ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Ενσωμάτωση Τεχνολογίας ETL για Αποτελεσματική και Συνεργατική Επεξεργασία Δεδομένων Κυβερνοασφάλειας: Μία Μελέτη Περίπτωσης** **Integration of ETL Technology for Effective and Collaborative Processing of Cybersecurity Data: A Case Study**
Ονοματεπώνυμο Φοιτητή	**Γεώργιος Λεβαντής**
Πατρώνυμο	**Θεόδωρος Λεβαντής**
Αριθμός Μητρώου	**ΜΠΚΕΔ2216**
Επιβλέπων	**Κοτζανικολάου Παναγιώτης, Καθηγητής**

Ημερομηνία Παράδοσης **Σεπτέμβριος 2024**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Δημήτριος Αποστόλου Καθηγητής	Δρ. Γρηγόριος Κορωνάκος Διδάσκων ΠΜΣ	Παναγιώτης Κοτζανικολάου Καθηγητής

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη χρήση σύγχρονων τεχνολογιών επεξεργασίας δεδομένων, όπως είναι η τεχνολογία ETL (Extract, Transform, Load) και συστημάτων ελέγχου έκδοσης (version control), για τη διαχείριση και επεξεργασία δεδομένων κυβερνοασφάλειας. Αναλύονται οι τρεις βασικές φάσεις της διαδικασίας ETL — εξαγωγή, μετασχηματισμός και φόρτωση δεδομένων — καθώς και η ενσωμάτωση εργαλείων ελέγχου έκδοσης, όπως το Git και το GitHub, για την υποστήριξη της συνεργασίας και την εξασφάλιση της ακεραιότητας των δεδομένων. Επιπλέον, διερευνάται η χρήση των προτύπων STIX/TAXII για την ασφαλή ανταλλαγή πληροφοριών απειλών και η εφαρμογή του Power BI για την οπτικοποίηση αυτών των δεδομένων. Η εργασία καταδεικνύει τη σημασία της αυτοματοποίησης και της βελτιστοποίησης των ροών εργασίας για την αποτελεσματικότερη διαχείριση δεδομένων και τη λήψη στρατηγικών αποφάσεων.

Abstract

This thesis focuses on the use of modern data processing technologies, such as ETL (Extract, Transform, Load) and version control systems for managing and processing cybersecurity data. It provides an analysis of the three core stages of the ETL process—data extraction, transformation, and loading—along with the integration of version control tools like Git and GitHub to support collaboration and ensure data integrity. Additionally, the study explores the application of STIX/TAXII standards for secure threat information sharing and the use of Power BI for data visualization. The thesis highlights the importance of automation and workflow optimization to enhance the efficiency of data management and support strategic decision-making.

1. Εισαγωγή

Στο σημερινό ψηφιακό περιβάλλον, η κυβερνοασφάλεια αποτελεί έναν από τους πιο κρίσιμους τομείς για κάθε οργανισμό, ανεξαρτήτως μεγέθους ή κλάδου. Οι απειλές στον κυβερνοχώρο εξελίσσονται συνεχώς, με αποτέλεσμα η ανίχνευση, η ανάλυση και η απόκριση σε αυτές να είναι πιο δύσκολες και απαιτητικές από ποτέ. Η διαδικασία ETL (Extract, Transform, Load) αποτελείται από τρία βασικά στάδια: την εξαγωγή δεδομένων από διάφορες πηγές, τον μετασχηματισμό τους σε χρήσιμη μορφή μέσω καθαρισμού και εμπλουτισμού, και τέλος τη φόρτωσή τους σε μια αποθήκη δεδομένων. Οι τεχνολογίες ETL (Extract, Transform, Load) παίζουν κεντρικό ρόλο σε αυτήν την προσπάθεια, διευκολύνοντας τη διαχείριση και επεξεργασία μεγάλων όγκων δεδομένων που σχετίζονται με απειλές, επιθέσεις και γενικότερα με την ασφάλεια των πληροφοριών. Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία επιχειρεί να αναπτύξει και να υλοποιήσει μια ολοκληρωμένη λύση ETL που όχι μόνο θα καλύπτει τις απαιτήσεις της κυβερνοασφάλειας, αλλά θα επιτρέπει και τη συνεργατική επεξεργασία και οπτικοποίηση των δεδομένων.

1.1 Περιγραφή του υπό μελέτη ερωτήματος

Στον τομέα της κυβερνοασφάλειας, η ικανότητα να ανταποκρίνεται κανείς γρήγορα και αποτελεσματικά σε νέες απειλές είναι κρίσιμη. Οι κυβερνοεπιθέσεις αυξάνονται τόσο σε συχνότητα όσο και σε πολυπλοκότητα, καθιστώντας αναγκαία την ύπαρξη ισχυρών εργαλείων για την ανάλυση και επεξεργασία δεδομένων σε πραγματικό χρόνο. Ωστόσο, η υπάρχουσα τεχνολογία ETL συχνά δεν είναι επαρκής για την κάλυψη των αυξανόμενων αναγκών της ανάλυσης δεδομένων κυβερνοασφάλειας. Η ραγδαία αύξηση των δεδομένων και η ανάγκη για γρήγορη ανάλυση και ανταπόκριση σημαίνει ότι οι παραδοσιακές μέθοδοι ETL, που σχεδιάστηκαν για πιο σταθερά και προβλέψιμα περιβάλλοντα, μπορεί να μην είναι κατάλληλες για τις σύγχρονες απαιτήσεις της κυβερνοασφάλειας.

Τα βασικά προβλήματα που αντιμετωπίζονται σε αυτό το πλαίσιο περιλαμβάνουν:

1. Δυσκολίες στην κοινή χρήση πληροφοριών απειλών:

Η ανταλλαγή πληροφοριών απειλών μεταξύ οργανισμών είναι ζωτικής σημασίας, αλλά παραμένει δύσκολη λόγω της έλλειψης κοινών προτύπων και αποτελεσματικών μηχανισμών διαμοιρασμού. Οι πληροφορίες για απειλές μπορούν να προέρχονται από διάφορες πηγές, όπως διεθνείς οργανισμοί, εθνικές αρχές, ή ακόμα και ιδιωτικές εταιρείες ασφάλειας. Το STIX/TAXII έχει προταθεί ως πρότυπο για την επίλυση αυτού του προβλήματος, αλλά η ενσωμάτωση του σε υπάρχουσες διαδικασίες ETL παραμένει πρόκληση, ειδικά λόγω της πολυπλοκότητας των απαιτήσεων και της ανάγκης για διασύνδεση με διάφορα συστήματα. Η αποτελεσματική ανταλλαγή πληροφοριών απειλών είναι κρίσιμη για την ταχεία απόκριση σε νέες απειλές, αλλά χωρίς κατάλληλη τεχνολογική υποδομή, αυτή η ανταλλαγή μπορεί να καθυστερήσει ή να είναι αναξιόπιστη. Οι οργανισμοί χρειάζονται λύσεις που να τους επιτρέπουν να ενσωματώνουν και να μοιράζονται πληροφορίες απειλών με ασφάλεια και ταχύτητα, διασφαλίζοντας την αποτελεσματική προστασία των συστημάτων τους.

2. Ανάγκη για συνεργατική ανάπτυξη:

Η αποτελεσματική συνεργασία σε έργα επεξεργασίας δεδομένων είναι καθοριστική, ειδικά όταν εμπλέκονται πολλαπλές ομάδες ή οργανισμοί. Η ανάλυση και η ανταπόκριση σε απειλές απαιτούν συχνά τη συνεργασία μεταξύ διαφόρων ομάδων, όπως οι ομάδες ανάλυσης δεδομένων, οι ειδικοί στην κυβερνοασφάλεια, και οι διαχειριστές δικτύων. Ωστόσο, η έλλειψη οργανωμένων εργαλείων και διαδικασιών για τη συνεργατική ανάπτυξη μπορεί να οδηγήσει σε ασυνεπή και ασαφή αποτελέσματα. Η χρήση εργαλείων όπως το Git και το GitHub μπορεί να βελτιώσει την

οργάνωση και τη διαφάνεια σε τέτοιες συνεργασίες, αλλά απαιτείται προσεκτικός σχεδιασμός και προσαρμογή στις συγκεκριμένες ανάγκες του έργου. Οι οργανισμοί που δεν διαθέτουν τις κατάλληλες υποδομές για συνεργατική ανάπτυξη κινδυνεύουν να αντιμετωπίσουν προβλήματα συντονισμού και ποιότητας, τα οποία μπορούν να επηρεάσουν αρνητικά την αποτελεσματικότητα των προσπαθειών κυβερνοασφάλειας.

3. Ελλιπής οπτικοποίηση δεδομένων:

Η δυνατότητα να οπτικοποιούνται τα δεδομένα είναι απαραίτητη για την κατανόηση των απειλών και την ανάπτυξη στρατηγικών απόκρισης. Παρόλο που υπάρχουν πολλά εργαλεία για την οπτικοποίηση δεδομένων, η ενσωμάτωσή τους στις διαδικασίες ETL συχνά παρουσιάζει δυσκολίες, δημιουργώντας κενά στην αλυσίδα επεξεργασίας δεδομένων. Η οπτικοποίηση των δεδομένων μπορεί να βοηθήσει τους αναλυτές να κατανοήσουν περίπλοκα μοτίβα και σχέσεις, τα οποία μπορεί να μην είναι άμεσα εμφανή από τα ακατέργαστα δεδομένα. Ωστόσο, χωρίς κατάλληλα εργαλεία και τεχνικές για την οπτικοποίηση, οι οργανισμοί μπορεί να χάσουν σημαντικές πληροφορίες ή να μην αντιδράσουν εγκαίρως σε αναδυόμενες απειλές. Η ενσωμάτωση των εργαλείων οπτικοποίησης στις διαδικασίες ETL είναι απαραίτητη για να εξασφαλιστεί ότι τα δεδομένα μπορούν να παρουσιαστούν με τρόπο που να διευκολύνει τη λήψη γρήγορων και τεκμηριωμένων αποφάσεων.

Αυτή η εργασία στοχεύει να αντιμετωπίσει αυτά τα προβλήματα μέσω της ανάπτυξης ενός ολοκληρωμένου πλαισίου ETL σε επίπεδο μεταπτυχιακής εργασίας και τοπικού περιβάλλοντος, το οποίο θα υποστηρίζει την αποδοτική και συνεργατική επεξεργασία δεδομένων κυβερνοασφάλειας. Το πλαίσιο αυτό θα ενσωματώνει τα πρότυπα STIX/TAXII για την ανταλλαγή πληροφοριών απειλών και θα προτείνει βέλτιστες πρακτικές για τη συνεργατική ανάπτυξη με τη χρήση εργαλείων όπως το Git και το GitHub, επιτυγχάνοντας ταυτόχρονα και μια καλύτερη οπτικοποίηση των δεδομένων για την ανάλυση των απειλών χρησιμοποιώντας το Microsoft Power BI. Η χρήση του Power BI, ενός ισχυρού εργαλείου για την ανάλυση και οπτικοποίηση δεδομένων, θα επιτρέψει στους αναλυτές να δημιουργούν προσαρμοσμένα dashboards και αναφορές που θα βοηθούν στην καλύτερη κατανόηση και διαχείριση των απειλών.

1.2 Στόχοι

Ο βασικός στόχος της παρούσας μεταπτυχιακής διατριβής είναι να διερευνήσει και να υλοποιήσει μια ολοκληρωμένη λύση για την αποτελεσματική και συνεργατική επεξεργασία και οπτικοποίηση δεδομένων κυβερνοασφάλειας, χρησιμοποιώντας τεχνολογίες ETL. Οι τεχνολογίες αυτές είναι ζωτικής σημασίας για τη διαχείριση των δεδομένων που παράγονται σε πραγματικό χρόνο από διάφορες πηγές, όπως τα συστήματα παρακολούθησης δικτύων, οι βάσεις δεδομένων καταγραφής συμβάντων και οι εξειδικευμένες πλατφόρμες ανάλυσης απειλών. Η εργασία αυτή επιδιώκει να αναπτύξει μια λύση που θα μπορεί να ενσωματωθεί σε υπάρχουσες υποδομές, παρέχοντας ταυτόχρονα δυνατότητες για ανάλυση και οπτικοποίηση των δεδομένων με τρόπο που να υποστηρίζει τη λήψη αποφάσεων.

Ειδικότερα, οι επιμέρους στόχοι περιλαμβάνουν:

1. Διερεύνηση των τεχνολογιών ETL και της εφαρμογής τους στην κυβερνοασφάλεια:

Ένας από τους πρωταρχικούς στόχους είναι η ανάλυση του πώς οι τεχνολογίες ETL μπορούν να προσαρμοστούν στις ειδικές απαιτήσεις των δεδομένων που σχετίζονται με την κυβερνοασφάλεια. Η διαχείριση δεδομένων στον τομέα της κυβερνοασφάλειας απαιτεί όχι μόνο τη συλλογή και συσχέτιση δεδομένων από πολλαπλές πηγές, αλλά και την ικανότητα να τα μετασχηματίζει και να τα φορτώνει σε κεντρικά συστήματα, όπου μπορούν να αναλυθούν σε πραγματικό χρόνο. Αυτό περιλαμβάνει την εξέταση της διαχείρισης και ανάλυσης δεδομένων που

αφορούν απειλές και επιθέσεις, με σκοπό την ανάπτυξη μεθοδολογιών που να επιτρέπουν την αποδοτική επεξεργασία μεγάλου όγκου δεδομένων σε πραγματικό χρόνο. Είναι απαραίτητο να κατανοηθεί πώς οι τεχνολογίες ETL μπορούν να συνδυαστούν με άλλα συστήματα, όπως οι μηχανές ανάλυσης συμπεριφοράς και οι βάσεις δεδομένων πληροφοριών απειλών, για να παρέχουν μια ολοκληρωμένη εικόνα της κατάστασης ασφαλείας ενός οργανισμού.

2. Ενσωμάτωση του προτύπου STIX/TAXII:

Η εργασία στοχεύει επίσης στην ενσωμάτωση των προτύπων STIX (Structured Threat Information Expression) [12,13] και TAXII (Trusted Automated Exchange of Indicator Information) [12,13] σε ένα περιβάλλον ETL. Τα πρότυπα STIX/TAXII έχουν σχεδιαστεί για να διευκολύνουν την τυποποίηση και την ανταλλαγή πληροφοριών απειλών μεταξύ διαφορετικών οργανισμών και συστημάτων. Η ενσωμάτωση αυτή αναμένεται να υποστηρίξει τη διαδικασία κοινής χρήσης και επεξεργασίας πληροφοριών απειλών, διευκολύνοντας την ανταλλαγή κρίσιμων δεδομένων μεταξύ διαφορετικών οργανισμών και ομάδων ασφαλείας. Αυτή η ανταλλαγή πληροφοριών είναι κρίσιμη για τη βελτίωση της συλλογικής άμυνας κατά των απειλών στον κυβερνοχώρο, επιτρέποντας σε οργανισμούς να ανταλλάσσουν πληροφορίες για νέες απειλές και να αντιδρούν πιο γρήγορα σε αυτές [12,13].

3. Ανάπτυξη συνεργατικών μεθόδων επεξεργασίας δεδομένων*:*

Ένας άλλος σημαντικός στόχος είναι η διερεύνηση των τρόπων με τους οποίους εργαλεία όπως το Git και το GitHub [21,23] μπορούν να χρησιμοποιηθούν για την προώθηση της συνεργατικής ανάπτυξης σε έργα ETL. Η συνεργασία είναι καθοριστική για την αποτελεσματική διαχείριση των δεδομένων κυβερνοασφάλειας, καθώς πολλές ομάδες και οργανισμοί χρειάζεται να συνεργαστούν για την ανάλυση και την αντιμετώπιση των απειλών. Η ανάγκη για συνεργατική επεξεργασία δεδομένων γίνεται ιδιαίτερα έντονη σε περιβάλλοντα όπου οι απειλές εξελίσσονται γρήγορα και απαιτούν άμεση απόκριση. Η εργασία αυτή θα εξετάσει πώς τα εργαλεία ανάπτυξης λογισμικού μπορούν να ενσωματωθούν σε έργα ETL για να διευκολύνουν τη συνεργασία μεταξύ γεωγραφικά διασκορπισμένων ομάδων, μειώνοντας ταυτόχρονα τα σφάλματα και βελτιώνοντας την ποιότητα των δεδομένων.

4. Πρακτική υλοποίηση μιας ολοκληρωμένης λύσης ETL:

Τέλος, η εργασία θα εστιάσει στην υλοποίηση μιας ολοκληρωμένης λύσης ETL σε ένα τοπικό περιβάλλον (local PC), η οποία θα αποδεικνύει τη θεωρητική ανάλυση και θα προσφέρει μια λειτουργική αρχιτεκτονική για την επεξεργασία δεδομένων κυβερνοασφάλειας. Η λύση αυτή θα καλύπτει όλο το φάσμα της επεξεργασίας, από την εξαγωγή και τον μετασχηματισμό των δεδομένων μέχρι τη φόρτωση και την τελική ανάλυση τους. Επιπλέον, η υλοποίηση θα περιλαμβάνει την οπτικοποίηση των δεδομένων χρησιμοποιώντας εργαλεία όπως το Power BI, τα οποία θα επιτρέπουν στους χρήστες να αναλύουν και να παρουσιάζουν τα δεδομένα με τρόπους που διευκολύνουν τη λήψη τεκμηριωμένων αποφάσεων. Αυτή η πρακτική υλοποίηση θα δώσει τη δυνατότητα στους οργανισμούς να αξιολογήσουν την αποτελεσματικότητα της λύσης σε πραγματικές συνθήκες και να προσαρμόσουν τις διαδικασίες τους αναλόγως.

1.3 Συνεισφορά της Διατριβής

Η συμβολή της παρούσας διπλωματικής εργασίας επικεντρώνεται στην ανάπτυξη και εφαρμογή μιας ολοκληρωμένης λύσης για την επεξεργασία δεδομένων κυβερνοασφάλειας, με έμφαση στη συνεργασία και την ενσωμάτωση πρότυπων τεχνολογιών. Η εργασία αυτή προσφέρει νέες προοπτικές και λύσεις που μπορούν να βοηθήσουν τους οργανισμούς να βελτιώσουν την ασφάλειά τους και να ανταποκριθούν πιο αποτελεσματικά στις απειλές.

Δημιουργία ενός πλαισίου ETL για κυβερνοασφάλεια:

Η εργασία προτείνει και υλοποιεί ένα ειδικά διαμορφωμένο πλαίσιο ETL (Extract, Transform, Load) που είναι προσαρμοσμένο στις ανάγκες της κυβερνοασφάλειας. Το πλαίσιο αυτό σχεδιάστηκε για να υποστηρίξει τη διαχείριση και ανάλυση δεδομένων απειλών σε ένα περιβάλλον που απαιτεί αυστηρές διαδικασίες και ακρίβεια. Αντί να επικεντρώνεται σε μεγάλα δεδομένα ή αναλύσεις σε πραγματικό χρόνο, το πλαίσιο δίνει προτεραιότητα στην ασφάλεια, την ακρίβεια και την αξιοπιστία κατά τη διαδικασία μετασχηματισμού και φόρτωσης των δεδομένων, εξασφαλίζοντας ότι τα δεδομένα απειλών μπορούν να επεξεργαστούν, να οπτικοποιηθούν και να διανεμηθούν με συνέπεια για περαιτέρω χρήση.

Επιπλέον, αυτή η λύση αποτελεί μια προσαρμοσμένη πρόταση που μπορεί να επεκταθεί και να εφαρμοστεί σε μεγαλύτερα και πιο πολύπλοκα σύνολα δεδομένων (datasets) από εκείνα που χρησιμοποιούνται στο συγκεκριμένο case study. Αυτή η προσαρμοστικότητα καλύπτει πολλές από τις ανάγκες που έχουν οι εταιρείες στις υλοποιήσεις έργων ETL, επιτρέποντας τους να επεξεργάζονται και να διαχειρίζονται με αποτελεσματικότητα τα δεδομένα τους σε ένα περιβάλλον με αυξημένες απαιτήσεις. Η ευελιξία και η δυνατότητα προσαρμογής του πλαισίου σημαίνει ότι οι οργανισμοί μπορούν να αντιμετωπίσουν τις μεταβαλλόμενες απαιτήσεις της αγοράς και να προσαρμόσουν τις διαδικασίες τους ανάλογα με τις νέες απειλές και τις τεχνολογικές εξελίξεις.

Ενσωμάτωση των προτύπων STIX/TAXII:

Ένα από τα κύρια σημεία συμβολής της εργασίας είναι η ενσωμάτωση των προτύπων STIX (Structured Threat Information Expression) και TAXII (Trusted Automated Exchange of Indicator Information) στο προτεινόμενο πλαίσιο ETL. Αυτή η ενσωμάτωση διευκολύνει τη δομημένη και ασφαλή ανταλλαγή πληροφοριών απειλών μεταξύ οργανισμών και ομάδων κυβερνοασφάλειας. Η εργασία αναπτύσσει μια προσέγγιση που επιτρέπει την απρόσκοπτη ενσωμάτωση αυτών των προτύπων, δίνοντας στους οργανισμούς τη δυνατότητα να ανταλλάσσουν πληροφορίες με αποτελεσματικότητα και χωρίς να διακινδυνεύουν την ασφάλεια ή την ακεραιότητα των δεδομένων.

Η ενσωμάτωση των προτύπων αυτών στο πλαίσιο ETL επιτρέπει τη δημιουργία μιας συνεκτικής και ολοκληρωμένης λύσης που μπορεί να ανταποκριθεί στις ανάγκες της σύγχρονης κυβερνοασφάλειας. Μέσω αυτής της προσέγγισης, οι οργανισμοί μπορούν να ενισχύσουν τη συνεργασία τους και να ανταλλάσσουν πληροφορίες με μεγαλύτερη ταχύτητα και ακρίβεια, βελτιώνοντας έτσι τη συνολική τους ασφάλεια.

Ενίσχυση της συνεργατικότητας μέσω εργαλείων ανάπτυξης λογισμικού:

Η εργασία εξετάζει τη χρήση εργαλείων όπως το Git και το GitHub για τη διαχείριση και τον συντονισμό συνεργατικών έργων στην κυβερνοασφάλεια. Η συμβολή της εδώ έγκειται στην παροχή κατευθυντήριων γραμμών και βέλτιστων πρακτικών για τη χρήση αυτών των εργαλείων, ώστε να βελτιωθεί η συνεργασία μεταξύ διαφορετικών ομάδων που εργάζονται σε κοινά έργα.

Η εργασία αναλύει πώς τα εργαλεία αυτά μπορούν να ενσωματωθούν στο πλαίσιο ETL, επιτρέποντας την αποτελεσματική διαχείριση των αλλαγών, την παρακολούθηση της εξέλιξης του έργου και τη διασφάλιση της ποιότητας του κώδικα μέσα από διαδικασίες όπως τα pull requests. Επιπλέον, η χρήση αυτών των εργαλείων βοηθά στη δημιουργία ενός δομημένου και αποδοτικού περιβάλλοντος συνεργασίας, όπου οι ομάδες μπορούν να εργάζονται ταυτόχρονα και να ανταλλάσσουν ιδέες και λύσεις σε πραγματικό χρόνο.

Υλοποίηση μιας ολοκληρωμένης λύσης σε τοπικό περιβάλλον:

Μία από τις σημαντικές συνεισφορές αυτής της εργασίας είναι η πρακτική υλοποίηση της προτεινόμενης λύσης σε ένα τοπικό περιβάλλον (local PC). Η υλοποίηση αυτή δείχνει πώς

μπορούν να ενσωματωθούν και να λειτουργήσουν όλες οι παραπάνω τεχνολογίες και προσεγγίσεις σε ένα πραγματικό περιβάλλον εργασίας.

Αυτή η προσέγγιση δίνει στους ερευνητές και τους επαγγελματίες την ευκαιρία να δοκιμάσουν και να αξιολογήσουν την αποτελεσματικότητα της λύσης σε ασφαλείς και ελεγχόμενες συνθήκες, καθιστώντας την λύση αυτή ιδιαίτερα χρήσιμη για οργανισμούς που χρειάζονται να επεκτείνουν τις δυνατότητές τους σε πιο σύνθετα έργα και μεγαλύτερα datasets. Με τη συγκεκριμένη προσέγγιση, η εργασία αυτή καλύπτει σημαντικές ανάγκες των εταιριών που δραστηριοποιούνται στην ανάπτυξη και υλοποίηση έργων ETL, προσφέροντας μια προσαρμοσμένη και ευέλικτη λύση που μπορεί να επεκταθεί ανάλογα με τις αυξανόμενες απαιτήσεις.

Η υλοποίηση αυτής της λύσης σε τοπικό περιβάλλον επιτρέπει στους οργανισμούς να κατανοήσουν πώς μπορούν να ενσωματώσουν τις νέες τεχνολογίες και μεθόδους στις υπάρχουσες υποδομές τους, χωρίς να χρειάζεται να αναπτύξουν πολύπλοκες και δαπανηρές λύσεις από το μηδέν. Αυτό μειώνει το κόστος και τον χρόνο υλοποίησης, ενώ παράλληλα βελτιώνει την αποτελεσματικότητα και την αποδοτικότητα των διαδικασιών κυβερνοασφάλειας.

1.4 Δομή της Διατριβής

Η παρούσα διπλωματική εργασία είναι δομημένη σε πέντε κύρια κεφάλαια, τα οποία περιγράφουν με λεπτομέρεια την προσέγγιση που ακολουθήθηκε για την επίτευξη των στόχων, την ανάλυση των αποτελεσμάτων και τα συμπεράσματα που εξάγονται. Κάθε κεφάλαιο έχει σχεδιαστεί για να παρέχει στον αναγνώστη μια σαφή κατανόηση της διαδικασίας και των ευρημάτων της εργασίας.

Ακολουθεί η περιγραφή της δομής της διπλωματικής εργασίας:

- **Κεφάλαιο 1: Εισαγωγή:**

  Το πρώτο κεφάλαιο παρέχει μια γενική εικόνα του θέματος της εργασίας και καθορίζει τους στόχους της. Ξεκινά με την παρουσίαση του προβλήματος που εξετάζεται, θέτοντας το πλαίσιο μέσα στο οποίο η εργασία λαμβάνει χώρα. Εδώ περιγράφεται η σημαντικότητα του θέματος της κυβερνοασφάλειας και γιατί η ανάπτυξη ενός προσαρμοσμένου πλαισίου ETL είναι κρίσιμη για την αντιμετώπιση των σύγχρονων προκλήσεων. Στη συνέχεια, γίνεται αναφορά στη σχετική εργασία που έχει γίνει στον τομέα, ώστε να εντοπιστούν τα κενά που η παρούσα εργασία επιχειρεί να καλύψει. Τέλος, παρουσιάζεται η συμβολή της διπλωματικής εργασίας και περιγράφεται η δομή της, παρέχοντας μια καθοδηγητική εικόνα για το τι θα ακολουθήσει στα επόμενα κεφάλαια.

- **Κεφάλαιο 2: Επισκόπηση Βιβλιογραφίας:**

  Το δεύτερο κεφάλαιο είναι αφιερωμένο στην ανασκόπηση της υπάρχουσας βιβλιογραφίας. Σε αυτό το τμήμα, εξετάζονται οι εξελίξεις στην τεχνολογία ETL και οι παραδοσιακές μεθοδολογίες που έχουν χρησιμοποιηθεί μέχρι σήμερα. Γίνεται αναφορά στους περιορισμούς αυτών των παραδοσιακών μεθόδων και στις προκλήσεις που προκύπτουν όταν εφαρμόζονται σε περιβάλλοντα κυβερνοασφάλειας. Στη συνέχεια, αναλύονται οι σύγχρονες προσεγγίσεις και οι νέες αρχιτεκτονικές λύσεις που έχουν προταθεί στην βιβλιογραφία, με έμφαση στην ενσωμάτωση τεχνολογιών και προτύπων που βελτιώνουν την αποτελεσματικότητα της διαδικασίας ETL. Ιδιαίτερη έμφαση δίνεται στην ενσωμάτωση των προτύπων STIX/TAXII για την κοινή χρήση πληροφοριών απειλών και στη σημασία της συνεργατικής ανάπτυξης σε έργα ETL. Το κεφάλαιο αυτό λειτουργεί ως θεωρητική βάση για την πρακτική υλοποίηση που ακολουθεί.

- **Κεφάλαιο 3: Μεθοδολογία:**

Το τρίτο κεφάλαιο περιγράφει τη μεθοδολογία που ακολουθήθηκε για την υλοποίηση της προτεινόμενης λύσης. Αρχικά, παρουσιάζεται η διαδικασία δημιουργίας συνθετικών δεδομένων, τα οποία χρησιμοποιήθηκαν για τις δοκιμές των διαδικασιών ETL. Αυτά τα δεδομένα προσομοιώνουν τις πραγματικές συνθήκες εργασίας σε περιβάλλον κυβερνοασφάλειας, προσφέροντας μια αξιόπιστη βάση για την αξιολόγηση της λύσης. Στη συνέχεια, αναλύεται η υλοποίηση των διαδικασιών ETL χρησιμοποιώντας SQL Layers και Stored Procedures, με στόχο τη βελτιστοποίηση της επεξεργασίας των δεδομένων. Το κεφάλαιο αυτό εξετάζει επίσης την ενσωμάτωση των προτύπων STIX/TAXII και περιγράφει πώς αυτά ενσωματώθηκαν στο πλαίσιο ETL για να διευκολύνουν την ανταλλαγή πληροφοριών απειλών. Τέλος, γίνεται αναφορά στη χρήση των εργαλείων Git και GitHub για τη συνεργατική ανάπτυξη, προσφέροντας μια ολοκληρωμένη εικόνα της μεθοδολογίας που ακολουθήθηκε.

- **Κεφάλαιο 4: Μελέτη Περίπτωσης – Αρχιτεκτονική και Υλοποίηση:**

  Το τέταρτο κεφάλαιο αποτελεί τον πυρήνα της διπλωματικής εργασίας, καθώς παρουσιάζει μια λεπτομερή περιγραφή της αρχιτεκτονικής και της υλοποίησης του έργου ETL που αναπτύχθηκε στο πλαίσιο της εργασίας. Περιγράφονται τα βήματα που ακολουθήθηκαν για την εισαγωγή, μετασχηματισμό και φόρτωση των δεδομένων, αναδεικνύοντας τις προκλήσεις και τις λύσεις που δόθηκαν. Ιδιαίτερη έμφαση δίνεται στην ενσωμάτωση του STIX/TAXII, όπου αναλύεται πώς αυτά τα πρότυπα υποστηρίζουν την ασφαλή και αποτελεσματική ανταλλαγή πληροφοριών απειλών. Το κεφάλαιο εξετάζει επίσης τη χρήση των εργαλείων Git και GitHub, παρουσιάζοντας συγκεκριμένα παραδείγματα και πρακτικές εφαρμογές που ενισχύουν τη συνεργατικότητα στο έργο. Τέλος, περιγράφεται η χρήση του Power BI για την οπτικοποίηση των δεδομένων κυβερνοασφάλειας, προσφέροντας ένα ισχυρό εργαλείο για την ανάλυση και κατανόηση των αποτελεσμάτων.

- **Κεφάλαιο 5: Συμπεράσματα:**

  Το πέμπτο και τελευταίο κεφάλαιο συνοψίζει τα κύρια ευρήματα της διπλωματικής εργασίας και προτείνει κατευθύνσεις για μελλοντική έρευνα. Αναλύονται τα αποτελέσματα που επιτεύχθηκαν μέσα από την εφαρμογή του προτεινόμενου πλαισίου ETL και συζητούνται οι δυνατότητες περαιτέρω βελτιώσεων και εξελίξεων στον τομέα της επεξεργασίας δεδομένων κυβερνοασφάλειας. Το κεφάλαιο αυτό προσφέρει επίσης προτάσεις για την ενσωμάτωση νέων τεχνολογιών και μεθόδων, οι οποίες θα μπορούσαν να ενισχύσουν ακόμα περισσότερο την αποτελεσματικότητα και την αποδοτικότητα των διαδικασιών ETL σε περιβάλλοντα κυβερνοασφάλειας.

Αυτή η δομή εξασφαλίζει μια ομαλή και λογική ροή πληροφοριών, ξεκινώντας από τη θεωρητική θεμελίωση του προβλήματος και φτάνοντας μέχρι την πρακτική εφαρμογή και αξιολόγηση της προτεινόμενης λύσης. Με αυτόν τον τρόπο, ο αναγνώστης αποκτά μια ολοκληρωμένη κατανόηση του θέματος και των αποτελεσμάτων της εργασίας, καθώς και των πιθανών μελλοντικών κατευθύνσεων για περαιτέρω έρευνα και ανάπτυξη.

2. Επισκόπηση Βιβλιογραφίας

2.1 Ιστορική Προοπτική για την Εξέλιξη της Τεχνολογίας ETL

Η τεχνολογία ETL (Extract, Transform, Load) έχει τις ρίζες της στις αρχές της δεκαετίας του 1970, όταν οι επιχειρήσεις άρχισαν να αναζητούν μεθόδους για τη συλλογή και ενσωμάτωση δεδομένων από διαφορετικές πηγές με σκοπό την ανάλυση και την υποστήριξη αποφάσεων. Τα πρώτα συστήματα ETL ήταν σχετικά απλά, δεδομένου ότι ο όγκος και η πολυπλοκότητα των δεδομένων ήταν περιορισμένα. Η κύρια πρόκληση σε αυτή τη φάση ήταν η συλλογή δεδομένων από ετερογενείς πηγές, όπως βάσεις δεδομένων, αρχεία και άλλα επιχειρησιακά συστήματα, και η ενσωμάτωσή τους σε κεντρικές αποθήκες δεδομένων (data warehouses).[1,2,3,4]

Καθώς οι ανάγκες των επιχειρήσεων εξελίχθηκαν, η τεχνολογία ETL αναπτύχθηκε περαιτέρω για να ανταποκριθεί στις απαιτήσεις για πιο περίπλοκες διαδικασίες μετασχηματισμού και καθαρισμού δεδομένων [2,3]. Στη δεκαετία του 1980 και του 1990, τα ETL εργαλεία απέκτησαν μεγαλύτερη λειτουργικότητα, επιτρέποντας πιο σύνθετους μετασχηματισμούς και τη δυνατότητα χειρισμού μεγαλύτερων όγκων δεδομένων. Τα εργαλεία αυτά άρχισαν να ενσωματώνουν χαρακτηριστικά όπως ο καθαρισμός δεδομένων, η ενοποίηση και η διαχείριση εξαιρέσεων, προσφέροντας πιο ολοκληρωμένες λύσεις για τη διαχείριση των δεδομένων.[2,4,14,15]

Η σχετική βιβλιογραφία [1,2,3,4,14,15] επιβεβαιώνει την εξέλιξη των παραδοσιακών μεθοδολογιών ETL και αναδεικνύει τα όριά τους. Στις αρχικές φάσεις, οι διαδικασίες ETL βασίζονταν σε batch επεξεργασία, που αν και κατάλληλη για σταθερές επιχειρηματικές ανάγκες, παρουσίαζε προβλήματα όταν οι ανάγκες γίνονταν πιο δυναμικές. Ειδικά στον τομέα της κυβερνοασφάλειας, όπου η ανάλυση δεδομένων πρέπει να πραγματοποιείται σε πραγματικό χρόνο, οι παραδοσιακές ETL διαδικασίες δεν ήταν επαρκείς. Η καθυστέρηση στην επεξεργασία δεδομένων μπορεί να οδηγήσει σε καθυστέρηση στον εντοπισμό κυβερνοαπειλών, δίνοντας στους επιτιθέμενους πολύτιμο χρόνο για να προκαλέσουν ζημιές.

Με την έλευση του Διαδικτύου και την ανάπτυξη των δικτύων, η ανάγκη για επεξεργασία δεδομένων σε πραγματικό χρόνο έγινε πιο επιτακτική. Οι παραδοσιακές διαδικασίες ETL, που βασίζονταν σε batch επεξεργασία, δεν μπορούσαν να ανταποκριθούν επαρκώς σε αυτές τις νέες απαιτήσεις.[1] Αυτό οδήγησε στη δημιουργία νέων μεθόδων και εργαλείων που μπορούσαν να επεξεργαστούν δεδομένα σε πραγματικό χρόνο, αλλάζοντας το τοπίο της επεξεργασίας δεδομένων.

Στα τέλη της δεκαετίας του 2000 και στις αρχές του 2010, οι εξελίξεις στις τεχνολογίες υπολογιστικού νέφους (cloud computing) και τα μεγάλα δεδομένα (big data) έφεραν μια νέα εποχή στην τεχνολογία ETL. Οι πλατφόρμες ETL που βασίζονται στο cloud επιτρέπουν την κλιμάκωση της επεξεργασίας δεδομένων, ενώ η υιοθέτηση των αρχών του DevOps και των μικροϋπηρεσιών επέτρεψε την αυτοματοποίηση και την ευελιξία στις διαδικασίες ETL [6,7]. Αυτές οι εξελίξεις καθιστούν το ETL μια κρίσιμη τεχνολογία για τη σύγχρονη επιχειρησιακή νοημοσύνη και την ανάλυση δεδομένων, προσφέροντας λύσεις που μπορούν να προσαρμοστούν στις ανάγκες των επιχειρήσεων, ανεξαρτήτως μεγέθους και κλάδου.[6,7]

Η σχετική έρευνα [5,7,8,9,10,11,12] υποδεικνύει τη σημασία των σύγχρονων τεχνολογιών όπως οι in-memory βάσεις δεδομένων και η χρήση εργαλείων όπως το Apache Kafka, Apache Hadoop και Apache Spark για τη διαχείριση ροών δεδομένων σε πραγματικό χρόνο.[9] Αυτές οι λύσεις προτάθηκαν ως απάντηση στις προκλήσεις που αντιμετωπίζουν οι παραδοσιακές μέθοδοι ETL. Οι in-memory βάσεις δεδομένων επιταχύνουν σημαντικά τις διαδικασίες ETL, καθώς τα

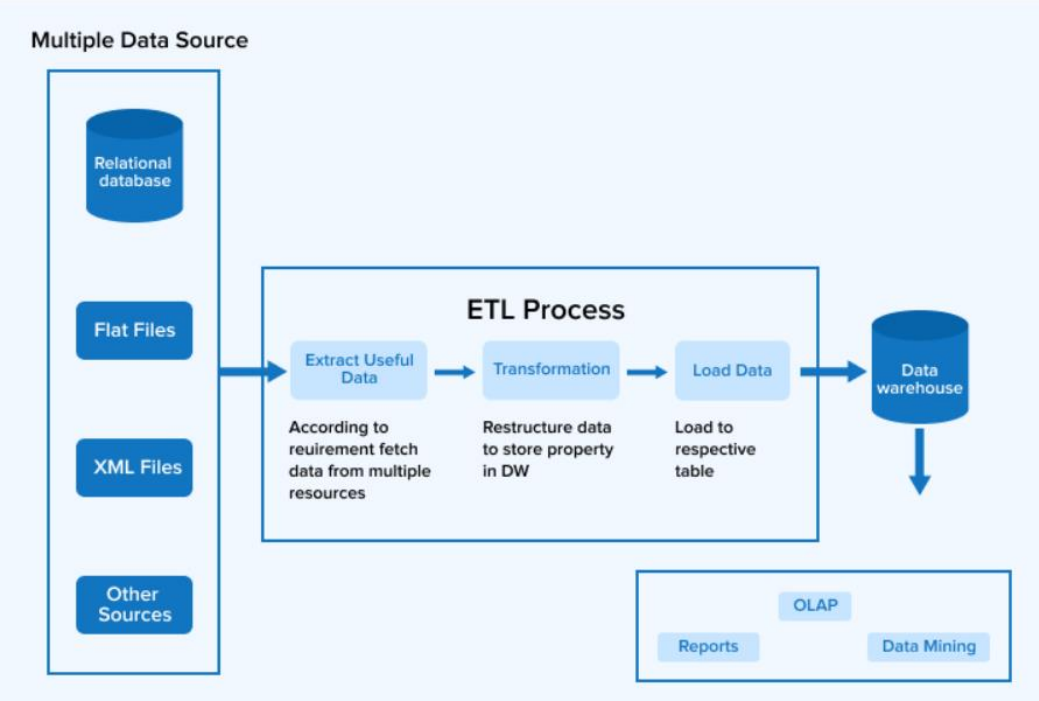
δεδομένα αναλύονται άμεσα, μόλις εισέλθουν στο σύστημα [10]. Η υιοθέτηση μικροϋπηρεσιών προσφέρει επιπλέον ευελιξία και δυνατότητα κλιμάκωσης στην επεξεργασία δεδομένων, επιτρέποντας την εύκολη ενσωμάτωση νέων λειτουργιών χωρίς να διαταράσσεται η συνολική διαδικασία [7].

Συμπερασματικά, η ιστορική εξέλιξη της τεχνολογίας ETL, σε συνδυασμό με την ανάλυση της σχετικής βιβλιογραφίας, δείχνει πώς οι επιχειρήσεις και οι οργανισμοί έχουν προσαρμοστεί στις αυξανόμενες ανάγκες επεξεργασίας και ανάλυσης δεδομένων.

2.2 Ανασκόπηση Παραδοσιακών Μεθοδολογιών ETL και Περιορισμοί

Οι παραδοσιακές μεθοδολογίες ETL έχουν χρησιμοποιηθεί ευρέως για την ενσωμάτωση δεδομένων από διάφορες πηγές σε αποθήκες δεδομένων, παρέχοντας τη βάση για αναλύσεις επιχειρησιακής νοημοσύνης [1,4]. Αυτές οι μεθοδολογίες, όμως, συνοδεύονται από περιορισμούς που καθίστανται εμφανείς ιδιαίτερα σε περιβάλλοντα όπου η ταχύτητα και η κλίμακα της επεξεργασίας δεδομένων είναι κρίσιμες [3,8].

Στις παραδοσιακές διαδικασίες ETL, τα δεδομένα εξάγονται συνήθως σε συγκεκριμένα χρονικά διαστήματα (batch processing) από διαφορετικές πηγές. Αυτά τα δεδομένα στη συνέχεια μετασχηματίζονται σύμφωνα με προκαθορισμένους κανόνες και φορτώνονται σε αποθήκες δεδομένων. Το μοντέλο αυτό είναι αποτελεσματικό όταν οι ανάγκες των επιχειρήσεων είναι σταθερές και προβλέψιμες, αλλά παρουσιάζει προβλήματα σε πιο δυναμικά περιβάλλοντα.



Εικόνα 1. Παραδοσιακή Διαδικασία ETL για Ενοποίηση Δεδομένων από Πολλαπλές Πηγές σε Αποθήκη Δεδομένων

Ένας από τους κύριους περιορισμούς των παραδοσιακών μεθοδολογιών ETL είναι η έλλειψη ευελιξίας [4,8,13]. Οι διαδικασίες είναι συχνά στατικές, με περιορισμένη δυνατότητα προσαρμογής σε νέες απαιτήσεις ή αλλαγές στις πηγές δεδομένων. Αυτό σημαίνει ότι η προσαρμογή σε νέες επιχειρησιακές ανάγκες μπορεί να απαιτεί σημαντική επανασχεδίαση των ροών εργασίας ETL, κάτι που είναι χρονοβόρο και δαπανηρό.

Ένας άλλος περιορισμός αφορά την απόδοση. Οι παραδοσιακές διαδικασίες ETL βασίζονται στη batch επεξεργασία, η οποία δεν μπορεί να ανταποκριθεί επαρκώς στις απαιτήσεις για ανάλυση σε πραγματικό χρόνο [6,9,12]. Καθώς οι επιχειρήσεις γίνονται πιο συνδεδεμένες και οι αποφάσεις πρέπει να λαμβάνονται γρήγορα, η καθυστέρηση που προκαλείται από τις παραδοσιακές μεθοδολογίες ETL μπορεί να αποτελεί εμπόδιο.

Επιπλέον, οι παραδοσιακές μεθοδολογίες ETL δεν είναι σχεδιασμένες για να χειριστούν μεγάλους όγκους δεδομένων (big data) ή πολύπλοκες δομές δεδομένων, όπως είναι τα μη δομημένα δεδομένα [3,12]. Η ανάπτυξη και η συντήρηση αυτών των διαδικασιών μπορεί να είναι πολύπλοκη και δαπανηρή, ειδικά όταν οι πηγές δεδομένων και οι απαιτήσεις της επιχείρησης συνεχώς εξελίσσονται.

2.3 Εξέταση Σύγχρονων Προσεγγίσεων ETL

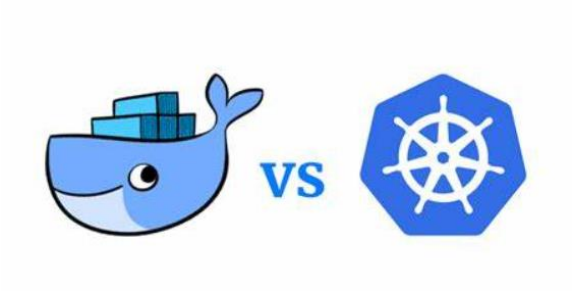
Οι σύγχρονες προσεγγίσεις ETL έχουν σχεδιαστεί για να αντιμετωπίσουν τις προκλήσεις που συνοδεύουν τις παραδοσιακές μεθοδολογίες, προσφέροντας μεγαλύτερη ευελιξία, ταχύτητα και δυνατότητα κλιμάκωσης. Αυτές οι νέες προσεγγίσεις αξιοποιούν τεχνολογίες αιχμής, όπως το υπολογιστικό νέφος (cloud computing), οι μικροϋπηρεσίες (microservices) και οι βάσεις δεδομένων in-memory, για να παρέχουν πιο δυναμικές και προσαρμοστικές λύσεις.

Μία από τις πιο σημαντικές σύγχρονες προσεγγίσεις είναι η χρήση τεχνολογιών υπολογιστικού νέφους για την κλιμάκωση των διαδικασιών ETL [7,10,16]. Οι πλατφόρμες cloud προσφέρουν δυνατότητες αποθήκευσης και επεξεργασίας δεδομένων που μπορούν να προσαρμοστούν στις ανάγκες της επιχείρησης, ανεξάρτητα από τον όγκο των δεδομένων ή την πολυπλοκότητα των μετασχηματισμών [6, 7,17]. Αυτή η προσέγγιση επιτρέπει στις επιχειρήσεις να εκμεταλλεύονται την ευελιξία και την απόδοση που προσφέρουν οι πόροι του cloud, μειώνοντας παράλληλα το κόστος και το χρόνο υλοποίησης. Οι μεγαλύτεροι πάροχοι cloud υπηρεσιών είναι η Microsoft, η Amazon και η Google.

Εικόνα 2. Οι Κορυφαίοι Πάροχοι Υπηρεσιών Cloud: Google Cloud, AWS, και Microsoft Azure

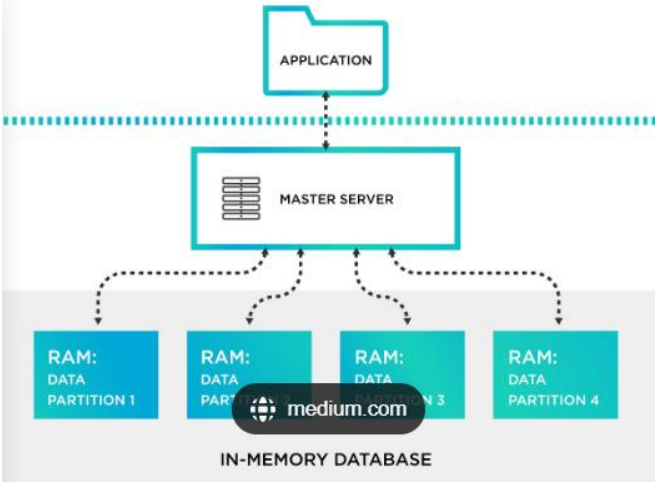
Οι μικροϋπηρεσίες αποτελούν άλλη μια σύγχρονη προσέγγιση που έχει υιοθετηθεί ευρέως στον χώρο των ETL [7,9]. Οι μικροϋπηρεσίες επιτρέπουν την ανάπτυξη ευέλικτων και ανεξάρτητων μονάδων λογισμικού που μπορούν να εκτελούνται και να αναβαθμίζονται αυτόνομα, χωρίς να επηρεάζουν άλλες υπηρεσίες [7,9]. Αυτή η προσέγγιση επιτρέπει την κλιμάκωση των διαδικασιών ETL ανάλογα με τις ανάγκες της επιχείρησης και την ενσωμάτωση νέων λειτουργιών ή την τροποποίηση των υπαρχουσών χωρίς να διαταράσσεται η συνολική λειτουργία του συστήματος. Για παράδειμα, αντί να υπάρχει μία μεγάλη και μονολιθική διαδικασία ETL που συλλέγει δεδομένα από διάφορες πηγές, μπορεί να υπάρχουν ξεχωριστές μικροϋπηρεσίες για κάθε τύπο πηγής δεδομένων (π.χ., μια υπηρεσία για δεδομένα από βάσεις δεδομένων SQL, μια άλλη για δεδομένα από αρχεία JSON, κ.λπ.). Αυτές οι υπηρεσίες μπορούν να λειτουργούν ανεξάρτητα και να ενημερώνονται ή να αναβαθμίζονται ξεχωριστά, χωρίς να επηρεάζουν το

υπόλοιπο σύστημα [8,9]. Εργαλεία για την οργάνωση και το πακετάρισμα μικρουπηρεσιών είναι το Kubernetes και το Docker.



Εικόνα 3. Docker και Kubernetes

Οι βάσεις δεδομένων in-memory είναι επίσης μια σημαντική τεχνολογία που έχει αλλάξει τον τρόπο που οι διαδικασίες ETL πραγματοποιούνται. Οι in-memory βάσεις δεδομένων αποθηκεύουν δεδομένα στη μνήμη αντί στον δίσκο, επιτρέποντας έτσι ταχύτερη πρόσβαση και επεξεργασία των δεδομένων. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε περιβάλλοντα όπου η ταχύτητα είναι κρίσιμη, όπως στην ανάλυση δεδομένων σε πραγματικό χρόνο για την ανίχνευση και απόκριση σε κυβερνοαπειλές. Μερικά εργαλεία και πλατφόρμες που αξιοποιούν την τεχνολογία in-memory βάσεων δεδομένων για την επιτάχυνση των ETL διαδικασιών είναι το SAP HANA και το Redis.



Εικόνα 4. In-Memory DB

Ένα άλλο σημαντικό χαρακτηριστικό των σύγχρονων προσεγγίσεων είναι η δυνατότητα ενοποίησης με πλατφόρμες ανάλυσης δεδομένων και εργαλεία επιχειρησιακής νοημοσύνης (BI), όπως το Power BI και το Tableau. Αυτή η ενσωμάτωση επιτρέπει στις επιχειρήσεις να οπτικοποιούν και να αναλύουν τα δεδομένα τους με τρόπους που διευκολύνουν τη λήψη αποφάσεων, βελτιώνοντας την ανταγωνιστικότητά τους στην αγορά.

Εικόνα 5. Οπτικοποίηση και Ανάλυση Επιθέσεων με PowerBI

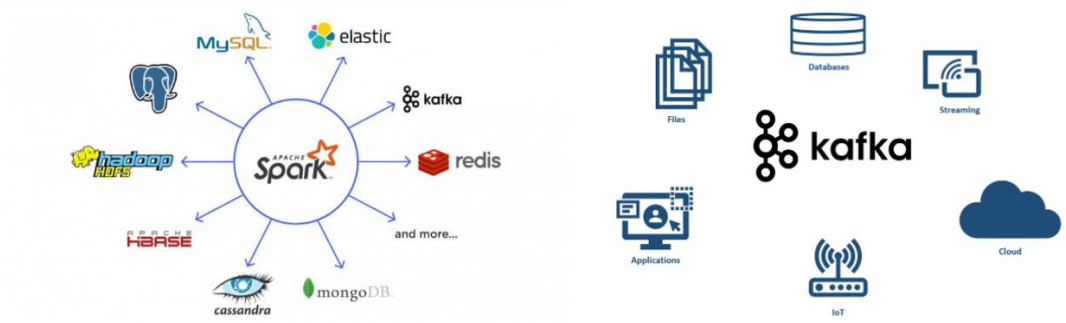
2.4 Συζήτηση για Νέα Αρχιτεκτονικά Παραδείγματα

Η ανάπτυξη νέων αρχιτεκτονικών παραδειγμάτων στον τομέα των ETL διαδικασιών έχει επιτρέψει την επίλυση πολλών από τα προβλήματα που συνδέονται με τις παραδοσιακές μεθοδολογίες. Τα παραδείγματα αυτά περιλαμβάνουν την αρχιτεκτονική βασισμένη σε συμβάντα (event-driven architecture), την υιοθέτηση υπηρεσιών RESTful και τη χρήση κατανεμημένων συστημάτων για την επεξεργασία δεδομένων. Το RESTful είναι ένας όρος που χρησιμοποιείται για να περιγράψει τις υπηρεσίες web που ακολουθούν την αρχιτεκτονική αρχή του REST (Representational State Transfer). Το REST είναι ένα αρχιτεκτονικό στυλ για το σχεδιασμό δικτυακών εφαρμογών, το οποίο χρησιμοποιεί το πρωτόκολλο HTTP και βασίζεται σε μια σειρά αρχών και περιορισμών. Ο όρος RESTful δηλώνει ότι η υπηρεσία που περιγράφεται ακολουθεί αυτές τις αρχές

Η αρχιτεκτονική βασισμένη σε συμβάντα αποτελεί ένα παράδειγμα που έχει προσελκύσει μεγάλο ενδιαφέρον τα τελευταία χρόνια. Σε αυτή την αρχιτεκτονική, τα δεδομένα υποβάλλονται σε επεξεργασία καθώς παράγονται ή διατίθενται, επιτρέποντας την επεξεργασία σε πραγματικό χρόνο. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε περιβάλλοντα όπου οι επιχειρήσεις πρέπει να αντιδρούν άμεσα σε αλλαγές, όπως στην ανίχνευση κυβερνοαπειλών.

Η υιοθέτηση υπηρεσιών (Periyasamy, 2024)για την επικοινωνία μεταξύ των στοιχείων του συστήματος είναι ένα άλλο σημαντικό αρχιτεκτονικό παράδειγμα. Οι RESTful υπηρεσίες επιτρέπουν την ευέλικτη και επεκτάσιμη επικοινωνία μεταξύ των συστημάτων, διευκολύνοντας την ενσωμάτωση διαφορετικών εργαλείων και τεχνολογιών στις διαδικασίες ETL.

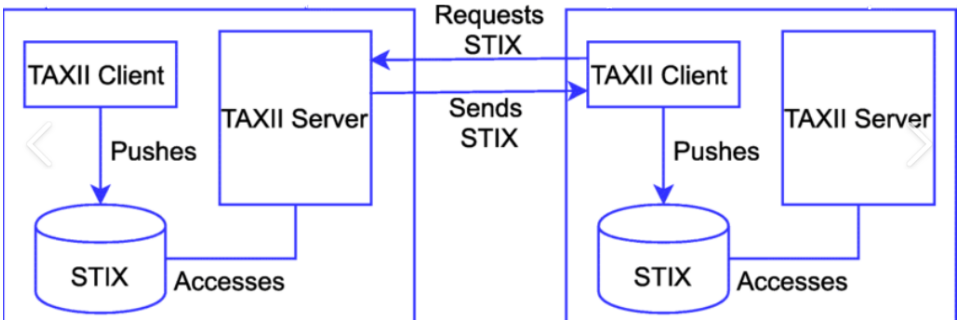
Τέλος, τα κατανεμημένα συστήματα επιτρέπουν την επεξεργασία μεγάλων όγκων δεδομένων σε κλίμακα, βελτιώνοντας την αποδοτικότητα και την απόδοση των διαδικασιών ETL. Οι πλατφόρμες όπως το Apache Hadoop και το Apache Spark επιτρέπουν την επεξεργασία και την ανάλυση δεδομένων σε τεράστια κλίμακα, παρέχοντας εργαλεία για τη διαχείριση πολύπλοκων ETL ροών εργασίας.



Εικόνα 6. Ενσωμάτωση και Συνεργασία Apache Spark και Apache Kafka με Διάφορες Τεχνολογίες και Πηγές Δεδομένων

2.5 Ενσωμάτωση του STIX/TAXII για Κοινή Χρήση Πληροφοριών Απειλών

Η ενσωμάτωση των προτύπων STIX (Structured Threat Information Expression) και TAXII (Trusted Automated Exchange of Indicator Information) στις διαδικασίες ETL είναι κρίσιμη για την αποτελεσματική ανταλλαγή πληροφοριών απειλών στον τομέα της κυβερνοασφάλειας [18,19,20]. Το STIX είναι ένα πρότυπο που χρησιμοποιείται για την αναπαράσταση πληροφοριών σχετικά με απειλές, όπως δείκτες συμβάντων, προφίλ επιτιθέμενων και λεπτομέρειες επιθέσεων [19,20]. Το TAXII είναι το πρωτόκολλο μεταφοράς που χρησιμοποιείται για την ασφαλή και αξιόπιστη ανταλλαγή αυτών των πληροφοριών μεταξύ οργανισμών. [18,19,20]



Εικόνα 7.Διαδικασία Ανταλλαγής Πληροφοριών για Κυβερνοαπειλές μέσω TAXII και STIX

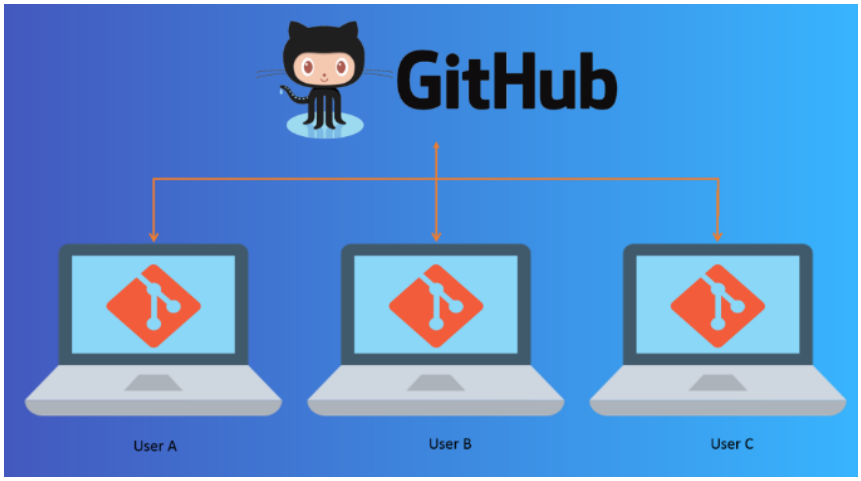
Η ενσωμάτωση αυτών των προτύπων σε ένα σύστημα ETL επιτρέπει την αυτοματοποιημένη συλλογή, επεξεργασία και ανταλλαγή πληροφοριών απειλών, βελτιώνοντας έτσι την αποτελεσματικότητα της απόκρισης στις απειλές.[19,20] Οι οργανισμοί μπορούν να ενσωματώσουν πληροφορίες από διάφορες πηγές, όπως εθνικά κέντρα ασφάλειας και διεθνείς οργανισμούς, και να τις χρησιμοποιούν για την ενίσχυση των μέτρων ασφαλείας τους.[19]

2.6 Συνεργατική Ανάπτυξη σε Έργα ETL

Η συνεργατική ανάπτυξη είναι απαραίτητη για την επιτυχή υλοποίηση έργων ETL, ειδικά σε περιβάλλοντα όπου πολλές ομάδες και οργανισμοί πρέπει να συνεργαστούν για την ανάπτυξη και τη συντήρηση των διαδικασιών. Τα εργαλεία όπως το Git και το GitHub επιτρέπουν τη διαχείριση του κώδικα και των αλλαγών με τρόπο που διευκολύνει τη συνεργασία και την παρακολούθηση της προόδου. [21,22,23]

Η χρήση αυτών των εργαλείων επιτρέπει στις ομάδες να εργάζονται ταυτόχρονα σε διαφορετικά μέρη ενός έργου, ενώ οι διαδικασίες όπως τα pull requests και οι έλεγχοι κώδικα

συμβάλλουν στη διασφάλιση της ποιότητας και της συνέπειας του κώδικα. Η υιοθέτηση αυτών των πρακτικών είναι ιδιαίτερα σημαντική σε έργα που αφορούν τη κυβερνοασφάλεια, καθώς εξασφαλίζει ότι οι διαδικασίες ETL είναι αξιόπιστες και ασφαλείς. [21,22,23]



Εικόνα 8. Συνεργατική Ανάπτυξη Κώδικα με το GitHub: Διαχείριση και Συγχρονισμός Εργασιών Χρηστών

3. Μεθοδολογία

3.1 Περιγραφή της Μεθοδολογίας

Η μεθοδολογία που ακολουθήθηκε σε αυτή τη διπλωματική εργασία έχει σχεδιαστεί με γνώμονα την επίτευξη μιας αποτελεσματικής και συνεργατικής προσέγγισης για την επεξεργασία και διαχείριση δεδομένων κυβερνοασφάλειας. Η υλοποίηση αυτής της μεθοδολογίας αξιοποιεί τις τεχνολογίες ETL (Extract, Transform, Load) και τα πρότυπα STIX/TAXII, τα οποία είναι βασικά για την ασφαλή ανταλλαγή πληροφοριών απειλών μεταξύ συστημάτων και οργανισμών. Επιπλέον, το Git χρησιμοποιήθηκε ως εργαλείο ελέγχου εκδόσεων για τη διαχείριση του κώδικα και την προώθηση της συνεργασίας, ενώ το Power BI για την αποτελεσματική οπτικοποίηση των δεδομένων, διευκολύνοντας την ανάλυση και την εξαγωγή συμπερασμάτων από τα αποτελέσματα της επεξεργασίας.

3.1.1 Στόχοι της Μεθοδολογίας

Ο πρωταρχικός στόχος της μεθοδολογίας είναι η ανάπτυξη ενός ολοκληρωμένου συστήματος που να μπορεί να διαχειρίζεται αποτελεσματικά δεδομένα που αφορούν το επίπεδο κυβερνοασφάλειααπό τη φάση της εξαγωγής τους, μέσω της μετατροπής και εμπλουτισμού τους, έως και την τελική αξιοποίησή τους για αναλύσεις και αναφορές. Τα δεδομένα αυτά σχετίζονται με συμβάντα κυβερνοασφάλειας που κατατάσσονται σε διαφορετικές κατηγορίες ανάλογα με τη σοβαρότητά τους (χαμηλή, μεσαία, υψηλή). Τα δεδομένα αποθηκεύονται αρχικά σε μια βάση δεδομένων Microsoft SQL, από όπου και υποβάλλονται σε επεξεργασία και ανάλυση. Η μεθοδολογία χωρίζεται σε τρία κύρια στάδια: την αρχικοποίηση και επεξεργασία δεδομένων, την ενσωμάτωση τεχνολογιών, και την παρακολούθηση και βελτιστοποίηση της διαδικασίας μέσω εργαλείων συνεργασίας.

Το σύστημα που αναπτύχθηκε ακολουθεί μια πολυεπίπεδη αρχιτεκτονική που διασφαλίζει την αποδοτικότητα και την ευελιξία στη διαχείριση των δεδομένων. Η προσέγγιση αυτή επιτρέπει την ευκολία προσαρμογής σε μελλοντικές ανάγκες και επεκτάσεις, καθώς και τη διασφάλιση της ακρίβειας και της ασφάλειας στη διαχείριση των πληροφοριών απειλών.

3.1.2 Διαδικασία Αρχικοποίησης και Επεξεργασίας Δεδομένων

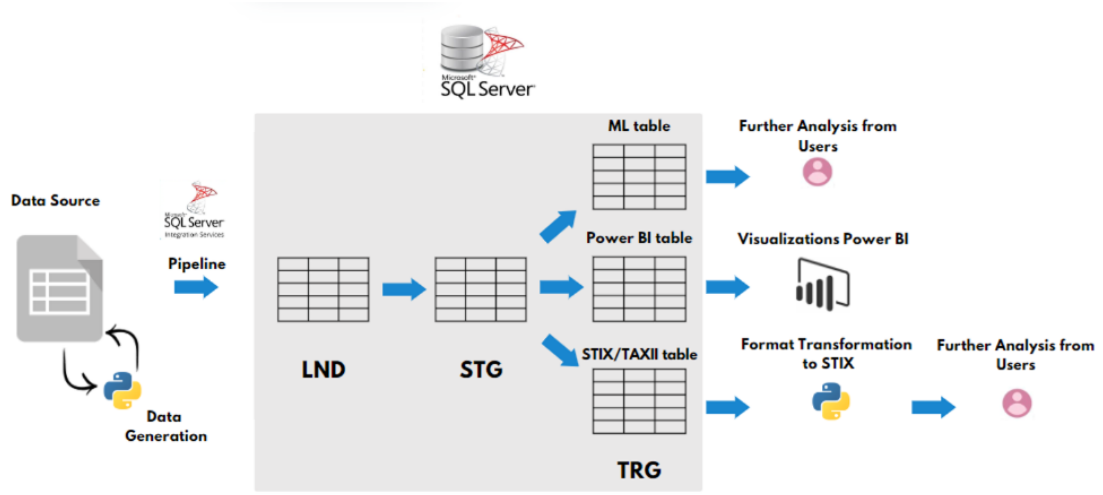
Το πρώτο στάδιο της μεθοδολογίας περιλαμβάνει τη δημιουργία και την αρχικοποίηση των συνθετικών δεδομένων, τα οποία χρησιμοποιούνται για τη δοκιμή και αξιολόγηση των διαδικασιών ETL. Τα δεδομένα αυτά αρχικά αποθηκεύονται σε ένα αρχείο Excel, το οποίο περιλαμβάνει 25 στήλες και 40.000 γραμμές. Κάθε γραμμή του αρχείου αντιπροσωπεύει ένα συμβάν κυβερνοασφάλειας, με πληροφορίες όπως IP διευθύνσεις, timestamps, severity levels, και τύπους επιθέσεων.

Αυτή η προσέγγιση επιτρέπει την εύκολη διαχείριση μεγάλων όγκων δεδομένων και την προετοιμασία τους για εισαγωγή στη βάση δεδομένων. Η χρήση του Excel για την αρχική αποθήκευση των δεδομένων παρέχει μια απλή και ευέλικτη μέθοδο για τη δημιουργία και την επεξεργασία των συνθετικών δεδομένων, προτού αυτά μεταφερθούν σε πιο προηγμένα συστήματα διαχείρισης βάσεων δεδομένων.

Η εισαγωγή των δεδομένων στη βάση δεδομένων της Microsoft SQL πραγματοποιείται μέσω του εργαλείου SSIS (SQL Server Integration Services), το οποίο επιτρέπει την αυτοματοποίηση της διαδικασίας εισαγωγής δεδομένων από το Excel στο SQL. Αυτή η διαδικασία εξασφαλίζει την αποδοτική και επεκτάσιμη διαχείριση των δεδομένων, μειώνοντας τον χρόνο και τις απαιτήσεις πόρων που απαιτούνται για τη μεταφορά και την αποθήκευση των δεδομένων.

Η βάση δεδομένων έχει σχεδιαστεί με τρία διακριτά επίπεδα (layers): το Landing schema, το Staging schema και το Target schema.

- *Landing schema*: Σε αυτό το επίπεδο, τα δεδομένα εισάγονται στην ακατέργαστη μορφή τους (raw), όπως έρχονται από το Excel. Αυτό το επίπεδο χρησιμεύει ως το αρχικό σημείο αποθήκευσης των δεδομένων, εξασφαλίζοντας ότι τα αρχικά δεδομένα παραμένουν διαθέσιμα για πιθανή μελλοντική χρήση ή αναφορά.
- *Staging schema*: Το δεύτερο επίπεδο είναι το σημείο όπου τα δεδομένα υποβάλλονται σε διαδικασίες καθαρισμού και προετοιμασίας. Αυτή η φάση περιλαμβάνει τον καθαρισμό των δεδομένων από λάθη, την αφαίρεση διπλότυπων εγγραφών και την προετοιμασία τους για περαιτέρω ανάλυση. Το Staging schema λειτουργεί ως μια ενδιάμεση βάση, όπου τα δεδομένα είναι πλέον έτοιμα να μεταφερθούν στο τελικό στάδιο.
- *Target schema*: Στο τρίτο και τελευταίο επίπεδο, τα δεδομένα αποθηκεύονται σε τρεις ξεχωριστούς πίνακες,ο καθένας προορισμένος για διαφορετική χρήση. Ο πρώτος πίνακας είναι σχεδιασμένος για χρήση από αναλυτές δεδομένων, όπου τα δεδομένα μπορούν να αναλυθούν και να χρησιμοποιηθούν για μηχανική μάθηση (ML). Ο δεύτερος πίνακας προορίζεται για αναλυτές, οι οποίοι το χρησιμοποιούν για τη δημιουργία αναφορών και την ενημέρωση του Power BI. Ο τρίτος πίνακας χρησιμοποιείται από ένα Python script, το οποίο μετατρέπει τα δεδομένα σε μορφή STIX/TAXII, επιτρέποντας την ασφαλή και αποτελεσματική ανταλλαγή πληροφοριών με τα ενδιαφερόμενα μέρη.



Εικόνα 9.Αρχιτεκτονική Διαδικασίας ETL για Χρήση σε ML, Power BI και STIX/TAXII

Η διαίρεση σε αυτά τα τρία επίπεδα διασφαλίζει ότι κάθε στάδιο της επεξεργασίας δεδομένων είναι σαφώς οριοθετημένο και μπορεί να ελεγχθεί και να βελτιωθεί ανεξάρτητα. Επιπλέον, η χρήση των διαφορετικών schemas επιτρέπει την εύκολη προσαρμογή των δεδομένων ανάλογα με τις ανάγκες των χρηστών και τις απαιτήσεις των εφαρμογών.

3.1.3 Ενσωμάτωση Σύγχρονων Τεχνολογιών

Στο δεύτερο στάδιο της μεθοδολογίας, ενσωματώνονται τα πρότυπα STIX και TAXII, τα οποία είναι κρίσιμα για τη δομημένη ανταλλαγή πληροφοριών απειλών. Αυτή η ενσωμάτωση πραγματοποιείται μέσω ενός Python script που διαβάζει τα δεδομένα από το Target schema και τα μετατρέπει στη σωστή μορφή για να διανεμηθούν στα ενδιαφερόμενα μέρη.

Η ενσωμάτωση των STIX/TAXII επιτρέπει την αυτοματοποίηση της διαδικασίας μετατροπής των δεδομένων, μειώνοντας έτσι τον κίνδυνο ανθρώπινων λαθών και αυξάνοντας την αποδοτικότητα. Τα πρότυπα STIX (Structured Threat Information Expression) και TAXII (Trusted Automated Exchange of Indicator Information) έχουν σχεδιαστεί για να παρέχουν μια κοινή γλώσσα και ένα πλαίσιο για την περιγραφή, την ανταλλαγή, και την ανάλυση πληροφοριών απειλών.

Η χρήση αυτών των προτύπων επιτρέπει τη διαλειτουργικότητα μεταξύ διαφορετικών συστημάτων ασφαλείας και οργανισμών, βελτιώνοντας την ικανότητα των οργανισμών να ανταποκρίνονται σε απειλές με γρήγορο και συντονισμένο τρόπο. Η διαδικασία αυτή υποστηρίζει τη δομημένη ανταλλαγή πληροφοριών, διασφαλίζοντας ότι τα δεδομένα είναι τυποποιημένα και μπορούν να χρησιμοποιηθούν από οποιοδήποτε σύστημα ή εργαλείο που υποστηρίζει τα πρότυπα STIX/TAXII.

3.1.4 Παρακολούθηση και Βελτιστοποίηση

Το τρίτο στάδιο της μεθοδολογίας περιλαμβάνει τη χρήση εργαλείων όπως το Git και το GitHub για την παρακολούθηση των αλλαγών και τη διασφάλιση της ποιότητας της ανάπτυξης. Αυτά τα εργαλεία είναι απαραίτητα για την αποτελεσματική συνεργασία και τη διαχείριση των εκδόσεων του κώδικα, ειδικά σε περιβάλλοντα όπου πολλές ομάδες ανάπτυξης εργάζονται σε διαφορετικά μέρη του ίδιου έργου.

Το Git είναι ένα διανεμημένο σύστημα ελέγχου εκδόσεων που επιτρέπει στους προγραμματιστές να καταγράφουν τις αλλαγές στον κώδικά τους, να διατηρούν ιστορικό των τροποποιήσεων, και να επιστρέφουν σε προηγούμενες εκδόσεις όταν απαιτείται. Ένα από τα κύρια πλεονεκτήματα του Git είναι η δυνατότητα για παράλληλη ανάπτυξη μέσω των "branches". Κάθε branch μπορεί να θεωρηθεί ως μια ξεχωριστή έκδοση του κώδικα, όπου οι προγραμματιστές μπορούν να αναπτύξουν νέες δυνατότητες ή να διορθώσουν σφάλματα, χωρίς να επηρεάζουν την κύρια έκδοση (main branch).

Η διαδικασία συγχώνευσης (merge) επιτρέπει την ενσωμάτωση αλλαγών από διαφορετικά branches στην κύρια έκδοση του έργου, διασφαλίζοντας ότι όλες οι τροποποιήσεις ενοποιούνται με τάξη. Επιπλέον, το Git επιτρέπει την αναγνώριση των συγγραφέων των αλλαγών και την παρακολούθηση της εξέλιξης του κώδικα μέσω λεπτομερών "commits".

Το GitHub, από την άλλη πλευρά, είναι μια διαδικτυακή πλατφόρμα που φιλοξενεί αποθετήρια Git και επιτρέπει την απομακρυσμένη συνεργασία. Μέσω του GitHub, οι προγραμματιστές μπορούν να συνεργάζονται απομακρυσμένα, να αναθεωρούν και να συζητούν τον κώδικα μέσω "pull requests" (PRs). Τα pull requests επιτρέπουν σε άλλους προγραμματιστές να αναθεωρήσουν τον κώδικα, να συζητήσουν πιθανές βελτιώσεις και τελικά να εγκρίνουν ή να απορρίψουν τις αλλαγές.

Αυτή η διαδικασία αναθεώρησης είναι κρίσιμη για τη διασφάλιση της ποιότητας του κώδικα, καθώς επιτρέπει την αναγνώριση και την επίλυση προβλημάτων πριν ενσωματωθούν στον κύριο κώδικα. Στο πλαίσιο αυτής της διπλωματικής εργασίας, η χρήση του Git και του GitHub διασφαλίζει ότι κάθε συνεισφορά στον κώδικα είναι τεκμηριωμένη, ελεγχόμενη και επανεξετασμένη από άλλους συντελεστές πριν γίνει μέρος του τελικού προϊόντος.

Η διαδικασία αυτή υλοποιείται μέσω του Visual Studio Code, το οποίο χρησιμοποιείται για τη συγγραφή και την παρακολούθηση του κώδικα. Το Visual Studio Code είναι ένας ισχυρός επεξεργαστής κώδικα που προσφέρει υποστήριξη για πολλαπλές γλώσσες προγραμματισμού και ενσωματώνεται άμεσα με το Git και το GitHub, διευκολύνοντας έτσι τη διαχείριση του κώδικα και τη συνεργασία μεταξύ των μελών της ομάδας.

Συνολικά, η μεθοδολογία αυτή επιτρέπει την ανάπτυξη μιας ευέλικτης και επεκτάσιμης λύσης για την επεξεργασία δεδομένων κυβερνοασφάλειας, διασφαλίζοντας ότι τα δεδομένα μπορούν να διαχειριστούν, να επεξεργαστούν και να διανεμηθούν αποτελεσματικά, ενώ παράλληλα παρέχεται η δυνατότητα για συνεχή παρακολούθηση και βελτίωση των διαδικασιών.

3.2 Δημιουργία Συνθετικών Δεδομένων για Δοκιμές Διαδικασιών ETL

Η διαδικασία δημιουργίας συνθετικών δεδομένων αποτελεί μια κρίσιμη φάση στην υλοποίηση αυτής της διπλωματικής εργασίας, καθώς επιτρέπει την προσομοίωση πραγματικών συμβάντων κυβερνοασφάλειας μέσω της παραγωγής νέων, τυχαία δημιουργημένων δεδομένων με βάση πάντα τα αρχικά δεδομένα. Η προσέγγιση αυτή εξασφαλίζει ότι το σύστημα που αναπτύσσεται μπορεί να δοκιμαστεί και να αξιολογηθεί σε ένα περιβάλλον που προσομοιώνει πραγματικές συνθήκες. Η ανάπτυξη και υλοποίηση αυτής η εφαρμογής έγινε σε γλώσσα python και ο κώδικας είναι προσβάσιμος στο αποθετήριο [24]

3.2.1 Στόχοι της Δημιουργίας Συνθετικών Δεδομένων

Η κύρια επιδίωξη της δημιουργίας συνθετικών δεδομένων είναι η διασφάλιση ότι οι διαδικασίες ETL μπορούν να διαχειριστούν με επιτυχία διαφορετικά σενάρια και περιπτώσεις επίθεσης. Τα συνθετικά δεδομένα επιτρέπουν την προσομοίωση μιας ευρείας ποικιλίας απειλών και επιθέσεων, διασφαλίζοντας ότι το σύστημα μπορεί να ανταποκριθεί με ακρίβεια και αποτελεσματικότητα σε αυτές τις προκλήσεις.

Η χρήση συνθετικών δεδομένων είναι ιδιαίτερα σημαντική για την προσομοίωση σπάνιων ή ακραίων σεναρίων που μπορεί να μην περιλαμβάνονται στα πραγματικά δεδομένα. Επιπλέον, τα συνθετικά δεδομένα παρέχουν τη δυνατότητα για επαναλαμβανόμενες δοκιμές, επιτρέποντας την αξιολόγηση της ανθεκτικότητας και της ακρίβειας των διαδικασιών σε διαφορετικά περιβάλλοντα και συνθήκες.

3.2.2 Διαδικασία Δημιουργίας Συνθετικών Δεδομένων

Η διαδικασία δημιουργίας συνθετικών δεδομένων ξεκινά με την ανάγνωση των απαραίτητων πληροφοριών από αρχεία κειμένου. Αυτές οι πληροφορίες περιλαμβάνουν τις τοποθεσίες γεωγραφικών δεδομένων, τις πληροφορίες συσκευών, και τις πληροφορίες χρηστών. Αυτές οι πληροφορίες χρησιμοποιούνται αργότερα για την παραγωγή τυχαίων δεδομένων που θα προσομοιώσουν διάφορα σενάρια κυβερνοασφάλειας.

Η ανάγνωση των δεδομένων από τα αρχεία κειμένου διασφαλίζει ότι τα συνθετικά δεδομένα βασίζονται σε ρεαλιστικές τιμές, καθιστώντας τα κατάλληλα για τις ανάγκες δοκιμών. Αυτή η φάση είναι σημαντική για τη διασφάλιση της ποιότητας και της συνέπειας των δεδομένων που παράγονται, καθώς τα δεδομένα πρέπει να αντικατοπτρίζουν πραγματικές συνθήκες και να είναι χρήσιμα για την αξιολόγηση των διαδικασιών ETL.

Στη συνέχεια, η διαδικασία διαβάζει τα υπάρχοντα δεδομένα από ένα αρχείο CSV, το οποίο περιέχει ήδη καταγεγραμμένα συμβάντα κυβερνοασφάλειας. Αυτά τα δεδομένα αποτελούν τη βάση για τη δημιουργία νέων δεδομένων, διατηρώντας τη συνοχή και τη δομή των αρχικών καταγραφών. Με την ανάγνωση των δεδομένων από το αρχείο CSV, επιτυγχάνεται η διατήρηση της δομής και της μορφής των δεδομένων, κάτι που είναι κρίσιμο για τη συνέχεια της διαδικασίας ETL.

Η δημιουργία νέων συνθετικών δεδομένων πραγματοποιείται μέσω μιας σειράς τυχαίων επιλογών και κατανομών, οι οποίες προσθέτουν ποικιλία και ρεαλισμό στα δεδομένα. Κάθε νέο δεδομένο περιλαμβάνει πληροφορίες όπως τυχαία δημιουργημένες διευθύνσεις IP, τυχαίες πόρτες δικτύου, είδος κυκλοφορίας (HTTP, DNS, FTP), και τύπους επιθέσεων (DDoS, Intrusion,

Malware). Οι κατανομές πιθανότητας που χρησιμοποιούνται για την επιλογή του πρωτοκόλλου, του τύπου επίθεσης και άλλων παραμέτρων εξασφαλίζουν ότι τα δεδομένα αντικατοπτρίζουν πιθανά σενάρια που μπορεί να συμβούν σε ένα περιβάλλον κυβερνοασφάλειας.

Η διαδικασία περιλαμβάνει επίσης τη δημιουργία τυχαίων timestamps για κάθε νέο δεδομένο. Αυτά τα timestamps δημιουργούνται μέσα στην τρέχουσα ημερομηνία, εξασφαλίζοντας ότι τα δεδομένα είναι σύγχρονα και σχετίζονται με τον τρέχοντα χρόνο. Αυτό είναι σημαντικό, καθώς επιτρέπει τη δημιουργία ενός σεναρίου που προσομοιώνει τις καθημερινές δραστηριότητες και επιθέσεις που μπορούν να συμβούν σε ένα δίκτυο.

Ένας σημαντικός παράγοντας στη διαδικασία δημιουργίας των δεδομένων είναι η προσαρμογή των επιπέδων σοβαρότητας των επιθέσεων (severity) και των ενεργειών που λαμβάνονται (action taken). Για παράδειγμα, τα δεδομένα με υψηλή σοβαρότητα (High severity) είναι πιο πιθανό να σχετίζονται με σημαντικές ενέργειες όπως η απόρριψη ή η καταγραφή της επίθεσης, και συνήθως συνδέονται με υψηλότερες βαθμολογίες ανωμαλιών (anomaly scores). Αυτές οι διαβαθμίσεις επιτρέπουν την προσομοίωση διαφόρων επιπέδων απειλών και αντιδράσεων μέσα στο σύστημα, δημιουργώντας ένα πιο ρεαλιστικό σύνολο δεδομένων για τις δοκιμές.

3.2.3 Δημιουργία Νέου Αρχείου Δεδομένων

Αφού δημιουργηθούν τα νέα συνθετικά δεδομένα, αποθηκεύονται σε ένα νέο αρχείο CSV. Κάθε αρχείο CSV φέρει ένα μοναδικό όνομα που αντιστοιχεί στην ημερομηνία δημιουργίας του, διασφαλίζοντας έτσι την οργανωμένη αποθήκευση και την εύκολη ανάκληση των δεδομένων για μελλοντική χρήση.

Αυτό το αρχείο μπορεί να χρησιμοποιηθεί για την τροφοδότηση του συστήματος ETL και την αξιολόγηση της αποτελεσματικότητας και της ακρίβειας των διαδικασιών επεξεργασίας δεδομένων. Η οργανωμένη αποθήκευση των δεδομένων επιτρέπει την εύκολη αναφορά και χρήση των δεδομένων για μελλοντικές δοκιμές και αξιολογήσεις, διασφαλίζοντας ότι το σύστημα μπορεί να αναπαράγει και να αξιολογεί τα αποτελέσματα με συνέπεια και ακρίβεια.

Η δημιουργία συνθετικών δεδομένων με αυτήν τη μέθοδο επιτρέπει την προσομοίωση μιας μεγάλης ποικιλίας σεναρίων κυβερνοασφάλειας, συμβάλλοντας στην αξιολόγηση και τη βελτιστοποίηση των διαδικασιών ETL που αναπτύσσονται στο πλαίσιο αυτής της διπλωματικής εργασίας. Τα δεδομένα αυτά είναι απαραίτητα για τη δοκιμή της ανθεκτικότητας του συστήματος και την επαλήθευση ότι οι διαδικασίες που έχουν αναπτυχθεί μπορούν να ανταποκριθούν σε πραγματικές συνθήκες και απαιτήσεις.

3.3 Υλοποίηση ETL με SQL Layers και Stored Procedures

Η υλοποίηση διαδικασιών ETL (Extract, Transform, Load) είναι μία από τις κρίσιμες φάσεις σε κάθε έργο ανάλυσης δεδομένων, καθώς αφορά την εξαγωγή δεδομένων από διάφορες πηγές, τη μετατροπή τους σε μια πιο χρήσιμη μορφή και τελικά τη φόρτωσή τους σε μια βάση δεδομένων ή σε άλλο αποθηκευτικό χώρο για περαιτέρω επεξεργασία και ανάλυση. Στο πλαίσιο αυτής της εργασίας, η διαδικασία ETL έχει υλοποιηθεί χρησιμοποιώντας SQL layers και stored procedures για την αποτελεσματική διαχείριση των δεδομένων σε ένα σύστημα Microsoft SQL Server.

3.3.1 Σχεδιασμός και Υλοποίηση της Διαδικασίας ETL

Η διαδικασία ETL στο έργο αυτό έχει σχεδιαστεί ώστε να αξιοποιεί πλήρως τη δομή της βάσης δεδομένων, χρησιμοποιώντας τρία επίπεδα (schemas) για την αποθήκευση και μετασχηματισμό των δεδομένων*: Landing, Staging*, και *Target*. Κάθε ένα από αυτά τα επίπεδα έχει συγκεκριμένο ρόλο στη διαχείριση των δεδομένων.

- *Landing Layer (LND):* Σε αυτό το επίπεδο, τα δεδομένα αποθηκεύονται όπως έχουν ληφθεί από την πηγή τους, χωρίς καμία μετατροπή ή καθαρισμό. Αυτή η αρχική αποθήκευση διασφαλίζει ότι υπάρχει πάντα ένα σημείο αναφοράς για τα ακατέργαστα δεδομένα.
- *Staging Layer (STG):* Στο στάδιο αυτό, πραγματοποιείται ο πρώτος καθαρισμός και μετασχηματισμός των δεδομένων. Εδώ, τα δεδομένα είναι έτοιμα για περαιτέρω επεξεργασία και διασφάλιση ποιότητας πριν φορτωθούν στο τελικό επίπεδο.
- *Target Layer (TRG):* Το τελευταίο στάδιο περιλαμβάνει τη φόρτωση των δεδομένων στο τελικό τους προορισμό. Τα δεδομένα σε αυτό το επίπεδο είναι πλήρως καθαρισμένα και μετασχηματισμένα, έτοιμα για χρήση από τους τελικούς χρήστες και τις εφαρμογές, όπως οι αναλύσεις μηχανικής μάθησης ή οι αναφορές επιχειρησιακής ευφυΐας (Business Intelligence).

3.3.2 Υλοποίηση της Διαδικασίας Μετασχηματισμού

Η διαδικασία μετασχηματισμού των δεδομένων υλοποιείται μέσω stored procedures, οι οποίες εκτελούν συγκεκριμένες εργασίες στο πλαίσιο της ETL. Οι κύριες διαδικασίες είναι οι ακόλουθες:

1. *[STG].[LND_TO_STG]:* Αυτή η stored procedure (LND_TO_STG.sql) είναι υπεύθυνη για τη μεταφορά και το μετασχηματισμό των δεδομένων από το Landing Layer στο Staging Layer. Η διαδικασία χρησιμοποιεί δύο Common Table Expressions (CTEs) για να καθορίσει ποια δεδομένα πρέπει να μεταφερθούν. Ο πρώτος CTE, initial_cte, ελέγχει και προσθέτει μια ημερομηνία δημιουργίας (create_date) στα δεδομένα, ενώ ο δεύτερος, filter_new_data_cte, φιλτράρει τα δεδομένα ώστε να περιλαμβάνει μόνο νέα ή ενημερωμένα αρχεία που δεν έχουν ήδη μεταφερθεί στο Staging Layer. Τέλος, τα φιλτραρισμένα δεδομένα εισάγονται στο [STG].[dataset01].
2. *[TRG].[STG_TO_TRG_ML]:* Αυτή η stored procedure (STG_TO_TRG_ML.sql) αναλαμβάνει να μεταφέρει τα δεδομένα από το Staging Layer στο Target Layer που προορίζεται για αναλύσεις μηχανικής μάθησης (Machine Learning). Η διαδικασία χρησιμοποιεί παρόμοια προσέγγιση με την προηγούμενη, φιλτράροντας τα νέα δεδομένα και μετασχηματίζοντάς τα ώστε να συμπεριλαμβάνουν πρόσθετες στήλες που είναι απαραίτητες για την ανάλυση, όπως οι κατηγοριοποιημένες και κωδικοποιημένες στήλες (one-hot encoded columns). Αυτές οι πρόσθετες στήλες επιτρέπουν την προετοιμασία των δεδομένων για πιο πολύπλοκες αναλύσεις από τα μοντέλα μηχανικής μάθησης.
3. *[TRG].[STG_TO_TRG_PBI]:* Αυτή η stored procedure (STG_TO_TRG_PBI.sql) εστιάζει στην προετοιμασία των δεδομένων για αναφορές Power BI. Τα δεδομένα μεταφέρονται από το Staging Layer στο Target Layer που προορίζεται για αναλύσεις επιχειρησιακής ευφυΐας (Business Intelligence). Η διαδικασία περιλαμβάνει επίσης τη φιλτράρισμα των νέων δεδομένων και την κατηγοριοποίηση γεωγραφικών και άλλων δεδομένων για καλύτερη οπτικοποίηση στις αναφορές.

Η χρήση των stored procedures επιτρέπει την αυτοματοποίηση των διαδικασιών μετασχηματισμού και τη βελτίωση της αποδοτικότητας. Οι stored procedures εξασφαλίζουν ότι οι διαδικασίες μπορούν να επαναληφθούν με συνέπεια και ακρίβεια, ενώ παράλληλα επιτρέπουν την προσαρμογή των διαδικασιών ανάλογα με τις απαιτήσεις των δεδομένων και των χρηστών.

3.3.3 Τεχνικές Λεπτομέρειες και Βέλτιστες Πρακτικές

Στην υλοποίηση αυτών των διαδικασιών ETL, ακολουθούνται κάποιες βέλτιστες πρακτικές που εξασφαλίζουν την αποδοτικότητα και την ακρίβεια:

1. Χρήση Common Table Expressions (CTEs): Οι CTEs επιτρέπουν την εύκολη και κατανοητή διαχείριση των δεδομένων εντός των stored procedures, διευκολύνοντας την ανάγνωση και τη συντήρηση του κώδικα.
2. Καθορισμός create_date: Η χρήση της στήλης create_date διασφαλίζει ότι μόνο νέα ή ενημερωμένα δεδομένα μεταφέρονται στα επόμενα στάδια της διαδικασίας ETL, μειώνοντας τον όγκο των δεδομένων που επεξεργάζονται κάθε φορά και βελτιώνοντας την απόδοση.
3. Κωδικοποίηση Δεδομένων (One-Hot Encoding): Για τις ανάγκες της μηχανικής μάθησης, η διαδικασία περιλαμβάνει την κωδικοποίηση κατηγοριών σε δυαδικές μορφές (1/0), επιτρέποντας την ευκολότερη χρήση τους από τα μοντέλα μηχανικής μάθησης.
4. Γεωγραφική Κατηγοριοποίηση: Για τις ανάγκες των αναλύσεων Power BI, τα δεδομένα κατηγοριοποιούνται γεωγραφικά, κάτι που διευκολύνει την οπτικοποίηση και την κατανόηση των αναφορών.

Η εφαρμογή αυτών των βέλτιστων πρακτικών διασφαλίζει ότι οι διαδικασίες ETL είναι αποδοτικές, ακριβείς, και εύκολα διαχειρίσιμες. Επιπλέον, η χρήση των stored procedures επιτρέπει την εύκολη συντήρηση και την προσαρμογή των διαδικασιών ανάλογα με τις ανάγκες του έργου.

3.4 Ενσωμάτωση του STIX/TAXII για Κοινή Χρήση Πληροφοριών Απειλών

Η σύγχρονη διαχείριση και ανταλλαγή πληροφοριών απειλών αποτελεί θεμελιώδες στοιχείο για την αποτελεσματική άμυνα ενάντια σε κυβερνοαπειλές. Η χρήση προτύπων όπως το STIX (Structured Threat Information Expression) και το TAXII (Trusted Automated Exchange of Indicator Information) παρέχει έναν τυποποιημένο τρόπο για την έκφραση, αποθήκευση, και ανταλλαγή πληροφοριών απειλών σε μορφή που είναι κατανοητή τόσο από ανθρώπους όσο και από συστήματα.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, εξετάστηκε και υλοποιήθηκε η ενσωμάτωση αυτών των προτύπων για την αποτελεσματικότερη διαχείριση και κοινή χρήση πληροφοριών απειλών. Η ενσωμάτωση των STIX/TAXII επιτρέπει την αυτοματοποίηση της διαδικασίας μετατροπής των δεδομένων, μειώνοντας έτσι τον κίνδυνο ανθρώπινων λαθών και αυξάνοντας την αποδοτικότητα.

3.4.1 Δημιουργία STIX Objects

Το πρώτο βήμα για την ενσωμάτωση του STIX αφορά τη μετατροπή των υπαρχόντων δεδομένων απειλών σε STIX objects. Για το σκοπό αυτό, χρησιμοποιήθηκε ένα αρχείο CSV, το οποίο περιλαμβάνει δεδομένα από τη βάση δεδομένων, όπως πληροφορίες για διευθύνσεις IP, πρωτόκολλα δικτύου, τύπους απειλών, και άλλες σχετικές παραμέτρους. Το CSV αρχείο αυτό επεξεργάζεται μέσω ενός Python script (TRG_to_STIX.ipynb), το οποίο δημιουργεί τα κατάλληλα STIX Indicator και ObservedData objects.

Τα Indicator objects περιλαμβάνουν πληροφορίες σχετικά με απειλές, όπως οι κακόβουλες IP διευθύνσεις, και συνοδεύονται από χαρακτηριστικά όπως το pattern_type και το valid_from, που ορίζουν τη μορφή και τη χρονική ισχύ του δείκτη απειλής. Παράλληλα, τα ObservedData objects καταγράφουν τις παρατηρήσεις της δραστηριότητας δικτύου,

περιλαμβάνοντας τα δεδομένα που συλλέχθηκαν από τα συστήματα ανίχνευσης, όπως οι IP διευθύνσεις προέλευσης και προορισμού, τα ports, και τα χρησιμοποιούμενα πρωτόκολλα.

Αυτά τα δεδομένα αποθηκεύονται σε ένα JSON αρχείο STIX Bundle, το οποίο περιέχει όλες τις απαραίτητες πληροφορίες για την ανάλυση και τη διάδοση των απειλών. Η χρήση των STIX objects παρέχει έναν τυποποιημένο τρόπο για την περιγραφή των απειλών, επιτρέποντας την εύκολη ανταλλαγή και χρήση αυτών των δεδομένων από διάφορα συστήματα και οργανισμούς.

Αν και το τρέχον έργο επικεντρώνεται στη δημιουργία και αποθήκευση αυτών των STIX objects, η αποστολή τους μέσω ενός TAXII server δεν έχει ενσωματωθεί στη διαδικασία. Παρά ταύτα, η δυνατότητα αυτή υπάρχει και μπορεί να υλοποιηθεί σε μελλοντική εργασία, προσφέροντας επιπλέον λειτουργικότητα και ασφάλεια στη διαχείριση των πληροφοριών απειλών.

3.4.2 Προοπτική Ενσωμάτωσης TAXII

Η ενσωμάτωση του TAXII για την ασφαλή ανταλλαγή αυτών των πληροφοριών απειλών θα αποτελέσει το επόμενο φυσικό βήμα για τη βελτίωση της παρούσας λύσης. Το TAXII παρέχει ένα σύνολο από API που επιτρέπουν την αυτοματοποιημένη και ελεγχόμενη ανταλλαγή πληροφοριών απειλών μέσω δικτυακών υπηρεσιών.

Με αυτήν την προσέγγιση, το STIX Bundle που δημιουργήθηκε θα μπορούσε να σταλεί σε έναν TAXII server, ο οποίος θα διαχειρίζεται την αποθήκευση και την κοινή χρήση των πληροφοριών με εξουσιοδοτημένους χρήστες ή οργανισμούς. Η διασύνδεση αυτή θα επέτρεπε στους οργανισμούς να μοιράζονται έγκαιρα και με α (Standards, 2021)κρίβεια κρίσιμες πληροφορίες απειλών, ενισχύοντας τη συλλογική ασφάλεια.

Ο TAXII server θα λειτουργούσε ως μια κεντρική αποθήκη που δέχεται, διαχειρίζεται, και προωθεί τις πληροφορίες αυτές προς τους ενδιαφερόμενους, διασφαλίζοντας ότι όλοι οι συμμετέχοντες έχουν πρόσβαση στις πιο πρόσφατες και σχετικές πληροφορίες απειλών. Η υλοποίηση αυτής της δυνατότητας σε μελλοντική εργασία θα επέτρεπε μια αυτοματοποιημένη και ασφαλή διαδικασία ανταλλαγής πληροφοριών, η οποία θα μπορούσε να ενισχύσει περαιτέρω την ικανότητα ενός οργανισμού να αντιδράσει αποτελεσματικά σε νέες και εξελισσόμενες κυβερνοαπειλές.

3.5 Χρήση του Git και του GitHub για Συνεργατική Ανάπτυξη

Στη σύγχρονη ανάπτυξη λογισμικού, η συνεργατική ανάπτυξη κώδικα είναι ζωτικής σημασίας για την αποτελεσματικότητα και την ποιότητα των παραγόμενων έργων. Η χρήση του Git, ενός διανεμημένου συστήματος ελέγχου εκδόσεων, σε συνδυασμό με το GitHub, μια πλατφόρμα φιλοξενίας αποθετηρίων Git, επιτρέπει στους προγραμματιστές να συνεργάζονται αρμονικά, ανεξάρτητα από την τοποθεσία τους.

3.5.1 Βασικές Αρχές του Git

Το Git είναι ένα εργαλείο που επιτρέπει στους προγραμματιστές να καταγράφουν τις αλλαγές στον κώδικά τους, να διατηρούν ιστορικό των τροποποιήσεων και να επιστρέφουν σε προηγούμενες εκδόσεις όταν χρειάζεται. Ένα από τα κύρια πλεονεκτήματα του Git είναι η δυνατότητα για παράλληλη ανάπτυξη μέσω των branches. Κάθε branch μπορεί να θεωρηθεί ως μια ξεχωριστή έκδοση του κώδικα, όπου οι προγραμματιστές μπορούν να αναπτύξουν νέες δυνατότητες ή να διορθώσουν σφάλματα, χωρίς να επηρεάζουν την κύρια έκδοση (main branch).

Η διαδικασία συγχώνευσης (merge) επιτρέπει την ενσωμάτωση αλλαγών από διαφορετικά branches στην κύρια έκδοση του έργου, διασφαλίζοντας ότι όλες οι τροποποιήσεις ενοποιούνται με τάξη. Επιπλέον, το Git επιτρέπει την αναγνώριση των συγγραφέων των αλλαγών και την παρακολούθηση της εξέλιξης του κώδικα μέσω λεπτομερών commits.

3.5.2 Χρήση του GitHub για Απομακρυσμένη Συνεργασία

Το GitHub προσφέρει μια διαδικτυακή πλατφόρμα όπου αποθετήρια Git μπορούν να φιλοξενηθούν, επιτρέποντας στους προγραμματιστές να συνεργάζονται απομακρυσμένα. Ένα από τα βασικά χαρακτηριστικά του GitHub είναι η δυνατότητα για συνεργασία μέσω "pull requests" (PRs). Μέσω ενός pull request, ένας προγραμματιστής μπορεί να ζητήσει την ενσωμάτωση των αλλαγών του από το branch του στην κύρια έκδοση του έργου. Άλλοι συνεισφέροντες μπορούν να αναθεωρήσουν τον κώδικα, να συζητήσουν πιθανές βελτιώσεις και τελικά να εγκρίνουν ή να απορρίψουν το pull request.

Αυτός ο μηχανισμός ενισχύει τη συνεργατική ανάπτυξη, διασφαλίζοντας ότι όλες οι αλλαγές περνούν από έναν έλεγχο ποιότητας πριν ενσωματωθούν στον κύριο κώδικα. Στο πλαίσιο της παρούσας διπλωματικής, η διαδικασία αυτή διασφαλίζει ότι κάθε συνεισφορά στον κώδικα είναι τεκμηριωμένη, ελεγχόμενη και επανεξετασμένη από άλλους συντελεστές πριν γίνει μέρος του τελικού προϊόντος.

3.5.3 Διαχείριση του Κώδικα και Παρακολούθηση Προόδου

Το GitHub παρέχει επιπλέον εργαλεία για την παρακολούθηση θεμάτων (issues), την καταγραφή bugs και την οργάνωση της εργασίας μέσω projects και milestones. Αυτά τα χαρακτηριστικά καθιστούν το GitHub ένα κέντρο διαχείρισης έργων, διευκολύνοντας την ομαδική εργασία και τη συστηματική παρακολούθηση της προόδου του έργου.

Μέσω των pull requests, κάθε αλλαγή στον κώδικα υποβάλλεται σε αναθεώρηση, επιτρέποντας την έγκαιρη ανίχνευση και επίλυση προβλημάτων. Αυτή η πρακτική έχει αποδειχθεί πολύτιμη στην ανάπτυξη της διπλωματικής, καθώς επιτρέπει την ενσωμάτωση μόνο των αλλαγών που πληρούν τις απαιτήσεις ποιότητας.

4. Αρχιτεκτονική και Υλοποίηση

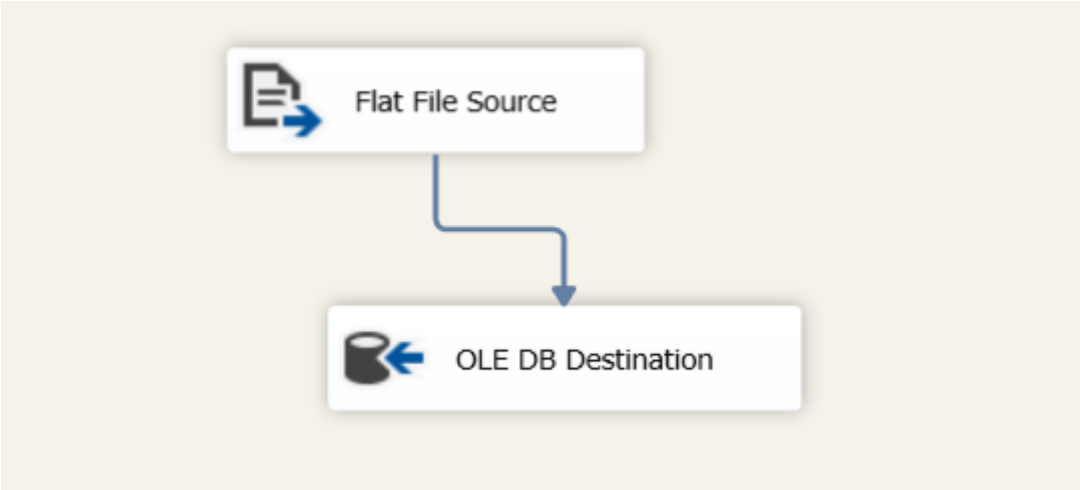
4.1 Λεπτομερής Περιγραφή του Έργου ETL

Η διαδικασία ETL (Extract, Transform, Load) που υλοποιήθηκε στο πλαίσιο αυτού του έργου αποτελεί τον πυρήνα της διαχείρισης και αξιοποίησης δεδομένων από ένα συνθετικό dataset κυβερνοασφάλειας. Στόχος αυτής της διαδικασίας είναι η διασφάλιση ότι τα δεδομένα μπορούν να μετασχηματιστούν, να εμπλουτιστούν και να προετοιμαστούν κατάλληλα για διάφορες χρήσεις, όπως μηχανική μάθηση, αναλύσεις και επιχειρηματική ευφυΐα. Στη συνέχεια περιγράφεται αναλυτικά η κάθε φάση της διαδικασίας ETL.

4.1.1 Αρχικοποίηση και Εισαγωγή Δεδομένων

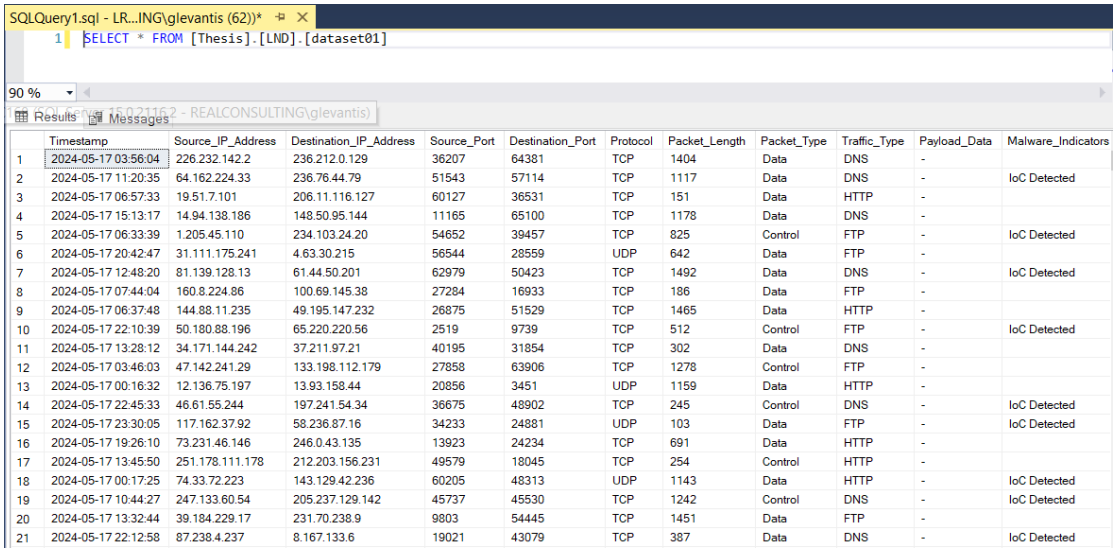
Το πρώτο στάδιο της διαδικασίας ETL περιλαμβάνει την αρχικοποίηση και την εισαγωγή των δεδομένων στη βάση δεδομένων. Το dataset που χρησιμοποιήθηκε περιέχει συνθετικά δεδομένα, αποτελείται από 40.000 γραμμές και 25 στήλες, και είναι σε μορφή .csv. Τα δεδομένα περιέχουν καταγεγραμμένα περιστατικά κυβερνοασφάλειας με διαβαθμίσεις σοβαρότητας (low, medium, high). Αυτά τα δεδομένα εισάγονται αρχικά στο Landing schema, το οποίο λειτουργεί ως η «ωμή» περιοχή αποθήκευσης δεδομένων, δηλαδή εκεί όπου τα δεδομένα αποθηκεύονται ακριβώς όπως εισάγονται, χωρίς καμία προηγούμενη επεξεργασία.

Η διαδικασία εισαγωγής υλοποιείται μέσω του εργαλείου SQL Server Integration Services (SSIS) στο Visual Code. Το SSIS είναι ένα ισχυρό εργαλείο για τη μεταφορά δεδομένων, το οποίο επιτρέπει την εξαγωγή δεδομένων από διάφορες πηγές (στην περίπτωση αυτή, από το .csv file), την εκτέλεση μετασχηματισμών και τη φόρτωσή τους στη βάση δεδομένων SQL Server.



Εικονα 10. Ροή Δεδομένων SSIS: Από Αρχείο Flat File σε OLE DB

Στο Landing schema, τα δεδομένα αποθηκεύονται στην ακατέργαστη μορφή τους, διασφαλίζοντας έτσι ότι τα αρχικά δεδομένα παραμένουν ανέπαφα για μελλοντική αναφορά ή περαιτέρω επεξεργασία. Η διατήρηση των πρωτογενών δεδομένων είναι κρίσιμη, καθώς εξασφαλίζει ότι υπάρχει πάντα μια αξιόπιστη πηγή δεδομένων, στην οποία μπορεί να επιστρέψει ο αναλυτής αν χρειαστεί.


SQLQuery1.sql - LR...ING\glevantis (62))*

```
SELECT * FROM [Thesis].[LND].[dataset01]
```

90 %

Results Messages

	Timestamp	Source_IP_Address	Destination_IP_Address	Source_Port	Destination_Port	Protocol	Packet_Length	Packet_Type	Traffic_Type	Payload_Data	Malware_Indicators
1	2024-05-17 03:56:04	226.232.142.2	236.212.0.129	36207	64381	TCP	1404	Data	DNS	-	
2	2024-05-17 11:20:35	64.162.224.33	236.76.44.79	51543	57114	TCP	1117	Data	DNS	-	IoC Detected
3	2024-05-17 06:57:33	19.51.7.101	206.11.116.127	60127	36531	TCP	151	Data	HTTP	-	
4	2024-05-17 15:13:17	14.94.138.186	148.50.95.144	11165	65100	TCP	1178	Data	DNS	-	
5	2024-05-17 06:33:39	1.205.45.110	234.103.24.20	54652	39457	TCP	825	Control	FTP	-	IoC Detected
6	2024-05-17 20:42:47	31.111.175.241	4.63.30.215	56544	28559	UDP	642	Data	FTP	-	
7	2024-05-17 12:48:20	81.139.128.13	61.44.50.201	62979	50423	TCP	1492	Data	DNS	-	IoC Detected
8	2024-05-17 07:44:04	160.8.224.86	100.69.145.38	27284	16933	TCP	186	Data	FTP	-	
9	2024-05-17 06:37:48	144.88.11.235	49.195.147.232	26875	51529	TCP	1465	Data	HTTP	-	
10	2024-05-17 22:10:39	50.180.88.196	65.220.220.56	2519	9739	TCP	512	Control	FTP	-	IoC Detected
11	2024-05-17 13:28:12	34.171.144.242	37.211.97.21	40195	31854	TCP	302	Data	DNS	-	
12	2024-05-17 03:46:03	47.142.241.29	133.198.112.179	27858	63906	TCP	1278	Control	FTP	-	
13	2024-05-17 00:16:32	12.136.75.197	13.93.158.44	20856	3451	UDP	1159	Data	HTTP	-	
14	2024-05-17 22:45:33	46.61.55.244	197.241.54.34	36675	48902	TCP	245	Control	DNS	-	IoC Detected
15	2024-05-17 23:30:05	117.162.37.92	58.236.87.16	34233	24881	UDP	103	Data	FTP	-	IoC Detected
16	2024-05-17 19:26:10	73.231.46.146	246.0.43.135	13923	24234	TCP	691	Data	HTTP	-	
17	2024-05-17 13:45:50	251.178.111.178	212.203.156.231	49579	18045	TCP	254	Control	HTTP	-	
18	2024-05-17 00:17:25	74.33.72.223	143.129.42.236	60205	48313	UDP	1143	Data	HTTP	-	IoC Detected
19	2024-05-17 10:44:27	247.133.60.54	205.237.129.142	45737	45530	TCP	1242	Control	DNS	-	IoC Detected
20	2024-05-17 13:32:44	39.184.229.17	231.70.238.9	9803	54445	TCP	1451	Data	FTP	-	
21	2024-05-17 22:12:58	87.238.4.237	8.167.133.6	19021	43079	TCP	387	Data	DNS	-	IoC Detected

Εικονα 11. Ακατέργαστα Δεδομένα στο Landing Schema

4.1.2 Μετασχηματισμός Δεδομένων (Transform)

Μετά την εισαγωγή στο Landing schema, η διαδικασία ETL προχωρά στο στάδιο του μετασχηματισμού, όπου τα δεδομένα καθαρίζονται, εμπλουτίζονται και προετοιμάζονται για την τελική τους αποθήκευση και χρήση στο Staging schema. Αυτή η φάση είναι κρίσιμη για την εξασφάλιση της ποιότητας και της ακρίβειας των δεδομένων, καθώς οποιαδήποτε ασυνέπεια ή λάθος μπορεί να επηρεάσει σημαντικά τα αποτελέσματα των αναλύσεων και τις αποφάσεις που βασίζονται σε αυτά τα δεδομένα.

4.1.2.1 Καθαρισμός Δεδομένων

Ο καθαρισμός των δεδομένων είναι το πρώτο βήμα σε αυτή τη φάση. Στόχος του καθαρισμού είναι να διασφαλιστεί ότι τα δεδομένα είναι συνεπή και χωρίς σφάλματα. Αυτό περιλαμβάνει την αφαίρεση διπλότυπων εγγραφών, τη διόρθωση τυχόν σφαλμάτων στην εισαγωγή δεδομένων και τη διασφάλιση ότι οι στήλες περιέχουν τα κατάλληλα δεδομένα για την προοριζόμενη ανάλυση.

Για παράδειγμα, αν κάποια από τις στήλες περιέχει μη έγκυρες τιμές (π.χ. μια ημερομηνία σε λάθος μορφή), αυτές διορθώνονται ή απομακρύνονται κατά τη διάρκεια του καθαρισμού. Η αφαίρεση αυτών των προβλημάτων είναι κρίσιμη για να αποτραπούν τυχόν σφάλματα στις επόμενες φάσεις ανάλυσης ή μετασχηματισμού των δεδομένων.

4.1.2.2 Μεταφορά Δεδομένων από το Landing Schema στο Staging Schema

Η μεταφορά των δεδομένων από το Landing schema στο Staging schema πραγματοποιείται μέσω της αποθηκευμένης διαδικασίας (Stored Procedure) [STG].[LND_TO_STG]. Αυτή η διαδικασία είναι η καρδιά του σταδίου μετασχηματισμού και περιλαμβάνει διάφορες βασικές μετατροπές και επιπλέον, βελτιστοποιεί τη μεταφορά των δεδομένων επεξεργάζοντας μόνο τις νέες εγγραφές που έχουν προστεθεί μετά το αρχικό ιστορικό load.:

Μετατροπή Ημερομηνιών και Timestamps: Οι χρονικές πληροφορίες (timestamps) μετατρέπονται σε μορφή date, που υποστηρίζεται πλήρως από τη βάση δεδομένων SQL. Αυτή η μετατροπή εξασφαλίζει την ακρίβεια και την αποτελεσματική χρήση των χρονικών δεδομένων σε αναλύσεις που απαιτούν ακριβείς χρονικές σημάνσεις, όπως η ανάλυση των τάσεων των επιθέσεων κυβερνοασφάλειας με την πάροδο του χρόνου.

```
SELECT
timestamp,
year,
month,
month_name,
day,
day_name
FROM [Thesis].[STG].[dataset01]
```

90 %

Results | Messages

	timestamp	year	month	month_name	day	day_name
1	2021-08-13 16:56:20.0000000	2021	8	August	13	Friday
2	2020-01-12 17:11:15.0000000	2020	1	January	12	Sunday
3	2023-10-09 20:07:53.0000000	2023	10	October	9	Monday
4	2020-11-10 22:48:59.0000000	2020	11	November	10	Tuesday
5	2023-01-13 11:23:53.0000000	2023	1	January	13	Friday
6	2022-12-12 19:57:10.0000000	2022	12	December	12	Monday
7	2020-03-01 05:28:43.0000000	2020	3	March	1	Sunday
8	2020-09-07 21:47:24.0000000	2020	9	September	7	Monday
9	2023-04-20 00:08:16.0000000	2023	4	April	20	Thursday
10	2020-08-04 06:50:47.0000000	2020	8	August	4	Tuesday

Εικονα 12 Μετατροπή Timestamps σε Μορφή Date

Δημιουργία Μοναδικού Κλειδιού: Για κάθε εγγραφή, δημιουργείται ένα μοναδικό κλειδί (unique key) μέσω hash, χρησιμοποιώντας συνδυασμό της χρονικής σήμανσης, της διεύθυνσης IP πηγής και της διεύθυνσης IP προορισμού. Αυτό εξασφαλίζει ότι κάθε εγγραφή μπορεί να αναγνωριστεί και να ανακληθεί με ακρίβεια κατά τη διάρκεια αναλύσεων ή περαιτέρω επεξεργασιών.

```
SELECT key_column FROM [Thesis].[STG].[dataset01]
```

90 %

Results | Messages

	key_column
1	0001B5F76C07EE2DB980AEFBAF027932
2	0001C69037A85E611A71415B85004262
3	0002AAEB0D8F50D0CE77FD7C3E497DCE
4	0003BCB5A187620EC74D443917022FD0
5	0004D94CA7EF741E090CEE578E42E40E
6	0007358C74921E9B0C234FAA3A533940
7	000A857AFC3F3EEF537D3C27DD64DE78
8	000C61BA101F6196D5F5BBE3FCD9B476
9	000D5141637B48332F7DE6B622E5A98F

Εικονα 13 Δημιουργία Μοναδικού Κλειδιού (Unique Key) για Αναγνώριση Εγγραφών μέσω Hash

Ορισμός Τύπου Δεδομένων Στηλών: Οι στήλες μετατρέπονται σε μορφή που διευκολύνει την ανάλυση. Για παράδειγμα, οι στήλες που περιέχουν διευθύνσεις IP, αριθμούς port, τύπους πρωτοκόλλων, και κατηγορίες επιθέσεων μετατρέπονται σε τυποποιημένες μορφές με τη χρήση αρκετών built-in συναρτήσεων συνδιαστικά (CONVERT), που είναι ευκολότερες στη χρήση για μελλοντικές αναλύσεις. Αυτή η κανονικοποίηση διασφαλίζει ότι τα δεδομένα είναι συνεπή και συγκρίσιμα.

Εμπλουτισμός Δεδομένων: Τα δεδομένα εμπλουτίζονται με επιπρόσθετες πληροφορίες, όπως γεωγραφικά δεδομένα. Οι τοποθεσίες (*from_location* και *to_location*) μετατρέπονται σε συγκεκριμένες χώρες και ηπείρους μέσω μιας σειράς κανόνων και συνθηκών που βασίζονται στις ήδη υπάρχουσες πληροφορίες. Αυτός ο εμπλουτισμός των δεδομένων καθιστά ευκολότερη την ανάλυση και κατανόηση των γεωγραφικών προτύπων στις επιθέσεις.

SQLQuery1.sql - LR...ING\glevantis (62))*

```
1  SELECT Geo_location_Data FROM [Thesis].[LND].[dataset01]
```

90 %

Results | Messages

	Geo_location_Data
1	Rajpur Sonarpur.Kerala
2	Wuhan.Vienna
3	Shanghai.Sikkim
4	Amsterdam.Paris
5	Hangzhou.Latur
6	Chongqing.Vienna
7	Mexico City.Paris
8	Shenzhen.Madrid
9	Shenzhen.Vienna
10	Guangzhou.London

Εικονα 13 Εμπλουτισμός Δεδομένων με Γεωγραφικές Πληροφορίες

```
2  SELECT [from_location],[to_location] FROM [Thesis].[STG].[dataset01]
```

90 %

Results | Messages

	from_location	to_location
1	Bathinda	Uttarakhand
2	Fatehpur	Rajasthan
3	Sambalpur	Haryana
4	Kamarhati	Meghalaya
5	Tezpur	Nagaland
6	Madanapalle	Gujarat
7	Kumbakonam	Gujarat
8	Agra	Punjab
9	Jorhat	Telangana
10	Tirupati	Nagaland
11	Davanagere	Gujarat

Εικονα 14. Ανάλυση Γεωγραφικών Τοποθεσιών: Από και Προς Τοποθεσίες για Κυβερνοεπιθέσεις

Χρήση Common Table Expression (CTE) για Φιλτράρισμα Νέων Δεδομένων: Με τη χρήση Common Table Expression (CTE), η Stored Procedure ανιχνεύει και φιλτράρει τα δεδομένα που έχουν προστεθεί από την τελευταία ενημέρωση. Αυτό εξασφαλίζει ότι μόνο οι νέες επιθέσεις ή τα νέα περιστατικά που έχουν καταγραφεί στη βάση δεδομένων μεταφέρονται στο Staging schema, μειώνοντας έτσι σημαντικά τον χρόνο επεξεργασίας και βελτιώνοντας την απόδοση της διαδικασίας ETL.Η λογική αυτή όχι μόνο διευκολύνει την καθημερινή ροή δεδομένων, αλλά επίσης διασφαλίζει ότι η μεταφορά δεδομένων παραμένει γρήγορη και αποτελεσματική, ακόμη και σε περιπτώσεις μεγάλων αυξήσεων του όγκου δεδομένων. Οποιαδήποτε νέα δεδομένα εντοπίζονται και μεταφέρονται χωρίς να επηρεάζονται τα ήδη επεξεργασμένα ιστορικά δεδομένα.

```
WITH initial_cte AS (
    SELECT
        *,
        CASE
            WHEN CONVERT(date, CONVERT(datetime, CONVERT(varchar(19), [Timestamp]), 120)) = CONVERT(date, getdate()) THEN CONVERT(date, getdate())
            ELSE '1900-01-01'
        END AS create_date
    FROM
        [LND].[dataset01]
),
filter_new_data_cte AS (
    SELECT *
    FROM initial_cte
    WHERE create_date > (
            SELECT COALESCE(MAX(create_date), '1900-01-01') AS cdate
            FROM [STG].[dataset01]
        )
    /*OR create_date = '1900-01-01'*/ --uncomment for initial data
    )
```

Εικονα 15. Φιλτράρισμα Νέων Δεδομένων με Χρήση Common Table Expression (CTE) στη Διαδικασία ETL

4.1.2.3 Μετασχηματισμός Δεδομένων από το Staging Schema στο Target Schema

Μετά την ολοκλήρωση των μετασχηματισμών στο Staging schema, τα δεδομένα μεταφέρονται στο Target schema. Σε αυτό το στάδιο, τα δεδομένα προετοιμάζονται για συγκεκριμένες εφαρμογές και χρήσεις, μέσω τριών αποθηκευμένων διαδικασιών που δημιουργούν αντίστοιχα tables στο Target schema:

- *[TRG].[STG_TO_TRG_ML]:* Αυτή η αποθηκευμένη διαδικασία δημιουργεί το table που προορίζεται για αναλύσεις μηχανικής μάθησης. Τα δεδομένα, πέρα από την απλή μεταφορά τους, υφίστανται κωδικοποίηση κατηγοριών (one-hot encoding), όπου για κάθε κατηγορία δημιουργείται μια ξεχωριστή στήλη με δυαδικές τιμές (1 ή 0). Αυτός ο τρόπος κωδικοποίησης είναι απαραίτητος για την απρόσκοπτη χρήση των δεδομένων σε μοντέλα μηχανικής μάθησης, καθώς επιτρέπει την εύκολη ανάλυση κατηγορικών δεδομένων από τα μοντέλα.

```
SELECT [protocol_TCP]
      ,[protocol_UDP]
      ,[protocol_ICMP]
      ,[packet_type_Control]
      ,[packet_type_Data]
      ,[service_FTP]
      ,[service_HTTP]
```

90 %

Results | Messages

	protocol_TCP	protocol_UDP	protocol_ICMP	packet_type_Control	packet_type_Data	service_FTP	service_HTTP	service_DNS
1	1	0	0	1	0	1	0	0
2	0	1	0	0	1	1	0	0
3	1	0	0	1	0	0	1	0
4	1	0	0	1	0	0	1	0
5	0	1	0	0	1	1	0	0
6	0	0	1	1	0	0	0	1
7	1	0	0	1	0	0	0	1
8	0	1	0	0	1	0	1	0
9	0	1	0	0	1	0	1	0
10	0	1	0	0	1	0	1	0

Εικονα 16. Κωδικοποίηση Κατηγοριών (One-Hot Encoding) για Ανάλυση Μηχανικής Μάθησης

- *[TRG].[STG_TO_TRG_PBI]:* Αυτή η διαδικασία προετοιμάζει τα δεδομένα για χρήση σε αναφορές Power BI. Τα δεδομένα κατηγοριοποιούνται και εμπλουτίζονται με γεωγραφικές πληροφορίες, διευκολύνοντας τις οπτικοποιήσεις και αναλύσεις στο Power BI. Ειδικότερα, οι στήλες from_location και to_location μετατρέπονται σε ονόματα χωρών και ηπείρων, καθιστώντας ευκολότερη την ανάλυση γεωγραφικών προτύπων και τη συσχέτισή τους με συμβάντα κυβερνοασφάλειας.

```
SELECT from_location,
from_country,
from_continent,
to_location,
to_country,
to_continent
FROM [Thesis].[TRG].[dataset01_ml]
```

90 %

Results Messages

	from_location	from_country	from_continent	to_location	to_country	to_continent
1	Beijing	China	Asia	Amsterdam	Netherlands	Europe
2	Bareilly	India	Asia	New York City	USA	North America
3	Houston	USA	North America	Athens	Greece	Europe
4	Berlin	Germany	Europe	Atlanta	USA	North America
5	Madrid	Spain	Europe	Toronto	Canada	North America
6	Paris	France	Europe	Madrid	Spain	Europe
7	Toronto	Canada	North America	Madrid	Spain	Europe
8	Montreal	Canada	North America	Los Angeles	USA	North America
9	New York City	USA	North America	Amsterdam	Netherlands	Europe
10	Los Angeles	USA	North America	Brussels	Belgium	Europe

Εικονα 17. Προετοιμασία Δεδομένων με Γεωγραφική Πληροφορία για Αναφορές Power BI

- *[TRG].[STG_TO_TRG_STIX_TAXII]:* Αυτή η αποθηκευμένη διαδικασία δημιουργεί ένα table που προορίζεται για εξαγωγή δεδομένων σε μορφή STIX/TAXII, τα οποία χρησιμοποιούνται για την αυτόματη δημιουργία αντικειμένων STIX και τη διανομή τους μέσω του προτύπου TAXII. Αυτό επιτρέπει την ασφαλή και δομημένη ανταλλαγή πληροφοριών απειλών με άλλους οργανισμούς ή συστήματα.

4.1.2.4 Εξαγωγή και Διανομή Δεδομένων

Το τελικό στάδιο της διαδικασίας ETL είναι η εξαγωγή των δεδομένων από το Target schema για την τελική τους χρήση. Ανάλογα με το table στο οποίο αποθηκεύονται, τα δεδομένα μπορούν να χρησιμοποιηθούν σε διάφορες εφαρμογές και αναλύσεις:

1. *Μηχανική Μάθηση:* Τα δεδομένα που έχουν προετοιμαστεί στο table ML χρησιμοποιούνται από data scientists για την εκπαίδευση και βελτιστοποίηση μοντέλων μηχανικής μάθησης. Αυτά τα μοντέλα μπορούν να προβλέψουν μελλοντικές απειλές, να εντοπίσουν ανωμαλίες ή να αναγνωρίσουν μοτίβα επιθέσεων που μπορεί να μην είναι άμεσα εμφανή με συμβατικές μεθόδους ανάλυσης.

2. *Power BI Αναφορές*: Τα δεδομένα στο table PBI χρησιμοποιούνται για τη δημιουργία αναφορών και οπτικοποιήσεων στο Power BI. Αναλυτές μπορούν να εκμεταλλευτούν αυτές τις αναφορές για να κατανοήσουν καλύτερα τις γεωγραφικές κατανομές των επιθέσεων, τη συχνότητά τους, καθώς και τα χαρακτηριστικά των απειλών, παρέχοντας πολύτιμες πληροφορίες για τη λήψη αποφάσεων.

3. *STIX/TAXII Αντικείμενα:* Τα δεδομένα από το table STIX/TAXII εξάγονται και μετατρέπονται σε αντικείμενα STIX, τα οποία μπορούν να διανεμηθούν μέσω ενός TAXII server. Αυτή η δομημένη ανταλλαγή πληροφοριών απειλών επιτρέπει στους οργανισμούς να ενισχύσουν τη συνεργασία τους και να ανταποκριθούν αποτελεσματικότερα σε απειλές κυβερνοασφάλειας.

4.2 Ενσωμάτωση του STIX/TAXII για Κοινή Χρήση Πληροφοριών Απειλών

4.2.1 Εισαγωγή στο STIX/TAXII

Το STIX (*Structured Threat Information Expression*) και TAXII (*Trusted Automated Exchange of Indicator Information*) είναι δύο σημαντικά πρότυπα για την ανταλλαγή πληροφοριών απειλών στον τομέα της κυβερνοασφάλειας. Αναπτύχθηκαν από την κοινότητα του OASIS (*Organization for the Advancement of Structured Information Standards*) και αποτελούν τη βάση για την αυτοματοποιημένη ανταλλαγή πληροφοριών ασφαλείας μεταξύ οργανισμών.

Το STIX παρέχει μια τυποποιημένη γλώσσα για την αναπαράσταση πληροφοριών απειλών, όπως δείκτες (indicators), τακτικές, τεχνικές, διαδικασίες (TTPs), συμβάντα και παρατηρήσεις. Αυτό επιτρέπει στους οργανισμούς να περιγράφουν με ακρίβεια τα στοιχεία των απειλών, να μοιράζονται πληροφορίες και να αντιδρούν αποτελεσματικά. Τα δεδομένα που περιλαμβάνονται σε ένα STIX αντικείμενο μπορεί να περιέχουν δείκτες κακόβουλων διευθύνσεων IP, hash αρχείων, υπογραφές επιθέσεων, και άλλες κρίσιμες πληροφορίες που βοηθούν στην κατανόηση και την ανάλυση των απειλών.

Το TAXII, από την άλλη πλευρά, είναι το πρωτόκολλο που χρησιμοποιείται για τη μεταφορά των STIX δεδομένων μεταξύ διαφορετικών συστημάτων και οργανισμών. Το TAXII υποστηρίζει πολλούς τρόπους μεταφοράς, όπως push και pull, και επιτρέπει την ασφαλή και αξιόπιστη διανομή των πληροφοριών απειλών. Με τη χρήση του TAXII, οι οργανισμοί μπορούν να δημιουργήσουν κανάλια επικοινωνίας για τη συνεχή ενημέρωση και διάδοση πληροφοριών απειλών, ενισχύοντας έτσι τη συνεργασία στον τομέα της κυβερνοασφάλειας.

Στη σημερινή ψηφιακή εποχή, η ικανότητα να μοιράζεται κανείς αξιόπιστες και επικαιροποιημένες πληροφορίες απειλών είναι ζωτικής σημασίας για την προστασία των υποδομών και των δεδομένων από τις αυξανόμενες κυβερνοεπιθέσεις. Η ενσωμάτωση των προτύπων STIX και TAXII στις διαδικασίες διαχείρισης πληροφοριών απειλών επιτρέπει στους οργανισμούς να αυξήσουν την ακρίβεια, την ταχύτητα και την αποτελεσματικότητα της αντίδρασης τους σε πραγματικό χρόνο.

Τα πρότυπα αυτά όχι μόνο βελτιώνουν την επικοινωνία μεταξύ των διαφορετικών μονάδων ενός οργανισμού αλλά και διευκολύνουν τη συνεργασία με εξωτερικούς εταίρους, όπως κυβερνητικούς φορείς, ιδιωτικές εταιρείες και διεθνείς οργανισμούς. Η χρήση του STIX/TAXII παρέχει έναν κοινό γλωσσάρι για την καταγραφή και ανταλλαγή πληροφοριών, διασφαλίζοντας ότι όλες οι εμπλεκόμενες πλευρές έχουν την ίδια αντίληψη για τις απειλές και τις ενέργειες που απαιτούνται.

Επιπλέον, η αυτοματοποίηση που παρέχουν τα STIX/TAXII μειώνει την ανάγκη για χειροκίνητες διαδικασίες, οι οποίες είναι συχνά επιρρεπείς σε λάθη και καθυστερήσεις. Με την τυποποίηση της επικοινωνίας και την αυτοματοποιημένη ανταλλαγή πληροφοριών, οι οργανισμοί μπορούν να ανταποκρίνονται γρηγορότερα στις απειλές, βελτιώνοντας έτσι τη συνολική τους ασφάλεια.

4.2.2 Εργαλεία και Τεχνολογίες

Για την υλοποίηση και διαχείριση των STIX/TAXII, απαιτείται η χρήση εξειδικευμένων εργαλείων και τεχνολογιών που διευκολύνουν την επεξεργασία, την ανάλυση και την ανταλλαγή των πληροφοριών απειλών. Η επιλογή των κατάλληλων εργαλείων εξαρτάται από τις ανάγκες του οργανισμού, το επίπεδο τεχνικής εξειδίκευσης και τους στόχους που πρέπει να επιτευχθούν.

Παρακάτω αναλύονται μερικά από τα πιο διαδεδομένα εργαλεία και τεχνολογίες που χρησιμοποιούνται για τη διαχείριση STIX/TAXII.

1. *STIX2 Python Library:* Η βιβλιοθήκη stix2 για Python είναι ένα από τα πιο δημοφιλή εργαλεία για την εργασία με STIX δεδομένα. Επιτρέπει στους προγραμματιστές να δημιουργούν, να τροποποιούν και να αναλύουν STIX αντικείμενα εύκολα, χρησιμοποιώντας την Python. Η βιβλιοθήκη παρέχει μια πλούσια γκάμα από έτοιμες κλάσεις και μεθόδους που καλύπτουν όλες τις πτυχές της προτυποποίησης STIX 2.1, όπως Indicators, Observed Data, TTPs και άλλες. Είναι ιδιαίτερα χρήσιμη για την αυτοματοποίηση της δημιουργίας STIX δεδομένων και την ενσωμάτωση τους σε υπάρχοντα συστήματα ανάλυσης απειλών.

2. *TAXII Servers:* Οι TAXII servers είναι το κεντρικό σημείο για την ανταλλαγή STIX δεδομένων μεταξύ οργανισμών. Παραδείγματα τέτοιων servers είναι οι OpenTAXII, EclecticIQ και Soltra Edge. Αυτοί οι servers παρέχουν τη δυνατότητα δημιουργίας και διαχείρισης καναλιών TAXII, όπου μπορούν να ανταλλάσσονται πληροφορίες απειλών με ασφάλεια και αποδοτικότητα. Οι TAXII servers υποστηρίζουν διάφορα μοντέλα μεταφοράς δεδομένων, επιτρέποντας τόσο τη συνεχή ροή πληροφοριών (push) όσο και την αίτηση πληροφοριών κατ' απαίτηση (pull).

3. *Cortex XSOAR και άλλες SOAR πλατφόρμες:* Οι πλατφόρμες Security Orchestration, Automation, and Response (SOAR) όπως η Cortex XSOAR ενσωματώνουν δυνατότητες STIX/TAXII για την αυτοματοποίηση της απόκρισης σε κυβερνοαπειλές. Αυτές οι πλατφόρμες επιτρέπουν τη συλλογή πληροφοριών από διάφορες πηγές, τη δημιουργία STIX αντικειμένων και τη διάδοση αυτών μέσω TAXII. Η δυνατότητα ενσωμάτωσης με άλλα εργαλεία ασφάλειας, όπως SIEMs (Security Information and Event Management), καθιστά τις SOAR πλατφόρμες ιδανικές για τη διαχείριση ολοκληρωμένων στρατηγικών ασφάλειας.

4. *Threat Intelligence Platforms (TIPs):* Οι TIPs, όπως το MISP (Malware Information Sharing Platform) και το ThreatConnect, χρησιμοποιούνται για τη συγκέντρωση, την ανάλυση και την ανταλλαγή πληροφοριών απειλών σε επίπεδο οργανισμού ή κοινότητας. Αυτές οι πλατφόρμες υποστηρίζουν την ενσωμάτωση με STIX/TAXII και επιτρέπουν την εύκολη διαχείριση των πληροφοριών απειλών, τη δημιουργία αναφορών, και τη διανομή αυτών σε πραγματικό χρόνο. Οι TIPs επιτρέπουν στους οργανισμούς να μοιράζονται πληροφορίες με τους συνεργάτες τους και να λαμβάνουν επικαιροποιημένα δεδομένα απειλών από εξωτερικές πηγές.

5. *Συστήματα SIEM (Security Information and Event Management):* Τα SIEMs, όπως το Splunk, ArcSight, και QRadar, συνήθως ενσωματώνουν δυνατότητες STIX/TAXII για τη συλλογή και ανάλυση δεδομένων από πολλαπλές πηγές. Τα SIEMs χρησιμοποιούν αυτές τις πληροφορίες για την ανίχνευση και την απόκριση σε περιστατικά ασφαλείας σε πραγματικό χρόνο. Η ικανότητά τους να επεξεργάζονται μεγάλα όγκους δεδομένων και να ανιχνεύουν ανωμαλίες τα καθιστά κεντρικά εργαλεία για την ασφάλεια του δικτύου και τη διαχείριση περιστατικών.

Η επιτυχής ενσωμάτωση του STIX/TAXII απαιτεί την καλή γνώση των εργαλείων και των τεχνολογιών που είναι διαθέσιμα, καθώς και την κατανόηση των αναγκών της κάθε άσκησης. Η επιλογή της κατάλληλης τεχνολογίας μπορεί να κάνει τη διαφορά στην αποτελεσματικότητα της ανταλλαγής πληροφοριών και την ταχύτητα της αντίδρασης σε απειλές.

4.2.3 Πρακτική Εφαρμογή της Εξαγωγής Δεδομένων σε Μορφή STIX

Η εξαγωγή δεδομένων σε μορφή STIX αποτελεί μια κρίσιμη διαδικασία για την αποτελεσματική διανομή πληροφοριών απειλών. Σε αυτό το πλαίσιο, το παρακάτω Python script (TRG_to_STIX.ipynb) χρησιμοποιήθηκε για την αυτόματη δημιουργία STIX αντικειμένων από δεδομένα κυβερνοασφάλειας, τα οποία στη συνέχεια εξάγονται σε ένα αρχείο JSON.

Η χρήση της βιβλιοθήκης stix2 της Python διευκολύνει την όλη διαδικασία, επιτρέποντας την αυτοματοποίηση της δημιουργίας των STIX αντικειμένων.

Βήματα Εξαγωγής:

1. *Ανάλυση και Προετοιμασία των Δεδομένων:* Το script αρχικά διαβάζει τα δεδομένα κυβερνοασφάλειας από ένα αρχείο ή μια βάση δεδομένων και τα επεξεργάζεται για να εξάγει τις απαραίτητες πληροφορίες. Κάθε σειρά δεδομένων περιλαμβάνει στοιχεία όπως IP διευθύνσεις, τύπους επιθέσεων, timestamps, πρωτόκολλα δικτύου και άλλα χαρακτηριστικά που είναι απαραίτητα για την δημιουργία των STIX αντικειμένων.
2. *Δημιουργία STIX Αντικειμένων:* Το script δημιουργεί δύο κύρια είδη STIX αντικειμένων:
    - *Indicators*: Αυτά τα αντικείμενα περιγράφουν στοιχεία που μπορούν να αναγνωριστούν ως ενδείξεις κακόβουλης δραστηριότητας, όπως κακόβουλες IP διευθύνσεις ή hash αρχείων. Το script χρησιμοποιεί τη συνάρτηση create_indicator για να δημιουργήσει αυτά τα αντικείμενα βασισμένο στα δεδομένα που αναλύονται.
    - *Observed Data:* Αυτά τα αντικείμενα καταγράφουν παρατηρήσεις σχετικά με τη δικτυακή κυκλοφορία, όπως IP διευθύνσεις πηγής και προορισμού, πόρτες και πρωτόκολλα. Η συνάρτηση create_observed_data δημιουργεί αυτά τα αντικείμενα, τα οποία περιλαμβάνουν τις τεχνικές πληροφορίες της παρατηρούμενης δραστηριότητας.
3. *Δημιουργία του Bundle:* Μετά τη δημιουργία των STIX αντικειμένων, όλα τα αντικείμενα συλλέγονται σε ένα STIX bundle. Το bundle είναι μια συλλογή STIX αντικειμένων που μπορούν να διανεμηθούν ως ένα ενιαίο σύνολο δεδομένων. Το script δημιουργεί το bundle χρησιμοποιώντας την κλάση Bundle της βιβλιοθήκης stix2.
4. *Εξαγωγή σε JSON:* Το τελικό στάδιο είναι η εξαγωγή του STIX bundle σε μορφή JSON, η οποία είναι η τυπική μορφή για την αποθήκευση και διανομή STIX δεδομένων. Το script αποθηκεύει το εξαγόμενο JSON σε ένα αρχείο (stix_bundle.json), το οποίο μπορεί να χρησιμοποιηθεί για την ανταλλαγή πληροφοριών με άλλους οργανισμούς μέσω ενός TAXII server ή άλλων μεθόδων.

Το εξαγόμενο αρχείο stix_bundle.json περιλαμβάνει μια σειρά από STIX αντικείμενα, όπως δείκτες (indicators) και παρατηρήσεις (observed data). Κάθε αντικείμενο περιέχει κρίσιμες πληροφορίες για την ανάλυση και την κατανόηση των απειλών, όπως IP διευθύνσεις, timestamps, και πρωτόκολλα. Παρακάτω παρατίθεται ένα παράδειγμα από το παραγόμενο JSON:

{

"type": "bundle",

"id": "bundle--428ae43e-e142-4c67-a5ba-dde76565a04e",

"objects": [

```
{
    "type": "indicator",
    "spec_version": "2.1",
    "id": "indicator--d848a00d-2ef8-4a2a-a98c-fd6785a7d1e4",
    "created": "2024-08-24T15:10:19.096553Z",
    "modified": "2024-08-24T15:10:19.096553Z",
    "name": "Indicator for Intrusion",
    "pattern": "[ipv4-addr:value = '154.100.150.171']",
    "pattern_type": "stix",
    "pattern_version": "2.1",
    "valid_from": "2021-08-13T00:00:00Z"
},
{
    "type": "observed-data",
    "spec_version": "2.1",
    "id": "observed-data--d529cd41-5545-4515-942c-f8f86b000738",
    "created": "2024-08-24T15:10:19.096553Z",
    "modified": "2024-08-24T15:10:19.096553Z",
    "first_observed": "2021-08-13T00:00:00Z",
    "last_observed": "2021-08-13T00:00:00Z",
    "number_observed": 1,
    "objects": {
        "ipv4-addr--484f5e6a-5937-5867-8b38-ca15870418e9": {
            "type": "ipv4-addr",
            "spec_version": "2.1",
            "id": "ipv4-addr--484f5e6a-5937-5867-8b38-ca15870418e9",
            "value": "154.100.150.171"
        },
        "ipv4-addr--debe961c-3f5c-56b9-a63f-b729f1c5f263": {
            "type": "ipv4-addr",
            "spec_version": "2.1",
            "id": "ipv4-addr--debe961c-3f5c-56b9-a63f-b729f1c5f263",
```

```
            "value": "47101"
          },
          "network-traffic--84076f9c-9b1b-587b-ab6e-c8c08e641ebb": {
            "type": "network-traffic",
            "spec_version": "2.1",
            "id": "network-traffic--84076f9c-9b1b-587b-ab6e-c8c08e641ebb",
            "src_ref": "ipv4-addr--484f5e6a-5937-5867-8b38-ca15870418e9",
            "dst_ref": "ipv4-addr--debe961c-3f5c-56b9-a63f-b729f1c5f263",
            "src_port": 10478,
            "protocols": [
              "213"
            ]
          }
        }
      }
    ]
}
```


Αυτό το παράδειγμα δείχνει πώς τα δεδομένα μπορούν να οργανωθούν και να διαμοιραστούν σε μια μορφή που είναι ευρέως κατανοητή και χρησιμοποιήσιμη από άλλα συστήματα κυβερνοασφάλειας. Η χρήση του STIX JSON επιτρέπει τη γρήγορη ενσωμάτωση των πληροφοριών απειλών σε εργαλεία ανάλυσης, πλατφόρμες ανταλλαγής πληροφοριών και άλλα συστήματα ασφαλείας.

Προκλήσεις και Λύσεις:

Κατά την πρακτική εφαρμογή της εξαγωγής δεδομένων σε μορφή STIX, μπορεί να προκύψουν διάφορες προκλήσεις, όπως η ακρίβεια των δεδομένων, η συμβατότητα με άλλες πλατφόρμες και η ανάγκη για συνεχή ενημέρωση των πληροφοριών. Για την αντιμετώπιση αυτών των προκλήσεων, είναι κρίσιμο να διασφαλίζεται η ποιότητα των δεδομένων εισόδου, να χρησιμοποιούνται εργαλεία που υποστηρίζουν τα τελευταία πρότυπα STIX/TAXII, και να υπάρχει συνεχής συνεργασία με άλλους οργανισμούς για την ανταλλαγή πληροφοριών.

4.2.4 Οφέλη από τη Χρήση του STIX/TAXII για τη Διανομή Πληροφοριών Απειλών

Η χρήση του STIX/TAXII για τη διανομή πληροφοριών απειλών παρέχει πολυάριθμα οφέλη που ενισχύουν την ασφάλεια και την ανθεκτικότητα των οργανισμών σε κυβερνοαπειλές. Τα βασικά οφέλη περιλαμβάνουν:

- *Βελτιωμένη Συνεργασία και Συντονισμός:* Το STIX/TAXII επιτρέπει την αποτελεσματική ανταλλαγή πληροφοριών απειλών μεταξύ διαφορετικών οργανισμών, βιομηχανιών και κυβερνητικών φορέων. Αυτή η συνεργασία είναι κρίσιμη για την ανίχνευση και την

αντιμετώπιση των απειλών σε πραγματικό χρόνο. Ομοίως, η κοινή χρήση πληροφοριών επιτρέπει στους οργανισμούς να ενημερώνονται για νέες απειλές και να προσαρμόζουν τις άμυνές τους ανάλογα.

- *Αυτοματοποίηση και Ακρίβεια:* Μέσω της αυτοματοποίησης της συλλογής και της ανταλλαγής πληροφοριών, οι οργανισμοί μπορούν να μειώσουν τις ανθρώπινες παρεμβάσεις και τα σφάλματα. Η χρήση των προτύπων STIX/TAXII διασφαλίζει ότι οι πληροφορίες ανταλλάσσονται σε μια κοινή μορφή, βελτιώνοντας την ακρίβεια και τη συνέπεια των δεδομένων.
- *Κλιμάκωση και Επεκτασιμότητα:* Τα πρότυπα STIX/TAXII είναι σχεδιασμένα να υποστηρίζουν μεγάλους όγκους δεδομένων και να επεκτείνονται σύμφωνα με τις ανάγκες του οργανισμού. Αυτό επιτρέπει τη συνεχή προσαρμογή και αναβάθμιση των συστημάτων ασφαλείας, εξασφαλίζοντας την ικανότητα αντιμετώπισης μελλοντικών απειλών.
- *Ενίσχυση της Αντίληψης για τις Απειλές:* Με την ανταλλαγή πληροφοριών απειλών μέσω STIX/TAXII, οι οργανισμοί αποκτούν καλύτερη αντίληψη για το τοπίο των απειλών και μπορούν να εντοπίζουν πρότυπα και τάσεις που ίσως δεν θα ήταν εμφανή μέσω μεμονωμένων αναλύσεων. Αυτό βοηθά στη βελτίωση της στρατηγικής ασφάλειας και της προληπτικής προστασίας.
- *Εναρμόνιση με Διεθνή Πρότυπα:* Η χρήση STIX/TAXII συμμορφώνεται με τα διεθνή πρότυπα ασφαλείας και την νομοθεσία περί προστασίας δεδομένων. Αυτό καθιστά τα πρότυπα ιδανικά για οργανισμούς που δραστηριοποιούνται σε διεθνές επίπεδο και απαιτούν συμμόρφωση με πολλαπλά ρυθμιστικά πλαίσια.

Σε γενικές γραμμές, η υιοθέτηση των προτύπων STIX/TAXII προσφέρει στους οργανισμούς μια ολοκληρωμένη και ισχυρή λύση για την αντιμετώπιση των κυβερνοαπειλών, εξασφαλίζοντας ότι είναι πάντα προετοιμασμένοι για να ανταποκριθούν στις προκλήσεις που προκύπτουν στο σύγχρονο ψηφιακό περιβάλλον.

4.3 Συνεργατική Ανάπτυξη με Git και GitHub

4.3.1 Εισαγωγή στο Git και το GitHub

Το Git και το GitHub είναι δύο από τα πιο σημαντικά εργαλεία στην ανάπτυξη λογισμικού και τη διαχείριση κώδικα. Το Git είναι ένα κατανεμημένο σύστημα ελέγχου έκδοσης που επιτρέπει στους προγραμματιστές να παρακολουθούν τις αλλαγές στον κώδικα, να συνεργάζονται αποτελεσματικά και να διαχειρίζονται διάφορες εκδόσεις του έργου τους. Το Git δημιουργήθηκε από τον Linus Torvalds το 2005, και από τότε έχει γίνει το πρότυπο στον τομέα της διαχείρισης κώδικα.

Το GitHub, από την άλλη πλευρά, είναι μια πλατφόρμα που βασίζεται στο Git και προσφέρει υπηρεσίες φιλοξενίας αποθετηρίων Git μαζί με μια σειρά από επιπλέον χαρακτηριστικά που διευκολύνουν τη συνεργασία μεταξύ προγραμματιστών. Το GitHub παρέχει εργαλεία όπως pull requests, code reviews, και issues tracking, που καθιστούν την ανάπτυξη λογισμικού πιο διαφανή και οργανωμένη. Με τη χρήση του GitHub, οι ομάδες ανάπτυξης μπορούν να συνεργάζονται σε έργα ανεξάρτητα από την τοποθεσία τους, να μοιράζονται τον κώδικά τους με την ευρύτερη κοινότητα, και να χρησιμοποιούν εργαλεία αυτοματοποίησης για τη συνεχή ενσωμάτωση και ανάπτυξη (CI/CD).

Επιπλέον, το GitHub προσφέρει χαρακτηριστικά όπως τα GitHub Actions, τα οποία επιτρέπουν την αυτοματοποίηση workflows απευθείας μέσα στο αποθετήριο, και τα GitHub Pages, που επιτρέπουν τη φιλοξενία ιστοσελίδων απευθείας από το αποθετήριο. Έτσι, το GitHub δεν είναι απλώς μια πλατφόρμα φιλοξενίας κώδικα, αλλά ένα πλήρες οικοσύστημα για τη διαχείριση και την ανάπτυξη λογισμικού.

Η χρήση του Git και του GitHub είναι απαραίτητη για οποιοδήποτε σύγχρονο έργο ανάπτυξης λογισμικού, καθώς προσφέρουν τα μέσα για την ασφαλή και οργανωμένη διαχείριση του κώδικα, την αποτελεσματική συνεργασία μεταξύ προγραμματιστών, και την ενσωμάτωση σύγχρονων πρακτικών ανάπτυξης όπως η συνεχής ενσωμάτωση και η συνεχής παράδοση (CI/CD).

4.3.2 Οφέλη από τη Χρήση του Git και του GitHub σε έργα ETL

Η χρήση του Git και του GitHub σε έργα ETL (Extract, Transform, Load) προσφέρει πολυάριθμα οφέλη που μπορούν να βελτιώσουν την ποιότητα, την απόδοση και την αποτελεσματικότητα των έργων αυτών. Τα έργα ETL είναι σύνθετα και περιλαμβάνουν τη διαχείριση δεδομένων από διάφορες πηγές, την επεξεργασία και τον καθαρισμό τους, και την αποθήκευση τους σε κατάλληλες βάσεις δεδομένων. Η χρήση ενός συστήματος ελέγχου έκδοσης όπως το Git σε συνδυασμό με την πλατφόρμα συνεργασίας του GitHub μπορεί να κάνει αυτή τη διαδικασία πιο αποδοτική και ασφαλή.

1. *Ιχνηλασιμότητα και Έλεγχος Έκδοσης:* Το Git επιτρέπει στους προγραμματιστές να διατηρούν πλήρη ιστορικότητα των αλλαγών στον κώδικα. Κάθε αλλαγή μπορεί να συνδεθεί με ένα συγκεκριμένο commit, επιτρέποντας την ακριβή παρακολούθηση των τροποποιήσεων και την εύκολη επιστροφή σε προηγούμενες εκδόσεις σε περίπτωση προβλημάτων. Στα έργα ETL, όπου οι μετατροπές δεδομένων και οι αλλαγές στη λογική της επεξεργασίας είναι συχνές, η δυνατότητα αυτή είναι κρίσιμη για τη διασφάλιση της ακεραιότητας των δεδομένων και της λειτουργικότητας του συστήματος.

2. *Συνεργατική Ανάπτυξη*: Το GitHub προσφέρει εργαλεία που διευκολύνουν τη συνεργασία μεταξύ προγραμματιστών. Τα pull requests και τα code reviews επιτρέπουν την αναθεώρηση του κώδικα πριν από την ενσωμάτωση του στην κύρια βάση κώδικα, διασφαλίζοντας έτσι την ποιότητα του κώδικα. Επιπλέον, το GitHub επιτρέπει την ταυτόχρονη εργασία σε πολλαπλά branches, διευκολύνοντας την ανάπτυξη νέων χαρακτηριστικών και τη διόρθωση σφαλμάτων χωρίς να επηρεάζεται ο κύριος κώδικας του έργου.

3. *Αυτοματοποίηση και CI/CD*: Το GitHub παρέχει ενσωματωμένα εργαλεία για την αυτοματοποίηση workflows, όπως τα GitHub Actions. Σε έργα ETL, αυτό μπορεί να χρησιμοποιηθεί για την αυτοματοποίηση της διαδικασίας ενοποίησης κώδικα, τη συνεχή εκτέλεση δοκιμών, και την αυτόματη ανάπτυξη αλλαγών. Αυτό μειώνει το χρόνο παράδοσης και αυξάνει την αξιοπιστία των συστημάτων, καθώς οι αλλαγές μπορούν να δοκιμαστούν και να αναπτυχθούν ταχύτερα και με λιγότερες ανθρώπινες παρεμβάσεις.

4. *Διαχείριση Εκδόσεων και Branching Strategy*: Το Git και το GitHub προσφέρουν ισχυρές δυνατότητες για τη διαχείριση εκδόσεων μέσω της χρήσης branches. Σε έργα ETL, μπορεί να δημιουργηθούν διαφορετικά branches για την ανάπτυξη, τη δοκιμή, και την παραγωγή. Αυτή η στρατηγική διασφαλίζει ότι οι αλλαγές δοκιμάζονται εκτενώς πριν ενσωματωθούν στο κύριο σύστημα, μειώνοντας τον κίνδυνο εμφάνισης προβλημάτων σε παραγωγικά περιβάλλοντα.

5. *Διαφάνεια και Συνεργασία με Τρίτους*: Το GitHub προσφέρει τη δυνατότητα δημόσιων ή ιδιωτικών αποθετηρίων, επιτρέποντας τη συνεργασία με εξωτερικούς εταίρους, όπως προμηθευτές ή πελάτες. Στα έργα ETL, όπου μπορεί να απαιτείται συνεργασία με πολλαπλούς οργανισμούς ή ομάδες, η χρήση του GitHub διευκολύνει την κοινή χρήση του κώδικα, των scripts, και των διαδικασιών ETL με έναν ασφαλή και ελεγχόμενο τρόπο.

6. *Ασφάλεια και Διαχείριση Δικαιωμάτων*: Το GitHub επιτρέπει τη λεπτομερή διαχείριση δικαιωμάτων πρόσβασης στο αποθετήριο, εξασφαλίζοντας ότι μόνο εξουσιοδοτημένα άτομα έχουν πρόσβαση σε κρίσιμες λειτουργίες, όπως το merging σε κύρια branches. Αυτό είναι ιδιαίτερα σημαντικό σε έργα ETL, όπου η ακεραιότητα των δεδομένων και των διαδικασιών είναι κρίσιμη.

7. *Κοινοτική Υποστήριξη και Ενσωμάτωση με Άλλα Εργαλεία:* Το GitHub έχει μια μεγάλη κοινότητα χρηστών και υποστηρίζει την ενσωμάτωση με πληθώρα άλλων εργαλείων και πλατφορμών, όπως Jenkins, Jira, και Docker. Αυτό επιτρέπει την εύκολη ενσωμάτωση του GitHub σε υπάρχοντα οικοσυστήματα εργαλείων σε ένα έργο ETL, διευκολύνοντας την ολοκληρωμένη διαχείριση του έργου και την παρακολούθηση της προόδου.

8. *Διαχείριση Τεκμηρίωσης και Wiki:* Το GitHub παρέχει δυνατότητες για τη δημιουργία και τη φιλοξενία τεκμηρίωσης απευθείας μέσα στο αποθετήριο, μέσω της χρήσης markdown αρχείων ή μέσω του GitHub Wiki. Σε έργα ETL, η σωστή τεκμηρίωση είναι κρίσιμη για τη διασφάλιση της συντηρησιμότητας και της κατανόησης των διαδικασιών ETL από όλους τους εμπλεκόμενους. Η δυνατότητα διατήρησης της τεκμηρίωσης στο ίδιο μέρος με τον κώδικα διευκολύνει τη συνεχή ενημέρωσή της και την προσβασιμότητα από την ομάδα ανάπτυξης.

9. *Ιστορικό Αλλαγών και Αναφορές Σφαλμάτων:* Το GitHub επιτρέπει την παρακολούθηση των αλλαγών στον κώδικα και την αναφορά σφαλμάτων μέσω του issue tracking. Σε έργα ETL, όπου οι αλλαγές στα δεδομένα και τις διαδικασίες μπορεί να έχουν σημαντικό αντίκτυπο, η δυνατότητα παρακολούθησης των αλλαγών και των σφαλμάτων είναι κρίσιμη για τη διατήρηση της ποιότητας και της ακρίβειας του έργου.

Συνοψίζοντας, η χρήση του Git και του GitHub σε έργα ETL προσφέρει ένα ασφαλές, οργανωμένο και συνεργατικό περιβάλλον που βελτιώνει τη διαχείριση του κώδικα, επιταχύνει τις διαδικασίες ανάπτυξης και δοκιμής, και διασφαλίζει την ακεραιότητα και την ποιότητα του τελικού αποτελέσματος. Αυτά τα εργαλεία επιτρέπουν στις ομάδες να εργάζονται αποτελεσματικά, μειώνοντας τον κίνδυνο λαθών και επιτρέποντας την κλιμάκωση του έργου καθώς οι απαιτήσεις αυξάνονται.

4.3.3 Βέλτιστες Πρακτικές για τη Χρήση του Git και του GitHub σε έργα ETL

Η επιτυχής χρήση του Git και του GitHub σε έργα ETL απαιτεί την υιοθέτηση βέλτιστων πρακτικών που διασφαλίζουν την ομαλή ροή εργασίας, την ποιότητα του κώδικα και τη συνεργασία μεταξύ των μελών της ομάδας. Ακολουθούν μερικές από τις πιο σημαντικές βέλτιστες πρακτικές:

1. *Συνεπής Ονοματολογία Branches και Commits:* Είναι σημαντικό να διατηρείται μια συνεπής ονοματολογία για τα branches και τα commits. Τα branches θα πρέπει να ακολουθούν ένα πρότυπο ονοματολογίας που υποδεικνύει τον σκοπό τους, όπως feature/branch-name για νέα χαρακτηριστικά ή bugfix/branch-name για διορθώσεις. Τα μηνύματα των commits πρέπει να είναι περιγραφικά και να εξηγούν την αλλαγή που έγινε, διευκολύνοντας έτσι την κατανόηση του ιστορικού των αλλαγών.

2. *Χρήση Pull Requests για Ενσωμάτωση Αλλαγών:* Τα pull requests (PRs) είναι ο ιδανικός τρόπος για την ενσωμάτωση αλλαγών στον κύριο κώδικα. Μέσω των PRs, άλλοι προγραμματιστές μπορούν να αναθεωρήσουν τον κώδικα, να κάνουν σχόλια και να προτείνουν βελτιώσεις πριν από την ενσωμάτωση. Αυτό εξασφαλίζει ότι ο κώδικας που εισέρχεται στο κύριο branch είναι υψηλής ποιότητας και ελαχιστοποιεί τον κίνδυνο εισαγωγής σφαλμάτων.

3. *Τακτική Συγχώνευση των Branches:* Είναι σημαντικό να γίνεται τακτική συγχώνευση των branches για να διασφαλιστεί ότι οι αλλαγές από διαφορετικά branches είναι συμβατές μεταξύ τους. Η τακτική συγχώνευση μειώνει επίσης τον κίνδυνο συγκρούσεων κώδικα, οι οποίες μπορεί να είναι δύσκολο να επιλυθούν αν δεν αντιμετωπιστούν εγκαίρως.

4. *Δημιουργία και Διαχείριση Tags:* Τα tags είναι ένας τρόπος για να δημιουργήσετε σταθερές εκδόσεις του κώδικα σε σημαντικά στάδια ανάπτυξης. Σε έργα ETL, μπορείτε να χρησιμοποιήσετε tags για να σημειώσετε σημαντικές εκδόσεις του έργου, όπως την έκδοση ενός νέου ETL pipeline ή την ολοκλήρωση μιας σημαντικής αναβάθμισης.

5. *Ενσωμάτωση Συνεχούς Ενσωμάτωσης (CI):* Η συνεχής ενσωμάτωση είναι μια πρακτική που συνίσταται στην αυτόματη εκτέλεση δοκιμών και άλλων ελέγχων κάθε φορά που γίνεται μια αλλαγή στον κώδικα. Αυτό διασφαλίζει ότι τα νέα commits δεν προκαλούν προβλήματα στον κώδικα και ότι το έργο είναι πάντα σε κατάσταση λειτουργίας.

6. *Αυτοματοποιημένες Δοκιμές και Έλεγχος Ποιότητας:* Είναι καλή πρακτική να έχετε αυτοματοποιημένες δοκιμές που εκτελούνται κάθε φορά που υπάρχει ένα νέο commit. Σε έργα ETL, αυτό μπορεί να περιλαμβάνει δοκιμές που ελέγχουν την ακρίβεια των δεδομένων, την ακεραιότητα των μετατροπών, και την απόδοση του pipeline. Ο έλεγχος ποιότητας μέσω linting tools μπορεί επίσης να διασφαλίσει ότι ο κώδικας ακολουθεί συγκεκριμένα πρότυπα και είναι εύκολος στη συντήρηση.

7. *Τεκμηρίωση και Wiki:* Η καλή τεκμηρίωση είναι κρίσιμη σε έργα ETL, καθώς οι διαδικασίες μπορεί να είναι σύνθετες και απαιτούν σαφήνεια. Η χρήση του GitHub Wiki ή των markdown αρχείων στο αποθετήριο για την τεκμηρίωση των διαδικασιών και των κανόνων εργασίας βοηθά στη διατήρηση της γνώσης μέσα στην ομάδα και εξασφαλίζει ότι όλοι εργάζονται με τον ίδιο τρόπο.

8. *Ασφαλής Διαχείριση Διαπιστευτηρίων και Μυστικών:* Σε έργα ETL, τα διαπιστευτήρια και τα μυστικά (π.χ., κωδικοί πρόσβασης, API keys) πρέπει να αποθηκεύονται με ασφάλεια. Ποτέ δεν πρέπει να περιλαμβάνονται απευθείας στον κώδικα. Αντίθετα, μπορούν να χρησιμοποιηθούν εργαλεία όπως το GitHub Secrets για την ασφαλή διαχείριση και τη χρήση τους σε pipelines.

9. *Στρατηγική Branching:* Η χρήση μιας στρατηγικής branching, όπως το GitFlow ή το GitHub Flow, είναι κρίσιμη για τη σωστή διαχείριση των αλλαγών και των εκδόσεων σε ένα έργο ETL. Αυτές οι στρατηγικές παρέχουν κατευθυντήριες γραμμές για το πότε και πώς να δημιουργούνται νέα branches, πώς να γίνεται το merging, και πότε να εκδίδονται νέες εκδόσεις.

10. *Ανασκόπηση Κώδικα:* Η ανασκόπηση κώδικα από άλλους μέλη της ομάδας είναι μια βασική πρακτική για τη διασφάλιση της ποιότητας του κώδικα. Οι ανασκοπήσεις βοηθούν στην ανεύρεση σφαλμάτων, τη βελτίωση της λογικής του κώδικα, και την εκπαίδευση των μελών της ομάδας.

Η τήρηση αυτών των βέλτιστων πρακτικών διασφαλίζει ότι τα έργα ETL που χρησιμοποιούν Git και GitHub είναι οργανωμένα, εύκολα στη διαχείριση και ανθεκτικά σε αλλαγές και επεκτάσεις.

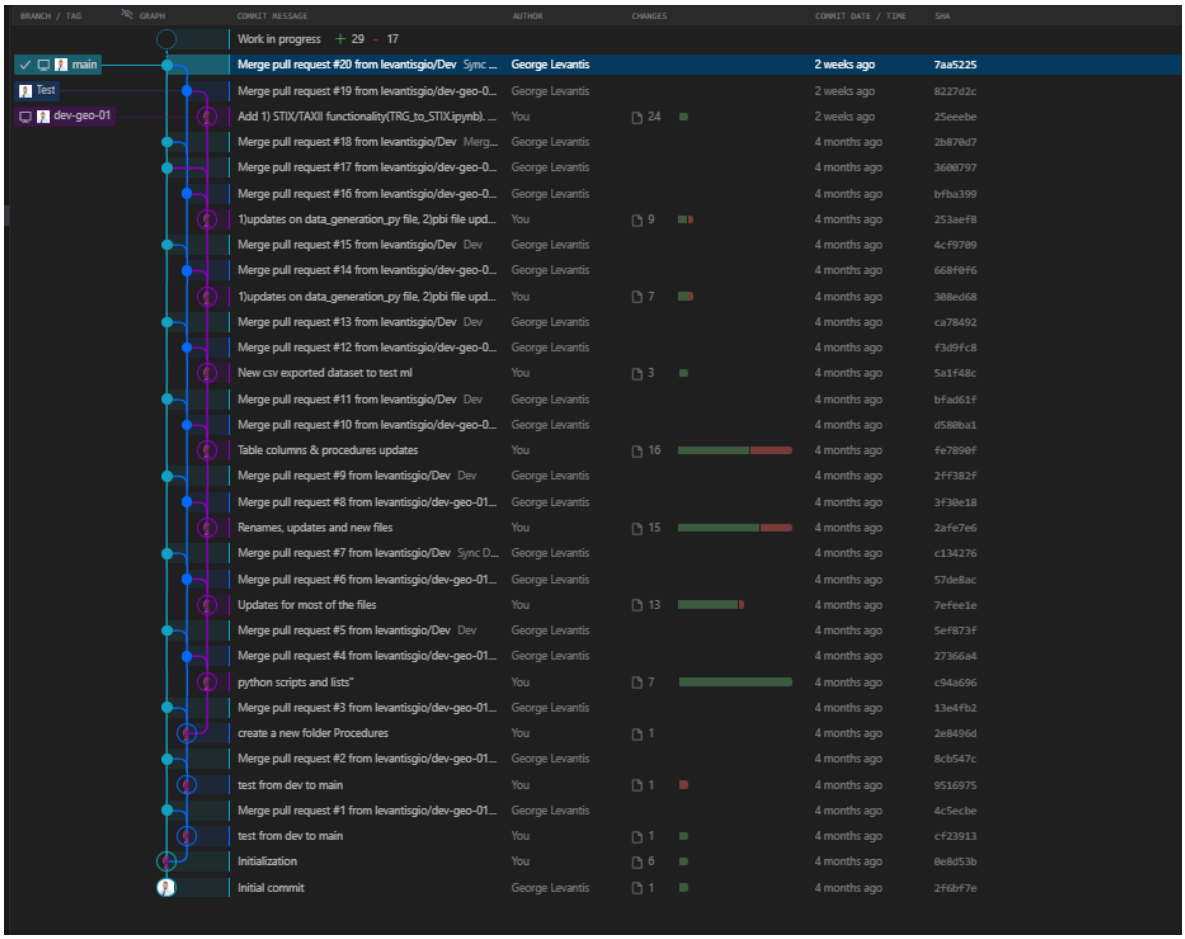
4.3.4 Πρακτική Εφαρμογή του Git και του GitHub στο Έργο

Στο πλαίσιο της ανάπτυξης του έργου μας, ακολουθήσαμε μια στρατηγική χρήσης του Git και του GitHub που περιλάμβανε τη δημιουργία και τη διαχείριση τριών κεντρικών branches: dev, test, και main. Κάθε ένα από αυτά τα branches είχε έναν συγκεκριμένο ρόλο στην ανάπτυξη και την παράδοση του έργου.

- **Dev Branch**: Το *dev-geo-01* branch είναι το κύριο branch ανάπτυξης. Όλες οι νέες δυνατότητες και αλλαγές στον κώδικα αναπτύσσονται σε ξεχωριστά feature branches και στη συνέχεια συγχωνεύονται στο remote dev-geo-01. Αυτό διασφαλίζει ότι το dev-geo-01 branch περιέχει τον πιο ενημερωμένο κώδικα, αλλά και ότι ο κώδικας αυτός έχει δοκιμαστεί και ελεγχθεί πριν προχωρήσει σε άλλα περιβάλλοντα.

- **Test Branch**: Το *Test* branch χρησιμοποιείται για δοκιμές. Αφού ολοκληρωθούν οι αλλαγές στο dev-geo-01 branch, δημιουργείται ένα pull request (PR) για τη συγχώνευση στο Test. Σε αυτό το στάδιο, ο κώδικας υποβάλλεται σε εντατικές δοκιμές για να διασφαλιστεί ότι λειτουργεί σωστά και ότι δεν υπάρχουν σφάλματα ή προβλήματα συμβατότητας. Το test branch είναι επίσης χρήσιμο για να δει η ομάδα πώς θα λειτουργήσουν οι αλλαγές σε ένα περιβάλλον που προσομοιάζει την παραγωγή.

- **Main Branch**: Το *main* branch είναι το branch παραγωγής. Μόλις οι αλλαγές περάσουν από το Test branch και επικυρωθούν, δημιουργείται ένα νέο PR για τη συγχώνευση στο main. Αυτό διασφαλίζει ότι μόνο ο δοκιμασμένος και σταθερός κώδικας εισέρχεται στο branch παραγωγής, ελαχιστοποιώντας τον κίνδυνο προβλημάτων στο live περιβάλλον.



Εικονα 17. Ιστορικό Συγχωνεύσεων και Αιτημάτων Pull Request

Συγκεκριμένα, η *Εικόνα 17* απεικονίζει το ιστορικό των *pull requests* και των συγχωνεύσεων (merges) που έγιναν στο αποθετήριο.

Τα βασικά σημεία της εικόνας:

- **Κλαδιά (Branches)**:

Υπάρχουν τρία κύρια κλαδιά (branches) που εμφανίζονται στην αριστερή πλευρά της εικόνας: το main, το dev-geo-01, και το test. Αυτά τα κλαδιά αντιπροσωπεύουν διαφορετικά στάδια ανάπτυξης του έργου. Το main είναι συνήθως το σταθερό κλαδί που περιέχει την τελική έκδοση του κώδικα, ενώ τα υπόλοιπα κλαδιά είναι για δοκιμές ή νέα χαρακτηριστικά.

- **Pull Requests**:

Τα pull requests (π.χ., "Merge pull request #20 from levantsig/dev-sync") υποδεικνύουν προσπάθειες να ενσωματωθεί κώδικας από ένα δευτερεύον κλαδί στο κεντρικό κλαδί (main). Κατά την ανάπτυξη, οι προγραμματιστές κάνουν αλλαγές στα δευτερεύοντα κλαδιά και ζητούν να συγχωνευθούν αυτές οι αλλαγές στο κύριο κλαδί όταν ολοκληρωθούν και ελεγχθούν.

- **Συγγραφείς και Επεξεργασίες**:

Ο μοναδικός συνεισφέρων στο repository είναι ο "George Levantis", ο οποίος έχει πραγματοποιήσει πολλές από τις αλλαγές. Κάθε pull request περιέχει και πληροφορίες για το SHA (ένα μοναδικό αναγνωριστικό για κάθε commit) και τον αριθμό των αλλαγών (lines of code) που έγιναν σε κάθε commit.

- **Merge Workflow**:

Τα pull requests ελέγχονται, και στη συνέχεια γίνονται merge (συγχωνεύονται). Αυτό είναι σημαντικό γιατί διασφαλίζει ότι ο κώδικας που ενσωματώνεται στο κεντρικό κλαδί έχει περάσει από κάποιον έλεγχο και είναι συμβατός με τον υπόλοιπο κώδικα. Στην εικόνα εμφανίζονται αρκετά merge από το κλαδί levantsig/dev-geo-01 στο main, που δείχνει ότι ο κύριος κλάδος ενημερώνεται τακτικά με τις νέες αλλαγές.

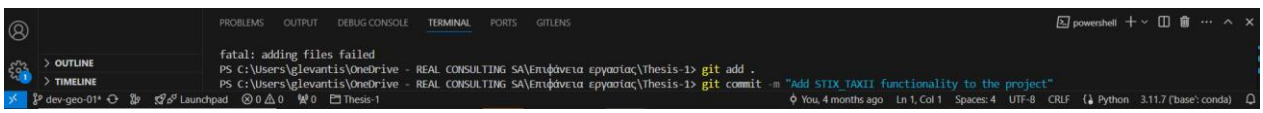
- **Pull Request Status**:

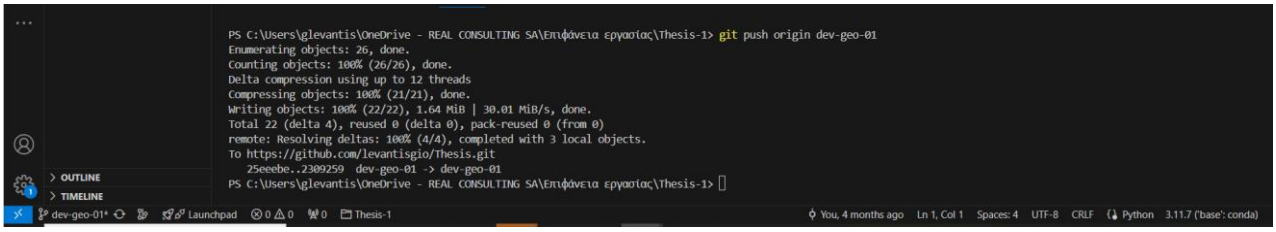
Στην εικόνα βλέπουμε ότι τα περισσότερα pull requests είναι ολοκληρωμένα ("Merge pull request #"). Μερικά από αυτά περιλαμβάνουν σημαντικές ενημερώσεις στον κώδικα, όπως προσθήκες στη λειτουργικότητα STIX/TAXII.

- **Γραφική Απεικόνιση της Ιστορίας των Κλαδιών**:

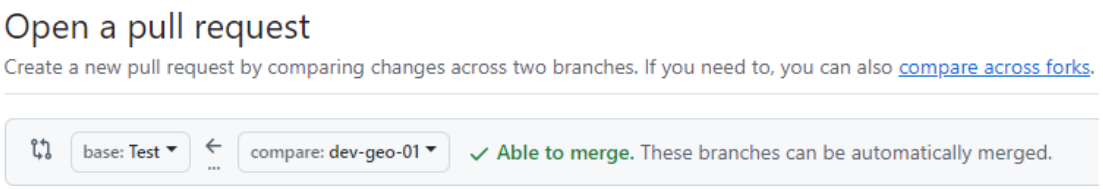
Η γραμμή στην αριστερή πλευρά δείχνει την πορεία των κλαδιών και τις συγχωνεύσεις τους. Οι κύκλοι και οι συνδέσεις μεταξύ των κλαδιών δείχνουν πότε έγιναν αλλαγές και πώς αυτές συνδέονται.

Η στρατηγική που ακολουθήσαμε για τα pull requests ήταν επίσης σημαντική για τη διαχείριση του κώδικα. Κάθε αλλαγή σε ένα feature branch υποβαλλόταν πρώτα σε αναθεώρηση μέσω PR. Οι αναθεωρήσεις αυτές επέτρεπαν σε άλλους προγραμματιστές να εξετάσουν τις αλλαγές, να προτείνουν βελτιώσεις, και να διασφαλίσουν ότι ο νέος κώδικας ήταν συμβατός με τον υπάρχοντα.

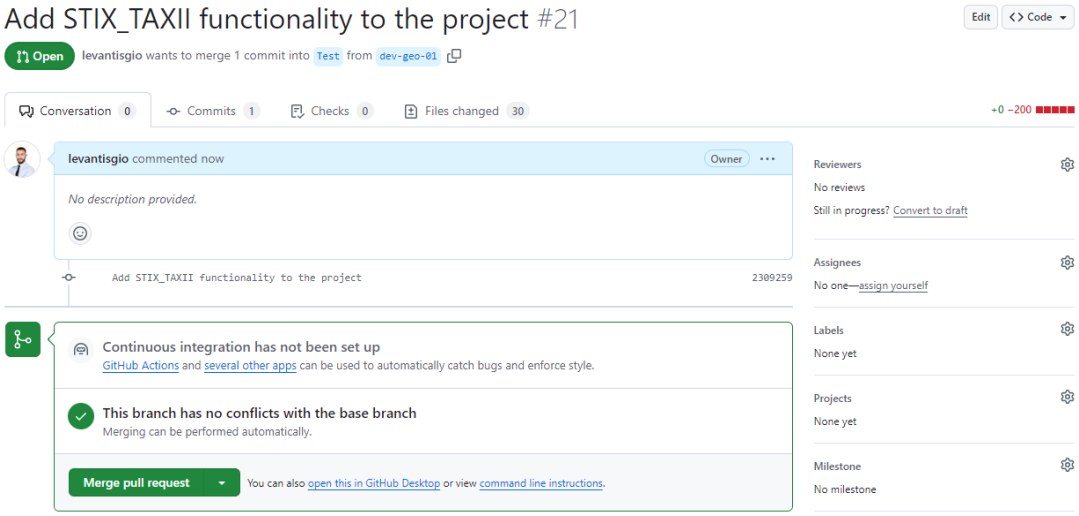




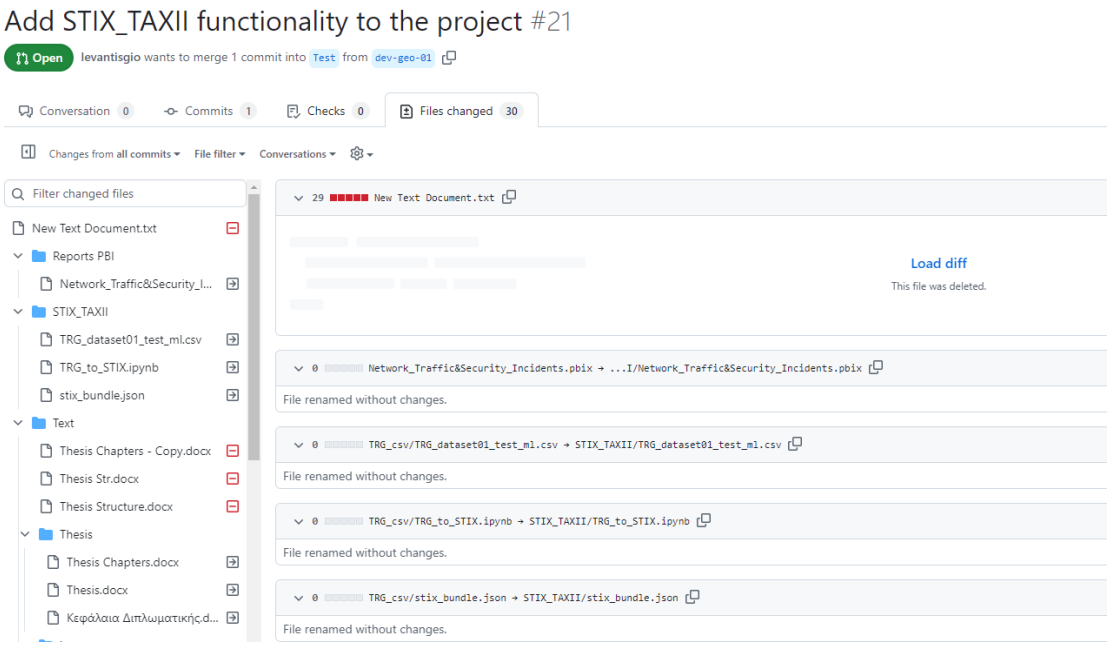
Εικονα 18. Εντολές Git Add, Commit, Push σε Repository με χρήση του branch dev-geo-01



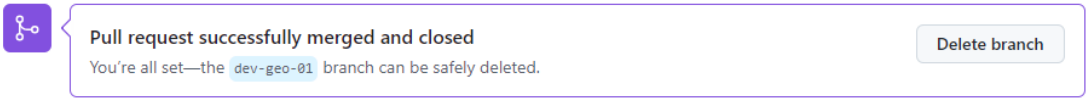
Εικονα 19. Δημιουργία Pull Request για Συγχώνευση του Branch dev-geo-01 με το Test



Εικονα 20. Pull Request για προσθήκη STIX_TAXII στο Έργο



Εικονα 21. Αλλαγές Αρχείων στο Pull Request για προσθήκη STIX_TAXII



Εικονα 22. Επιτυχής Συγχώνευση και Κλείσιμο του Pull Request

Μετά την έγκριση, το PR συγχωνευεται στο test branch, όπου ο κώδικας ενσωματωνεται με άλλες αλλαγές. Στη συνέχεια, οι αλλαγές αυτές προωθούνται σταδιακά και στο main με τον ίδιο τρόπο, εξασφαλίζοντας μια σταθερή και ελεγχόμενη ροή ανάπτυξης από το στάδιο της αρχικής ανάπτυξης έως την τελική παράδοση.

Αυτή η διαδικασία μας επιτρέπει να διατηρούμε έναν οργανωμένο και διαχειρίσιμο κύκλο ανάπτυξης, να εντοπίζουμε και να διορθώνουμε προβλήματα πριν φτάσουν στο production, και να διασφαλίζουμε ότι το έργο θα παραδοθεί με την υψηλότερη δυνατή ποιότητα.

4.4 Χρήση του Power BI για Οπτικοποίηση Δεδομένων

4.4.1 Εισαγωγή στο Power BI

Το Power BI είναι ένα ισχυρό εργαλείο της Microsoft για την ανάλυση και οπτικοποίηση δεδομένων, το οποίο επιτρέπει στους χρήστες να δημιουργούν διαδραστικές αναφορές και dashboards. Σχεδιάστηκε για να βοηθήσει τις επιχειρήσεις και τους επαγγελματίες να μετατρέπουν ακατέργαστα δεδομένα σε πολύτιμες πληροφορίες, παρέχοντας μια εύχρηστη πλατφόρμα για την κατανόηση και ανάλυση των δεδομένων.

Ένα από τα κύρια πλεονεκτήματα του είναι η ικανότητά του να συνδέεται με πολλές και διαφορετικές πηγές δεδομένων, όπως βάσεις δεδομένων SQL, αρχεία Excel, υπηρεσίες cloud, και άλλες πηγές δεδομένων. Μόλις τα δεδομένα εισαχθούν στο Power BI, μπορούν να υποστούν

επεξεργασία, να αναλυθούν και να παρουσιαστούν μέσω μιας σειράς από γραφήματα, πίνακες και άλλες οπτικοποιήσεις.

Το Power BI διακρίνεται για την ενσωμάτωση προηγμένων λειτουργιών ανάλυσης, όπως οι DAX (Data Analysis Expressions) για την πραγματοποίηση περίπλοκων υπολογισμών και την ανάλυση των δεδομένων σε βάθος. Επιπλέον, προσφέρει δυνατότητες για την εφαρμογή προσαρμοσμένων φίλτρων και την αλληλεπίδραση με τα δεδομένα σε πραγματικό χρόνο, γεγονός που ενισχύει την ευελιξία και την αποτελεσματικότητα της ανάλυσης.

Η χρήση του δεν περιορίζεται μόνο σε ειδικούς της πληροφορικής. Αντιθέτως, η φιλική προς τον χρήστη διεπαφή του το καθιστά προσιτό σε επαγγελματίες από διάφορους τομείς, επιτρέποντας σε άτομα χωρίς ιδιαίτερες τεχνικές γνώσεις να δημιουργούν αποτελεσματικές και εντυπωσιακές οπτικοποιήσεις δεδομένων. Αυτό το χαρακτηριστικό το καθιστά ένα δημοφιλές εργαλείο για τη λήψη αποφάσεων βάσει δεδομένων σε οργανισμούς κάθε μεγέθους.

Ένα άλλο σημαντικό χαρακτηριστικό του είναι η δυνατότητα κοινής χρήσης αναφορών και dashboards με άλλα μέλη της ομάδας ή του οργανισμού, είτε μέσω του cloud είτε μέσω ενσωμάτωσης σε άλλες εφαρμογές της Microsoft, όπως το Teams ή το SharePoint. Αυτό προωθεί τη συνεργασία και επιτρέπει την εύκολη πρόσβαση στις πληροφορίες που είναι σημαντικές για την επιχείρηση.

Επίσης παρέχει δυνατότητες για την ανάπτυξη προσαρμοσμένων λύσεων μέσω της χρήσης του Power BI API, που επιτρέπει την ενσωμάτωση του Power BI σε άλλες εφαρμογές και την αυτοματοποίηση των διαδικασιών ανάλυσης δεδομένων. Αυτή η δυνατότητα ενισχύει τη συνολική ευελιξία και προσαρμοστικότητα του εργαλείου, καθιστώντας το ιδανικό για την κάλυψη των αναγκών κάθε οργανισμού.

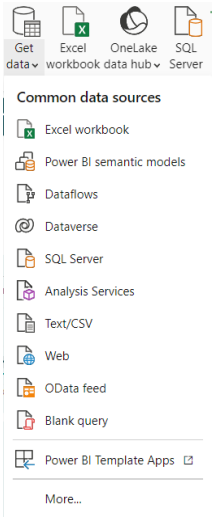
Σε γενικές γραμμές, το Power BI είναι ένα ισχυρό εργαλείο που μπορεί να χρησιμοποιηθεί από επιχειρήσεις για τη βελτιστοποίηση της διαδικασίας λήψης αποφάσεων, την ενίσχυση της κατανόησης των δεδομένων και τη βελτίωση της απόδοσης μέσω της οπτικοποίησης και ανάλυσης των δεδομένων.

4.4.2 Σύνδεση του Power BI με SQL Βάσεις Δεδομένων

Η σύνδεση του Power BI με SQL βάσεις δεδομένων είναι μια από τις πιο ισχυρές και ευρέως χρησιμοποιούμενες δυνατότητες του εργαλείου, επιτρέποντας στους χρήστες να αποκτούν πρόσβαση σε μεγάλα σύνολα δεδομένων και να τα αναλύουν με εύκολο και αποδοτικό τρόπο. Η διαδικασία σύνδεσης είναι απλή και μπορεί να πραγματοποιηθεί σε λίγα βήματα, διευκολύνοντας τους χρήστες να αρχίσουν να δουλεύουν με τα δεδομένα τους.

Για να συνδεθείτε σε μια SQL βάση δεδομένων στο Power BI, ακολουθήστε τα εξής βήματα:

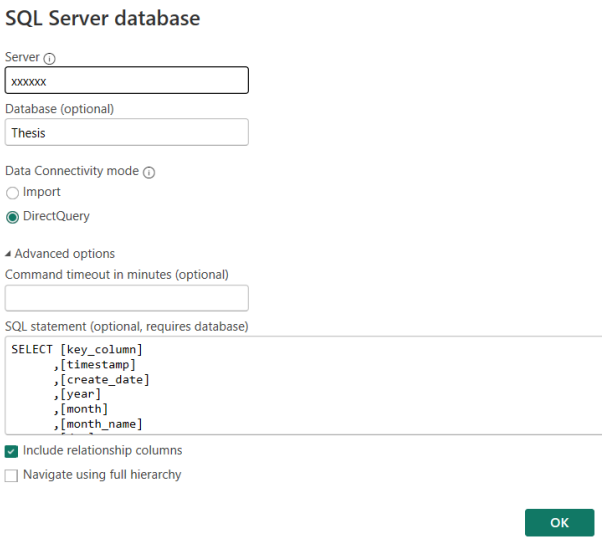
1. *Επιλογή Πηγής Δεδομένων*: Από την αρχική σελίδα του Power BI Desktop, επιλέξτε την επιλογή "Get Data" και στη συνέχεια επιλέξτε "SQL Server" από τη λίστα διαθέσιμων πηγών δεδομένων. Αυτή η ενέργεια θα ανοίξει ένα νέο παράθυρο όπου θα πρέπει να εισάγετε τις λεπτομέρειες της σύνδεσης.



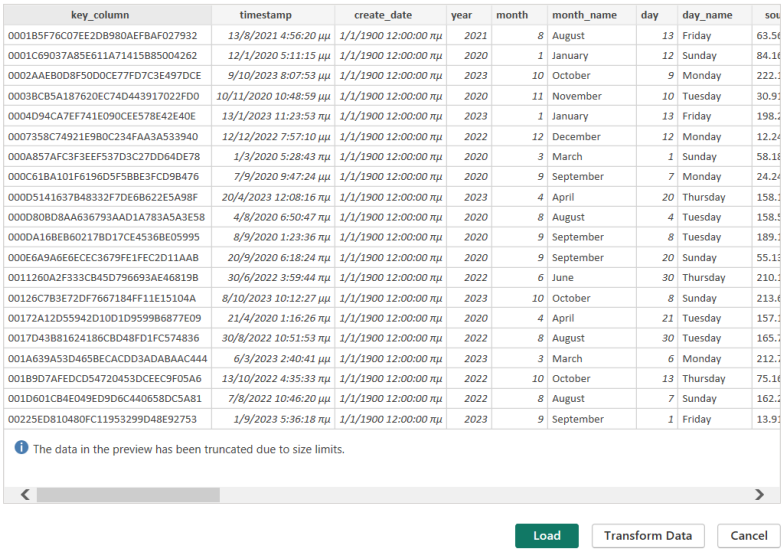
Εικονα 22. Επιλογές Πηγών Δεδομένων στο Power BI

2. *Εισαγωγή Στοιχείων Σύνδεσης:* Στο παράθυρο σύνδεσης, θα πρέπει να εισάγετε το όνομα του διακομιστή SQL (SQL Server Name) και το όνομα της βάσης δεδομένων (Database Name) στην οποία θέλετε να συνδεθείτε. Εάν ο διακομιστής SQL απαιτεί διαπιστευτήρια για την πρόσβαση, θα πρέπει να τα εισάγετε επίσης σε αυτό το στάδιο. Το Power BI υποστηρίζει τόσο τον έλεγχο ταυτότητας των Windows όσο και τον SQL Server έλεγχο ταυτότητας.

3. *Επιλογή Μεθόδου Σύνδεσης:* Αφού εισάγετε τα απαραίτητα στοιχεία, επιλέξτε τη μέθοδο σύνδεσης που επιθυμείτε. Μπορείτε να επιλέξετε είτε "Import" για να φορτώσετε τα δεδομένα στο Power BI και να τα επεξεργαστείτε τοπικά, είτε "DirectQuery" για να συνδεθείτε απευθείας στη βάση δεδομένων και να πραγματοποιείτε ερωτήματα σε πραγματικό χρόνο χωρίς να εισάγετε τα δεδομένα.



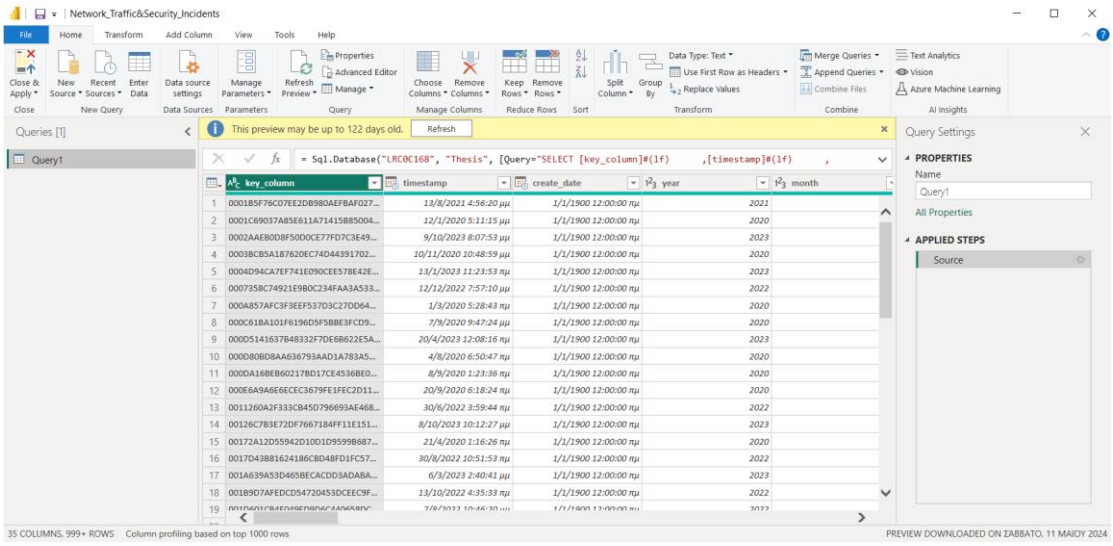
Εικονα 23. Σύνδεση με SQL Server



4. *Επιλογή Πινάκων και Δεδομένων:* Μετά τη σύνδεση, το Power BI θα εμφανίσει μια λίστα με τους διαθέσιμους πίνακες και τις προβολές στη βάση δεδομένων. Από αυτή τη λίστα, μπορείτε να επιλέξετε τους πίνακες και τα πεδία που θέλετε να εισάγετε στο Power BI για περαιτέρω ανάλυση και οπτικοποίηση.

5. *Δημιουργία Σχέσεων:* Αφού εισάγετε τα δεδομένα στο Power BI, μπορείτε να δημιουργήσετε σχέσεις μεταξύ των πινάκων, εάν αυτό είναι απαραίτητο, για να διασφαλίσετε ότι τα δεδομένα συνδέονται σωστά και μπορούν να αναλυθούν με ακρίβεια. Το Power BI παρέχει εργαλεία για την εύκολη δημιουργία και διαχείριση αυτών των σχέσεων.

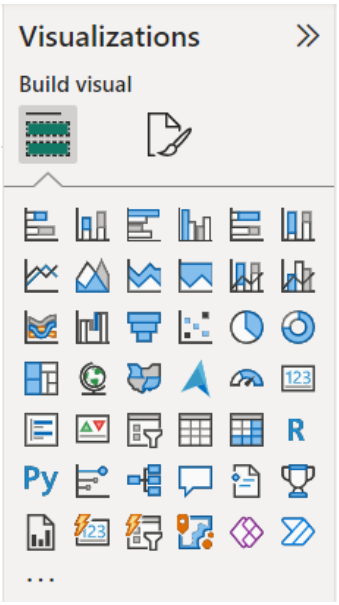
6. *Φιλτράρισμα και Μετασχηματισμός Δεδομένων:* Το Power BI προσφέρει εργαλεία για τον φιλτράρισμα και τον μετασχηματισμό των δεδομένων που εισάγονται από την SQL βάση δεδομένων. Μπορείτε να καθαρίσετε, να φιλτράρετε και να μετασχηματίσετε τα δεδομένα πριν από την ανάλυση, εξασφαλίζοντας έτσι ότι τα δεδομένα είναι έτοιμα για χρήση.



Εικονα 24. Προεπισκόπηση Δεδομένων στο Power Query Editor από SQL Server

7. *Δημιουργία Οπτικοποιήσεων:* Μετά την εισαγωγή και την επεξεργασία των δεδομένων, μπορείτε να αρχίσετε να δημιουργείτε οπτικοποιήσεις. Το Power BI προσφέρει μια

πληθώρα επιλογών για την οπτικοποίηση των δεδομένων, όπως γραφήματα, πίνακες, χάρτες, και πολλά άλλα, επιτρέποντας την εύκολη κατανόηση των δεδομένων.



Εικονα 25. Διαθέσιμες Οπτικοποιήσεις στο Power BI

Η σύνδεση του Power BI με SQL βάσεις δεδομένων επιτρέπει την αξιοποίηση μεγάλων όγκων δεδομένων και τη μετατροπή τους σε χρήσιμες πληροφορίες, συμβάλλοντας στην αποτελεσματική λήψη αποφάσεων και στη βελτίωση των επιχειρησιακών διαδικασιών.

4.4.3 Οπτικοποίηση Δεδομένων Κυβερνοασφάλειας στο Power BI

Η οπτικοποίηση δεδομένων κυβερνοασφάλειας είναι ένα κρίσιμο στοιχείο για την κατανόηση των απειλών, την αξιολόγηση των κινδύνων και τη λήψη αποφάσεων σχετικά με την ασφάλεια ενός οργανισμού. Το Power BI προσφέρει τα κατάλληλα εργαλεία και δυνατότητες για την ανάλυση και την οπτικοποίηση δεδομένων κυβερνοασφάλειας με τρόπο που βοηθά τους επαγγελματίες της ασφάλειας να αντιληφθούν τις τάσεις, τις ανωμαλίες και τις επιθέσεις.

Συλλογή και Προετοιμασία Δεδομένων:

Το πρώτο βήμα για την οπτικοποίηση των δεδομένων κυβερνοασφάλειας στο Power BI είναι η συλλογή των δεδομένων από διάφορες πηγές. Αυτές οι πηγές μπορεί να περιλαμβάνουν αρχεία καταγραφής (logs), αναφορές από εργαλεία ασφάλειας όπως SIEMs (Security Information and Event Management), δεδομένα από firewalls, IDS/IPS συστήματα, και άλλες πηγές δεδομένων σχετικών με την ασφάλεια. Οι πηγές δεδομένων μπορούν να συνδεθούν στο Power BI μέσω των ενσωματωμένων συνδέσμων, επιτρέποντας την εύκολη πρόσβαση και την εξαγωγή των δεδομένων είτε απευθειας στο εργαλείο είτε όπως στο case study, μέσω μίας βάσης δεδομένων.

Δημιουργία Οπτικοποιήσεων:

Μετά την προετοιμασία των δεδομένων, το επόμενο βήμα είναι η δημιουργία οπτικοποιήσεων που μπορούν να αναδείξουν τις κρίσιμες πληροφορίες και τις τάσεις στα δεδομένα κυβερνοασφάλειας. Το Power BI προσφέρει ένα ευρύ φάσμα από γραφήματα, πίνακες και άλλες μορφές οπτικοποίησης που επιτρέπουν στους χρήστες να παρουσιάζουν τα δεδομένα με τρόπο που είναι εύκολα κατανοητός και ερμηνεύσιμος.

Ανάλυση και Ερμηνεία των Δεδομένων:

Η ανάλυση των δεδομένων κυβερνοασφάλειας στο Power BI παρέχει στους οργανισμούς την ικανότητα να κατανοούν τις απειλές και να λαμβάνουν αποφάσεις για την προστασία των συστημάτων τους. Μέσω της παρακολούθησης των τάσεων, της ανίχνευσης ανωμαλιών και της ανάλυσης των προτύπων επίθεσης, οι υπεύθυνοι ασφάλειας μπορούν να αναπτύξουν στρατηγικές άμυνας που είναι προσαρμοσμένες στις συγκεκριμένες ανάγκες τους.

Το Power BI επιτρέπει επίσης την εφαρμογή προβλεπτικής ανάλυσης, όπου οι χρήστες μπορούν να χρησιμοποιούν τα ιστορικά δεδομένα για να προβλέψουν μελλοντικές επιθέσεις και να εντοπίσουν περιοχές που απαιτούν πρόσθετη προστασία. Αυτό προσφέρει στους οργανισμούς ένα σημαντικό πλεονέκτημα, καθώς μπορούν να προετοιμαστούν για πιθανές απειλές πριν αυτές εκδηλωθούν.

Η ερμηνεία των δεδομένων που παράγονται από τις οπτικοποιήσεις βοηθά στην επικοινωνία των ευρημάτων με τα ανώτερα στελέχη και τα υπόλοιπα μέλη του οργανισμού, διευκολύνοντας τη λήψη αποφάσεων βάσει δεδομένων. Με αυτόν τον τρόπο, το Power BI συνεισφέρει σημαντικά στην ενίσχυση της κυβερνοασφάλειας.

5. Μελέτη Περίπτωσης

5.1. Εισαγωγή

Η παρούσα μελέτη περίπτωσης εστιάζει στην ανάπτυξη και υλοποίηση μιας διαδικασίας ETL (Extract, Transform, Load) για δεδομένα κυβερνοασφάλειας. Η διαδικασία αυτή χρησιμοποιείται για τη συλλογή, μετασχηματισμό και φόρτωση δεδομένων από ένα συνθετικό σύνολο δεδομένων με περιστατικά κυβερνοεπιθέσεων. Το έργο αποσκοπεί στη δημιουργία μιας βάσης δεδομένων που μπορεί να υποστηρίξει αναλύσεις μηχανικής μάθησης, αναφορές επιχειρηματικής ευφυΐας και τη διανομή πληροφοριών μέσω των προτύπων STIX/TAXII.

Το έργο επιδιώκει τη δημιουργία μιας κεντρικής βάσης δεδομένων που να υποστηρίζει την κατανόηση των επιθέσεων και τη διαχείριση απειλών με ακρίβεια. Μέσω της ανάλυσης και της οπτικοποίησης στο Power BI, όπου και η μελέτη περίπτωσης επικεντρώνεται, οι αναλυτές ασφάλειας μπορούν να λάβουν άμεσα πληροφορίες για τις τάσεις, τη σοβαρότητα και την κατανομή των επιθέσεων, προσφέροντας καλύτερη λήψη αποφάσεων και γρήγορη ανταπόκριση σε περιστατικά ασφαλείας.

5.2. Ανάλυση του Προβλήματος

Στο πλαίσιο των σημερινών απειλών στον κυβερνοχώρο, η ανάγκη για αυτοματοποιημένη διαχείριση μεγάλων ποσοτήτων δεδομένων ασφαλείας είναι ιδιαίτερα έντονη. Τα δεδομένα που προέρχονται από συστήματα κυβερνοασφάλειας, όπως IDS (Intrusion Detection Systems), firewalls, και logs από διάφορες πηγές, είναι συχνά πολύπλοκα και ογκώδη. Η φύση αυτών των δεδομένων απαιτεί αποτελεσματικά εργαλεία ETL για την επεξεργασία και αξιοποίησή τους, προκειμένου να διευκολυνθεί η αναγνώριση μοτίβων επιθέσεων, η κατηγοριοποίηση απειλών και η λήψη ενημερωμένων αποφάσεων για την ασφάλεια.

Η πρόκληση έγκειται στη διαχείριση και ανάλυση αυτών των δεδομένων σε πραγματικό χρόνο ή σχεδόν σε πραγματικό χρόνο, ειδικά όταν προέρχονται από πολλαπλές πηγές και σε διαφορετικές μορφές. Τα δεδομένα μπορεί να περιλαμβάνουν χρονικές σημάνσεις, διευθύνσεις IP, τύπους επιθέσεων, καθώς και πληροφορίες για τη σοβαρότητα των περιστατικών. Χωρίς μια δομημένη και αυτοματοποιημένη διαδικασία ETL, τα δεδομένα αυτά παραμένουν ακατέργαστα και δύσκολα στην αξιοποίηση, περιορίζοντας την αποτελεσματικότητα τόσο στην εκπαίδευση μοντέλων μηχανικής μάθησης όσο και στις επιχειρησιακές αναλύσεις.

Τέλος, υπάρχει η ανάγκη για οπτικοποίηση αυτών των δεδομένων σε εργαλεία όπως το Power BI, ώστε οι αναλυτές να μπορούν να δουν σε πραγματικό χρόνο τα μοτίβα επιθέσεων, να εντοπίσουν γεωγραφικά hot-spots και να πάρουν άμεσα αποφάσεις για την άμυνα των συστημάτων τους. Αυτό καθιστά την οπτικοποίηση δεδομένων απαραίτητο στοιχείο στην έγκαιρη αντίδραση και αποτροπή απειλών, ειδικά σε οργανισμούς με μεγάλο αριθμό περιστατικών ασφάλειας καθημερινά.

5.3. Λύση

Η λύση περιλαμβάνει την ανάπτυξη ενός συστήματος ETL, που αποτελείται από τρία κύρια στάδια:

- Εισαγωγή Δεδομένων (Extract Phase): Τα δεδομένα εισάγονται από ένα αρχείο CSV που περιλαμβάνει 40.000 εγγραφές και 25 στήλες, με πληροφορίες όπως διευθύνσεις IP, τύπους επιθέσεων και σοβαρότητες περιστατικών. Αυτά αποθηκεύονται στο "Landing schema" για μελλοντική αναφορά χωρίς επεξεργασία.

- Μετασχηματισμός Δεδομένων (Transform Phase): Τα δεδομένα μετασχηματίζονται και περνουν στο "Staging schema", μέσω διαδικασιών καθαρισμού και εμπλουτισμού. Η φάση αυτή περιλαμβάνει την αφαίρεση διπλότυπων, τη διόρθωση σφαλμάτων και την προσθήκη γεωγραφικών δεδομένων. Επιπλέον, δημιουργούνται μοναδικά κλειδιά για κάθε περιστατικό με χρήση αλγορίθμων hash.

- Φόρτωση Δεδομένων (Load Phase): Τα δεδομένα που μετασχηματίστηκαν αποθηκεύονται σε ένα "Target schema", έτοιμα για χρήση σε εφαρμογές όπως το Power BI.

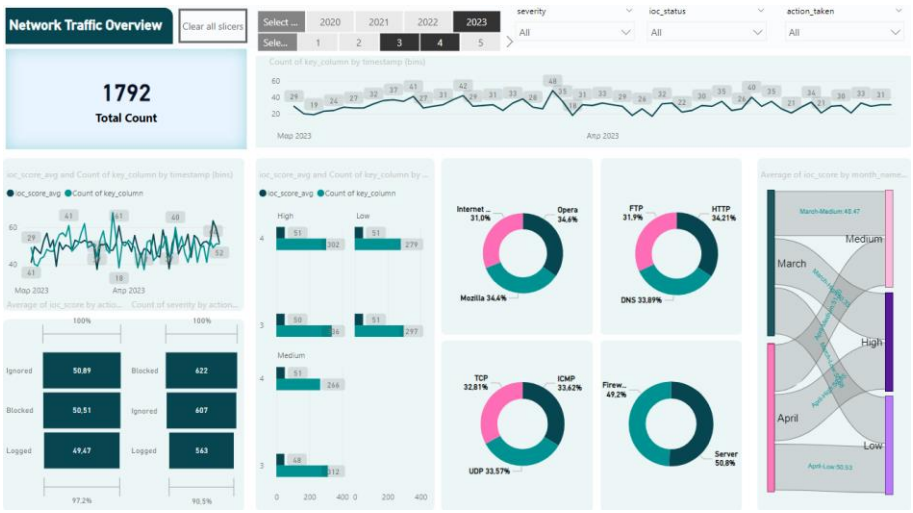
5.4. Χρήση του Power BI για Οπτικοποίηση Δεδομένων

Το Power BI αποτέλεσε το βασικό εργαλείο για την οπτικοποίηση των δεδομένων κυβερνοασφάλειας. Μέσω διαδραστικών dashboards, υπάρχει δυνατότητα να αναλυθούν μεγάλες ποσότητες δεδομένων με αποτελεσματικότητα.

5.4.1 Διαδραστικά Dashboards

Τα παρακάτω παραδείγματα παρουσιάζουν τις δυνατότητες της οπτικοποίησης στο Power BI:

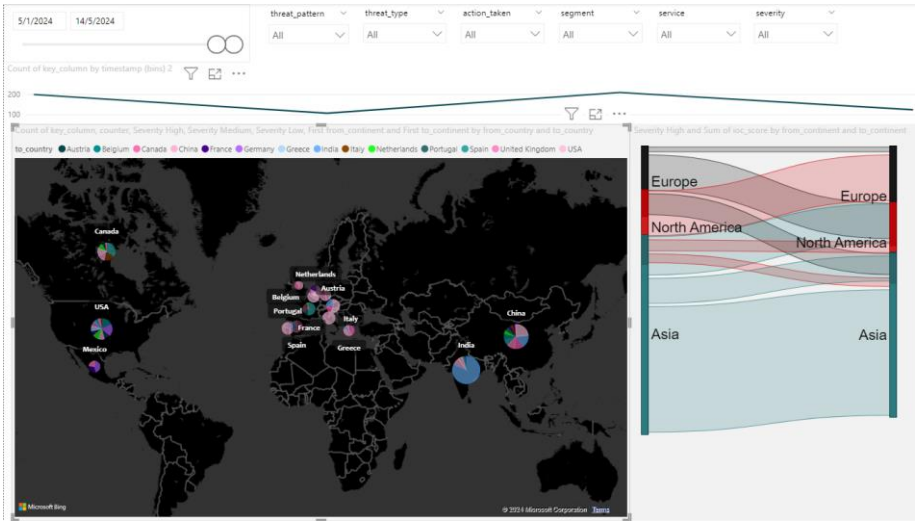
1. *Γραφήματα Ανάλυσης Τάσεων*: Στην *Εικόνα 26*, στο πρώτο dashboard, παρουσιάζει μια συνολική εικόνα του δικτυακού traffic και τον αριθμό και την κατανομή των περιστατικών με βάση τη σοβαρότητα και άλλες μεταβλητές, όπως το action taken (π.χ. blocked, ignored). Επίσης, φαίνονται τάσεις traffic ανά μήνα, καθώς και αναλύσεις για τα διάφορα πρωτόκολλα που εμπλέκονται (TCP, ICMP, UDP). Το dashboard αυτό επιτρέπει στους αναλυτές ασφαλείας να εντοπίζουν αυξήσεις στη δραστηριότητα δικτύου με βάση τον χρόνο και να παρακολουθούν την αποτελεσματικότητα των δράσεων (blocked, ignored, logged). Οι πίνακες με τα πρωτόκολλα βοηθούν στον εντοπισμό πιθανών τρωτών σημείων (π.χ. αυξημένη χρήση UDP ή TCP μπορεί να συνδεθεί με συγκεκριμένες επιθέσεις).



Εικονα 26. Ανάλυση Περιστατικών Ασφάλειας: Συνολικός Αριθμός και Κατανομή Απειλών

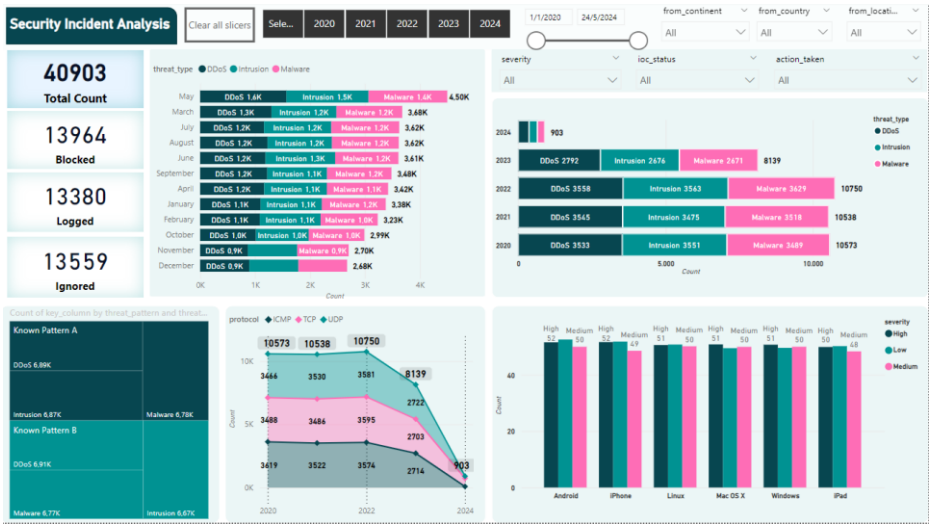
2. **Χάρτες Γεωγραφικής Κατανομής Επιθέσεων**: Στην *Εικόνα 27*, παρατηρούμε την κατανομή των περιστατικών ασφάλειας σε γεωγραφικό επίπεδο. Στον χάρτη απεικονίζονται οι χώρες και οι ήπειροι όπου σημειώθηκαν οι περισσότερες επιθέσεις, ενώ

στο δεξί μέρος υπάρχει ένα διάγραμμα ροής που δείχνει τις εισερχόμενες και εξερχόμενες ροές traffic ανά ήπειρο.



Εικονα 27 Γεωγραφική Κατανομή Επιθέσεων και Ροές Κυβερνοαπειλών ανά Ήπειρο

3. **Πίνακες KPIs**: Στην *Εικόνα 28*, Το τρίτο dashboard επικεντρώνεται στον αριθμό των περιστατικών ασφαλείας, διαχωρισμένων ανά τύπο επίθεσης (π.χ. DDoS, Intrusion, Malware) και μετρώντας δράσεις όπως blocked, logged, ignored. Δείχνει επίσης την κατανομή των περιστατικών με βάση συσκευές (Android, iPhone, Linux κτλ.) και πρωτόκολλα. Εδώ φαίνονται οι κορυφαίες απειλές που αντιμετωπίζονται (π.χ. intrusion, malware), και πώς αυτές καταγράφονται ή απορρίπτονται (blocked, ignored).



Εικονα 28. Ανάλυση Απειλών Κυβερνοασφάλειας

5.4.2 Οφέλη της Οπτικοποίησης Δεδομένων

Τα οφέλη από τη χρήση του Power BI σε αυτή την εργασία περιλαμβάνουν:

- *Ενισχυμένη Κατανόηση των Απειλών:* Τα διαδραστικά dashboards παρέχουν ένα σαφές και ολοκληρωμένο overview των δεδομένων, διευκολύνοντας την κατανόηση των απειλών που αντιμετωπίζει ο οργανισμός.

- *Ταχεία Ανίχνευση και Απόκριση σε Απειλές:* Με την ικανότητα άμεσης ενημέρωσης των δεδομένων, οι υπεύθυνοι ασφάλειας μπορούν να ανιχνεύουν και να αντιδρούν σε απειλές με ταχύτητα και ακρίβεια.

- *Δυνατότητες Προβλεπτικής Ανάλυσης:* Η ενσωμάτωση ιστορικών δεδομένων και η συνεχή ενημέρωσή τους επιτρέπει την ανάπτυξη προβλεπτικών μοντέλων για τον εντοπισμό μελλοντικών απειλών.

- *Βελτίωση της Συνεργασίας:* Η δυνατότητα κοινής χρήσης των dashboards και των αναφορών μέσω του Power BI Service διευκολύνει τη συνεργασία μεταξύ των τμημάτων του οργανισμού, διασφαλίζοντας ότι όλοι εργάζονται με τα ίδια δεδομένα και πληροφορίες. Η προσέγγιση αυτή ενισχύει την αποτελεσματικότητα της ανάλυσης δεδομένων κυβερνοασφάλειας, παρέχοντας στους οργανισμούς τη δυνατότητα να αντιδρούν γρήγορα στις κυβερνοαπειλές και να βελτιώνουν συνεχώς τις στρατηγικές τους.

5.5. Αποτελέσματα

Η υλοποίηση της λύσης ETL σε συνδυασμό με το Power BI παρέχει πολλαπλά οφέλη στους υπεύθυνους ασφάλειας και στους αναλυτές δεδομένων, βελτιώνοντας σημαντικά τη διαδικασία ανίχνευσης, ανάλυσης και απόκρισης σε κυβερνοαπειλές. Τα αποτελέσματα αυτής της προσέγγισης χωρίζονται σε τρεις βασικές κατηγορίες:

Χρονική Ανάλυση Απειλών:

Οι αναφορές αυτές επιτρέπουν την άμεση παρακολούθηση της εξέλιξης των απειλών σε συγκεκριμένες περιόδους, όπως φαίνεται από τα χρονικά γραφήματα. Αυτό παρέχει τη δυνατότητα εντοπισμού αιχμών δραστηριότητας ή ξαφνικών αυξήσεων σε περιστατικά ασφάλειας.

Ανάλυση με βάση τη Σοβαρότητα:

Η δυνατότητα φιλτραρίσματος περιστατικών με βάση τη σοβαρότητά τους (π.χ. high, medium, low) επιτρέπει στους υπεύθυνους να εστιάσουν γρήγορα στα περιστατικά υψηλής σοβαρότητας που απαιτούν άμεση παρέμβαση και ανάλυση. Για παράδειγμα, στο dashboard εμφανίζεται η κατανομή των περιστατικών ανά σοβαρότητα, επιτρέποντας την ιεράρχηση των ενεργειών αντιμετώπισης.

Οπτικοποίηση με Γεωγραφικά Δεδομένα:

Οι χάρτες γεωγραφικής κατανομής, όπως αυτοί που παρουσιάστηκαν στο παραπάνω παράδειγμα δίνουν στους αναλυτές μια σαφή εικόνα των επιθέσεων που προέρχονται από διάφορες χώρες ή περιοχές, διευκολύνοντας την εντολή εστιασμένων αντιμέτρων και την ανάπτυξη στρατηγικών άμυνας ανά γεωγραφική περιοχή.

6. Συμπεράσματα

6.1 Περίληψη των Κύριων Ευρημάτων

Στο πλαίσιο αυτής της εργασίας, πραγματοποιήθηκε ανάλυση των ιστορικών και σύγχρονων λύσεων ETL με έμφαση στα δεδομένα κυβερνοασφάλειας. Η μελέτη επικεντρώθηκε ιδιαίτερα στην ανάλυση δεδομένων που σχετίζονται με απειλές, όπως πληροφορίες από περιστατικά επιθέσεων, καταγραφές σε συστήματα ασφαλείας, και γεωγραφικά δεδομένα των επιθέσεων. Το αποθετήριο στο οποίο έχουν αναρτηθεί όλα τα script κώδικα (repository) βρίσκεται στο GitHub [17]. Μέσα από την ανάπτυξη και εφαρμογή συγκεκριμένων μεθοδολογιών, όπως η ενσωμάτωση των προτύπων STIX/TAXII, η χρήση του Git και του GitHub για τη διαχείριση έργων ETL, και η οπτικοποίηση δεδομένων με το Power BI, προέκυψαν σημαντικά ευρήματα που αξίζει να συνοψιστούν.

Ευελιξία της 3-layer αρχιτεκτονικής:

Ένα από τα κύρια σημεία που αναδείχθηκαν κατά την ανάλυση ήταν η ευελιξία που παρέχει η 3-layer αρχιτεκτονική στη διαδικασία ETL. Αυτή η προσέγγιση χωρίζει τη ροή των δεδομένων σε τρία στάδια: Landing schema, Staging schema και Target schema, διασφαλίζοντας την αποτελεσματική αποθήκευση, επεξεργασία και διανομή των δεδομένων. Η ευελιξία αυτής της αρχιτεκτονικής επιτρέπει στους οργανισμούς να διαχειρίζονται μεγάλες ποσότητες δεδομένων με διαφάνεια, ενώ τους προσφέρει τη δυνατότητα να προσαρμόσουν και να βελτιώσουν τη ροή δεδομένων με βάση τις επιχειρησιακές τους ανάγκες.

Οπτικοποίηση δεδομένων με το Power BI:

Η ενσωμάτωση του Power BI προσέφερε σημαντικά πλεονεκτήματα στον τομέα της οπτικοποίησης και της κατανόησης των δεδομένων κυβερνοασφάλειας. Η χρήση διαδραστικών dashboards δίνει τη δυνατότητα στους υπεύθυνους ασφάλειας να αναλύουν μεγάλα και πολύπλοκα σύνολα δεδομένων. Για παράδειγμα, μέσα από την οπτικοποίηση των απειλών και της γεωγραφικής κατανομής τους, έγινε σχετικά εύκολος ο εντοπισμός των περιοχών με αυξημένη δραστηριότητα και απειλές.

Η δυνατότητα άμεσου φιλτραρίσματος και ανάλυσης σε βάθος δίνει στις ομάδες ασφάλειας τη δυνατότητα να εντοπίζουν τάσεις και να λαμβάνουν ενημερωμένες αποφάσεις γρήγορα. Αυτό όχι μόνο βοηθά στην κατανόηση των υπαρχουσών απειλών, αλλά επίσης επιτρέπει την αναγνώριση νέων μοτίβων επιθέσεων, βελτιώνοντας την πρόληψη μελλοντικών απειλών.

Επιπλέον, η χρήση διαδραστικών γραφημάτων και χαρτών διευκολύνει τη δημιουργία αναφορών που μπορούν εύκολα να κατανοηθούν και να μοιραστούν με τα ανώτερα στελέχη του οργανισμού. Αυτή η δυνατότητα διευκολύνει τη λήψη αποφάσεων, βασισμένων σε δεδομένα, από όλα τα επίπεδα της διοίκησης.

Ενσωμάτωση των Προτύπων STIX/TAXII:

Η αποτελεσματικότητα της χρήσης των προτύπων STIX/TAXII για την ανταλλαγή πληροφοριών απειλών. Η χρήση του STIX επέτρεψε την τυποποίηση και τη δομημένη αναπαράσταση πληροφοριών απειλών, ενώ το TAXII θα διευκόλυνε την ασφαλή μεταφορά αυτών των πληροφοριών μεταξύ συστημάτων και οργανισμών. Οι οργανισμοί που ενσωματώνουν αυτά τα πρότυπα μπορούν να επιτύχουν καλύτερη συνεργασία, ταχύτερη απόκριση σε απειλές και βελτιωμένη ακρίβεια στην ανάλυση δεδομένων απειλών. Τα παραδείγματα από την πρακτική εφαρμογή αυτών των προτύπων στο πλαίσιο του case study, όπως η αυτόματη δημιουργία STIX

αντικειμένων μέσω του python script, ανέδειξαν τη σημαντική συμβολή της αυτοματοποίησης στη βελτίωση της αποτελεσματικότητας.

Συνεργατική Ανάπτυξη με Git και GitHub:

Η χρήση του Git και του GitHub για τη διαχείριση των έργων ETL απέδειξε τη σημασία της συνεργασίας και της αποτελεσματικής διαχείρισης εκδόσεων στον τομέα της κυβερνοασφάλειας. Μέσα από την εφαρμογή μιας καλά δομημένης στρατηγικής branching, με κεντρικά branches (dev, test, main) και διαδικασίες pull request για τη συγχώνευση αλλαγών.

6.2 Προτάσεις για Περαιτέρω Έρευνα

Για να συνεχιστεί η ερευνητική προσπάθεια που ξεκίνησε με την παρούσα διπλωματική εργασία, υπάρχουν αρκετές κατευθύνσεις που αξίζει να εξερευνηθούν περαιτέρω. Η τεχνολογία και οι μεθοδολογίες που αναπτύχθηκαν μέχρι τώρα παρέχουν ένα ισχυρό θεμέλιο, αλλά η εξέλιξη της κυβερνοασφάλειας απαιτεί συνεχή έρευνα και καινοτομία. Στην ενότητα αυτή προτείνονται δύο βασικές περιοχές στις οποίες μπορεί να εστιαστεί η μελλοντική έρευνα: η προχωρημένη οπτικοποίηση δεδομένων με τη χρήση τεχνητής νοημοσύνης (AI) και μηχανικής μάθησης (Machine Learning), καθώς και η επέκταση της έρευνας σε διαφορετικούς τομείς, όπως η υγειονομική περίθαλψη και οι χρηματοοικονομικές υπηρεσίες. Η επιπλέον διερεύνηση αυτών των θεμάτων έχει τη δυνατότητα να προσφέρει σημαντικές βελτιώσεις στην πρόληψη και την απόκριση σε κυβερνοαπειλές, ενισχύοντας την ανθεκτικότητα των οργανισμών

1. In-memory Databases και Real-Time Streaming:

Η υφιστάμενη αρχιτεκτονική βασίζεται στη σταδιακή εισαγωγή, μετασχηματισμό και φόρτωση (ETL) δεδομένων από αρχεία σε μια βάση δεδομένων SQL. Μια προτεινόμενη κατεύθυνση για περαιτέρω ανάπτυξη είναι η ενσωμάτωση τεχνολογιών in-memory databases και real-time streaming. Οι βάσεις δεδομένων in-memory επιτρέπουν τη διαχείριση και ανάλυση μεγάλων συνόλων δεδομένων σε πραγματικό χρόνο, επιταχύνοντας σημαντικά τις διαδικασίες ανάλυσης και αντίδρασης σε κυβερνοεπιθέσεις.

Επιπλέον, η ενσωμάτωση πλατφορμών για streaming δεδομένων, όπως το Apache Kafka, θα επιτρέψει τη συνεχή ροή δεδομένων σε πραγματικό χρόνο. Με αυτόν τον τρόπο, οι κυβερνοαπειλές μπορούν να εντοπιστούν και να αναλυθούν σχεδόν αμέσως μετά την εμφάνισή τους, καθιστώντας πιο αποτελεσματική την ανίχνευση και απόκριση στις επιθέσεις. Η ενσωμάτωση αυτών των τεχνολογιών θα ενισχύσει την επεκτασιμότητα της λύσης και θα διευκολύνει την προσαρμογή της σε περιβάλλοντα με υψηλές απαιτήσεις σε πραγματικό χρόνο.

2. *Προχωρημένη Οπτικοποίηση Δεδομένων με Χρήση AI και Machine Learning:*

Η οπτικοποίηση δεδομένων κυβερνοασφάλειας στο Power BI παρείχε αξιόπιστα αποτελέσματα, αλλά η ενσωμάτωση προηγμένων τεχνικών AI και machine learning μπορεί να προσφέρει ακόμα πιο πλούσιες και προληπτικές αναλύσεις. Προτείνεται η έρευνα γύρω από τη χρήση μοντέλων πρόβλεψης και ανάλυσης συναισθημάτων που μπορούν να ενσωματωθούν στα Power BI dashboards. Με αυτόν τον τρόπο, οι οργανισμοί θα μπορούσαν να προβλέπουν και να προλαμβάνουν κυβερνοεπιθέσεις με μεγαλύτερη ακρίβεια.

3. *Επέκταση της Έρευνας σε Διαφορετικούς Τομείς:*

Τέλος, προτείνεται η επέκταση της έρευνας σε διαφορετικούς τομείς, όπως η υγειονομική περίθαλψη, οι χρηματοοικονομικές υπηρεσίες, και η εκπαίδευση. Ο κάθε τομέας έχει μοναδικές ανάγκες και προκλήσεις όσον αφορά την κυβερνοασφάλεια, και η εφαρμογή των τεχνολογιών και

μεθοδολογιών που αναπτύχθηκαν σε αυτή την έρευνα θα μπορούσε να προσφέρει νέα ευρήματα και βελτιώσεις.

Αυτές οι προτάσεις για περαιτέρω έρευνα παρέχουν κατευθύνσεις για την επόμενη φάση της επιστημονικής και τεχνολογικής ανάπτυξης στον τομέα της κυβερνοασφάλειας. Η περαιτέρω έρευνα και εφαρμογή αυτών των ιδεών θα μπορούσε να προσφέρει σημαντικές βελτιώσεις στη διαχείριση και ανταπόκριση σε απειλές, συμβάλλοντας έτσι στην ενίσχυση της συνολικής ανθεκτικότητας των οργανισμών απέναντι σε κυβερνοεπιθέσεις.

Αναφορές

1. Alkis Simitsis, S. S. (n.d.). The History, Present, and Future of ETL Technology.

2. Alkis Simitsis, U. D. (n.d.). Data Integration Flows for Business Intelligence.

3. Vassiliadis, P., S. A. (2002). Conceptual modeling for ETL processes.

4. Ponniah, P. (2010). Data Warehousing Fundamentals for IT Professionals.

5. Brown, P. (2023). awesome-etl - list of ETL tools. Link: https://github.com/pawl/awesome-etl

6. Blend B (2021). Cloud Computing and Big Data Analytics: A Survey on ETL Approaches

7. Zhang, T. DevOps and Microservices in Cloud-Based ETL

8. Carlo Batini, C. C. (2009). Methodologies for data quality assessment and improvement.

9. Md. Delowar Hossain, T. S. (2023). The role of microservice approach in edge computing: Opportunities, challenges, and research directions.

10. Michael Armbrust, T. D. (2020). Delta Lake: High-Performance ACID.

11. N. Ahmed, A. L. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench.

12. N. Marz, W. J. (2015). Big Data: Principles and best practices of scalable real-time data systems

13. B.Chambers, M.Z (2018) The Definitive Guide - Big data processing made simple

14. P. Vassiliadis, A. S. (2008). Near Real Time ETL. In Springer Annals of Information Systems.

15. J. Leskovec, R.A., U.J.(2020). Mining of Massive Datasets.

16. Periyasamy, R. (2024). ETL Architecture: Design and Best Practices. Link: https://peliqan.io/blog/etl-architecture/

17. Sonia Bergamaschi, F. G. (2011). A semantic approach to ETL technologies.

18. OASIS STIX/TAXII Standards(2021). https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.pdf.

19. National Institute of Standards and Technology (NIST). Link: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf

20. Getting Started with STIX 2.1 Link: https://oasis-open.github.io/cti-documentation/stix/gettingstarted.html

21. Baudis, P. (2009). Current Concepts in Version Control Systems

22. McDonald, N, K. B. (2014). Modeling Distributed Collaboration on Github

23. Gitlab Link: https://about.gitlab.com/topics/version-control/

24. Αποθετήριο Github: Link: https://github.com/levantisgio/Thesis