέχουν στην συστοιχία.
- Εγκατάσταση του Cloudera Manager Server. Θα γίνει εγκατάσταση όλων των πακέτων στον διακομιστή διαχείρισης
- Εγκατάσταση βάσεων δεδομένων. Στο στάδιο αυτό δημιουργούνται οι βάσεις δεδομένων που είναι απαραίτητες για τη λειτουργία Cloudera Manager. Ως κύρια βάση χρησιμοποιήθηκε βάση δεδομένων mysql (PostgreSQL, MariaDB και Oracle είναι άλλες επιλογές) μετά την παραμετροποίηση του σχετικού αρχείου /etc/mysql/my.cnf. Στην συνέχεια δημιουργούμε βάσεις δεδομένων και ορίζουμε δικαιώματα πρόσβασης για τις βάσεις που χρησιμοποιούνται από τις παρακάτω υπηρεσίες ή δομικά στοιχεία της συστοιχίας: Cloudera Manager Server, Activity Monitor, Reports Manager, Hue, Hive Metastore Server,Sentry Server, Cloudera Navigator Audit Server, Cloudera Navigator Metadata Server και Oozie.
- Εγκατάσταση βάσης δεδομένων του Cloudera Manager (CM). Δημιουργία και παραμετροποίηση της βάση δεδομένων που θα χρησιμοποιεί ο CM, με τη χρήση αυτοματοποιημένης δέσμης εντολών.
- Εγκατάσταση CDH και λοιπού λογισμικού. Κάνουμε εκκίνηση του CM μέσω γραμμής εντολών και έχουμε πρόσβαση στη διεπαφή διαχείρισης, όπου ακολουθώντας τον οδηγό

εγκατάστασης της συστοιχίας, ορίζουμε την έκδοση CM, τους κόμβους που θα συμμετέχουν σε αυτήν, το αποθετήριο εγκατάστασης, τους κωδικούς πρόσβασης κ.α

- Εγκατάσταση της συστοιχίας. Με την ολοκλήρωση του προηγούμενου βήματος ακολουθούμε αυτόματα τον οδηγό παραμετροποίησης της συστοιχίας, όπου καθορίζονται οι υπηρεσίες που θα εγκατασταθούν, τους ρόλους που θα έχει ο κάθε κόμβος και θα δηλώσουμε τα στοιχεία των βάσεων δεδομένων που έχουμε δημιουργήσει σε προηγούμενα στάδια. Κάνουμε επισκόπηση των επιλογών και ολοκληρώνουμε τη διαδικασία.

Ιδιαίτερη προσοχή πρέπει να δοθεί στη συμβατότητα των εκδόσεων του λογισμικού που θα χρησιμοποιηθεί σε όλα τα βήματα. Κατά τη διάρκεια της εγκατάστασης επιλύουμε τυχόν προβλήματα ανατρέχοντας με προσοχή σε πληροφορίες της εγκατάστασης ή στον ιστότοπο υποστήριξης της Cloudera.

Με την ολοκλήρωση της ανωτέρω διαδικασίας, έχουμε δημιουργήσει μια συστοιχία Hadoop, διαχειριζόμενη από Cloudera Manager, όπως φαίνεται στην παρακάτω εικόνα.



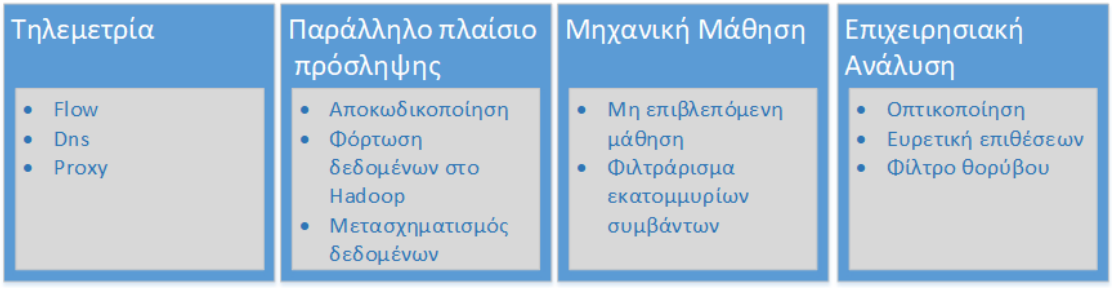
Εικόνα 6: Cloudera Manager - Εγκατεστημένες υπηρεσίες και λειτουργική κατάσταση

Πρόσβαση στην διαχειριστική διεπαφή του HDFS έχουμε http://namenode:50070 και μπορούμε να συνδεθούμε στο API του HDFS στην δ/ση http://namenode:8020.

Κεφάλαιο 4 - Apache Spot

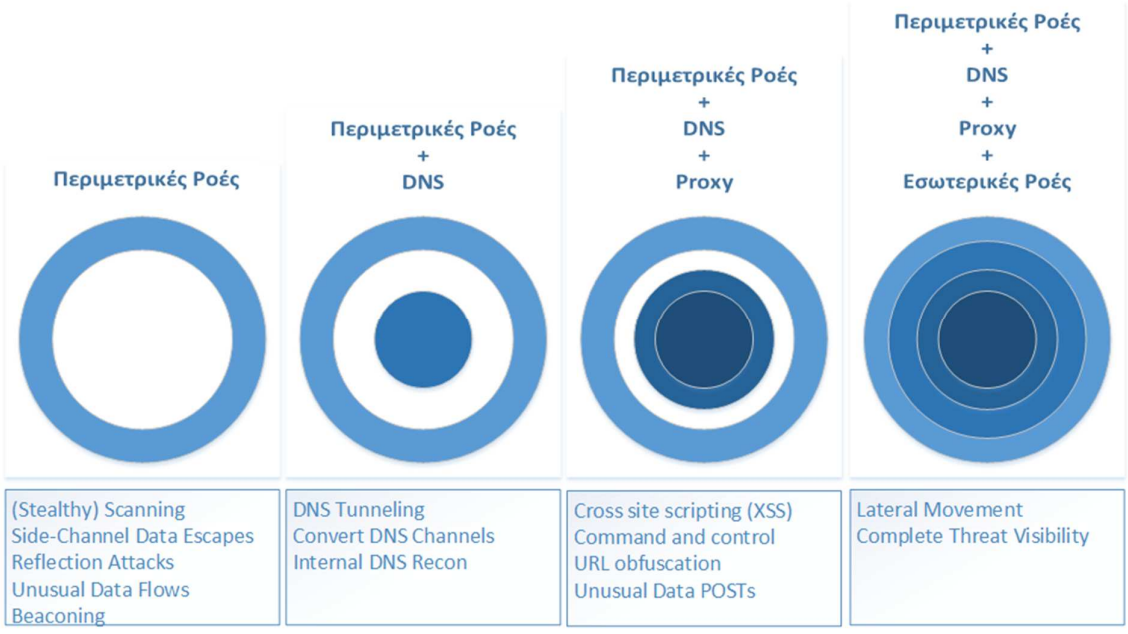
4.1 Εισαγωγή

Το Apache Spot αξιοποιώντας τις ισχυρές τεχνολογίες Μεγάλων δεδομένων και Επιστημονικής Υπολογιστικής μπορεί να είναι χρήσιμο στην επίλυση δύσκολων θεμάτων της ασφάλειας δικτύων [49]. Συνοπτικά η λειτουργία του και τα χαρακτηριστικά του παρουσιάζονται στην παρακάτω εικόνα:



Εικόνα 7: Χαρακτηριστικά Apache Spot

Τα δεδομένα που επεξεργάζεται το Apache Spot στην παρούσα φάση ανάπτυξής του είναι ροές δικτυακών συνδέσεων NetFlow (περιμετρικών ή και εσωτερικών του δικτύου), αρχεία καταγραφής DNS και αρχεία καταγραφής Proxy. Η πρόσληψη των δεδομένων, που ανάλογα με την εγκατάσταση, μπορεί να περιλαμβάνει και τις 3 πήγες ή κάποιες από αυτές, γίνεται παράλληλα. Η κατηγορία των απειλών που εντοπίζεται εξαρτάται από τον τύπο των δεδομένων που επεξεργάζεται. Η συσχέτιση των ανιχνεύσιμων απειλών με τις πηγές των δεδομένων φαίνεται στην Εικόνα 8.



Εικόνα 8: Τύποι δεδομένων και ανιχνεύσιμες απειλές

Αναλυτικότερα τα αρχεία που μπορεί να προσλάβει ως είσοδο είναι:

- Αρχεία Netflow σε μορφή nfcapd. Παράγονται από δικτυακές συσκευές που υποστηρίζουν το εν λόγω καθιερωμένο πρωτόκολλο της Cisco Systems και καταγράφονται από συλλέκτες Netflow όπως το nfdump [50].
- Αρχεία καταγραφής DNS σε μορφή .pcap. Τα αρχεία DNS .pcap παράγονται από καταγραφή της κίνησης του εν λόγω πρωτοκόλλου με τη χρήση εργαλείων όπως το tshark [51].
- Αρχεία proxy bluecoat .log. Είναι αρχεία κείμενου που εξάγονται από διακομιστές μεσολάβησης της εταιρίας BlueCoat Systems (σήμερα τμήμα της Symantec)

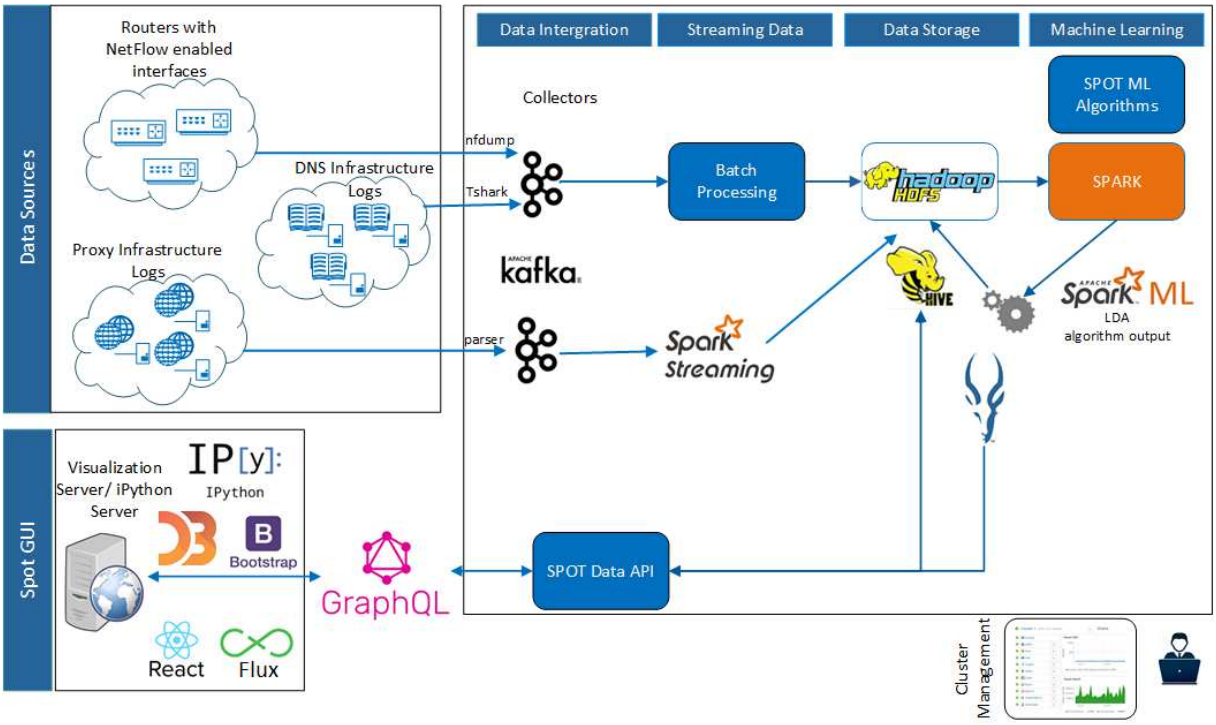
4.2 Αρχιτεκτονική

Η διαδικασία ανάλυσης για τον εντοπισμό ύποπτων συνδέσεων περιλαμβάνει 3 βασικά αυτόνομα στάδια τα οποία αποτελούν την αλυσίδα εκτέλεσης του Spot:

- Πρόσληψη και αποθήκευση των δεδομένων
- Εφαρμογή αλγόριθμου μηχανικής και ανάλυση ύποπτων συνδέσεων
- Προσθήκη εννοιολογικού πλαισίου και οπτικοποίηση δεδομένων

Κάθε στάδιο της διαδικασίας ενεργοποιείται αυτόνομα από τη γραμμή εντολών και σε μια τυπική εγκατάσταση υπάρχουν διαφορετικοί κόμβοι για την εκτέλεσή τους. Οι κόμβοι αυτοί, σε αντιστοιχία με την παραπάνω λίστα είναι ο κόμβος Ingest, κόμβος ML και ο κόμβος OA.

Η αρχιτεκτονική του Apache Spot παρουσιάζεται στην παρακάτω εικόνα:



Εικόνα 9: Αρχιτεκτονική του Apache Spot

Η συλλογή των δεδομένων γίνεται από διεργασίες που ονομάζονται Collectors, οι οποίες παρακολουθούν συγκεκριμένες διαδρομές στο σύστημα αρχείων (μια για κάθε τύπο δεδομένων) και συλλέγουν αρχεία. Τα αρχεία αυτά είτε παράγονται τη στιγμή εκείνη από τις αντίστοιχες πηγές είτε είναι αποθηκευμένα και απλά, μεταφέρονται στις συγκεκριμένες θέσεις του συστήματος αρχείων. Τα αρχεία - εκτός των αρχείων των διακομιστών μεσολάβησης τα οποία είναι αρχείου κειμένου- πρέπει να μετατραπούν από τους Collectors σε κείμενο, με τη χρήση των εργαλείων ανάλυσης nfdump (για NetFlow) και tshark (για DNS) και στη συνέχεια αποθηκεύονται. Η αποθήκευση γίνεται στο HDFS στην αρχική τους μορφή, ώστε να υπάρχει δυνατότητα χρήσης τους σε περίπτωση εγκληματολογικής ανάλυσης και σε Hive σε μορφή Avro-parquet για ανάλυση με γλώσσα HiveQL.

Με την ολοκλήρωση της μεταφοράς ενός αρχείου στις διαδρομές που αναφερθήκαμε πριν, αν το αρχείο είναι μεγαλύτερο από 1 MB, τότε το όνομα και η θέση του αποστέλλεται στο Kafka ενώ, αν είναι μικρότερο, αποστέλλεται το ίδιο το αρχείο. Στην πρώτη περίπτωση ακολουθεί ασυνεχής επεξεργασία από το Spot, ενώ στην δεύτερη ακολουθεί επεξεργασία από την βιβλιοθήκη Spark Streaming.

Όπως περιγράφεται κατά τη διαδικασία της πρόσληψης των δεδομένων, μεσολαβεί το Apache Kafka (βλέπε Εικόνα 9) το οποίο με την αξιόπιστη και κατανεμημένη αρχιτεκτονική του, εξασφαλίζει την υψηλή αξιοπιστία της διαδικασίας. Έτσι η δημιουργία μεγάλων ουρών κατά την επεξεργασία των αρχείων ή η τυχόν κατάρρευση κάποιου Collector δεν προκαλούν απώλεια των δεδομένων. Για κάθε τύπο δεδομένων δημιουργείται διαφορετικό topic και ο διαμερισμός κάθε θέματος καθορίζεται από τον αριθμό των Spot Workers. Οι Spot Workers είναι οι διεργασίες οι οποίες είναι συνδρομητές στα topics του Kafka και είναι αυτοί που χειρίζονται τα δεδομένα (ανάγνωση, συντακτική ανάλυση, αποθήκευση). Οι Spot Workers είναι δυο κατηγοριών: πολυνηματικοί εργάτες σε γλώσσα Python και εργάτες Spark-Streaming.

Με την ολοκλήρωση της εισαγωγής των δεδομένων μπορούμε να προχωρήσουμε στον εντοπισμό των ύποπτων συνδέσεων. Το Spot για τον διαχωρισμό της κανονικής από την κίνηση με ύποπτη συμπεριφορά, χρησιμοποιεί θεματική μοντελοποιήση, διαδικασία που χρησιμοποιείται συχνά στην επεξεργασία φυσικής γλώσσας. Ορίζοντας σαν κείμενο τις καταγραφές που σχετίζονται με μια συγκεκριμένη IP δ/ση, στη συνέχεια μπορεί να εφαρμοστεί ο αλγόριθμος μηχανικής μάθησης LDA που υποστηρίζεται από την βιβλιοθήκη SparkML, όπου θα γίνει σημασιολογική ανάλυση βαθμολογώντας τις εγγραφές. Οι εγγραφές με την μικρότερη βαθμολογία θα χαρακτηριστούν ύποπτες.

Με τη διαδικασία της λειτουργικής ανάλυσης τα αποτελέσματα εμπλουτίζονται με γεωγραφικά στοιχεία και από στοιχεία που παρέχονται από υπηρεσίες φήμης σχετικά με απειλές, όπως το Facebook Threat Exchange. Τα τελικά αποτελέσματα παρουσιάζονται με χρήση σύγχρονων τεχνολογιών βασισμένες σε JavaScript. Η πρόσβαση στα δεδομένα είναι εύκολη με την τεχνολογία GraphQL[15]. Για το περιβάλλον χρήστη χρησιμοποιείται ReactJS πάνω σε αρχιτεκτονική Flux[16], ενώ η οπτικοποίηση των δεδομένων γίνεται με βιβλιοθήκη D3[17]. Τέλος για την περαιτέρω επεξεργασία των δεδομένων με την ανάπτυξη κώδικα από τον αναλυτή ασφαλείας, υπάρχει δυνατότητα χρήσης iPython[18].

Το Apache Spot είναι ένα project που βασίζεται στο Hadoop, αλλά δύναται να εγκατασταθεί και σε Cloud υπηρεσίες με τις οποίες υπάρχει συμβατότητα, όπως η υπηρεσία Amazon Web Services (AWS)

[15] https://graphql.org/
[16] https://facebook.github.io/flux/
[17] https://d3js.org/
[18] https://ipython.org/

που υποστηρίζει τις βασικές βοηθητικές υπηρεσίες του Spot όπως Kafka, Spark, Hive και Impala. Υπάρχουν επίσης εγκαταστάσεις, όπου έχει χρησιμοποιηθεί το StreamSets[19], αντί του Kafka, για την πρόσληψη των δεδομένων ενώ η βιβλιοθήκη μηχανικής μάθησης του MLib του Spark υποστηρίζει μεγάλο πλήθος αλγόριθμων κατάταξης και ομαδοποίησης. Το Apache Spot μπορεί, επίσης, να τροποποιηθεί για την ανάλυση επιπλέον τύπων δεδομένων, όπως καταγραφές από : Web server, Λειτουργικά συστήματα, Τοίχους προστασίας, IDS/IPS, Kerberos , Syslog, AWS Virtual Private Cloud Flows κ.α

4.3 Αξιολόγηση συνδέσεων με μηχανική μάθηση

Στον πυρήνα της λειτουργίας του Spot βρίσκεται ο αλγόριθμός μηχανικής μάθησης Latent Dirichlet Allocation (LDA) ο οποίος εντάσσεται στην κατηγορία αλγόριθμων μη επιβλεπόμενης μάθησης. Αρχικά χρησιμοποιήθηκε στην πληθυσμιακή γενετική για τον εντοπισμό γενετικών μοτίβων και λίγο αργότερα εφαρμόστηκε στην επεξεργασία φυσικής γλώσσας. Είναι ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους στη θεματική μοντελοποίηση, δηλαδή τη διαδικασία εντοπισμού θεμάτων σε ένα σύνολο κειμένων. Με βάση τον ορισμό του, είναι ένα παραγωγικό πιθανολογικό μοντέλο για μια συλλογή εγγράφων, τα οποία αναπαρίστανται ως μείγμα λανθανόντων θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μια κατανομή λέξεων. Πρόκειται για ένα ιεραρχικό Μπαγεσιανό (Bayesian) μοντέλο τριών επιπέδων [52]. Είσοδος του αλγορίθμου είναι ένα σύνολο κειμένων (η δομή των οποίων είναι αδιάφορη) και ο αριθμός των κρυμμένων θεμάτων που το μοντέλο θα αναγνωρίσει. Η έξοδός του είναι για κάθε κείμενο μια κατανομή πιθανότητας των θεμάτων και για κάθε θέμα μια κατανομή πιθανότητας των λέξεων.

Αναλυτικότερα η λειτουργεία του αλγόριθμου είναι η παρακάτω:

Η πιθανότητα να υπάρχει κάποιο θέμα z σε ένα συγκεκριμένο κείμενο είναι $p(z|\theta_d)$ και η πιθανότητα να υπάρχει μια λέξη σε ένα συγκριμένο θέμα είναι p(w|z). Αν υποθέσουμε ότι η θεματική μείξη των λέξεων είναι ανεξάρτητη του κειμένου που εξετάζουμε, η εκτίμηση της πιθανότητας που υπολογίζει ο αλγόριθμος για τη λέξη w ενός κείμενου d είναι:

$$\mathrm{p}(w|\theta_d) = \sum_z p(w|z)p(z|\theta_d)$$

[19] https://streamsets.com

Όπως αναφέραμε, ο αλγόριθμος LDA είναι μοντέλο μη επιβλεπόμενης μάθησης, αλλά στην υλοποίησή του στο SPOT, καθώς έχει συμπεριληφθεί ανάδραση από τον χρήστη, επηρεάζεται η εκτίμηση του μοντέλου για το τι είναι ύποπτο, μετατρέποντας την λειτουργία του σε ήμι-επιβλεπόμενη μάθηση. Αν ο αναλυτής θεωρήσει ότι ένα αποτέλεσμα δεν έπρεπε να συμπεριλαμβάνεται στις ύποπτες συνδέσεις και δώσει σχετική ανατροφοδότηση (βλέπε παρακάτω), εφεξής δραστηριότητα με τα ίδια χαρακτηριστικά δεν θα θεωρείται ύποπτη. Αυτό επιτυγχάνεται στις επόμενες εκτελέσεις του αλγορίθμου, εισάγοντας στην κανονική κίνηση πολλαπλές εγγραφές, όπως αυτή που έγινε η επισήμανση, ανεβάζοντας έτσι τεχνητά τη βαθμολόγησή της. Ο αριθμός των όμοιων αυτών εγγραφών που θα προστεθούν καθορίζεται από την παράμετρο DUPFACTOR του αρχείου παραμέτρων spot.conf.

Για να προσαρμόσουμε τα δεδομένα μας στην λειτουργία του αλγόριθμου κάνουμε την παρακάτω αντιστοίχιση:

Πίνακας 2: Αντιστοίχιση εννοιών αλγορίθμου LDA σε στοιχεία των αρχείων καταγραφής

Σώμα Κειμένου	Σύνολο αρχείων καταγραφής
Κείμενο	Εγγραφές καταγραφής κάθε IP διεύθυνσης
Λέξη	Εγγραφή σε αρχείο καταγραφής (επεξεργασμένη)
Θέμα	Προφίλ συνήθους συμπεριφοράς δικτύου

Επίσης, για την εφαρμογή τεχνικής επεξεργασίας φυσικής γλώσσας σημαντικός είναι ο ρόλος την επεξεργασίας των εγγραφών για τη μετατροπή τους σε λέξεις. Ο στόχος της επεξεργασίας είναι διττός: η παραγωγή λέξεων με αρκετή επικάλυψη μεταξύ των κειμένων, ώστε να μπορεί να γίνει συσχέτιση (κάθε εγγραφή περιέχει χρονοσφραγίδα και IP δ/νση κάνοντάς την σχεδόν μοναδική) και τα αποτελέσματα που προκύπτουν να αφορούν τον τύπο της κίνησης και όχι τον κόμβο που τη δημιουργεί. Καθώς κάθε τύπος δεδομένων έχει διαφορετικά στοιχεία στις εγγραφές του, υπάρχει συγκεκριμένος αλγόριθμος κωδικοποίησης για τη μετατροπή της εγγραφής σε λέξη. Στον παρακάτω πίνακα παρουσιάζονται τα πεδία των δεδομένων από τα αρχεία καταγραφής που αποθηκεύονται στο HDFS κατά τη διαδικασία της πρόσληψης. Με έντονα γράμματα έχουν επισημανθεί τα πεδία που συμμετέχουν στον σχηματισμό των λέξεων.

Πίνακας 3: Πεδία που αποθηκεύονται στο HDFS κατά την πρόσληψη των δεδομένων

Δεδομένα FLOW	Δεδομένα DNS	Δεδομένα Proxy
trhour: η ώρα της καταγραφής της ροής	frame_time: χρονοσφραγίδα του πλαισίου του ερωτήματος	p_date: ημέρα του ερωτήματος
sip: πηγαία διεύθυνση IP της ροής	unix_tstamp: χρονοσφραγίδα Unix του ερωτήματος	**p_time**: ώρα του ερωτήματος
dip: διεύθυνση IP προορισμού της ροής	**frame_len**: μήκος πλαισίου	clientip: διεύθυνση IP του πελάτη που κάνει τα αιτήματα στο διακομιστή
sport: πηγαία θύρα της ροής	ip_dst: διεύθυνση IP του πελάτη που κάνει το ερώτημα	host: κόμβος του αιτήματος
dport: θύρα προορισμού της ροής	**dns_qry_name**: όνομα του ερωτήματος DNS	**reqmethod**: μέθοδος του αιτήματος
proto: πρωτόκολλο που χρησιμοποιήθηκε από τη ροή	dns_qry_class: Κλάση του ερωτήματος DNS	**useragent**: πράκτορας χρήστη
ipkt: αριθμός πακέτων της ροης	**dns_qry_type**: τύπος του ερωτήματος DNS	**resconttype**: τύπος περιεχομένου της απάντησης
ibyt: αριθμός byte της ροής	**dns_qry_rcode**: κώδικας απάντησης του ερωτήματος DNS	**respcode**: κωδικός απάντησης
		fulluri: ολόκληρο το URI του αιτήματος

Σαν παράδειγμα εφαρμογής του αλγορίθμου, η δημιουργούμενη λέξη για δεδομένα ροής δικτύου (για την ακρίβεια είναι δύο λέξεις, μια ανά κατεύθυνση) για την παρακάτω εγγραφή:

```
Date first seen        Duration    Proto   Src IP Addr:Port        Dst IP Addr:Port          Flags  Tos  Packets
2016-01-28 19:58:56.588   57.512 TCP       x.x.37.98:52669 ->   x.x.155.142:18156          .A....   0      12

Bytes   pps     bps   Bpp Flows
904     0       125   75    1
```

είναι:

333333_TCP_19_11_5 και για τις δυο κατευθύνσεις.

4.4 Οπτικοποίηση και παρουσίαση αποτελεσμάτων

Εστιάζοντας στα στοιχεία που εξάγονται ως πληροφορία για τον αναλυτή ασφαλείας, το περιβάλλον χρήσης είναι δομημένο κατά αντιστοιχία των βημάτων που ακολουθεί μια διαδικασία αξιολόγησης απειλών και περιλαμβάνει:

- την εξέταση και αξιολόγηση των ύποπτων συνδέσεων, οι οποίες είναι εμπλουτισμένες με επιπλέον στοιχεία, για τον εντοπισμό πραγματικών απειλών.
- τη διερεύνηση των απειλών που εντοπίστηκαν, στο σύνολο των δεδομένων.
- τη συνολική παρουσίαση των αποτελεσμάτων.
- τη δυνατότητα αναζήτησης επιπλέον στοιχείων γράφοντας, κατά περίπτωση, κώδικα σε iPython.

4.4.1 Suspicious

Σε όλες τις περιπτώσεις το κεντρικό παράθυρο για την ανάλυση είναι το Suspicious (Ύποπτες συνδέσεις), το οποίο, ανάλογα με τον τύπο των δεδομένων που αναλύουμε, περιλαμβάνει τα παρακάτω στοιχεία:

Πίνακας 4: Στοιχεία που περιλαμβάνονται στις ύποπτες συνδέσεις

Δεδομένα Flow	Δεδομένα Dns	Δεδομένα Proxy
Time	Timestamp	Time
Source IP	Client IP	Client IP
Destination IP	Query	Host
Source Port	Query Class	Web Category
Destination Port	Query Type	Response Code
Protocol	Response Code	
Input Packets		
Input Bytes		
Output Packets		
Output Bytes		

Οι παράμετροι αυτές των συνδέσεων είναι τα κύρια χαρακτηριστικά με βάση τα οποία ο αναλυτής ασφαλείας θα αξιολογήσει κάθε σύνδεση. Τα στοιχεία έχουν προκύψει είτε άμεσα από τα δεδομένα που έχουμε εισάγει στο σύστημα είτε έμμεσα, αφού προστίθενται κατά τη λειτουργική ανάλυση (ΟΑ) των αποτελεσμάτων. Τέτοια στοιχεία είναι ο χαρακτηρισμός των IP διευθύνσεων ως εσωτερικές ή εξωτερικές δ/νσεις του δικτύου, η προσθήκη στοιχείων προσδιορισμού της γεωγραφικής θέσης της εξωτερικής IP (περιλαμβάνει και το όνομα του τομέα στον οποίο έχει

αποδοθεί η IP δ/νση) και η on line αξιολόγηση που έχει αυτή η IP από τις υπηρεσίες McAfee Global Threat Intelligence ή Facebook Threat Exchange.

Στο παράθυρο Scoring έχουμε το επιβλεπόμενο (supervised) μέρος του μοντέλου μηχανικής μάθησης. Εδώ ο αναλυτής ασφαλείας έχει τη δυνατότητα να αξιολογήσει τις ύποπτες συνδέσεις που το μοντέλο έχει αναγνωρίσει. Μπορεί να χαρακτηρίσει την επικινδυνότητα, ως High (1), Middle (2) ή Low (3) ανάλογα με την εικόνα που σχηματίζει για τη σύνδεση αυτή συνυπολογίζοντας και άλλες παραμέτρους, που άπτονται της ασφάλειας δικτύων. Επόμενες εκτελέσεις του μοντέλου μηχανικής μάθησης, θα λάβουν υπόψιν αυτές τις αξιολογήσεις, ώστε το μοντέλο, όσο χρησιμοποιείται, να γίνεται πιο αποδοτικό.

Επιλέγοντας κάποια συγκεκριμένη γραμμή στο παράθυρο των ύποπτων συνδέσεων έχουμε επισήμανση της σύνδεσης, όπως αυτή εμφανίζεται διάγραμμα του παράθυρου Network View. Στο παράθυρο Details, εμφανίζονται:

- για την ανάλυση δεδομένων NetFlow, όλες οι συνδέσεις που έγιναν το ίδιο λεπτό, μεταξύ των δύο δ/νσεων IP της υπό εξέταση ροής.
- για την ανάλυση δεδομένων DNS, λεπτομέρειες για τη συγκεκριμένη DNS εγγραφή.
- για την ανάλυση δεδομένων Proxy εμφανίζονται επιπλέον λεπτομέρειες που σχετίζονται με τη συγκεκριμένη εγγραφή, όπως User Agent, MIME Type, Proxy Server IP, Bytes.

Στο παράθυρο Network View έχουμε γραφική αναπαράσταση των ύποπτων συνδέσεων. Σε χρηστικό επίπεδο αυτό μας δίνει μια οπτικοποίηση των σχέσεων μεταξύ των ύποπτων συνδέσεων. Αν έχουμε προσδιορίσει τις εσωτερικές δ/νσεις του δικτύου, τότε οι εσωτερικοί κόμβοι εμφανίζονται ως ρόμβοι και οι εξωτερικοί ως κύκλοι.

Από το Network View μπορούμε εύκολα να δούμε:

- για την ανάλυση δεδομένων Netflow, επιλέγοντας κάποιο κόμβο, στο παράθυρο Detail, εμφανίζεται ένα διάγραμμα χορδών, το οποίο παρουσιάζει τις συνδέσεις του εν λόγω κόμβου με τους υπόλοιπους κόμβους που υπάρχουν μέσα στις ύποπτες συνδέσεις καθώς και τον όγκο των δεδομένων που αντάλλαξαν. Το δεξί κλικ πάνω σε κάποιο κόμβο λειτουργεί ως φίλτρο των ύποπτων συνδέσεων ως προς αυτόν.
- για την ανάλυση δεδομένων DNS, επιλέγοντας κάποιο κόμβο, στο παράθυρο Details εμφανίζει ένα δενδρόγραμμα που παρουσιάζει όλα τα ερωτήματα DNS που έγιναν από την

συγκεκριμένη δ/νση IP. Το αριστερό κλικ λειτουργεί ως φίλτρο με βάση τη συγκεκριμένη IP δ/νση, για τα περιεχόμενα των παραθύρων Suspicious και Details.

- για την ανάλυση δεδομένων Proxy, το διάγραμμα για την οπτικοποίηση είναι διάγραμμα τύπου hierarchical force με την παρακάτω δομή:

  Root Proxy Node

  Proxy Request Method

  Proxy Host

  Proxy Path

  Client IP Address

  Οι κόμβοι της ιεραρχικής δεντρικής δομής παρουσιάζονται με διαφορετικά χρώματα, ανάλογα με την αξιολόγηση που μας έχει δώσει η υπηρεσία αξιολόγησης υπόληψης της McAfee ή του Facebook (εφόσον αυτές έχουν ενεργοποιηθεί).

Τα πεδία που παρουσιάζονται στον αναλυτή ασφαλείας για την αξιολόγηση – χαρακτηρισμό (scoring) των συνδέσεων, ανά κατηγορία δεδομένων, είναι τα παρακάτω:

Πίνακας 5: Στοιχεία που παρουσιάζονται για την αξιολόγηση των ύποπτων συνδέσεων

Δεδομένα Flow	Δεδομένα Dns	Δεδομένα Proxy
Source IP	Client IP	URI
Destination IP	Query	
Source Port		
Destination Port		

4.4.2 Threat Investigation

Μετά την αξιολόγηση ακολουθεί η Threat Investigation (Διερεύνηση Απειλών), η οποία αφορά μόνο τις συνδέσεις που κατά το προηγούμενο βήμα χαρακτηρίστηκαν υψηλού κινδύνου (High). Για τις συνδέσεις αυτές και ανά κατηγορία δεδομένων θα δούμε τις παρακάτω επιπλέον λεπτομέρειες:

- για την ανάλυση δεδομένων NetFlow, για τη δραστηριότητα της συγκεκριμένης IP δ/νσης εμφανίζονται οι παρακάτω πίνακες με τα 20 πρώτα αποτελέσματα, που προκύπτουν από την αναζήτηση στο σύνολο των δεδομένων:

Top source IP per connections

Top destination IP per connections

Top source IP per bytes transferred

Top destination IP per bytes transferred

- για την ανάλυση δεδομένων DNS, εμφανίζονται τα 20 πρώτα αποτελέσματα που προέκυψαν κατά την αναζήτηση στο σύνολο των δεδομένων. Μας επιστρέφει τα ερωτήματα DNS της συγκεκριμένης IP δ/νσης, εφόσον αναζητούμε με βάση την IP ή τις IP δ/νσεις που έκαναν το ίδιο ερώτημα, αν αναζητούμε με βάση το ερώτημα.
- για την ανάλυση δεδομένων Proxy θα γίνει αναζήτηση, ώστε να βρεθεί επιπλέον δραστηριότητα σχετική με τη συγκεκριμένη proxy εγγραφή του διακομιστή. Τα αποτελέσματα επί του συνόλου των δεδομένων παρουσιάζονται και εδώ με τη μορφή πίνακα που περιλαμβάνει αναλυτικά στοιχεία κάθε εγγραφής.

Με την επισκόπηση κάθε απειλής, την προσθήκη των σχετικών σχολίων και την αποθήκευσή τους δημιουργούνται τα αρχεία, τα οποία σχετίζονται με την οπτικοποίηση των δεδομένων που θα παρουσιαστούν στο Storyboard. Η δημιουργία των αρχείων απαιτεί κάποιο χρόνο καθώς παίρνει αποτελέσματα από τα σχετικά ερωτήματα impala επί των δεδομένων.

4.4.3 Storyboard

Εδώ γίνεται μια συνολική αναφορά στις απειλές της συγκεκριμένης ημέρας και βασίζεται στα δυο προηγούμενα στάδια. Απευθύνεται κυρίως στην ιεραρχία για γρήγορη και περιεκτική ενημέρωση. Περιλαμβάνει το σύνολο των απειλών που βρέθηκαν με τα σχόλια επί αυτών. Παρουσιάζονται, επίσης, γραφήματα αναφορικά με τη χρονική σειρά των γεγονότων και γραφήματα με λεπτομέρειες για τις συνδέσεις που συνάδουν με τις απειλές που εντοπίστηκαν.

4.5 Εγκατάσταση

Μια τυπική εγκατάσταση περιλαμβάνει τρείς κόμβους που ανήκουν σε μια συστοιχία Hadoop, με διακριτούς ρόλους κατά αντιστοιχία των βασικών λειτουργιών του Spot χωρίς αυτό να είναι δεσμευτικό. Οι κόμβοι είναι: ο *ingest-node* (ή gateway) για την πρόσληψη των δεδομένων, ο *ml-node* για τη βαθμολόγηση με μηχανική μάθηση και oa-node (ή UI) για την επιχειρησιακή ανάλυση, οπτικοποίηση και τη διαχείριση από το Cloudera Manager. Στην Hadoop συστοιχία πρέπει να έχουν

εγκατασταθεί και παραμετροποιηθεί οι παρακάτω υπηρεσίες: HIVE, IMPALA, KAFKA, SPARK, YARN, Zookeeper. Η κατανομή των λειτουργιών και των ρόλων στους 3 κόμβους της παρούσας εγκατάστασης φαίνεται στον παρακάτω πίνακα:

Πίνακας 6: Κατανομή ρόλων στους 3 διακομιστές του οικοσυστήματος Hadoop

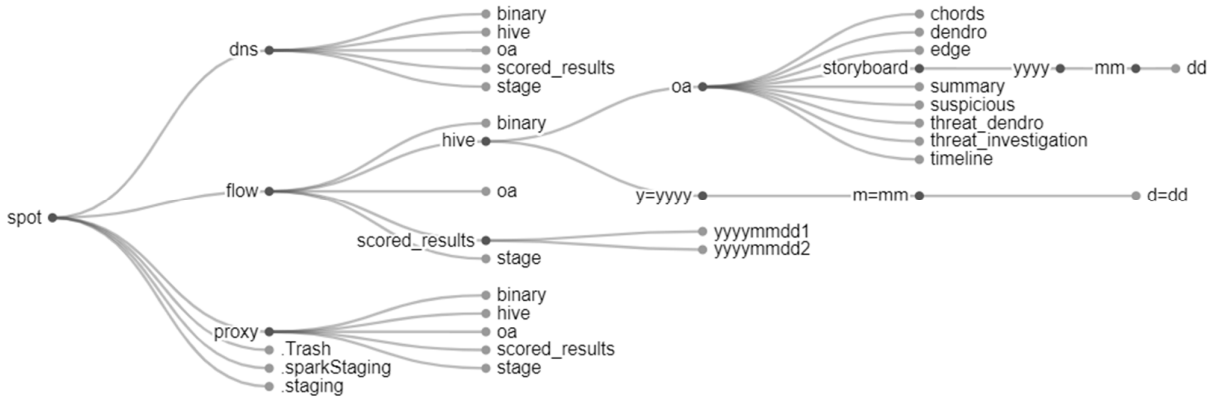
OA-Node: Cloud1	INGEST- Node: Cloud2	ML-Node: Cloud3
HDFS Gateway	HDFS Balancer	HDFS DataNode
	HDFS DataNode	HDFS SecondaryNameNode
	HDFS NameNode	Hive Gateway
	Hive Gateway	HiveServer2
	Hive Metastore Server	Impala Daemon
	HiveServer2	Impala StateStore
	Hue Load Balancer	Kafka Broker
	Hue Server	Spark 2 Gateway
	Impala Catalog Server	Spark Gateway
	Kafka Broker	Spark History Server
	Spark 2 Gateway	YARN (MR2 Included) NodeManager
	Spark 2 History Server	ZooKeeper Server
	Spark Gateway	
	YARN JobHistory Server	
	YARN ResourceManager	

Σε κάθε κόμβο θα εγκατασταθούν διαφορετικές λειτουργίες, όμως σε όλους απαιτείται η ύπαρξη του κεντρικού αρχείου παραμετροποίησης spot.conf. Το αρχείο περιέχει υποενότητες για την παραμετροποίηση των επιμέρους λειτουργιών. Πιο συγκεκριμένα περιλαμβάνει:

- τις IP δ/νσεις των κόμβων, το όνομα του χρήστη στο HDFS.
- την IP δ/νση του Hadoop NameNode, τις διαδρομές των φακέλων που θα γίνεται η αποθήκευση των δεδομένων στο Hive και η εξαγωγή των αποτελεσμάτων του αλγορίθμου.
- το όνομα της βάσης δεδομένων και τα στοιχεία του διακομιστή Impala.
- στοιχεία για την παραμετροποίηση της αυθεντικοποίησης kerberos (αν χρησιμοποιείται).

- όνομα του χρήστη στο linux με ssh πρόσβαση σε όλους τους κόμβους και διαδρομές για την αποθήκευση αποτελεσμάτων.
- παραμέτρους για τη λειτουργία του Spark, όπως αριθμός executor, μέγεθος μνήμης και αριθμός πυρήνων τους, μέγεθος μνήμης και άλλα στοιχεία του driver.
- παραμέτρους για τη λειτουργία του αλγορίθμου όπως Alpha, Beta, αριθμός θεμάτων, προεπιλεγμένη τιμή ορίου πιθανοτήτων για τον χαρακτηρισμό των συνδέσεων ως ύποπτες, τον αριθμό των καταγραφών που θα εισάγει για κάθε ανατροφοδότηση από τον χρήστη.

Κατά την εγκατάσταση το αρχείο spot.conf χρησιμοποιείται μεταξύ άλλων από το κεντρικό αρχείο εγκατάστασης hdfs_setup.sh το οποίο δημιουργεί τις βάσεις δεδομένων και τους φακέλους του HDFS. Στην Εικόνα 10 παρουσιάζεται η δομή των φακέλων που δημιουργούνται στο HDFS με τους υποφακέλους του flow σε πλήρη ανάπτυξη. Για την διαδικασία πρόσληψης δεδομένων στο αρχείο ingest_conf.json καθορίζονται τα στοιχεία των διακομιστών Kafka και Zookeeper, οι παράμετροι για το Spark-streaming καθώς και οι προσωρινοί φάκελοι που χρησιμοποιούνται κατά την πρόσληψη των δεδομένων (πριν την εισαγωγή τους στο HDFS) και οι τύποι των αρχείων που θα επεξεργάζονται τα εργαλεία ανάλυσης.



Εικόνα 10: Δομή φακέλων HDFS

Για την εγκατάσταση του Spot εκτός από τον πηγαίο κώδικα των φακέλων spot-ingest, spot-ml και spot-oa που πρέπει να μεταφερθούν στους αντίστοιχους κόμβους, τα κύρια δομικά στοιχεία λογισμικού που εγκαθίστανται σε κάθε κόμβο είναι:

Πίνακας 7: εγκατεστημένο λογισμικό στοιχεία ανά κόμβο

OA-Node: Cloud1	INGEST- Node: Cloud2	ML-Node: Cloud3
Python 2.7	pip	Scala
TLD	kafka-python	sbt
python-dev	watchdog	
ipython	nfdump (έκδοση για spot)	
npm	tshark	
browserify	screen	
uglify-js	spark-streaming	
mercurial		

4.6 Λειτουργία

Βήμα 1ο Πρόσληψη δεδομένων

Η πρόσληψη δεδομένων γίνεται ξεχωριστά για κάθε τύπο αρχείων δεδομένων που αναλύουμε και μπορεί να γίνει είτε με λήψη δεδομένων σε πραγματικό χρόνο (έχοντας ρυθμίσει σχετικά τον εξοπλισμό για την εξαγωγή τους) είτε μεταφέροντας αρχεία δεδομένων που έχουμε ήδη καταγράψει. Για παράδειγμα, για ροή δεδομένων δικτύου και έχοντας καθορίσει το φάκελο ~/flow στο αρχείο ingest_conf, η πρόσληψη μπορεί να γίνει είτε με εντολή (στον INGEST- Node) όπως: nfcapd -D -T all -p 9995 -w -t 120 -l ~/flow

είτε με μεταφορά (από τον κόμβο που έχουμε τα δεδομένα):

#scp nfcapd.201905121252 cloud2:~/ flow

Βήμα 2ο Έναρξη συλλέκτη (INGEST- Node)

$python master_collector.py -t flow -w 4

όπου flow(ή dns, ή proxy) ο τύπος των δεδομένων και 4 ο αριθμός των εργατών που θα δημιουργηθούν. Με την ενεργοποίηση του collector δημιουργείται το topic της ροής του Kafka. Το όνομα του topic εμφανίζεται στον χρήστη.

Βήμα 3ο Έναρξη εργατών (INGEST- Node)

Γνωρίζοντας το topic (πχ SPOT-INGEST-flow_16_07_19_851257) δημιουργούμε τους 4 workers:

$python worker.py -t flow -i 0 --topic SPOT-INGEST-flow_16_07_19_851257
$python worker.py -t flow -i 1 --topic SPOT-INGEST-flow_16_07_19_851257
$python worker.py -t flow -i 2 --topic SPOT-INGEST-flow_16_07_19_851257
$python worker.py -t flow -i 3 --topic SPOT-INGEST-flow_16_07_19_851257

Τα 2 προηγούμενα βήματα μπορούν να ενοποιηθούν με την παρακάτω δέσμη ενεργειών ingest_data.sh που δέχεται ως ορίσματα τον τύπο των δεδομένων και τον αριθμό των worker.

```
echo Executing:
echo "python master_collector.py -t $1 -w $2 &> collector.log"
python master_collector.py -t $1 -w $2 &> collector.log &
sleep 10
topic=$(cat collector.log | head -2 |  grep -oP "SPOT-INGEST-[^[:space:]]+")
echo Created topic $topic
echo Executing:
for ((i=0;i<=($2-1);i++));
do
     python worker.py -t $1 -i $i --topic $topic &> worker$i.log &
     echo "python worker.py -t $1 -i $i --topic $topic"
done
echo "Ready for $1 ingestion with $2 workers"
echo ""
echo "Check collector.log for ingestion details"
```

Βήμα 4ο Εκτέλεση αλγορίθμου μηχανικής μάθησης (ML-Node)

Μετά την ολοκλήρωση της πρόσληψης των δεδομένων και την καταχώρησή τους σε HDFS και Hive, μπορεί να γίνει η βαθμολόγηση των συνδέσεων με την εντολή:

./ml_ops.sh YYYMMDD <τύπος δεδομένων> <τιμή ορίου για χαρακτηρισμό ύποπτης> <μέγιστος αριθμός αποτελεσμάτων>

πχ: ./ml_ops2.sh 20160128 flow 1e-20 200

Βήμα 5ο Εφαρμογή επιχειρησιακής ανάλυσης (OA-Node)

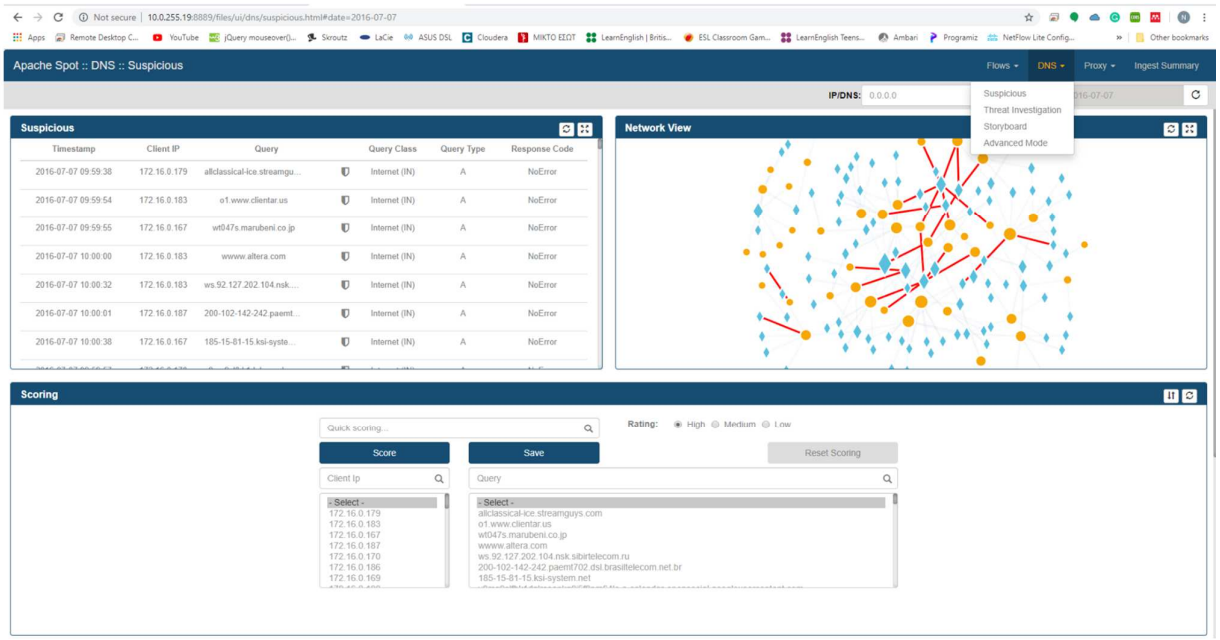
Για τον εμπλουτισμό των αποτελεσμάτων και την επιχειρησιακή ανάλυση εκτελούμε τον κώδικα του προγράμματος start_oa.py. Για το προηγούμενο παράδειγμα:

$sudo python start_oa.py -d 20160128 -t flow -l 200

όπου γίνεται ανάλυση των 200 πρώτων αποτελεσμάτων, δεδομένων flow για τη συγκεκριμένη ημέρα.

Για την φόρτωση της διεπαφής χρήσης του Spot (βλέπε *Εικόνα 11*)–αν δεν είναι ήδη ενεργοποιημένη- πρέπει να εκτελέσουμε την παρακάτω εντολή:

$ sudo ./runIpython.sh

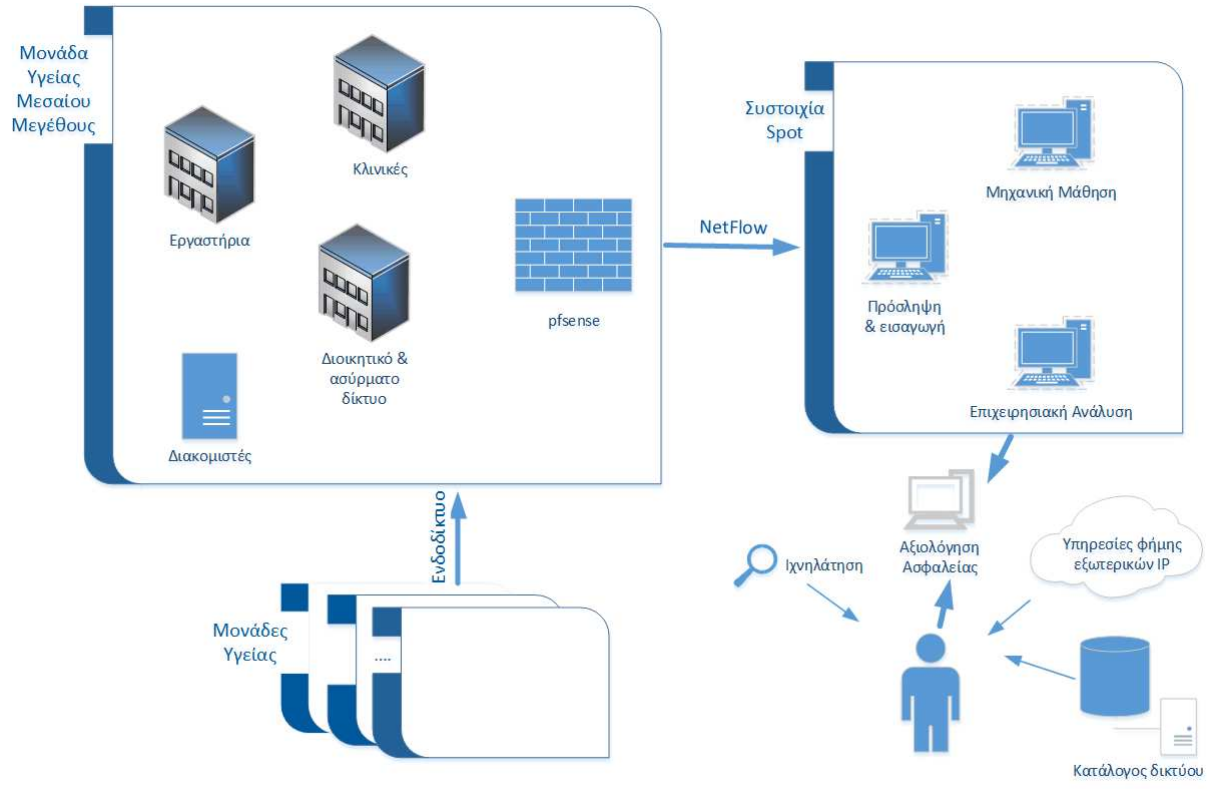


Εικόνα 11: Περιβάλλον χρήσης Apache Spot - Suspicious Connections

Κεφάλαιο 5 - Ανάλυση κίνησης δικτύου υποδομής στο χώρο της Υγείας

5.1 Εισαγωγή

Στα προηγούμενα κεφάλαια έγινε παρουσίαση των προκλήσεων και των εξελίξεων στον τομέα της έρευνας για την κυβερνοασφάλεια των ΥΖΣ. Παρουσιάστηκε, επίσης, αναλυτικά το Apache Spot ως το βασικό εργαλείο για την ανάλυση της κίνησης του δικτύου και το OpenVAS ως το εργαλείο για την αξιοποίηση των τελικών αποτελεσμάτων της προτεινόμενης μεθοδολογίας. Στο κεφάλαιο αυτό παρουσιάζουμε μια μελέτη περίπτωσης, την ανάλυση κίνησης δικτύου ΥΖΣ στον χώρο της Υγείας. Τα δεδομένα που αναλύονται είναι αρχεία καταγραφής δικτύου NetFow version 5, που προέρχονται από δίκτυο υποδομής Υγείας μεσαίου μεγέθους. Όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο (βλέπε Εικόνα 8), η ανάλυση ροών NetFlow με το Apache Spot μας δίνει τη δυνατότητα να ανιχνεύσουμε απειλές αόρατης σάρωσης, ανάκλασης, διαφυγή δεδομένων πλευρικών καναλιών, beaconing και γενικότερα ασυνήθιστες ροές δεδομένων.



Εικόνα 12: Δίκτυο ΥΖΣ και αξιολόγηση κίνησης

Η κίνηση του υπό εξέταση δικτύου περιλαμβάνει δίκτυο κλινικών, εργαστηρίων, διοικητικού και ασύρματου δικτύου για τη σύνδεση του προσωπικού (βλέπε Εικόνα 12). Περιλαμβάνει, επίσης, την

κίνηση των εξωτερικών (δημόσιων) IP δ/νσεων του οργανισμού και την κίνηση με το ενδοδίκτυο. Η κίνηση με το ενδοδίκτυο αφορά κίνηση με συνεργαζόμενες μονάδες εκτός του εξεταζόμενου οργανισμού. Σε κάποιες εσωτερικές IP δ/νσεις υπάρχουν υπηρεσίες οι οποίες είναι προσβάσιμες από το διαδίκτυο μέσω προώθησης θυρών. Είναι υπηρεσίες κυρίως ιστού, ηλεκτρονικής αλληλογραφίας και εικονικού ιδιωτικού δικτύου. Η καταχώρηση ροών ετερογενούς προελεύσεως (εσωτερικό, με το ενδοδίκτυο και εξωτερικών IP δ/νσεων) έχει το πλεονέκτημα της κεντρικής αποθήκευσης και αναζήτησης των ροών (βασική λειτουργία του Spot), αλλά καθώς το μοντέλο αξιολόγησης βασίζεται σε πιθανότητες, ενέχεται ο κίνδυνος ασυνήθιστες ροές με προέλευση ένα περισσότερο εκτεθειμένο δίκτυο να υπερσκελίζουν στατιστικά ασυνήθιστες ροές των άλλου δικτύου. Για τον λόγο αυτό, έγιναν δοκιμαστικές αναλύσεις στις ροές ορισμένων ημερών, έχοντας αφαιρέσει (φιλτράρει) την κίνηση με το ενδοδίκτυο και την κίνηση των δημοσίων IP δ/νσεων. Έχοντας δώσει επαρκή -για το μέγεθος του δικτύου- αριθμό αποτελεσμάτων (περίπου 250), δεν διαπιστώθηκε ουσιαστική επίδραση στην ποιότητα των αποτελεσμάτων από το φιλτράρισμα των ροών.

Αρχικά η εξαγωγή των δεδομένων προς τον υπολογιστή συλλέκτη των ροών, έγινε από κεντρικό μεταγωγέα επιπέδου 3 του δικτύου. Ο μεταγωγέας αυτός είχε δυνατότητα εξαγωγής δεδομένων sFlow, μια παραλλαγή του Netflow που πραγματοποιεί δειγματοληψία των ροών του δικτύου. Το Spot υποστηρίζει την ανάλυση ροών και σε μορφή sFlow αφού υπάρχει σχετική συμβατότητα με το NetFlow. Τα αποτελέσματα είναι αναμενόμενο να έχουν χαμηλότερη ακρίβεια καθώς δεν περιλαμβάνονται όλες οι ροές του δικτύου. Η ανάλυση ροών sFlow παρουσιάζει τεχνικό ενδιαφέρον καθώς στην πράξη, λόγω των χαμηλότερων απαιτήσεων σε υλισμικό, υποστηρίζεται από πολύ μεγάλο αριθμό συσκευών. Έγινε ρύθμιση του ρυθμού δειγματοληψίας στον μέγιστο που υποστηρίζει ο μεταγωγέας, αλλά η ποσότητα των δεδομένων που μπορούσε να συλλεχθεί ανά ημέρα δεν κρίθηκε επαρκής. Έτσι η εξαγωγή των προς επεξεργασία δεδομένων έγινε τελικά από το τείχος προστασίας ανοικτού κώδικα pfsense[20], στο οποίο εγκαταστάθηκε επί τούτου το πακέτο συλλέκτη NetFlow, softflowd.

5.2 Ρύθμιση παραμέτρων αλγορίθμου LDA

Όπως προαναφέρθηκε, ο αλγόριθμός LDA είναι αλγόριθμός μη επιβλεπόμενης μάθησης, δεν απαιτείται εκπαίδευσή του, αλλά πρέπει να γίνει ρύθμιση των παραμέτρων του ανάλογα με τα

[20] https://www.pfsense.org

δεδομένα που θα αναλύει. Έτσι κατά την εγκατάσταση του Spot και ανάλογα με το δίκτυο που θα χρησιμοποιηθεί, πρέπει να γίνει ορθή ρύθμιση των παραμέτρων με στόχο να έχουμε την καλύτερη ικανότητα ανίχνευσης ανωμαλιών στις ροές του. Οι παράμετροι του αλγορίθμου παρατίθενται στον παρακάτω πίνακα:

Παράμετρος	Γενική περιγραφή	Περιγραφή για την λειτουργία του Spot	Αρχικές Τιμές	Αρχείο ρύθμισης
Αριθμός επαναλήψεων (ldamaxiterations)	Αριθμός των επαναλήψεων του αλγορίθμου	Αριθμός των επαναλήψεων του αλγορίθμου	20	ml_ops.sh
Αριθμός θεμάτων (TOPIC_COUNT)	Αριθμός των θεμάτων	Αριθμός των φυσιολογικών προφίλ κίνησης του δικτύου	20	spot.conf
Παράμετρος alpha (LDA_ALPHA)	Κατανομή θεμάτων ανά κείμενο	Κατανομή των προφίλ στις καταγραφές κάθε IP διεύθυνσης	1,02	spot.conf
Παράμετρος beta (LDA_BETA)	Κατανομή λέξεων ανά θέμα	Κατανομή καταγραφών ανά προφίλ φυσιολογικής κίνησης	1,001	spot.conf

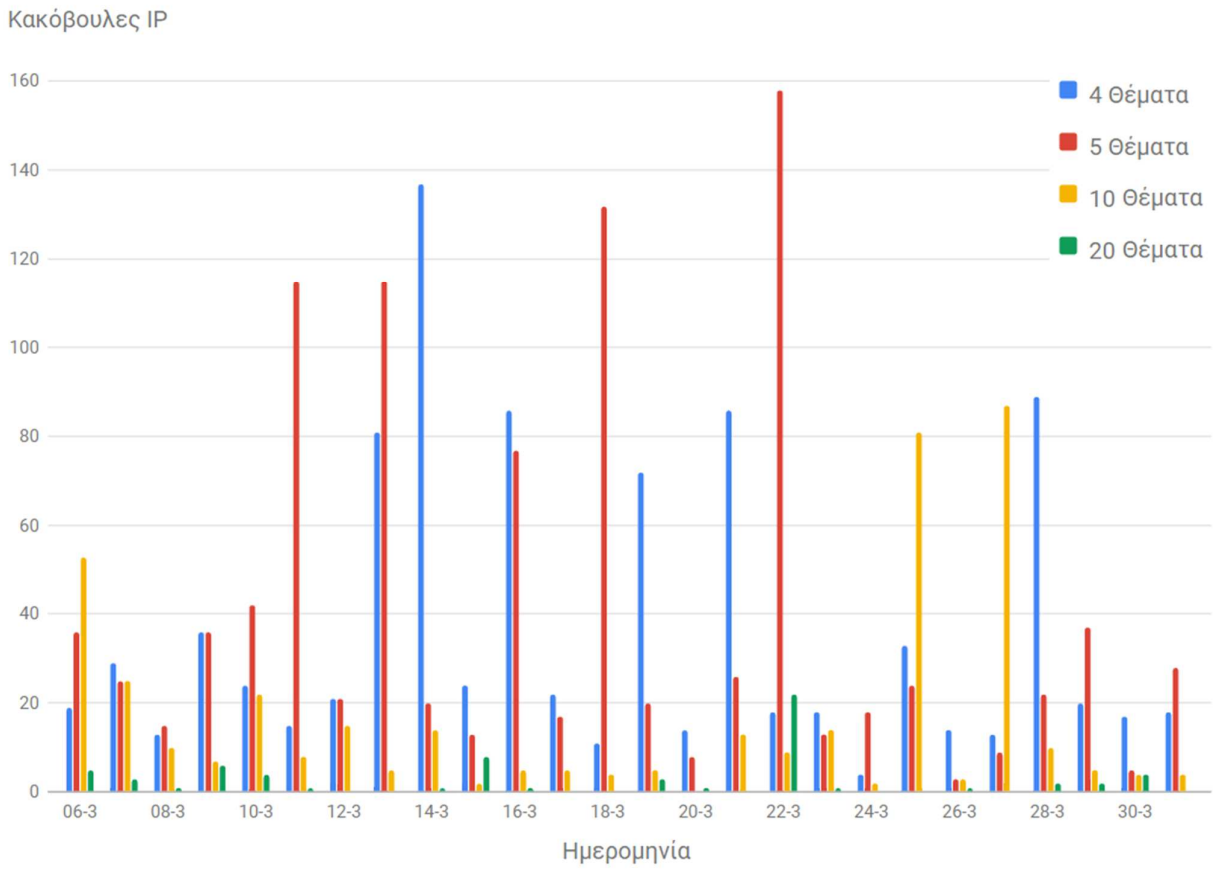
Πίνακας 8: Παράμετροι αλγορίθμου LDA

Η διαδικασία ρύθμισης των παραμέτρων – εκτός τον αριθμό των επαναλήψεων - είναι εμπειρική. Βασίζεται σε δοκιμές με διαφορετικές τιμές των παραμέτρων για την αναγνώριση όσο το δυνατόν περισσότερων γνωστών επιθέσεων, που συχνά γίνονται τεχνητά για τον σκοπό αυτό. Σε ένα πραγματικό δίκτυο, το οποίο δεν ελέγχουμε διαχειριστικά, όπως αυτό που εξετάζουμε, η πραγματοποίηση επιθέσεων δεν αποτέλεσε επιλογή. Η δυσκολία αυτή παρακάμφθηκε αναλύοντας και τις ροές προς ορισμένες εξωτερικές IP δ/νσης του οργανισμού, τις οποίες συλλέξαμε πριν το τοίχος προστασίας, συνεπώς στις ροές αυτές είναι καταγεγραμμένες επιθέσεις που είναι συνήθεις σε IP δ/νσεις που είναι εκτεθειμένες στο διαδίκτυο. Για την ταυτοποίηση των απειλών αυτών χρησιμοποιήσαμε την υπηρεσία φήμης OTX.

Ο αριθμός των επαναλήψεων πρέπει να είναι όσο το δυνατόν μεγαλύτερος, όμως η αύξησή του αυξάνει αντίστοιχα και τον χρόνο της επεξεργασίας για την εξαγωγή των ύποπτων ροών ενώ

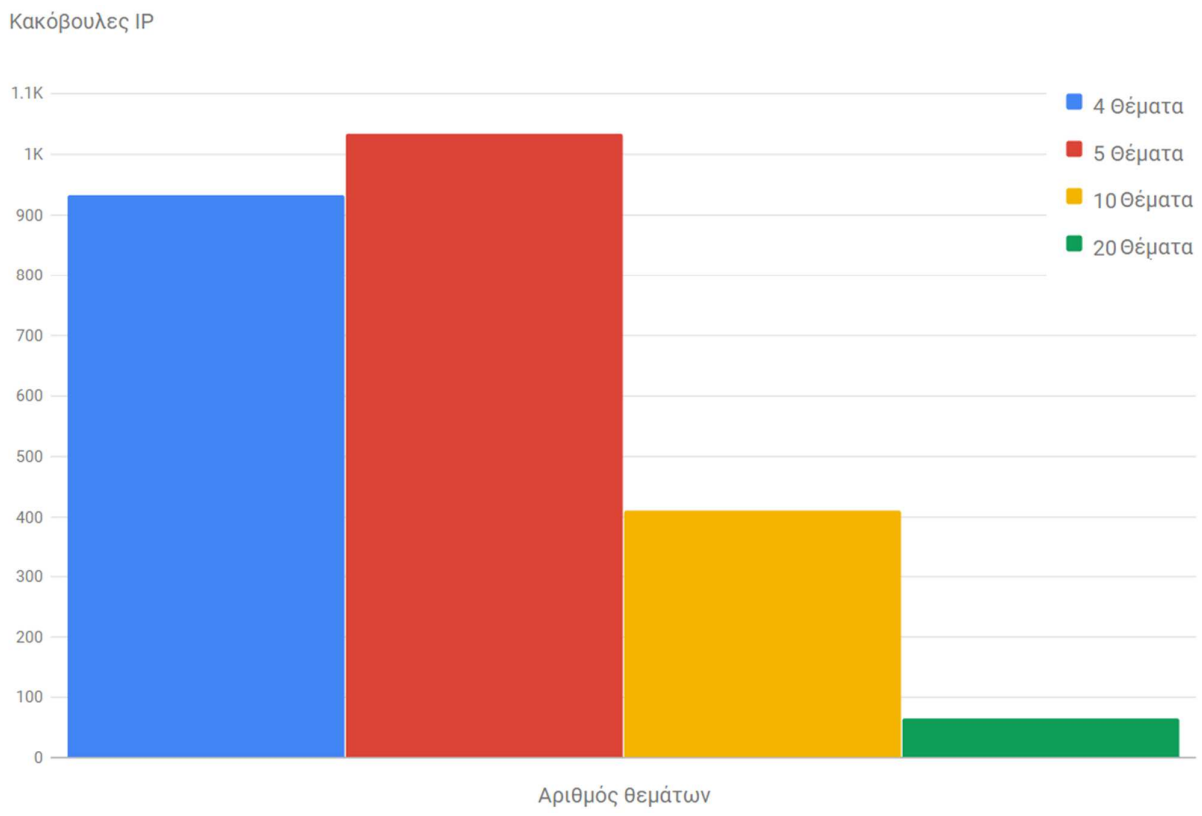
αυξάνονται και οι απαιτήσεις σε μνήμη. Αν η μνήμη δεν είναι επαρκής η διαδικασία τερματίζεται με σφάλματα χωρίς αποτελέσματα. Σύμφωνα με σχετικές συστάσεις για το SPOT, ο αριθμός των επαναλήψεων είναι επαρκής, όταν ξεπερνά τις 100. Για την ανάλυσή μας θέσαμε τον αριθμό των επαναλήψεων στο 120, αφού πρώτα είχαμε διαθέσει αρκετή μνήμη στις εικονικές μηχανές (20 GB) και έχοντας παραμετροποιήσει σχετικά την διαχείριση των πόρων. Η αύξηση του αριθμού των επαναλήψεων αύξησε τον χρόνο επεξεργασίας των ημερήσιων δεδομένων για την εκτέλεση του αλγορίθμου, από 15 περίπου λεπτά για 20 επαναλήψεις, στα 90 περίπου λεπτά για 120 επαναλήψεις (ανάλυση ροών όγκου 700 MB).

Για την επιλογή του αριθμού των θεμάτων -και στη συνέχεια των υπερπαραμέτρων alpha και beta- έγινε ανάλυση πολλών ημερών και επιλέχθηκε η τιμή με την οποία εντοπίστηκαν οι περισσότερες κακόβουλες IP. Γενικά προτείνονται πολλές αναλύσεις της ίδια ημέρας και επιλογή των παραμέτρων με τις οποίες προέκυψαν κατά μέσο όρο καλύτερα αποτελέσματα. Οι πολλές επαναλήψεις προτείνονται γιατί αναλύσεις με τις ίδιες παραμέτρους δεν έχουν πανομοιότυπα αποτελέσματα. Στην υλοποίησή μας, επιλέξαμε να μην αναλύσουμε πολλές φορές την ίδια ημέρα, αλλά να κάνουμε μια ανάλυση ανά ημέρα για πολλές ημέρες, καθώς η κίνηση του δικτύου ενδέχεται να έχει σημαντικές διαφοροποιήσεις (ημέρες εφημερίας, Σαββατοκύριακα κ.α). Έτσι έγιναν αναλύσεις για περίπου 4 εβδομάδες και επιλέξαμε τις παραμέτρους με τις οποίες εντοπίστηκαν οι περισσότερες κακόβουλες IP δ/νσεις στις ροές των αποτελεσμάτων. Ο συνολικός αριθμός των αναλύσεων για τον προσδιορισμό των παραμέτρων είναι αρκετά μεγάλος. Η διαδικασία είναι χρονοβόρα και έγινε αυτοματοποιημένα χωρίς τη χρήση της ενσωματωμένης επιχειρησιακής ανάλυσης του SPOT-OA, αλλά με δέσμες ενεργειών και πελάτη Python του API της υπηρεσίας OTX. Για την επιλογή του αριθμού των θεμάτων έγιναν αναλύσεις για τιμές 4, 5, 10 και 20. Τα αποτελέσματα ανά ημέρα για όλες τις ημέρες παρουσιάζονται στο γράφημα της Εικόνα 13 .



Εικόνα 13: Κακόβουλες IP δ/νσεις που ανιχνεύτηκαν ανά ημέρα και αριθμό θεμάτων

Όπως προκύπτει από το παραπάνω γράφημα δεν υπάρχει σταθερά καλύτερη απόδοση για όλες τις ημέρες για τις οποίες έγινε ανάλυση, με συγκεκριμένο αριθμό θεμάτων. Αυτός ήταν ένας επιπλέον λόγος που έγιναν αναλύσεις για πολλές ημέρες. Η εικόνα είναι πιο κατατοπιστική στο επόμενο γράφημα όπου είναι εμφανές ότι τα καλυτέρα αποτελέσματα λαμβάνονται με 5 θέματα.



Εικόνα 14: Κακόβουλες IP δ/νσεις που ανιχνεύτηκαν συνολικά με διάφορες τιμές θεμάτων

Με αντίστοιχη λογική και έχοντας ορίσει των αριθμό θεμάτων 5 έγιναν οι αναλύσεις για την επιλογή των υπερπαραμέτρων alpha και beta. Οι αναλύσεις έγιναν για διάφορους συνδυασμούς τιμών και τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στην Εικόνα 15. Όπως προκύπτει από το γράφημα, έχουμε βέλτιστα αποτελέσματα, όταν έχουμε τιμές των υπερπαραμέτρων alpha=1.02 και beta=1.001. Οι τιμές αυτές ταυτίζονται με τις προεπιλεγμένες.

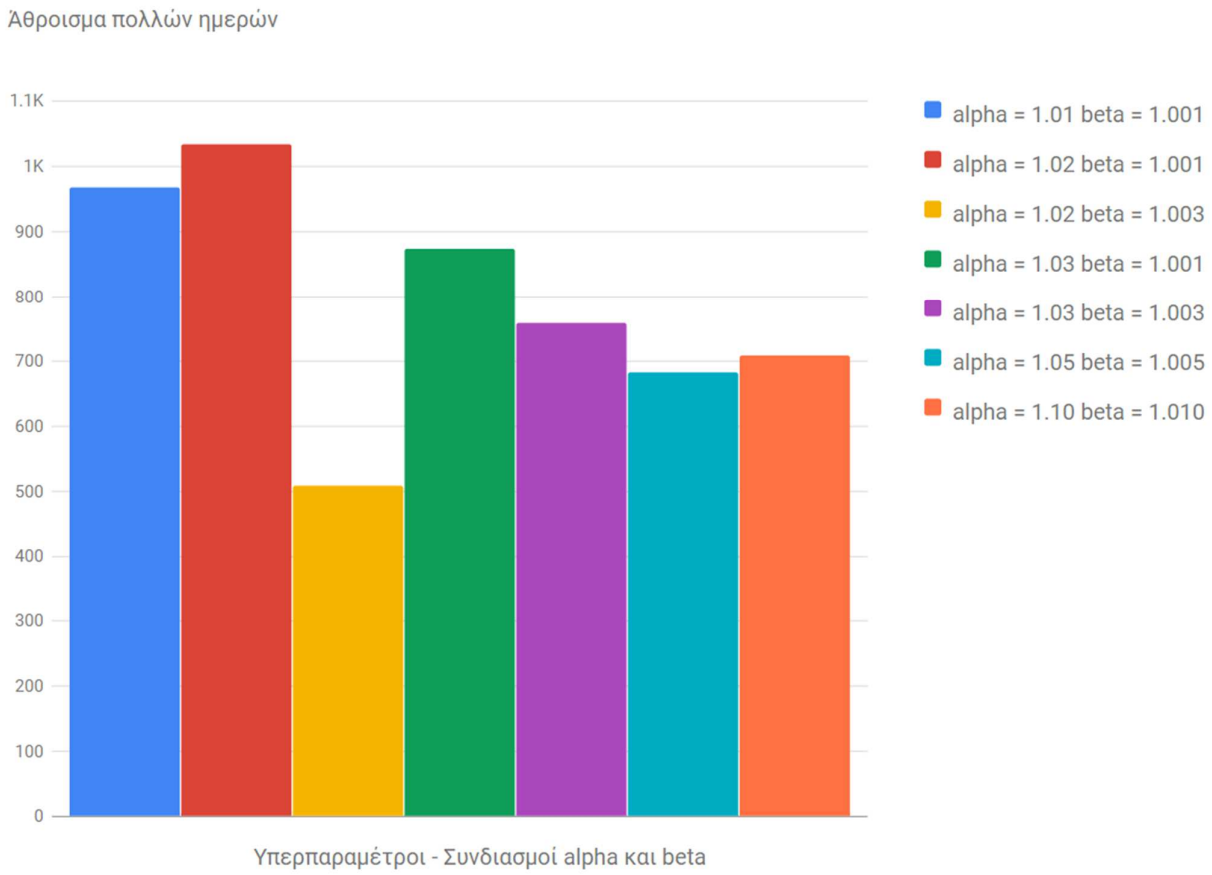
Συνοψίζοντας, οι παράμετροι αλγορίθμου για την ανάλυση ρυθμίστηκαν όπως παρακάτω:

Αριθμός επαναλήψεων: **120**

Αριθμός θεμάτων: **5**

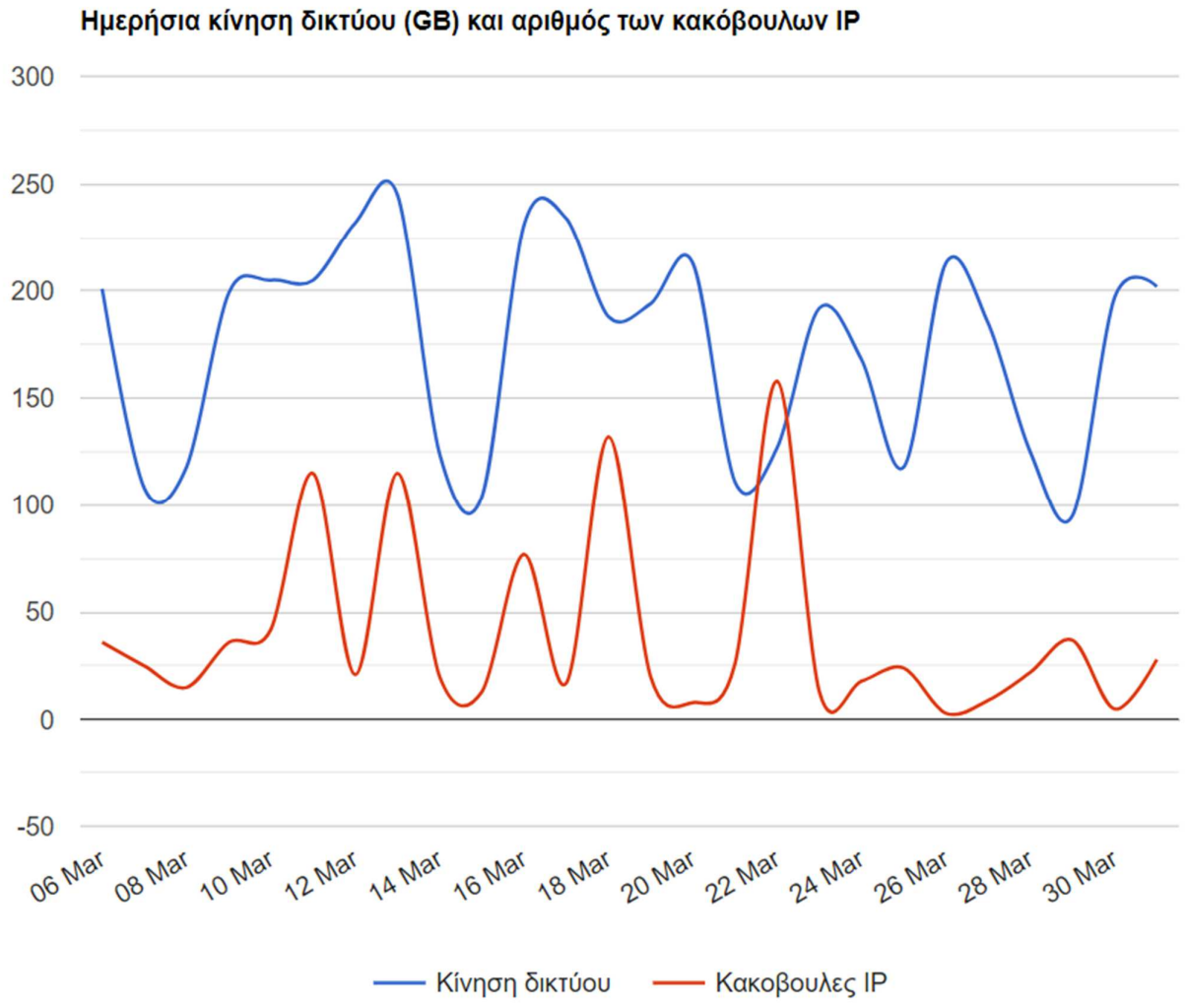
Παράμετρος alpha: **1.02**

Παράμετρος beta: **1.001**



Εικόνα 15: Κακόβουλες IPs δ/νσεις που αναγνωριστήκαν με διάφορες τιμές υπερπαραμέτρων

Έχοντας ολοκληρώσει τη ρύθμιση των παραμέτρων μπορούμε να απεικονίσουμε τα ημερήσια αποτελέσματα μαζί με τον ημερήσιο όγκο των δεδομένων.



5.3 Πλήρης ανάλυση αποτελεσμάτων μιας εβδομάδας

Εκτός από την αριθμητική καταγραφή των κακόβουλων IP που εντοπίστηκαν είναι χρήσιμο να προβούμε και σε μια πιο ενδελεχή καταγραφή και ανάλυση των αποτελεσμάτων. Η διαδικασία ανάλυσης δεν έγινε στο σύνολο των δεδομένων που συλλέχθηκαν, αλλά στα δεδομένα μιας εβδομάδας, χρόνος επαρκής για να έχουμε εικόνα των κακόβουλων συμπεριφορών που έχουν καταγραφεί και να σχηματίσουμε μια πληρέστερη εικόνα των αποτελεσμάτων.

Για την αποδοτική ανάλυση των ροών του δικτύου είναι απαραίτητη η ύπαρξη σαφούς εικόνας των κόμβων του υπό εξέταση δικτύου. Αυτό περιλαμβάνει την καταγραφή των εφαρμογών ανά κόμβο και το προφίλ των χρηστών που τους χρησιμοποιούν, την αποτύπωση των κεντρικών

πληροφοριακών δομών, τις διαδικτυακές εφαρμογές και υπηρεσίες, τις εφαρμογές ενδοδικτύου που χρησιμοποιούνται κ.α.

Στατιστικά στοιχεία των δεδομένων που συλλέχθηκαν και αναλύθηκαν παρουσιάζονται στον Πίνακας 9.

Πίνακας 9: Ποσοτικά χαρακτηριστικά δεδομένων που συλλέχθηκαν

Ημέρα Ανάλυσης	Ποσότητα Δεδομένων NetFlow	Αριθμός Ροών	Μοναδικές Εξωτερικές IP
1η	726 MB	9,434,057	22,080
2η	706 MB	8,941,456	22,717
3η	534 MB	6,836,570	16,558
4η	485 MB	6,314,026	15,942
5η	721 MB	9,251,215	23,110
6η	659 MB	8,521,860	22,282
7η	707 MB	9,145,638	23,684

Η ανάλυση του δικτύου έγινε αντλώντας από το Spot τα πρώτα 250 αποτελέσματα που έχουν πιθανότητα μικρότερη από 10^{-6}, με εκτέλεση του αλγορίθμου μια φορά, χωρίς ανατροφοδότηση. Η λειτουργία ανατροφοδότησης του αλγορίθμου μετά την εξαγωγή των πρώτων αποτελεσμάτων (που προστέθηκε στην τελευταία έκδοση), δεν χρησιμοποιήθηκε, καθώς, παρά τις προσπάθειες, λειτούργησε πλημμελώς. Με τη χρήση του Advanced Mode (περιβάλον Jupyter Notebook) και κώδικα iPython, έγινε διαλογή των εμπλεκόμενων IP δ/νσεων. Έτσι προκύπτουν οι μοναδικές εσωτερικές, εξωτερικές δ/νσεις και δ/νσεις ενδοδικτύου (βλέπε Πίνακας 10) ώστε η ιχνηλάτηση των απειλών να γίνει με διαφορετικό τρόπο.

Πίνακας 10: Κατανομή των IP διευθύνσεων στα ημερήσια αποτελέσματα

Περιγραφή	1η	2η	3η	4η	5η	6η	7η
Μοναδικές IP	187	267	243	136	155	200	240
Εσωτερικές	52	107	80	24	53	49	89
Εξωτερικές	121	149	158	109	96	142	141
Ενδοδίκτυου	10	6	3	2	4	7	6
Πολυεκπομπής	4	5	2	1	2	2	4
Επιβεβαιωμένα κακόβουλες IP	66	76	69	73	76	30	82

Εμπλεκόμενες εσωτερικές IP	4	5	3	4	6	4	4

Οι εξωτερικές δ/νσεις ελέγχονται μέσω των αντίστοιχων API από τις υπηρεσίες Open Threat Exchange (OTX) και IBM X-Force για τον εντοπισμό κακόβουλων IP δ/νσεων. Με τον όρο αυτό αναφερόμαστε σε δ/νσεις που έχουν καταγραφεί στο πρόσφατο παρελθόν να εκτελούν ενέργειες που δεν είναι αποδεκτές από πλευράς ασφάλειας. Στα αποτελέσματα που επιστρέφονται από την υπηρεσία OTX, οι κακόβουλες δ/νσεις έχουν την ένδειξη «1» ενώ στα αποτελέσματα που επιστρέφονται από την υπηρεσία IBM X-Force, οι κακόβουλες δ/νσεις έχουν βαθμολογία 2 έως 10, ανάλογα με τον βαθμό επικινδυνότητας. Ο βαθμός επικινδυνότητας είναι σύνθετος δείκτης που προκύπτει από τον τύπο της απειλής και τον βαθμό που η απειλή έχει επιβεβαιωθεί. Η υπηρεσία X-Force (μέσω του API) επιστρέφει και χαρακτηρισμό των κακόβουλων IP δ/νσεων καθώς τις κατατάσει σε μια ή περισσότερες κατηγορίες οι οποίες είναι: Anonymisation Services, Botnet Command and Control Server, Bots, Cryptocurrency Mining, Dynamic IPs, Malware, Scanning IPs και Spam. Στον Πίνακας 10 ως «Επιβεβαιωμένα κακόβουλες IP» αναγράφεται ο αριθμός των IPs που είναι καταγεγραμμένες από αυτές τις δυο υπηρεσίες. Οι «Εμπλεκόμενες εσωτερικές IP» είναι ο αριθμός των δ/νσεων IP που με βάση τα αποτελέσματα έχουν επικοινωνία με τις εν λόγω κακόβουλες IP δ/νσεις.

Η πλήρης λίστα των ανωτέρω IP δ/νσεων καταγράφεται σε αρχεία κειμένου και ακολουθείται η παρακάτω διαδικασία για την ιχνηλάτηση των απειλών:

Το αρχείο με τις εμπλεκόμενες εσωτερικές IPs οδηγείται στο OpenVas για αυτοματοποιημένη αξιολόγηση των ευπαθειών όλων των κόμβων.

Για τις δ/νσεις που έχουν καταγραφεί ως κακόβουλες εκτελούνται ερωτήματα SQL μέσω της διεπαφής του Hue ακολουθώντας τεχνικές ιχνηλάτησης, ώστε να γίνει έλεγχος για:

- Άλλες ροές από/προς την κακόβουλη IP δ/νση που κατεγράφη στα αποτελέσματα του Spot.
- Άλλες ροές από/προς το δίκτυο που ανήκει η κακόβουλη IP δ/νση.
- Ροές από/προς άλλες εσωτερικές IP δ/σεις του δικτύου μας και τα χαρακτηριστικά αυτών (χρονοσφραγίδα, θύρες, ποσότητα δεδομένων)
- Αντίστοιχες ροές που καταγράφηκαν άλλες ημερομηνίες.

Από τον παραπάνω έλεγχο και από τα σχετικά αποτελέσματα του OpenVas, ενδεχομένως, να προκύψει επιπλέον λίστα εσωτερικών κόμβων που χρήζει αξιολόγηση ευπάθειας.

Για τις ροές που χαρακτηρίστηκαν ύποπτες από το Spot, αλλά δεν έχει καταγραφεί ύποπτη συμπεριφορά από τις υπηρεσίες ανταλλαγής πληροφοριών για απειλές, γίνεται αξιολόγηση του ονόματος τομέα (όπως επεστράφη από το API του OTX), τη γεωγραφική περιοχή στην οποία ανήκουν και πότε δημιουργήθηκαν (όπου είναι διαθέσιμο). Κατά περίπτωση εκτελούνται SQL ερωτήματα όμοια με αυτά που περιγράφονται παραπάνω για τις κακόβουλες IP δ/νσεις. Επίσης, ανά περίπτωση αναζητούνται επιπλέον αξιολογήσεις από τις υπηρεσίες VirusTotal[21] της Google και Cisco Talos[22]. Στοιχεία για ευπάθειες που σχετίζονται με τις θύρες των υπό εξέταση ροών ελέγχονται στην ιστοσελίδα CVE του MITRE[23].

Η ανάλυση των ύποπτων ροών με δ/νσεις του ενδοδικτύου, επειδή δεν υπάρχει διαθέσιμη εξωτερική αξιολόγηση, γίνεται με βάση τα στοιχεία που καταγράφει το Spot και τη γνώση των εσωτερικών λεπτομερειών του δικτύου, όπως περιγράφηκε παραπάνω. Για συνδέσεις που υπάρχουν ερωτηματικά, εκτελούνται ερωτήματα SQL μέσω της διεπαφής του Hue για τη συγκέντρωση περισσότερων πληροφοριών σχετικά με τον αριθμό τυχόν αντίστοιχων ροών με τον ίδιο ή άλλους κόμβους, την ποσότητα των πακέτων και δεδομένων που ανταλλάχθηκαν, τη χρονική κατανομή τους την τρέχουσα ημέρα ή και άλλες ημέρες.

[21] https://www.virustotal.com/
[22] https://talosintelligence.com/
[23] https://cve.mitre.org/

Τα ευρήματα σε σχέση με τις κακόβουλες IP για τη διάρκεια της εβδομάδας που εξετάσαμε καταγράφονται στο Παράρτημα Α. Παρουσιάζονται ανά ημέρα και είναι πίνακες με αναλυτικά στοιχεία των κακόβουλων δ/νσεων με τις οποίες το δίκτυό μας έχει ροές. Στα στοιχεία αυτά, στη στήλη «Εταιρία» αναφέρεται είτε στην εταιρία στην οποία ανήκει η δ/νση είτε στον φορέα παροχής υπηρεσιών Διαδικτύου, στον οποίο έχει αποδοθεί το δίκτυο που περιλαμβάνει την εν λόγω IP δ/νση. Σημειώνεται ότι, σε περίπτωση που χρησιμοποιείται η τεχνική shared hosting, ο χαρακτηρισμός κακόβουλος αφορά μια ή περισσότερες φιλοξενούμενες ιστοσελίδες της συγκεκριμένης IP δ/νσης. Πέραν της περιγραφής της κακόβουλης δραστηριότητας από την υπηρεσία X-force παρατίθενται και οι παρατηρήσεις που προέκυψαν από τα ερωτήματα στο σύνολο των δεδομένων για όλες τις IP δ/νσεις που κατεγράφησαν στα αποτελέσματα.

Στην ανάλυση των δεδομένων του δικτύου, στα αποτελέσματα του Spot καταγράφονται συνολικά περίπου 916 εξωτερικές IP δ/νσεις (709 μοναδικές), από τις οποίες 472 (351 μοναδικές) χαρακτηρίζονται ως κακόβουλες, από μια τουλάχιστον εκ των υπηρεσιών OTX και X-Force. Στον Πίνακας 11 παρουσιάζεται ο αριθμός εμπλεκόμενων IP δ/νσεων ανά κατηγορία (εσωτερικών/εξωτερικών), που έχουν ροές με κακόβουλες IP και ο αριθμός των ροών αυτών ανά ημέρα.

Ημέρα	**1η**	**2η**	**3η**	**4η**	**5η**	**6η**	**7η**
Εμπλεκόμενες IP	4	5	3	6	4	4	4
Εμπλεκόμενες Εξωτερικές IP	2	2	1	2	2	2	2
Ροές με Εξωτερικές IP	96	115	102	210	109	112	100
Εμπλεκόμενες Εσωτερικές IP	2	3	2	4	2	4	2
Ροές με Εσωτερικές IP	3	4	8	13	2	9	3

Πίνακας 11: Αριθμός IP δ/νσεων και ροών που εμπλέκονται με επιβεβαιωμένα κακόβουλες δ/σνσεις

Από τα παραπάνω προκύπτει ότι στα αποτελέσματα:

- ο αριθμός των εμπλεκόμενων IP δ/νσεων σε ροές με επιβεβαιωμένα κακόβουλες δ/νσεις είναι μικρός
- ο αριθμός των ροών με επιβεβαιωμένα κακόβουλες δ/νσεις είναι αρκετά μεγάλος και σχετίζεται κυρίως με τις εξωτερικές IP δ/νσεις

Οι περιμετρικές ροές με τις εξωτερικές IP δ/νσεις που έχουν αναλυθεί από το Spot έχουν συλλεχθεί πριν από την εφαρμογή των κανόνων του τοίχου προστασίας και περιλαμβάνουν συνήθεις επιθέσεις σε κόμβους που είναι άμεσα εκτεθειμένοι στο διαδίκτυο. Τέτοιες ροές, εφόσον περιοριστούν από το τοίχος προστασίας, δεν αποτελούν πραγματικό κίνδυνο. Είναι όμως χρήσιμες για την ανάλυσή μας καθώς δείχνουν την ικανότητα του Spot να αναγνωρίζει επιθέσεις με βάση την απόκλιση της κίνησης από τα προφίλ της συνηθισμένης δραστηριότητας του δικτύου.

Συγκεντρωτικά, - όπως προκύπτει από το X-Force - οι κακόβουλες δραστηριότητες των IP δ/νσεων που καταγράφονται στα αποτελέσματα κατανέμονται, όπως φαίνεται στον Πίνακας 12. Σημειώνεται ότι αρκετές δ/νσεις ανήκουν σε περισσότερες από μια κατηγορία, με το χαρακτηρισμό «Dynamic IP» συνήθως να συνδυάζεται με κάποιο άλλο και κυρίως με «Scanning IP».

Κακόβουλη δραστηριότητα	Πλήθος IP δ/νσεων
Scanning IP	419
Anonymisation Services	4
Malware	3
Bots	8
Dynamic IP	61
Spam	21

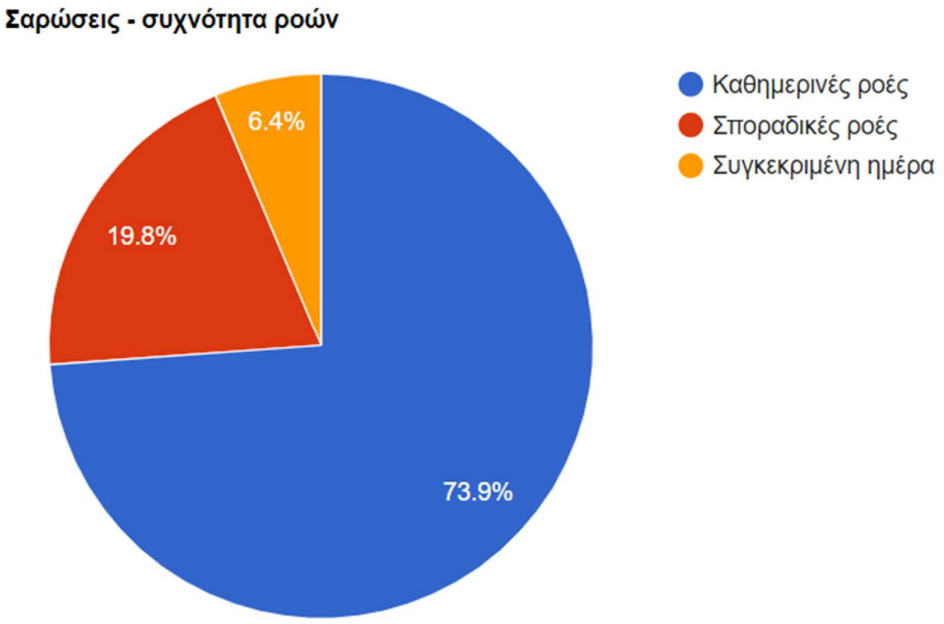
Πίνακας 12: Αριθμητική κατανομή κακόβουλων δραστηριότήτων

Οι σαρώσεις δικτύων είναι μια από τις βασικές και ευρέως χρησιμοποιούμενες τεχνικές στα πρώτα στάδια των κυβερνοεπιθέσεων, αποτελώντας ένα μεγάλο ποσοστό των επιθέσεων που εκτελούνται καθημερινά. Μπορούν να κατηγοριοποιηθούν ανάλογα με τη ροή της επίθεσης σε σχέση με την περίμετρο του δικτύου [53] ως εξής: από Εξωτερικό σε Εξωτερικό, από Εξωτερικό σε Εσωτερικό, από Εσωτερικό σε Εσωτερικό και από Εσωτερικό σε Εξωτερικό δίκτυο. Ως εσωτερικό δίκτυο θεωρούμε τους κόμβους που χρησιμοποιούν Μετάφραση Διεύθυνσης Δικτύου (NAT) για την επικοινωνία με το Διαδίκτυο (ή δεν έχουν άμεση πρόσβαση σε αυτό). Η πρώτη κατηγορία είναι αρκετά συνηθισμένη και η τελευταία αναφέρεται σε επιθέσεις που γίνονται από το εσωτερικό του δικτύου μας σε άλλα δίκτυα. Όταν η σάρωση είναι από Εσωτερικό σε Εσωτερικό δίκτυο, ο κακόβουλος κόμβος βρίσκεται εντός του δικτύου κάτι που είναι ιδιαίτερα επικίνδυνο καθώς έχουν παρακαμφθεί τα συστήματα περιμετρικής ασφάλειας ενώ είναι πλέον και πιο δύσκολο να εντοπιστεί. Το Apache Spot εφόσον έχει τροφοδοτηθεί και με την κίνηση του εσωτερικού δικτύου, μπορεί να εντοπίσει τέτοιες επιθέσεις. Οι επιθέσεις σάρωσης μπορούν να κατηγοριοποιηθούν ανάλογα με την τεχνική τους, ως σαρώσεις ευρείας περιοχής, συγκεκριμένου στόχου (κόμβου ή δικτύου), συγκεκριμένης ή

κατανεμημένης πηγής εκτέλεσης της σάρωσης. Οι σαρώσεις μπορούν επίσης να είναι κατακόρυφες δηλαδή σε μία ή περισσότερες θύρες ενός κόμβου, οριζόντιες, δηλαδή σε συγκεκριμένη θύρα πολλών κόμβων ή συνδυασμοί αυτών.

Αναλύοντας τη δραστηριότητα των κακόβουλων IP δ/νσεων και εξετάζοντας το σύνολο των ροών προς αυτό μέσω του Hue μπορούμε να αντλήσουμε χρήσιμα συμπεράσματα:

- σχεδόν το 90% ροών με κακόβουλες IP δ/νσεις αφορά σαρώσεις και ένα ποσοστό περίπου 10% είναι αποσπασματικές ροές χωρίς ιδιαίτερη βαρύτητα που δεν αξιολογούνται.
- σε περίπου 250 ροές από τις 472 εξεταζόμενες ροές προέκυψε ότι εκτελούν κακόβουλη δραστηριότητα και άλλες IP δ/νσεις (δύο ή περισσότερες) του ίδιου δικτύου.
- περίπου το 74% των σαρώσεων έχουν καθημερινές ροές και μόνο το 6.5% αφορά συγκεκριμένη ημέρα (βλέπε Εικόνα 16)
- περισσότερες από τις μισές ροές (237) που εμπλέκονται σε σαρώσεις θυρών των εξωτερικών δ/νσεων, βρέθηκε ότι έχουν ροές και προς ανοικτές θύρες εσωτερικών κόμβων.



Εικόνα 16: Ροές ανάλογα με την συχνότητα τους

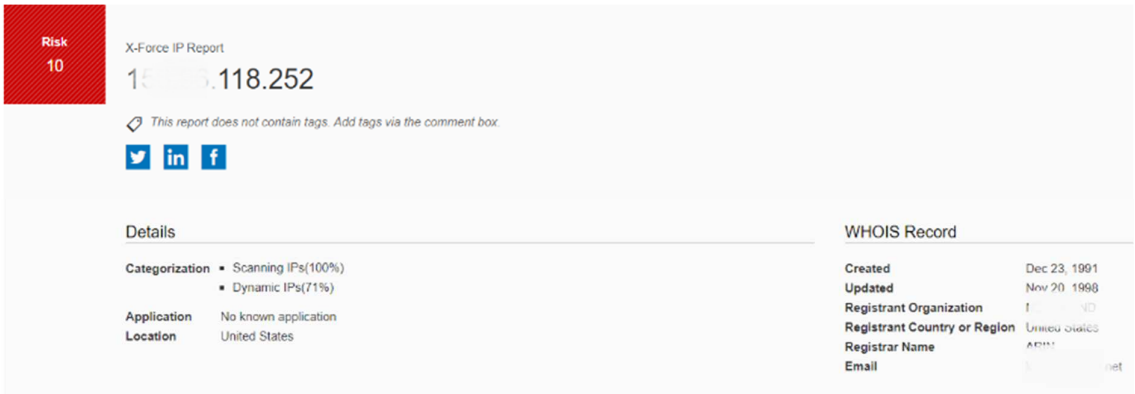
Περαιτέρω εξέταση των συμπερασμάτων που αναφέρθηκαν παραπάνω δείχνει ότι:

- οι ροές των κακόβουλων δ/νσεων με τις λίγες ανοικτές θύρες των εσωτερικών δ/νσεων που διαπιστώθηκαν με το Hue, αναγνωρίζονται σε πολύ μικρό βαθμό.

- Σαρώσεις με καθημερινή δραστηριότητα δεν αναγνωρίζονται πάντα στα αποτελέσματα διαφορετικών ημερών. Έτσι όπως προκύπτει από την ιχνηλάτηση, ενώ περίπου το 74% των σαρώσεων έχουν καθημερινές ροές, μόνο το 25% των ροών αυτών αναγνωρίζεται σε περισσότερες από μια ημέρες.

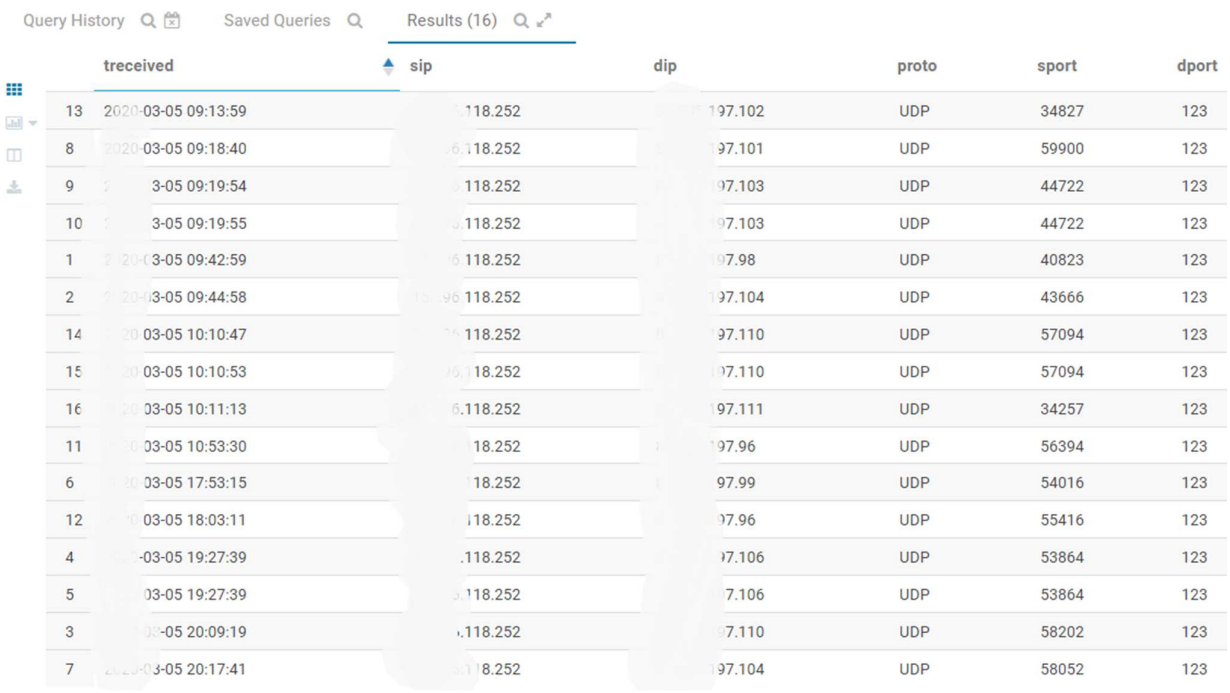
Ως παράδειγμα της διαδικασίας ιχνηλάτησης απειλής έχουμε την παρακάτω περίπτωση: στα αποτελέσματα της ανάλυσής μας, μεταξύ άλλων, περιλαμβάνεται η IP δ/νση x.x.118.252, η οποία μέσω του API του X-Force καταγράφεται ως κακόβουλη δ/νση, που πραγματοποιεί επιθέσεις αναγνώρισης.

Στον ιστότοπο X-Force εμφανίζονται οι σχετικές λεπτομέρειες.



Εικόνα 17: Απόσπασμα πληροφοριών από το X-force σχετικά με κακόβουλη IP δ/ση

Για να διαπιστώσουμε τι ακριβώς έχει συμβεί στο δίκτυό μας, αναζητούμε με την χρήση του Hue τις ροές από/προς την IP δ/νση, τη συγκεκριμένη ημερομηνία και έχουμε τα αποτελέσματα που εμφανίζονται στην Εικόνα 18, όπου βλέπουμε ότι οι εξωτερικές δ/νσεις του δικτύου σαρώνονται μια φορά έκαστη σε χρονικό διάστημα περίπου 12 ωρών. Η κακόβουλη δ/νση προσπαθεί να εντοπίσει κάποιον διακομιστή NTP. Οι διακομιστές NTP χρησιμοποιούνται συχνά για επιθέσεις ανάκλασης (CVE-2013-5211). Η σάρωση αυτή γίνεται με πολύ αργό ρυθμό και είναι ιδιαίτερα δύσκολο να εντοπιστεί.



Query History | Saved Queries | Results (16)

	treceived	sip	dip	proto	sport	dport
13	2020-03-05 09:13:59	118.252	97.102	UDP	34827	123
8	2020-03-05 09:18:40	118.252	97.101	UDP	59900	123
9	3-05 09:19:54	118.252	97.103	UDP	44722	123
10	3-05 09:19:55	118.252	97.103	UDP	44722	123
1	3-05 09:42:59	118.252	97.98	UDP	40823	123
2	3-05 09:44:58	118.252	97.104	UDP	43666	123
14	03-05 10:10:47	118.252	97.110	UDP	57094	123
15	03-05 10:10:53	118.252	97.110	UDP	57094	123
16	03-05 10:11:13	118.252	97.111	UDP	34257	123
11	03-05 10:53:30	18.252	97.96	UDP	56394	123
6	03-05 17:53:15	18.252	97.99	UDP	54016	123
12	03-05 18:03:11	118.252	97.96	UDP	55416	123
4	-03-05 19:27:39	.118.252	97.106	UDP	53864	123
5	03-05 19:27:39	118.252	7.106	UDP	53864	123
3	-05 20:09:19	.118.252	97.110	UDP	58202	123
7	3-05 20:17:41	8.252	97.104	UDP	58052	123

Εικόνα 18: Αποτελέσματα αναζήτησης ύποπτων ροών στο Hue

Στις ύποπτες ροές με εξωτερικές IP δ/νσεις που δεν είναι χαρακτηρισμένες κακόβουλες περιλαμβάνονται συχνά και εξερχόμενες συνδέσεις σε θύρα 5228 (Android Playstore), μηνύματα ICMP, συνδέσεις σε θύρα 6568 (AnyDesk) και GRE.

Στα αποτελέσματα του SPOT, στο σύνολο της εβδομάδας, καταγράφονται επίσης 38 IP δ/νσεις ενδοδικτύου. Στις ροές τους υπάρχει επικοινωνία με τις παρακάτω γνωστές θύρες:

80 (Http), 88 (Kerberos), 135 (Remote Procedure Call), 139 (NetBIOS), 389 (LDAP), 443 (Https), 445 (Microsoft-DS Active Directory, SMB)

και στις:

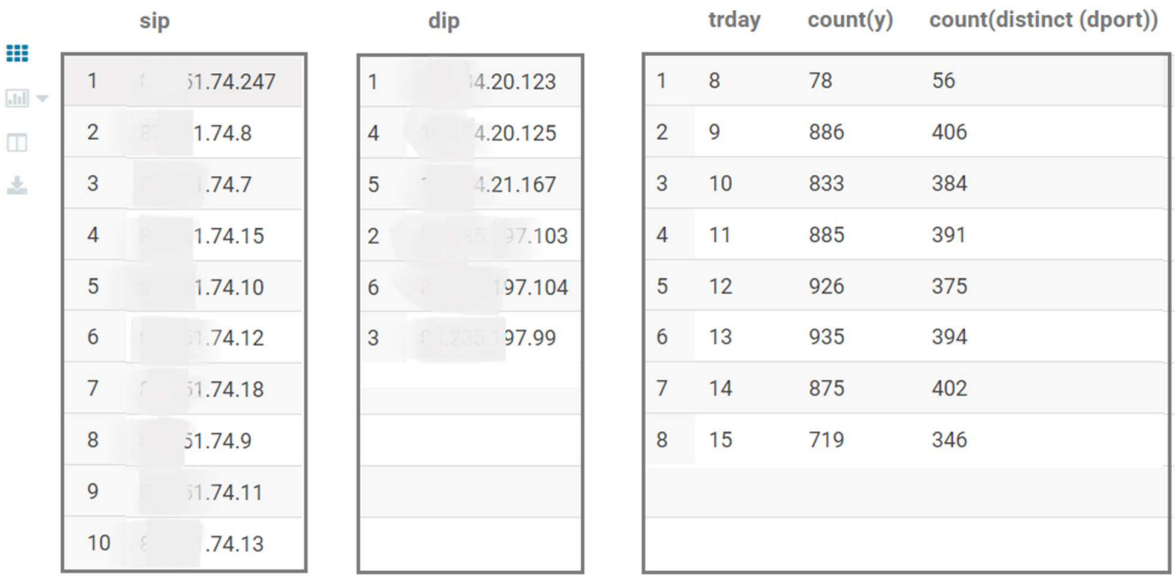
1521 (Oracle Database), 7680 (Windows Update Delivery Optimization)

Θεωρώντας ότι οι εξεταζόμενοι κόμβοι έχουν ενημερωμένα λειτουργικά συστήματα (αυτό αποτελεί στοιχειώδη προϋπόθεση ασφαλείας) και το δίκτυο χρησιμοποιεί Microsoft Active Directory και διακομιστές βάσης δεδομένων Oracle η δραστηριότητα αυτή δεν μπορεί να θεωρηθεί ύποπτη χωρίς επιπλέον ενδείξεις.

Αντίστοιχα είναι και τα αποτελέσματα των ροών με άλλες εσωτερικές IP δ/νσεις ενώ σαν ύποπτες καταγράφονται και κάποιες multicast συνδέσεις (IGMP, MultiCast DNS).

5.4 Ευρήματα

Για την διάρκεια των 2 μηνών, των οποίων τα δεδομένα αναλύσαμε, προέκυψαν ορισμένα αποτελέσματα που παρουσιάζουν ενδιαφέρον και αφορούν εσωτερικές δ/νσεις του δικτύου. Χαρακτηριστική εικόνα μιας σάρωσης που προέκυψε από ιχνηλάτηση ύποπτων ροών παρουσιάζεται στην Εικόνα 19. Αριστερά βλέπουμε τις διάφορες IP δ/νσεις του δικτύου που σαρώνουν τις εσωτερικές και εξωτερικές δ/νσεις (φαίνονται στο κέντρο) του υπό εξέταση δικτύου. Ακολουθεί ο αριθμός των σχετικών ροών και ο αριθμός των διαφορετικών θυρών προορισμού που έχουν καταγραφεί ανά ημέρα, που, όπως βλέπουμε, είναι αρκετά μεγάλος.



	sip
1	51.74.247
2	1.74.8
3	.74.7
4	1.74.15
5	1.74.10
6	1.74.12
7	1.74.18
8	51.74.9
9	1.74.11
10	.74.13

	dip
1	4.20.123
4	4.20.125
5	4.21.167
2	97.103
6	97.104
3	97.99

	trday	count(y)	count(distinct (dport))
1	8	78	56
2	9	886	406
3	10	833	384
4	11	885	391
5	12	926	375
6	13	935	394
7	14	875	402
8	15	719	346

Εικόνα 19: Στοιχεία χαρακτηριστικής σάρωσης

Η ιχνηλάτηση των σαρώσεων μπορεί να οδηγήσει σε εντοπισμό πιο σύνθετων επιθέσεων, όπως φαίνεται στην Εικόνα 20. Ο κακόβουλος κόμβος σαρώνει εξωτερικές δ/νσεις με αργό ρυθμό στην θύρα 1723, που χρησιμοποιείται από το πρωτόκολλο PPTP για τη δημιουργία συνδέσεων εικονικού ιδιωτικού δικτύου. Στη συνέχεια εντοπίζει μια IP δ/νση, η οποία απαντάει στη θύρα αυτή. Οι ροές προς αυτή την θύρα προωθούνται σε εσωτερικό κόμβο του δικτύου. Ακολουθεί μια σειρά ροών προς τον κόμβο αυτόν που έχει ενεργοποιημένη την υπηρεσία VPN, προχωρώντας στο επόμενο βήμα της επίθεσης, το οποίο συνήθως είναι είτε προσπάθειες σύνδεσης είτε ανίχνευση ευπαθειών.

	treceived	sip	dip	proto	sport	dport	ipkt	ibyt
1	0-04-10 23:06:13	24.171	197.104	TCP	53578	1723	1	46
2	-04-11 08:05:13	24.171	.197.103	TCP	43195	1723	1	46
3	-04-12 09:28:00	9.100	2.24.171	TCP	1723	38627	1	44
4	-04-12 09:28:00	24.171	4.19.100	TCP	38627	1723	2	92
5	-04-12 09:28:05	19.100	.24.171	TCP	1723	43354	4	216
6	-04-12 09:28:05	24.171	.19.100	TCP	43354	1723	5	563
7	-04-12 09:28:09	19.100	2.24.171	TCP	1723	48006	4	216
8	-04-12 09:28:09	4.171	4.19.100	TCP	48006	1723	5	563
9	-04-12 09:28:11	9.100	.24.171	TCP	1723	46503	4	216
10	-04-12 09:28:11	4.171	.19.100	TCP	46503	1723	4	419
11	04-12 09:28:17	4.171	19.100	TCP	40602	1723	5	471
12	4-12 09:28:17	9.100	24.171	TCP	1723	40602	4	216
13	4-12 09:28:20	9.100	.24.171	TCP	1723	52688	1	60
14	4-12 09:28:20	4.171	19.100	TCP	52688	1723	3	164
15	4-12 09:28:20	9.100	4.171	TCP	1723	42238	2	268
16	4-12 09:28:20	.4.171	9.100	TCP	42238	1723	5	424

Εικόνα 20: Ροές από επίθεση στην θύρα 1723

Όπως φαίνεται στον Πίνακας 12 η υπηρεσία X-Force χαρακτηρίζει κακόβουλες τις IP δ/νσεις που παρέχουν ανώνυμη πρόσβαση στο διαδίκτυο. Στα αποτελέσματά μας διαπιστώνουμε ότι έχει καταγραφεί επικοινωνία με κόμβο που παρέχει ανώνυμη πρόσβαση στο δίκτυο. Ο κόμβος έχει επίπεδο επικινδυνότητας 8.6 και οι ροές είναι προς την θύρα 9001, η οποία χρησιμοποιείται για σύνδεση στο σύστημα TOR. Παρόλο που τέτοιες συνδέσεις δεν αποδεικνύουν ότι υπάρχει επικοινωνία με κακόβουλους κόμβους, είναι γνωστό ότι το σύστημα TOR χρησιμοποιείται από διακομιστές C&C και συχνά οι πολιτικές ασφαλείας κατατάσσουν τις συνδέσεις αυτές στις κακόβουλες δραστηριότητες. Εύκολα μπορούμε να διαπιστώσουμε ότι υπάρχει μόνο ένας κόμβος στο δίκτυό μας με τέτοιου είδους δραστηριότητα. Ελέγχοντας τις σχετικές ροές (βλέπε Εικόνα 21) παρατηρούμε ότι είναι καθημερινές και εντοπίζονται κατά τη διάρκεια όλου του εικοσιτετραώρου. Το προφίλ του χρήστη και το ωράριο χρήσης του κόμβου, ενδεχομένως, να εξηγεί τη δραστηριότητα αυτή, διαφορετικά χρήζει περεταίρω ελέγχου.

	treceived	sip	dip	proto	sport	dport
4	-04-08 03:21:59	19.126	.66.236	TCP	52889	9001
5	-04-08 03:54:33	19.126	6.236	TCP	57058	9001
6	-04-08 04:36:48	19.126	66.236	TCP	57808	9001
7	-04-08 04:36:48	.66.236	19.126	TCP	9001	57808
8	-04-08 05:22:18	.66.236	19.126	TCP	9001	57058
9	04-08 05:43:36	19.126	66.236	TCP	57974	9001
10	04-08 05:43:36	.66.236	9.126	TCP	9001	57974
11	4-08 06:20:19	19.126	6.236	TCP	59366	9001
12	4-08 06:20:19	66.236	9.126	TCP	9001	59366
13	4-08 06:33:26	66.236	9.126	TCP	9001	59366
14	4-08 06:44:09	.19.126	6.236	TCP	59659	9001
15	4-08 07:53:06	19.126	.66.236	TCP	60862	9001
16	4-08 08:32:13	19.126	66.236	TCP	61678	9001
17	4-08 08:32:13	.66.236	.126	TCP	9001	61678
18	4-08 08:33:19	.66.236	.126	TCP	9001	59659
19	04-08 09:40:17	9.66.236	9.126	TCP	9001	60862
20	04-08 14:16:10	4.19.126	.66.236	TCP	65237	9001

Εικόνα 21: Συνδέσεις με διακομιστή ανώνυμης περιήγησης

Εξετάζοντας τις ύποπτες συνδέσεις του δικτύου εντοπίζουμε μια επίθεση στον διακομιστή αλληλογραφίας. Όπως φαίνεται στην Εικόνα 22, σε χρονικό διάστημα λιγότερο των 2 λεπτών μια IP δ/νση που είναι χαρακτηρισμένη ως Spam εμπλέκεται σε 400 περίπου ροές με τη θύρα 25 του διακομιστή.

Query History | Saved Queries | Results (409)

Number of rows

	treceived	sip		proto	sport	dport	ipkt	ibyt
1	2020-04-08 00:58:54	20.123	116.62	TCP	25	53343	9	627
2	2020-04-08 00:58:54	116.62	20.123	TCP	53343	25	8	452
3	2020-04-08 00:58:54	20.123	116.62	TCP	25	53377	6	433
4	2020-04-08 00:58:54	116.62	20.123	TCP	53377	25	6	293
5	2020-04-08 00:58:54	20.123	116.62	TCP	25	53397	6	433
6	2020-04-08 00:58:54	116.62	20.123	TCP	53397	25	5	248
7	2020-04-08 00:59:00	20.123	116.62	TCP	25	53507	6	433
398	2020-04-08 01:00:45	20.123	116.62	TCP	25	58009	6	433
399	2020-04-08 01:00:45	116.62	20.123	TCP	58009	25	6	295
400	2020-04-08 01:00:45	20.123	116.62	TCP	25	57926	6	433
401	2020-04-08 01:00:45	116.62	20.123	TCP	57926	25	6	293
402	2020-04-08 01:00:45	20.123	116.62	TCP	25	57939	6	433
403	2020-04-08 01:00:45	116.62	20.123	TCP	57939	25	6	292
404	2020-04-08 01:00:45	20.123	116.62	TCP	25	57966	6	433

Εικόνα 22: Πολλαπλές ροές στην θύρα 25 του διακομιστή αλληλογραφίας

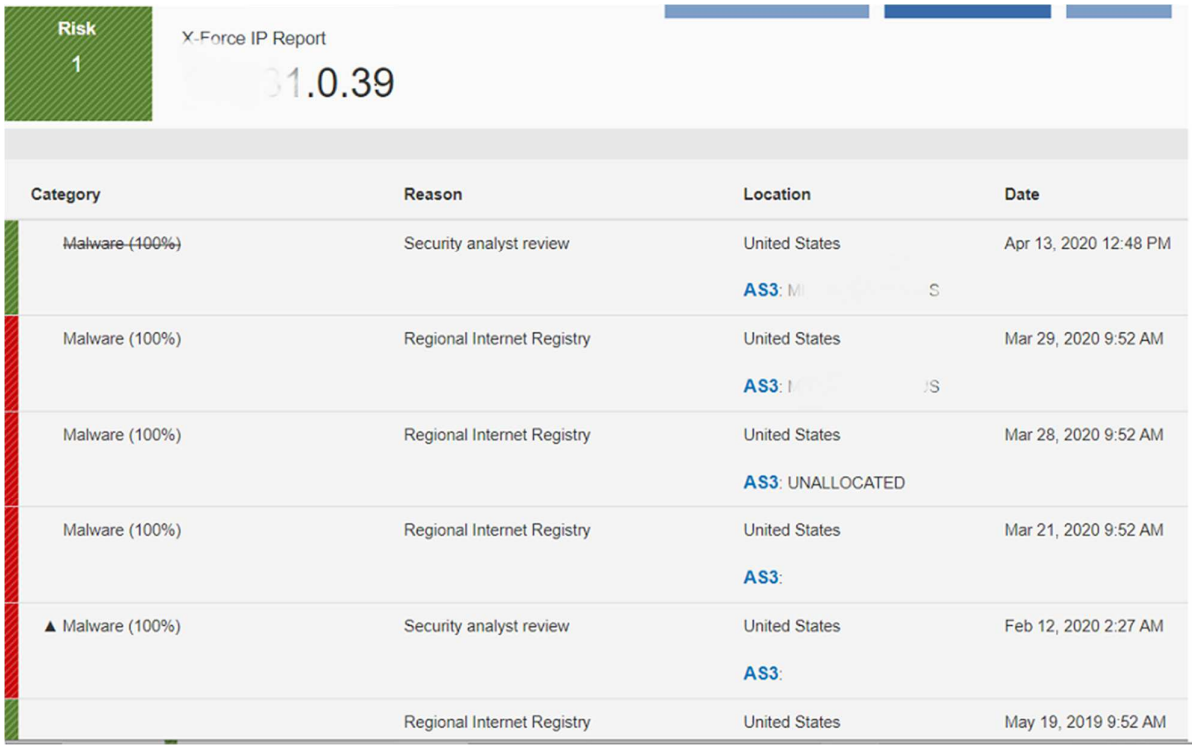
Ένα ιδιαίτερα ενδιαφέρον εύρημα προκύπτει από καταγραφή ροών του δικτύου με επικοινωνία σε μη ευρέως χρησιμοποιούμενες θύρες, όπως φαίνεται στην παρακάτω εικόνα.

Query History | Saved Queries | Results (4)

	treceived	sip	dip	proto	sport	dport	ipkt	ibyt
1	2020-03-25 23:52:51	19.88	0.39	TCP	52863	9131	196	7998
2	2020-03-25 23:52:51	0.39	19.88	TCP	9131	52863	453	636488
3	2020-03-26 00:18:53	19.88	0.39	TCP	53638	9101	457	67665
4	2020-03-26 00:18:53	0.39	19.88	TCP	9101	53638	939	1312570

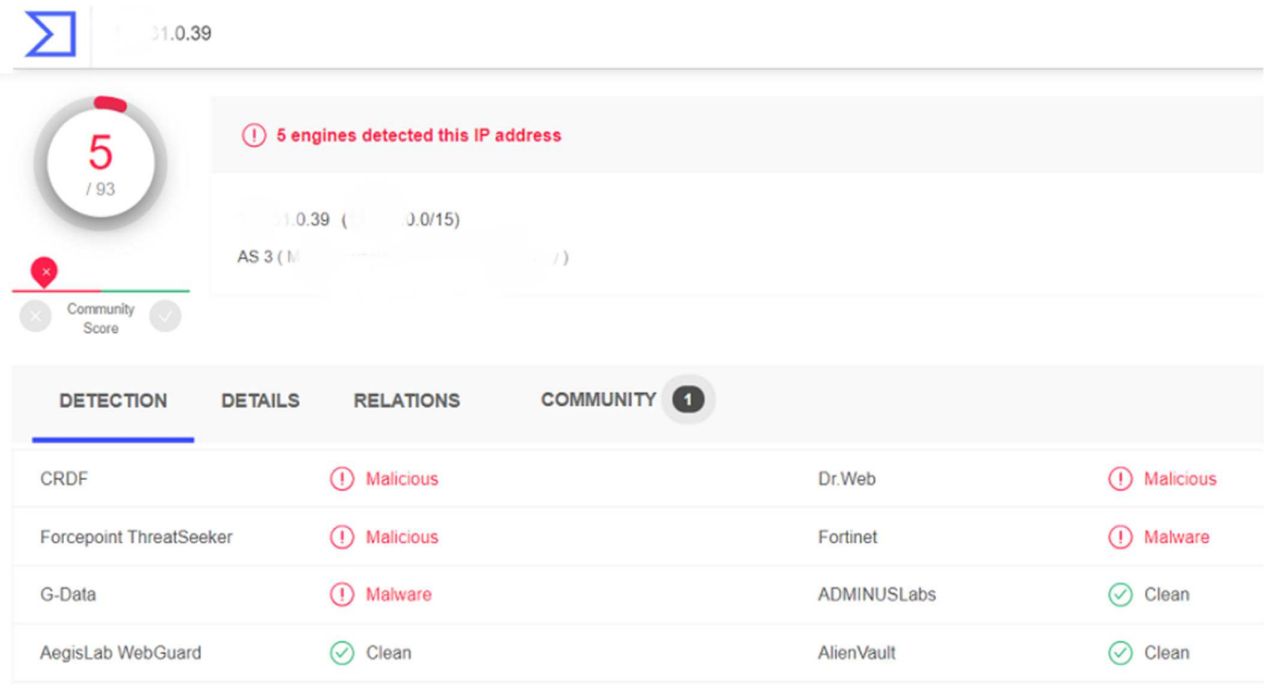
Εικόνα 23: Υποπτη εξερχόμενη σύνδεση

Στις δυο αυτές συνδέσεις που λαμβάνουν χώρα περίπου τα μεσάνυκτα σε ώρα UTC, δηλαδή περίπου στις 2 μ.μ. τοπική ώρα, μεταφέρονται περίπου 2 MB. Όπως φαίνεται από την Εικόνα 24 κατά το χρονικό διάστημα εντός του οποίου υπήρξε η εξερχόμενη σύνδεση με τον κόμβο x.x.0.39, ο κόμβος αυτός φιλοξενούσε κακόβουλο λογισμικό. Όπως είναι γνωστό, η εισαγωγή κακόβουλου λογισμικού σε ένα δίκτυο είναι ένα επικίνδυνο συμβάν που μπορεί να προκαλέσει σοβαρά προβλήματα στη λειτουργία ενός οργανισμού.



Εικόνα 24: Ιστορικό χαρακτηρισμού IP δ/νσης από το X-Force

Η κακόβουλη δραστηριότητα του κόμβου επιβεβαιώνεται και από άλλες πηγές του VirusTotal, όπως φαίνεται στην παρακάτω εικόνα.

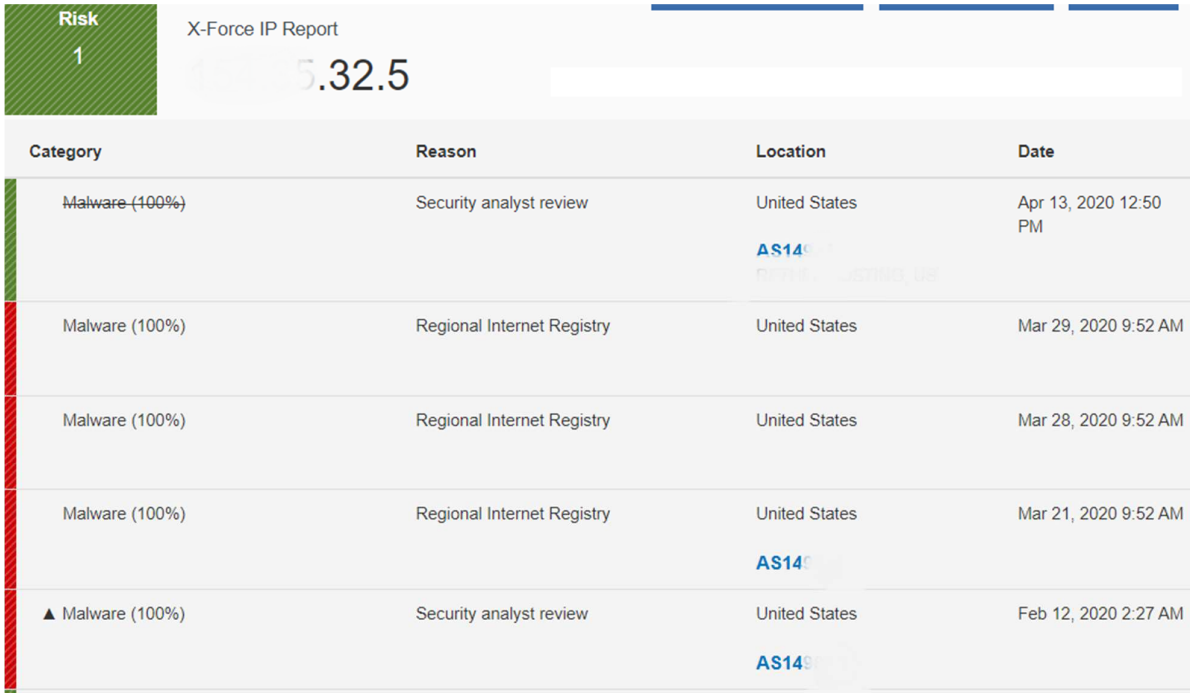


Εικόνα 25: Πληροφορίες για IP δ/νση από το VirusTotal

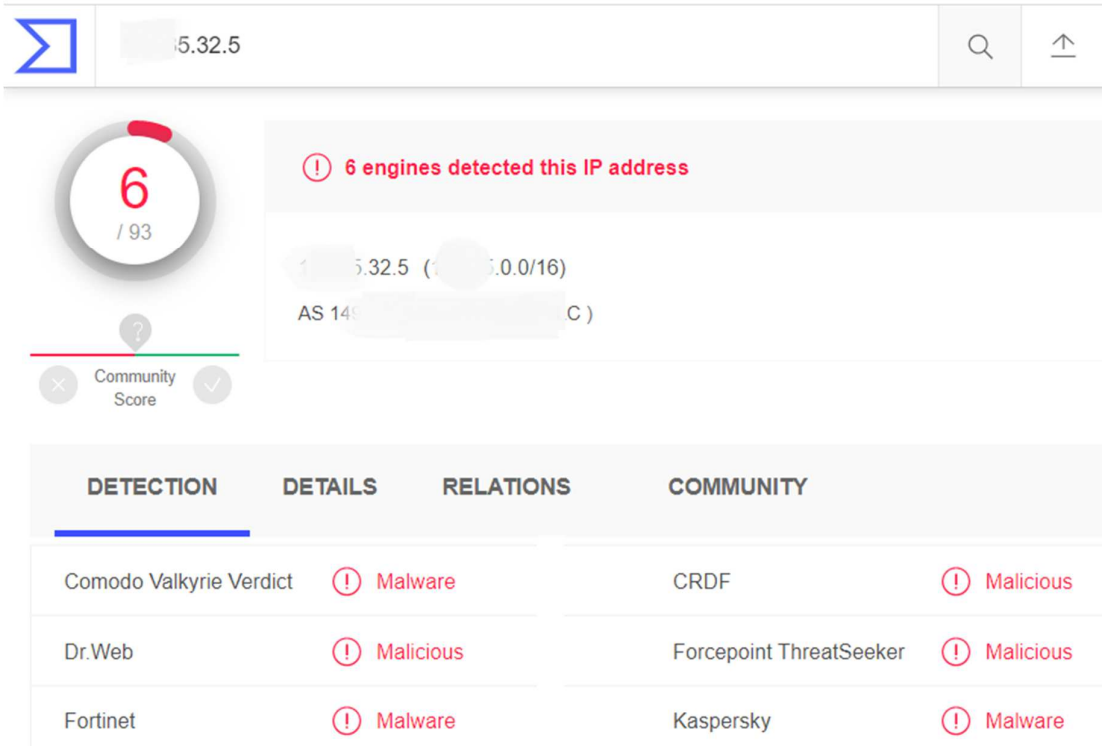
Διερευνώντας τις ροές του κόμβου, προς αναζήτηση άλλης επικίνδυνης δραστηριότητας, προκύπτει ότι, το ίδιο λεπτό με την ύποπτη ροή που κατέγραψε το Spot, υπάρχουν και άλλες ροές με κακόβουλες IP δ/νσεις. Όπως φαίνεται στην εικόνα που ακολουθεί, οι δ/νσεις που είναι σε κύκλο εμπλέκονται επίσης σε κακόβουλη δραστηριότητα.

	treceived	sip	dip	proto	sport	dport	ipkt	ibyt
6	03-25 23:52:48	9.206.212	04.19.88	TCP	80	52852	7	1744
7	03-25 23:52:48	4.19.88	5.32.5	TCP	52862	80	7	433
8	03-25 23:52:48	5.32.5	04.19.88	TCP	80	52862	10	1947
9	03-25 23:52:51	9.206.212	04.19.88	TCP	443	52865	5	232
10	03-25 23:52:51	4.19.88	09.206.212	TCP	52865	443	5	434
11	03-25 23:52:51	21.38	04.19.88	TCP	443	52864	5	233
12	03-25 23:52:51	4.19.88	21.38	TCP	52864	443	5	421
13	03-25 23:52:51	31.0.39	04.19.88	TCP	9131	52863	453	636488
14	03-25 23:52:51	4.19.88	.0.39	TCP	52863	9131	196	7998
15	03-25 23:53:03	4.19.88	5.193.9	TCP	52861	443	193	8336
16	03-25 23:53:03	5.193.9	.19.88	TCP	443	52861	381	567104
17	03-25 23:53:20	.16.163	19.88	TCP	443	52867	1	48
18	03-25 23:53:20	.19.88	7.16.163	TCP	52867	443	2	88

Ενδεικτικά η IP δ/νση x.x.32.5, η πρώτη από τις ύποπτες ροές, φαίνεται στις επόμενες εικόνες ότι έχει εντοπιστεί από πολλές πηγές ως κακόβουλη και είναι επίσης χαρακτηρισμένη ως Malware.



Εικόνα 26: Ιστορικό χαρακτηρισμού IP δ/νσης από το X-Force



Εικόνα 27: Πληροφορίες για IP δ/νση από το VirusTotal

Ομοίως, οι χαρακτηρισμοί των υπόλοιπών IP δ/νσεων είναι επίσης Malware και Anonymisation Services, συνδέοντας τις ροές αυτές σε ένα σημαντικό συμβάν ασφάλειας.

Στις ύποπτες ροές του Spot που αναλύσαμε, είναι συχνά καταγεγραμμένες ροές προς θύρες 137 και 139 εσωτερικών δ/σεων. Οι θύρες αυτές χρησιμοποιούνται από το NetBIOS για επικοινωνία σε τοπικά δίκτυα και η χρήση τους στο δημόσιο Διαδίκτυο δεν προτείνεται για λόγους ασφαλείας. Όπως φαίνεται στην Εικόνα 28, υπάρχουν επαναλαμβανόμενες ροές προς δημόσιες IP δ/νσεις σε μοτίβο που περιλαμβάνει και εξερχόμενες ροές προς τη θύρα 22 του κόμβου με τον οποίο γίνεται προσπάθεια σύνδεσης.

treceived	sip	dip	proto	sport	dport	ipkt	ibyt
20[illegible]-06-05 12:11:29	[illegible]5.149	[illegible]7.224.0	UDP	51301	137	2	156
[illegible]-06-05 12:11:29	[illegible]25.149	[illegible]7.224.0	TCP	51183	139	1	52
[illegible]-06-05 12:11:29	[illegible]224.0	[illegible].25.149	TCP	139	51183	1	46
[illegible]-06-05 12:11:32	[illegible]25.149	[illegible]7.224.0	TCP	51183	139	2	100
[illegible]-06-05 12:11:32	[illegible]224.0	[illegible].25.149	TCP	139	51183	2	92
[illegible]-06-05 12:28:34	[illegible]25.149	[illegible].224.0	TCP	51184	22	3	152
[illegible]-06-05 13:29:25	[illegible]25.149	[illegible]10.104	UDP	61923	137	3	234
[illegible]-06-05 13:29:29	[illegible]5.149	[illegible]0.104	UDP	61923	137	3	234
[illegible]-06-05 13:29:33	[illegible]5.149	[illegible]0.104	TCP	51258	139	3	152
[illegible]-06-05 13:29:33	[illegible]0.104	[illegible]25.149	TCP	139	51258	3	138
[illegible]-06-05 13:47:50	[illegible]25.149	[illegible]0.104	TCP	51259	22	3	152
[illegible]-06-05 17:17:25	[illegible]25.149	[illegible]239.255	UDP	60128	137	2	156
[illegible]-06-05 17:17:29	[illegible]25.149	[illegible]239.255	UDP	60128	137	4	312
[illegible]-06-05 17:17:31	[illegible]25.149	[illegible].239.255	TCP	51480	139	2	104
[illegible]-06-05 17:17:31	[illegible]239.255	[illegible].25.149	TCP	139	51480	2	92
[illegible]-06-05 17:17:32	[illegible]5.149	[illegible].239.255	TCP	51480	139	1	48
[illegible]-06-05 17:17:32	[illegible]239.255	[illegible].25.149	TCP	139	51480	1	46
[illegible]-06-05 17:37:54	[illegible]25.149	[illegible]239.255	TCP	51481	22	3	152

Εικόνα 28: Επαναλαμβανόμενες συνδέσεις σε θύρες 137, 139 και 22.

Οι ροές που ακολουθούν το μοτίβο αυτό γίνονται από τον ίδιο κόμβο προς διαφορετικές IP δ/νσεις που ανήκουν στον ίδιο οργανισμό, ο οποίος δεν σχετίζεται με τον χώρο της υγείας και πρόκειται για το υπουργείο Εθνικής Αμύνης του Ηνωμένου Βασιλείου. Αναζητώντας το ιστορικό του κόμβου βρίσκουμε ότι, ανά διαστήματα, υπάρχουν ροές προς διάφορες IP δ/σεις του εν λόγω οργανισμού.

Στις συνδέσεις προς τη θύρα 22 δεν υπάρχει απάντηση, γεγονός που δείχνει ότι δεν έχουν ενεργοποιημένη την υπηρεσία SSH.

Τέτοιες ροές δύσκολα μπορούν να χαρακτηριστούν κανονική συμπεριφορά ενός κόμβου και μπορούν να θεωρηθούν κακόβουλες ενέργειες από το δίκτυο που δέχεται τις ροές αυτές. Ένας τέτοιος χαρακτηρισμός της δημόσιας IP δ/νσης του δικτύου μας μπορεί να κοινοποιηθεί σε υπηρεσίες φήμης, όπως αυτές που χρησιμοποιούμε για την αξιολόγηση των ροών μας. Αυτές οι υπηρεσίες συχνά χρησιμοποιούνται από τοίχους προστασίας, ενημερώνοντάς τους, ώστε να αποτρέπουν την επικοινωνία με «κακόφημα» δίκτυα κάτι που, στην περίπτωσή μας, θα μπορούσε να προκαλέσει προβλήματα συνδεσιμότητας του δικτύου μας.

Συμπεράσματα

Το Apache Spot παρουσιάζει ιδιαίτερο ενδιαφέρον καθώς είναι ένα έργο που συνδυάζει τους τομείς των Μεγάλων Δεδομένων, της Μηχανικής Μάθησης και της Κυβερνοασφάλειας. Η εγκατάσταση, διαχείριση, παραμετροποίηση και λειτουργία του απαιτεί διαφορετικές εξειδικεύσεις. Το γεγονός αυτό καθώς και ότι είναι ένα έργο που δεν έχει ολοκληρωθεί πλήρως, καθιστά ακόμα και μια μικρή δοκιμαστική εγκατάσταση, όπως αυτή που αναπτύξαμε, αρκετά πολύπλοκη. Επίσης, δεν είναι ένα έργο ευρείας χρήσης με αποτέλεσμα να βελτιώνεται και να εξελίσσεται με αργό ρυθμό. Η λειτουργία του δεν προαπαιτεί τη ρύθμιση των κόμβων του δικτύου και μπορεί να χρησιμοποιηθεί εύκολα σε δίκτυα που περιλαμβάνουν κυβερνο-φυσικά συστήματα, συσκευές διαδικτύου των αντικειμένων και άλλο εξιδεικευμένο εξοπλισμό, όπως συσκευές ψηφιακής απεικόνισης.

Το Apache Spot δεν είναι ένα εργαλείο που προορίζεται για την ανίχνευση του κύριου όγκου των επιθέσεων ενός δικτύου. Δεν παράγει εντυπωσιακά αποτελέσματα, αλλά υποδεικνύει ύποπτες εγγραφές σε αρχεία καταγραφής. Όπως είδαμε από τα αποτελέσματα της ανάλυσής μας, πολυήμερες σαρώσεις δεν εντοπίστηκαν κάθε μέρα κάτι που, ίσως, δείχνει παράδοξο. Άλλες τεχνικές ανίχνευσης σαρώσεων θα είχαν μεγαλύτερη συνοχή στα αποτελέσματα. Για τον υπολογισμό της πιθανότητας, ώστε να θεωρηθεί ύποπτη κάποια καταγραφή, εμπλέκεται το σύνολο των καταγραφών της ημερήσιας κίνησης του δικτύου, αφού συμμετέχουν στη δημιουργία του προφίλ της φυσιολογικής κίνησης. Το προφίλ αυτό είναι φυσικό να έχει αποκλίσεις ακόμα και

μεταξύ συνεχόμενων ημερών ενώ διαφορικές ημέρες άλλες καταγραφές μπορεί να έχουν μεγαλύτερη απόκλιση από την φυσιολογική κίνηση. Το Apache Spot δεν προορίζεται για την επίλυση κοινών προβλημάτων - υπάρχουν πιο εξειδικευμένοι και αποτελεσματικοί τρόποι επίλυσής τους - αλλά για τον εντοπισμό «αόρατων» απειλών.

Κατά την ανάλυση του δικτύου εντοπίστηκε αμφίδρομη επικοινωνία από το εσωτερικό του δικτύου με κόμβο που φιλοξενούσε κακόβουλο λογισμικό. Η επικοινωνία αυτή ολοκληρώθηκε σε συνολικά μόλις τέσσερις ροές, όπου μεταφέρθηκαν ελάχιστα MB και είχε διαφύγει των υπόλοιπων μέτρων προστασίας του δικτύου. Αυτό είναι ένα άριστο παράδειγμα της χρησιμότητας του Apache Spot. Για να καταλήξουμε από το μεγάλο πλήθος των υποδεικνυόμενων εγγραφών στις αόρατες απειλές - που ενδεχομένως υπάρχουν – αν δεν έχουν ήδη εντοπιστεί από κάποια υπηρεσία φήμης ή ενεργούν στο εσωτερικό του δικτύου μας, είναι απαραίτητη η διερεύνηση των ροών. Το Spot αποτελεί εξαιρετικό εργαλείο για ιχνηλάτηση, αποτελώντας ένα κεντρικό σημείο συλλογής των αρχείων καταγραφής που παρέχει ευκολία στην εκτέλεση σύνθετων ερωτημάτων. Η ιχνηλάτηση που κάναμε για τις ύποπτες ροές αποκάλυψαν διάφορες επιθέσεις, κυρίως σαρώσεις, επίθεση σε θύρα 1723 (PTPP) και 25 (SMTP), χρήση υπηρεσιών ανώνυμης περιήγησης κ.α.

Το Apache Spot απαιτεί μεγάλο αριθμό πόρων, ειδικά αν συλλέγονται και δεδομένα από πηγές DNS και Proxy. Πέραν της πρόσληψης των δεδομένων, η επεξεργασία τους για την εξαγωγή αποτελεσμάτων απαιτεί αρκετό χρόνο. Σε αυτή την μικρή συστοιχία για επεξεργασία δεδομένων Netflow μιας ημέρας, μεγέθους λιγότερο από 1 GB χρειάζεται περίπου 90 λεπτά. Εκτός των σημαντικών πόρων για τη συλλογή και επεξεργασία των δεδομένων είναι απαραίτητη και η διάθεση ανθρώπινων πόρων με καλή γνώση του δικτύου για την ιχνηλάτηση των αποτελεσμάτων.

Αν η ανάλυση είναι συνεχής, σε βάθος χρόνου και περιλαμβάνει και τους άλλους τύπους δεδομένων που αναλύει, είναι αρκετές οι πιθανότητες να αποκαλυφθούν -εφόσον υπάρχουν- αόρατες ενέργειες που συνδέονται με σύνθετες και επικίνδυνες απειλές, όπως οι προχωρημένες μόνιμες απειλές. Το Apache Spot δείχνει ότι μπορεί να μειώσει τον μέσο χρόνο ανίχνευσης και τον μέσο χρόνο αντίδρασης σε ένα συμβάν κυβερνοασφάλειας. Οι σύγχρονες υποδομές ζωτικής σημασίας έχουν σαν προτεραιότητά τους την ελαχιστοποίηση των επιπτώσεων των κυβερνοεπίθεσεων και το Apache Spot είναι ένα εργαλείο που μπορεί να προστεθεί στην πολύ-επίπεδη προστασία τους έχοντας έναν ξεχωριστό ρόλο.

Βιβλιογραφία

[1] K. Bissell and L. Ponemon, “Ninth Annual Cost of Cybercrime Study Unlocking the Value of Improved Cybersecurity Protection the Cost of Cybercrime Contents,” p. 18, 2019.

[2] “Marriott says breach of Starwood guest database compromised info of up to 500 million.” [Online]. Available: https://www.nbcnews.com/tech/security/marriott-says-data-breach-compromised-info-500-million-guests-n942041. [Accessed: 04-Oct-2019].

[3] “All 3 Billion Yahoo Accounts Were Affected by 2013 Attack - The New York Times.” [Online]. Available: https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html. [Accessed: 04-Oct-2019].

[4] WEF, *The Global Risks Report 2019 - Insight Report*. 2019.

[5] International Information System Security Certification Consortium, “Cybersecurity Professionals Focus on Developing New Skills as Workforce Gap Widens Table of Contents,” 2018.

[6] IBM Security and Ponemon, “Cost of a Data Breach Report,” 2019.

[7] N. I. of S. and Technology, “Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1,” Gaithersburg, MD, Apr. 2018.

[8] M. Antonakakis *et al.*, “Understanding the Mirai Botnet,” *Proc. 26th USENIX Secur. Symp.*, pp. 1093–1110, 2017.

[9] K. W. Miller, J. Voas, and G. F. Hurlburt, “BYOD: Security and privacy considerations,” *IT Prof.*, vol. 14, no. 5, pp. 53–55, 2012.

[10] I. F. Mikhalevich and V. A. Trapeznikov, “Critical Infrastructure Security: Alignment of Views,” *2019 Syst. Signals Gener. Process. F. Board Commun. SOSG 2019*, pp. 1–5, 2019.

[11] “CPSSEC | Homeland Security.” [Online]. Available: https://www.dhs.gov/science-and-technology/cpssec. [Accessed: 19-Nov-2019].

[12] Department Of Health, “Investigation: WannaCry cyber attack and the NHS A picture of the National Audit Office logo,” no. April 2018, pp. 1–33, 2018.

[13] Cert-eu, “WannaCry Ransomware Campaign Exploiting SMB Vulnerability,” pp. 1–5, 2017.

[14] Robert M. Lee, Michael J. Assante, and Tim Conway, “Analysis of the Cyber Attack on the Ukrainian Power Grid Defense Use Case,” *Ics.Sans.Org*, pp. 2–11, 2016.

[15] N. Nelson, “The Impact of Dragonfly Malware on Industrial Control Systems,” *SANS Inst. InfoSec Read. Room*, pp. 1–25, 2016.

[16] T. Mahmood and U. Afzal, “Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools,” *Conf. Proc. - 2013 2nd Natl. Conf. Inf. Assur. NCIA 2013*, pp. 129–134, 2013.

[17] Y. Xin *et al.*, “Machine Learning and Deep Learning Methods for Cybersecurity,” *IEEE Access*, vol. 6, pp. 35365–35381, 2018.

[18] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, “On the effectiveness of machine and deep learning for cyber security,” *Int. Conf. Cyber Conflict, CYCON*, vol. 2018-May, pp. 371–389, 2018.

[19] A. A. Cardenas, P. K. Manadhata, and S. P. Rajan, “Big data analytics for security,” *IEEE Secur. Priv.*, vol. 11, no. 6, pp. 74–76, 2013.

[20] L. F. Dias and M. Correia, “Big Data Analytics for Intrusion Detection,” no. Ml, pp. 292–316, 2019.

[21] C. Zhang, X. Shen, X. Pei, and Y. Yao, “Applying Big Data Analytics into Network Security: Challenges, Techniques and Outlooks,” *Proc. - 2016 IEEE Int. Conf. Smart Cloud, SmartCloud 2016*, pp. 325–329, 2016.

[22] M. Belesioti *et al.*, “A new security approach in telecom infrastructures: The RESISTO Concept,” *Proc. - 15th Annu. Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2019*, pp. 212–218, 2019.

[23] Q. Zhu, Y. Zhao, L. Fei, and C. Zhou, “A Dynamic Decision-Making Approach for Cyber-Risk Reduction in Critical Infrastructure,” *8th Annu. IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst. CYBER 2018*, pp. 595–600, 2019.

[24] A. P. Fournaris, K. Lampropoulos, and O. Koufopavlou, “Hardware Security for Critical Infrastructures - The CIPSEC Project Approach,” *Proc. IEEE Comput. Soc. Annu. Symp. VLSI, ISVLSI*, vol. 2017-July, pp. 356–361, 2017.

[25] B. Satchidanandan and P. R. Kumar, “Dynamic watermarking: Active defense of networked cyber-physical systems,” *Proc. IEEE*, vol. 105, no. 2, pp. 219–240, 2017.

[26] D. Myers, S. Suriadi, K. Radke, and E. Foo, “Anomaly detection for industrial control systems using process mining,” *Comput. Secur.*, vol. 78, pp. 103–125, 2018.

[27] T. Alves, R. Das, and T. Morris, “Embedding Encryption and Machine Learning Intrusion Prevention Systems on Programmable Logic Controllers,” *IEEE Embed. Syst. Lett.*, vol. 10, no. 3, pp. 99–102, 2018.

[28] K. K. Venkatasubramanian, A. Banerjee, S. K. S. Gupta, and R. J. Walls, “A cyber-physical approach to trustworthy operation of health monitoring systems,” *2017 IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov. SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 -* , pp. 1–6, 2018.

[29] C. Service and P. Via, “MeDShare : Trust-less Medical Data Sharing Among,” *IEEE Access*, vol. 5, pp. 1–10, 2017.

[30] E. Markakis, Y. Nikoloudakis, E. Pallis, and M. Manso, “Security Assessment as a Service Cross-Layered System for the Adoption of Digital, Personalised and Trusted Healthcare,” *IEEE 5th World Forum Internet Things, WF-IoT 2019 - Conf. Proc.*, pp. 91–94, 2019.

[31] A. Rao, N. Carreon Rascon, R. Lysecky, and J. W. Rozenblit, “Probabilistic Security Threat

Detection for Risk Management in Cyber-Physical Medical Systems," *IEEE Softw.*, 2018.

[32] Y. Won *et al.*, "An attack-resilient CPS architecture for hierarchical control: A case study on train control systems," *Computer (Long. Beach. Calif).*, vol. 51, no. 11, pp. 46–55, 2018.

[33] M. Niraula, J. Graefe, R. Dlouhy, M. Layton, and M. Stevenson, "ATN/IPS security approach: Two-way mutual authentication, data integrity and privacy," *ICNS 2018 - Integr. Commun. Navig. Surveill. Conf.*, pp. 1A31-1A317, 2018.

[34] Y. Kim, J. Y. Jo, and S. Lee, "ADS-B vulnerabilities and a security solution with a timestamp," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 32, no. 11, pp. 52–61, 2017.

[35] N. Nicolaou, D. G. Eliades, C. Panayiotou, and M. M. Polycarpou, "Reducing vulnerability to cyber-physical attacks in water distribution networks," *Proc. - 2018 4th Int. Work. Cyber-Physical Syst. Smart Water Networks, CySWater 2018*, pp. 16–19, 2018.

[36] A. Mathur, "SecWater: A multi-layer security framework for water treatment plants," *Proc. - 2017 3rd Int. Work. Cyber-Physical Syst. Smart Water Networks, CySWATER 2017*, pp. 29–32, 2017.

[37] C. M. Mathas, "Evaluation of Apache Spot ' s machine learning capabilities in an SDN / NFV enabled environment Workshop paper," *Proc. 13th Int. Conf. Availability, Reliab. Secur. - ARES 2018*, pp. 1–10, 2018.

[38] A. Priovolos, G. Gardikis, D. Lioprasitis, and Costicoglou S., "Improving Apache Spot Using Autoencoders for Network Anomaly Detection," in *EuCNC 2020*, 2020.

[39] *Greenbone Security Manager with Greenbone OS 4*. Osnabrück Germany, 2018.

[40] "NVD - Home." [Online]. Available: https://nvd.nist.gov/. [Accessed: 05-Jan-2020].

[41] R. W. Kolb, "Apache Hadoop," in *The SAGE Encyclopedia of Business Ethics and Society*, 2018.

[42] Cloudera, "Apache Hadoop core components | Cloudera," 2018. [Online]. Available: https://www.cloudera.com/products/open-source/apache-hadoop/hdfs-mapreduce-yarn.html. [Accessed: 14-Dec-2019].

[43] "What is Apache MapReduce? | IBM." [Online]. Available: https://www.ibm.com/analytics/hadoop/mapreduce. [Accessed: 11-Dec-2019].

[44] "Untangling Apache Hadoop YARN, Part 1: Cluster and YARN Basics - Cloudera Blog." [Online]. Available: https://blog.cloudera.com/untangling-apache-hadoop-yarn-part-1-cluster-and-yarn-basics/. [Accessed: 11-Jan-2020].

[45] N. Narkhede, *Kafka the Definitive Guide*, 1st ed. O'Reilly Media, 2017.

[46] Cloudera, *Apache Spark Guide*. Palto Alto, CA: Cloudera, Inc., 2020.

[47] Cloudera, *Apache Hive Guide*. Palto Alto, CA: Cloudera, Inc., 2020.

[48] Cloudera, *Apache Impala Guide*. Palto Alto, CA: Cloudera, Inc., 2020.

[49] A. Spot, "Apache Spot," 2017. [Online]. Available: https://github.com/apache/incubator-spot.

[Accessed: 18-Dec-2019].

[50] P. Haag, *User Documentation nfdump & NfSen*. 2006.

[51] “tshark - The Wireshark Network Analyzer 3.2.1.” [Online]. Available: https://www.wireshark.org/docs/man-pages/tshark.html. [Accessed: 27-Jan-2020].

[52] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.

[53] E. Bou-Harb, M. Debbabi, and C. Assi, “Cyber scanning: A comprehensive survey,” *IEEE Commun. Surv. Tutorials*, vol. 16, no. 3, pp. 1496–1519, 2014.