UNIVERSITE HASSAN II DE CASABLANCA
Faculté des Sciences Ben M'Sik

DÉPARTEMENT DE MATHÉMATIQUES ET INFORMATIQUES

Projet Final Du Module Analyse Prédictive
MASTER DATA SCIENCE ET BIG DATA

Détection des Attaques En Temp Réel

Soutenu le : *24 / 02 / 2022*

Réalisé par : **LAGHZAOUI Hind**
BERRADA Anas
LAMHOUR Mohamed Akram
MSALEK Mohamed

Proposé par :

- **Pr. GHAZOUANI Mohamed**

Année Universitaire 2021-2022

Dédicaces

On dédie ce travail à toute nos familles. Notre parent, qui ont soutenu nous tout au long de notre vie, on vous remercier chaleureusement de notre cœur pour vos sacrifices et vos efforts.

Nos frères et sœurs, nous sommes profondément reconnaissant pour tous de l'encouragement que vous nous donnez.

A nos amis pour leur encouragement et pour tous les bons moments qu'on a vécus ensemble, on vous souhaite un avenir radieux.

A tous nos professeurs en témoignage de ma gratitude et profond respect.

Remerciements

On remercie tout d'abord ALLAH tout puissant de nous avoir donné le courage, la force et la patience d'achever ce modeste travail.

On tient à remercier notre professeur Monsieur ***GHAZOUANI Mohamed****, pour ses conseils pertinents et son encadrement.*

On remercie infiniment tous les enseignants et administrateurs à la faculté des Sciences Ben M'Sik ainsi les professeurs de département mathématique et informatique.

- *LAGHZAOUI Hind*
- *BERRADA Anas*
- *LAMHOUR Mohamed Akram*
- *MSALEK Mohamed*

Résumé

Internet et les réseaux informatiques sont devenus incontournables dans notre organisation et notre quotidien. À mesure que notre dépendance à l'égard des ordinateurs et des réseaux de communication augmente, les activités malveillantes deviennent plus fréquentes. Le trafic réseau doit être surveillé et analysé pour détecter les activités malveillantes et les attaques afin de garantir la fiabilité du réseau et la sécurité des informations des utilisateurs.

Récemment, des techniques d'apprentissage automatique ont été appliquées à la détection de cyberattaques. Les modèles d'apprentissage automatique sont capables d'extraire des similitudes et des modèles dans le trafic réseau. Contrairement aux méthodes basées sur les signatures, l'analyse manuelle n'est pas nécessaire pour extraire les schémas d'attaque. L'application d'algorithmes d'apprentissage automatique permet la construction automatisée de modèles prédictifs pour détecter les cyberattaques. Ce projet rapporte une analyse empirique de la détection des cyber-attaques à l'aide de méthodes d'apprentissage automatique.

À cette fin, nous étudions certains types courants d'attaques dans la détection des réseaux informatiques au niveau de la couche application. Pour former des modèles à l'aide de grands ensembles de données, nous collectons des données réseaux représentatifs de réseaux de production réels pour chaque type d'attaque. Notre analyse de chaque attaque comprend une étude

détaillée de la détection à l'aide de méthodes d'apprentissage automatique.

Nous travaillons également sur l'application omniprésente d'un outil très important dans plusieurs domaines, ELK stack, qui se compose de kibana, elasticsearch et logstach. Avec ses trois composants, nous pouvons réaliser une bonne visualisation des données ainsi qu'une détection des attaques, que nous exploiterons en détail tout au long du projet.

Par conséquent, ce projet fournit une analyse approfondie de la manière dont les techniques d'apprentissage automatique peuvent être utilisées pour détecter les cyberattaques lors du traitement de quantités massives de données, et c'est là que nous réalisons le rôle important que jouent les outils de big data dans plusieurs domaines. Au lieu d'un large éventail de cyberattaques, nous étudions l'application de différentes méthodes d'apprentissage automatique, y compris la classification, la détection d'anomalies au niveau de l'hôte et du réseau.

Abstract

The Internet and computer networks have become essential in our organization and our daily lives. As our dependence on computers and communication networks increases, malicious activity becomes more frequent. Network traffic should be monitored and analyzed for malicious activity and attacks to ensure network reliability and user information security.

Recently, machine learning techniques have been applied to the detection of cyberattacks. Machine learning models are able to extract similarities and patterns in network traffic. Unlike signature-based methods, manual analysis is not required to extract attack patterns. The application of machine learning algorithms enables the automated construction of predictive models to detect cyberattacks. This project reports an empirical analysis of the detection of cyber-attacks using machine learning methods.

To this end, we study some common types of attacks in detecting computer networks at the application layer. To train models using large datasets, we collect network data representative of real production networks for each type of attack. Our analysis of each attack includes a detailed study of detection using machine learning methods.

We are also working on the ubiquitous application of a very important tool in several fields, ELK stack, which consists of

kibana, elasticsearch and logstach. With its three components, we can achieve good data visualization as well as attack detection, which we will exploit in detail throughout the project.

Therefore, this project provides an in-depth analysis of how machine learning techniques can be used to detect cyberattacks when processing massive amounts of data, and this is where we realize the important role tools play. big data in several areas. Instead of a wide range of cyberattacks, we investigate the application of different machine learning methods, including classification, host and network level anomaly detection.

Table des matières

Introduction Générale

Les attaques dans un réseau informatique sont des actions volontaires et malveillantes menées au moyen d'un réseau informatique visant à causer un dommage aux informations et aux personnes qui les traitent. Et tout le monde peut en être la cible : les particuliers, les entreprises, les institutions, les services administratifs et de santé...

Le problème rencontré par la technologie actuelle de détection d'intrusion sur le réseau est de savoir comment réaliser une détection d'intrusion en temps réel et à grande vitesse. Avec la technologie d'aujourd'hui, les algorithmes d'apprentissage automatique touchent notre vie quotidienne à travers un large éventail d'applications. Il est nécessaire d'étudier l'utilisation de ses capacités interdisciplinaires pour détecter les attaques des réseaux informatiques.

En ce sens, nous développerons une solution de détection d'attaques en temps réel avec différents algorithmes d'apprentissage automatique pouvant être appliqués aux données du réseau pour distinguer les instances normales des instances attaquées, et visualiser les résultats via ElasticSearch, et une bonne maintenance des immenses données avec les outils de big data.

I. Contexte général du projet qui représente les termes principaux de la problématique de détection des attaques en mode réel avec leurs différents types.

II. Un état de l’art qui représente les principaux aspects de l'apprentissage automatique et les capacités actuelles d'Elk dans le

domaine de la cybersécurité, avec cette dernière caractéristique, selon des articles de recherches.

III. "Cybersecurity et Machine Learning" est consacré à l'étude de différentes techniques de création de solutions de détection d'attaques.

IV. Présentation du projet, définition des buts et objectifs et des exigences du projet et les principales mythologies de travail, l'implémentation contient des méthodes d'apprentissage automatiques et des implémentations d'outils de données ELK et BIG, et représente également les outils et les langages utilisés.

Chapitre I

Contexte général du projet

1. Cyberattaque

Les cyberattaques sont de plus en plus fréquentes et selon le rapport annuel sur la cybersécurité de Cisco, les attaquants peuvent déclencher des campagnes sans intervention humaine avec l'avènement des vers ransomware basés sur le réseau.

On parle de cyberattaque lorsqu'un individu ou une organisation tente délibérément et malicieusement de pénétrer dans le système d'information d'un autre individu ou organisation. Si l'objectif est généralement économique, certaines attaques récentes ont pour but la destruction de données.

Les pirates cherchent souvent à obtenir une rançon ou d'autres types de gains économiques, mais les attaques peuvent être perpétrées pour toute une série de motifs, notamment à des fins d'activisme politique.

2. Les types les plus courants d'attaques de cybersécurité

2.1 Malware

Le terme "malware" englobe différents types d'attaques, notamment les logiciels spyware, les virus et les worms. Les logiciels malveillants utilisent une vulnérabilité pour pénétrer dans un réseau lorsqu'un utilisateur clique sur un lien dangereux ou une pièce jointe de courrier électronique "plantés", qui sont utilisés pour installer un logiciel malveillant dans le système.

Les logiciels et les fichiers malveillants présents dans un système informatique peuvent :

- Refuser l'accès aux composants critiques du réseau
- Obtenir des informations en récupérant des données sur le disque dur
- Perturber le système, y compris le rendre inopérant.

Les logiciels malveillants sont très fréquents et il existe une grande variété de modus operandi. Les types les plus courants sont :

- **Les virus** : ils infectent les applications en s'attachant à la séquence d'initialisation. Le virus se réplique et infecte d'autres codes dans le système informatique. Les virus peuvent également s'attacher au code exécutable ou s'associer à un fichier en créant un fichier de virus portant le même nom mais avec une extension .exe, créant ainsi un leurre qui porte le virus.
- **Troyen** : un programme se cachant à l'intérieur d'un programme utile à des fins malveillantes. Contrairement aux virus, un cheval de Troie ne se reproduit pas et il est généralement utilisé pour créer une porte dérobée qui sera exploitée par les attaquants.
- **Les Worms** : contrairement aux virus, ils n'attaquent pas l'hôte, étant des programmes autonomes qui se propagent à travers les réseaux et les ordinateurs. Les worms sont souvent installés par le biais de pièces jointes à des courriels, envoyant une copie d'eux-mêmes à chaque contact de la liste de courriels de l'ordinateur infecté. Ils sont généralement utilisés pour surcharger un serveur de messagerie et réaliser une attaque par déni de service.
- **Ransomware** : un type de logiciel malveillant qui refuse l'accès aux données de la victime, menaçant de les publier ou de les supprimer si une rançon n'est pas payée. Les ransomwares

avancés utilisent l'extorsion cryptovirale, en chiffrant les données de la victime de sorte qu'il est impossible de les déchiffrer sans la clé de déchiffrement.

- **Spyware** : type de programme installé pour collecter des informations sur les utilisateurs, leurs systèmes ou leurs habitudes de navigation, et envoyer ces données à un utilisateur distant. L'attaquant peut ensuite utiliser ces informations à des fins de chantage ou télécharger et installer d'autres programmes malveillants à partir du web.

2.2 Phishing

Les attaques par hameçonnage sont extrêmement courantes et consistent à envoyer des quantités massives de courriels frauduleux à des utilisateurs peu méfiants, déguisés en provenance d'une source fiable. Les courriels frauduleux ont souvent l'apparence d'être légitimes, mais lient le destinataire à un fichier ou un script malveillant conçu pour permettre aux attaquants d'accéder à votre appareil afin de le contrôler ou de collecter des données de reconnaissance, d'installer des scripts/fichiers malveillants ou d'extraire des données telles que des informations sur l'utilisateur, des informations financières, etc.

Les attaques de phishing peuvent également avoir lieu via les réseaux sociaux et autres communautés en ligne, par le biais de messages directs d'autres utilisateurs ayant une intention cachée. Les hameçonneurs s'appuient souvent sur l'ingénierie sociale et d'autres sources d'informations publiques pour recueillir des informations sur votre travail, vos intérêts et vos activités, ce qui donne aux attaquants un avantage pour vous convaincre qu'ils ne sont pas ce qu'ils prétendent être.

Il existe plusieurs types d'attaques de phishing, dont les suivants :

- **Spear Phishing** : attaques ciblées visant des entreprises et/ou des personnes spécifiques.
- **Whaling** : attaques visant les cadres supérieurs et les parties prenantes d'une organisation.
- **Pharming** : l'empoisonnement du cache DNS est utilisé pour capturer les informations d'identification de l'utilisateur par le biais d'une fausse page de connexion.
- Les attaques de phishing peuvent également se faire par appel téléphonique (phishing vocal) et par message texte (phishing SMS).

2.3 Attaques de type "Man-in-the-Middle" (MitM)

Cela se produit lorsqu'un attaquant intercepte une transaction entre deux parties, en s'insérant au milieu. À partir de là, les cyberattaquants peuvent voler et manipuler des données en interrompant le trafic.

Ce type d'attaque exploite généralement les failles de sécurité d'un réseau, comme un réseau WiFi public non sécurisé, pour s'insérer entre l'appareil d'un visiteur et le réseau. Le problème de ce type d'attaque est qu'il est très difficile à détecter, car la victime pense que les informations vont vers une destination légitime. Les attaques de phishing ou de logiciels malveillants sont souvent exploitées pour mener une attaque MitM.

2.4 Attaque par déni de service (DOS)

Les attaques DoS fonctionnent en inondant les systèmes, les serveurs et/ou les réseaux de trafic pour surcharger les ressources et la bande passante. Le résultat rend le système incapable de traiter et de répondre aux demandes légitimes. Outre les attaques par déni de service (DoS), il existe également des attaques par déni de service distribué (DDoS).

Les attaques DoS saturent les ressources d'un système dans le but d'empêcher la réponse aux demandes de service. D'autre part, une attaque DDoS est lancée à partir de plusieurs machines hôtes infectées dans le but de parvenir à un déni de service et de mettre un système hors ligne, ouvrant ainsi la voie à une autre attaque pour pénétrer dans le réseau/l'environnement.

Les types d'attaques DoS et DDoS les plus courants sont l'attaque par inondation TCP SYN, l'attaque en forme de larme, l'attaque smurf, l'attaque ping-of-death et les botnets.

2.5 SQL Injections

Cela se produit lorsqu'un attaquant insère un code malveillant dans un serveur en utilisant le langage de requête du serveur (SQL), forçant le serveur à livrer des informations protégées. Ce type d'attaque consiste généralement à soumettre un code malveillant dans un commentaire ou un champ de recherche non protégé d'un site web. Les pratiques de codage sécurisées, telles que l'utilisation d'instructions préparées avec des requêtes paramétrées, constituent un moyen efficace de prévenir les injections SQL.

Lorsqu'une commande SQL utilise un paramètre au lieu d'insérer directement les valeurs, elle peut permettre au backend d'exécuter des requêtes malveillantes. De plus, l'interpréteur SQL utilise le paramètre uniquement comme une donnée, sans l'exécuter comme un code.

2.6 Exploitation de type "jour zéro" (Zero-day)

Une exploitation de type "jour zéro" consiste à exploiter une vulnérabilité du réseau lorsqu'elle est nouvelle et récemment annoncée, avant qu'un correctif ne soit publié et/ou mis en œuvre. Les attaquants du jour zéro sautent sur la vulnérabilité divulguée dans la petite fenêtre de temps où aucune solution/mesure préventive n'existe. La prévention des attaques de type "zero-day" exige donc une surveillance constante, une détection proactive et des pratiques agiles de gestion des menaces.

2.7 Attaque par mot de passe

Les mots de passe sont la méthode la plus répandue pour authentifier l'accès à un système d'information sécurisé, ce qui en fait une cible attrayante pour les cyberattaquants. En accédant au mot de passe d'une personne, un attaquant peut accéder à des données et des systèmes confidentiels ou critiques, y compris la capacité de manipuler et de contrôler ces données/systèmes.

Les attaquants de mots de passe utilisent une myriade de méthodes pour identifier un mot de passe individuel, y compris l'ingénierie sociale, l'accès à une base de données de mots de passe, le test de la connexion réseau pour obtenir des mots de passe non cryptés, ou simplement en devinant.

La dernière méthode mentionnée est exécutée d'une manière systématique connue sous le nom " brute-force attack". Une attaque par force brute utilise un programme qui essaie toutes les variantes et combinaisons possibles d'informations pour deviner le mot de passe.

Une autre méthode courante est l'attaque par dictionnaire, lorsque l'attaquant utilise une liste de mots de passe courants pour tenter d'accéder à l'ordinateur et au réseau d'un utilisateur. Les meilleures pratiques en matière de verrouillage de compte et l'authentification à deux facteurs sont très utiles pour prévenir une attaque par mot de passe. Les fonctions de verrouillage de compte peuvent bloquer le compte après un certain nombre de tentatives de mots de passe invalides et l'authentification à deux facteurs ajoute une couche de sécurité supplémentaire, en demandant à l'utilisateur qui se connecte de saisir un code secondaire disponible uniquement sur son ou ses dispositifs 2FA.

2.8 Cross-site Scripting

Une attaque par cross-site scripting envoie des scripts malveillants dans le contenu de sites web fiables. Le code malveillant se joint au contenu dynamique qui est envoyé au navigateur de la victime. Généralement, ce code malveillant consiste en un code Javascript exécuté par le navigateur de la victime, mais il peut inclure du Flash, du HTML et du XSS.

2.9 Rootkits

Les rootkits sont installés à l'intérieur d'un logiciel légitime, où ils peuvent obtenir le contrôle à distance et l'accès au niveau de l'administration d'un système. L'attaquant utilise ensuite le rootkit pour voler des mots de passe, des clés, des informations d'identification et récupérer des données critiques.

Comme les rootkits se cachent dans des logiciels légitimes, dès que vous autorisez le programme à apporter des modifications à votre système d'exploitation, le rootkit s'installe dans le système (hôte, ordinateur, serveur, etc.) et reste inactif jusqu'à ce que l'attaquant l'active ou qu'il soit déclenché par un mécanisme de persistance. Les rootkits se propagent généralement par le biais de pièces jointes d'e-mails et de téléchargements à partir de sites Web non sécurisés.

2.10 Attaques de l'Internet des objets (IoT)

Si la connectivité à l'internet de presque tous les appareils imaginables est pratique et facile pour les individus, elle offre également un nombre croissant et presque illimité de points d'accès que les attaquants peuvent exploiter et détruire. L'interconnexion des objets permet aux attaquants de s'introduire dans un point d'entrée et de l'utiliser comme porte d'entrée pour exploiter d'autres appareils du réseau.

Les meilleures méthodes pour prévenir une attaque IoT consistent à mettre à jour le système d'exploitation et à conserver un mot de passe fort pour chaque appareil IoT de votre réseau, et à changer souvent les mots de passe.

3. Besoin de la sécurité

La complexité et la variété des cyberattaques ne cessent de croître, avec un type d'attaque différent pour chaque objectif malveillant. Si les mesures de prévention de la cybersécurité diffèrent pour chaque type d'attaque, les bonnes méthodes de sécurité et l'hygiène informatique de base permettent généralement d'atténuer ces attaques.

Alors, comment les mesures de cybersécurité protègent-elles les utilisateurs et les systèmes ? Tout d'abord, la cybersécurité s'appuie sur des protocoles cryptographiques pour chiffrer les courriers électroniques, les fichiers et autres données critiques. Cela permet non seulement de protéger les informations en transit, mais aussi de se protéger contre la perte ou le vol.

En outre, end-user-security-software analysent les ordinateurs à la recherche de codes malveillants, les mettent en quarantaine, puis les suppriment de la machine. Les programmes de sécurité peuvent même détecter et supprimer le code malveillant caché dans le registre de démarrage primaire et sont conçus pour crypter ou effacer les données du disque dur de l'ordinateur.

Les protocoles de sécurité électronique se concentrent également sur la détection des logiciels malveillants en temps réel. Beaucoup utilisent l'analyse heuristique et comportementale pour surveiller le comportement d'un programme et de son code afin de se défendre contre les virus ou les chevaux de Troie qui changent de forme à chaque exécution (logiciels malveillants polymorphes et métamorphes). Les programmes de sécurité peuvent confiner les programmes potentiellement malveillants dans une bulle virtuelle distincte du réseau de l'utilisateur pour analyser leur comportement et apprendre à mieux détecter les nouvelles infections.

3.1 Conseils de Cybersécurité

Comment les entreprises et les particuliers peuvent-ils se prémunir contre les cybermenaces ? Voici nos principaux conseils de cybersécurité :

1- Mettez à jour vos logiciels et votre système d'exploitation : Vous bénéficiez ainsi des derniers correctifs de sécurité.
2- Utilisez un logiciel anti-virus : Les solutions de sécurité comme Kaspersky détectent et suppriment les menaces. Maintenez votre logiciel à jour pour bénéficier du meilleur niveau de protection.
3- Utilisez des mots de passe forts : Assurez-vous que vos mots de passe ne sont pas faciles à deviner.
4- N'ouvrez pas les pièces jointes des courriels provenant d'expéditeurs inconnus : Elles pourraient être infectées par des logiciels malveillants.
5- Ne cliquez pas sur les liens contenus dans les courriels provenant d'expéditeurs inconnus ou de sites Web inconnus : il s'agit d'un moyen courant de diffusion des logiciels malveillants.
6- Évitez d'utiliser des réseaux WiFi non sécurisés dans les lieux publics : Les réseaux non sécurisés vous rendent vulnérables aux attaques de type man-in-the-middle.

Chapitre II

L'état de l'art

Articles

Lors de nos recherches dans le cadre des " Détection des Attaques En Mode Réel", nous avons découvert de nombreux articles de projet sur différentes sources. Vous trouverez ci-dessous quelques articles sur ce type de problèmes, où on va exploiter ces connaissances pour construire le mien !

A SURVEY ON: "LOG ANALYSIS WITH ELK STACK TOOL" [1]

Le but de cet article était de suivre et d'analyser les logs afin de détecter un comportement anormal présenté par un système particulier. Pour réaliser cet objectif, ils ont opté comme outil ELK stack. '*Log analysis'* s'agit selon l'article d'un processus qui consiste à examiner, interpréter et comprendre les enregistrements générés par l'ordinateur, apporte des données essentielles sur comment le système est en train de fonctionner sous format d'un Log. Un log est constitué d'une série de messages en séquence temporelle qui décrivent les activités en cours dans un système. Dans un cas comme dans l'autre, l'analyse des Logs est l'art délicat d'examiner et d'interpréter ces messages afin de comprendre le fonctionnement interne du système.

Pour faire cela, ils ont choisi à travailler avec Elk Stack, présenté dans cet article par un arrangement total d'examen du journal. Il s'agit d'une combinaison de 3 composants :

- Elasticsearch : ELK utilise 'Elasticsearch pour la recherche profonde et l'investigation des informations. C'est le composant principal qui centralise les informations et y accède via une API Restful. Cet outil est utilisé par GitHub, Sound Cloud, Netflix et d'autres entreprises.
- Logstash : permet l'agrégation des données dans Elasticsearch. C'est un pipeline d'informations qui permet de rassembler, d'analyser, et d'enquêter sur un vaste assortiment d'informations et d'occasions organisées et non structurées produites à travers différents cadres,
- Kibana : Elk utilise Kibana pour la perception d'informations excellentes. Kibana aide à visualiser toute sorte d'informations organisées et non structurées rangées dans des enregistrements Elasticsearch.

Le système proposé décrit un environnement qui va vérifier toute sorte d'opérations malveillantes par mettre une interface appelée « Threat Intelligence » avec les outils d'ELK. C'était remarqué que Elk avait une très bonne performance à détecter ainsi que de bien visualiser « Security Logs » qui sont particulièrs, ce qui représente une bonne sécurité pour l'administration.

Ils ont aussi proposé comme une solution désignée à détecter les anomalies dans l'environnement de Hadoop, en implémentant quelques méthodes « weight based » avec des multiples nœuds, et en utilisant aussi Hive QL. Cette méthode a apporté une bonne visualisation des anomalies correctement détectés dans un diagramme graphique.

Comme un résultat final, ELK stack représente un outil important pour la détection des comportements anormales des systèmes, ainsi que les composants d'Elk peuvent fonctionner individuellement.

DISTRIBUTED DENIAL OF SERVICE ATTACKS DETECTION SYSTEM BY MACHINE LEARNING BASED ON DIMENSIONALITY REDUCTION [2]

L'idée principale de cet article s'agit d'utiliser les algorithmes du Data Mining afin de révéler et détecter les attaques de types DDOS. Les techniques du Data Mining permettent à l'utilisateur de distinguer entre les trafics normales et anormales avec une bonne précision.

L'article se focalise sur les attaques du type DDoS. Ce type s'agit d'un exemple des plus significatifs difficultés que l'internet confrontent. Ce sont utilisés dans les couches du réseau, transport et applications, qui utilisent des protocoles (HTTP, UDP et TCP). Les attaques DDoS influencent sur l'intégrité des données par perdre ces données, ou les représenter d'une façon incorrecte, la chose qui conduit à perdre la confiance des utilisateurs pour les secteurs de fondation. IDS (Intrusion Detection System) était développé pour examiner, identifier et arrêter ce genre d'action comme DDoS, et il est divisé en des sections : Misue Intrusion Detection (MIS) et Anomaly-Intrusion Detection (AID). IDS peut être classifié avec les différents algorithmes du Data Mining qui arrivent à reconnaître les nouvelles entrées sur le trafic du réseau.

Taxonomie des attaques DDoS consiste d'avoir stockés dans un jeu de données appelé CICDDoS2019 qui reflète des attaques implémentants aussi TCP/UDP dans la couche application, et ces attaques étaient divisés en deux catégories :

- Reflection-based où se trouve une grande difficulté à distinguer entre un utilisateur et un attaqueur.
- Exploitaton-based où les attaques TCP peuvent contenir flux de SYN tandis que les attaques UDP ont flux d'UDP et UDPLag.

Un modèle a été proposé qui vise à classifier le network-traffic à deux classes : Des paquets normales et des paquets d'attaqueur. Ce système proposé consiste 4 étapes :

- **Preprocessing :** où le jeu de données CICDDoS2019 était utilisé. Ce jeu comprte des valeurs numériques par convertir les symboles comme les adresses IP du source et destination, ID du flux, protocole utilisé par le réseau et les ports. Deux approches étaient utilisées durant ce processus :
  - Encoding : convertit les paquets originaux nominaux en des caractéristiques numériques
  - Log2 : Algorithmes Logarithmiques utilisés pour faire la standardisation
  - PCA : Appliqué 8 fois sur les différentes caractéristiques afin de réduire la dimension du jeu de données.

- **Detection Model :** Employer des modèles de classification des anomalies comme :
  - Random Forest (RF) algorithm : Implémenté pour l'extraction des patterns des données en classifiant les types des caractéristiques données dans le processus d'entraînement
  - Naive Bayes : utilisé pour classifier les données et faire une comparaison des résultats de ce modèle par rapport aux résultats du RF.
- **Classification :** Le résultat était testé par implémenter les données déjà entraînées

- **Evaluation du performance :** L'évaluation des modèles étaient faite par des mesures de classification comme Accuracy, Detection rate, False alarm, Positive Predictive Rate et F.Measure.

La plus grande précision a atteint 99% avec False alarm rate qui était prés à 0 en utilisant le modèle Random Forest. Avec Naive Bayes, les résultats étaient bons mais les valeurs variaient par rapport au pourcentage du dimension des données.

An Impact Analysis: Real Time DDoS Attack Detection and Mitigation using Machine Learning [3]

Dans cet article, ils ont parlé des attaques par déni de service distribué (DDoS) attaque la plus dévastatrice qui altère la fonctionnalité normale des services critiques dans la communauté Internet. Alors depuis l'usurpation le trafic partage les mêmes ressources que celui du légitime la détection et le filtrage deviennent très essentiels. La proposition modèle consiste en un système de surveillance en ligne (OMS), usurpé module de détection de trafic et limitation de débit basée sur l'interface (IBRL) algorithme. OMS fournit des mesures d'impact DDoS en temps réel temps en surveillant la dégradation de l'hôte et du réseau indicateurs de performance. Le module de détection de trafic usurpé intègre l'algorithme d'inspection du nombre de sauts (HCF) pour vérifier l'authenticité du paquet entrant au moyen de l'adresse IP source et ses sauts correspondants à la victime destinée. HCF couplé avec la machine à vecteurs de support (SVM) fournit une précision de 98,99 % avec réduit les faux positifs. Suivi de, l'algorithme IBRL restreint le trafic s'agrège au routeur victime lorsqu'il dépasse le système limites afin de fournir une bande passante suffisante pour rester les flux.

An Experimental Analysis of Attack Classification Using Machine Learning in IoT Networks [4]

Dans ce travail, les algorithmes ML sont comparés pour classification binaire et multi-classes sur l'ensemble de données Bot-IoT. En utilisant des méthodes d'apprentissage (ML) tel que k-nearest neighbour (KNN), support vector machine (SVM), decision tree (DT), naive Bayes (NB), random forest (RF), artificial neural network (ANN), et logistic regression (LR) qui peut être utilisée dans IDS, et basant sur plusieurs paramètres tels que accuracy, precision, recall, F1 score, et log loss, cet article comparent expérimentalement les éléments susmentionnés Algorithmes ML. Dans le cas d'une attaque par déni de service distribué HTTP (DDoS), la précision de RF est de 99 %. En outre, d'autres mesures de précision, de recall, de F1 score et de log loss basées sur les résultats de simulation révèlent que RF surpasse tous les types d'attaques en classification binaire.

Cependant, en multi-classe classification, KNN surpasse les autres algorithmes ML avec une précision de 99 %, soit 4 % de plus que RF.

Review on Efficient Log Analysis to Evaluate Multiple Honeypots using ELK [5]

La sécurité du réseau ne se contente pas de mettre en œuvre des protections telles que IPS/IDS, des pares-feux, mais il y a plusieurs façons d'améliorer la sécurité du réseau qui consiste à analyser les logs collectés à partir de différentes sources ce qui nous permettra de répondre à tant de questions.

Dans cet article ils ont examiné l'efficacité d'IDS/IPS et d'autres honeypots avec ELK pour aider à l'analyse des logs collectés. Elasticsearch, Logstash, Kibana « (ELK) stack » vont être utilisé comme serveur centralisé pour collecter, analyser, stocker, interroger et visualiser les résultats dans des graphiques, des barres, des secteurs.

Dans cet article ils ont utilisé Ubuntu 16.04 comme serveur pour installer ELK et tous les autres honeypots à l'aide de docker, pour permettre d'exécuter plusieurs Honeypots sur la même interface réseau. Docker encapsule les honeypots et les isole pour exécuter indépendamment et faciliter la mise à jour et la maintenance. Concernant la simulation d'attaque, ils ont utilisé Kali Linux, Windows XP et Windows 8 comme attaquant.

Pour réaliser leur projet ils ont suivi 4 étapes principales :

1- Créer des compteurs spéciaux pour compter tous les événements capturés par tous les honeypots individuellement et les démontrer.
2- Filtrer les logs contenant toutes les tentatives utilisées pour se connecter à distance.
3- Recouper les parties de logs qui pourraient être intéressantes vous pouvez remarquer quel le honeypot a détecté les tentatives de connexion à distance.
4- Filtrer les logs de tous les honeypot pour détecter uniquement les attaques qui ont ciblé notre système d'exploitation.

Après avoir simulé différents types d'attaques en utilisant différents types de systèmes d'exploitation contre le serveur pour évaluer les différents types de honeypot et système IDS pour comparer leurs capacités de détection des attaques et révéler des informations sur la source des attaques. Dans le tableau suivant montrera les détails sur les résultats trouvés :

Honeypot, IDS Name	Total Events captured	DoS attack	Attempts to remotely login	Detected OS of Attacker	Revealed Username/password used by hackers	Detected exploiting vulnerabilities
honeytrap	4464	No	No	No	No	No
dionaea	36	No	No	No	No	No
conpot	2	No	No	No	No	No
cowrie	37	No	Yes	No	Yes	No
elasticpot	2	No	No	No	No	No
Suricata	7817	Yes	No	Yes	No	Yes

Honeypot Deployment in Broadband Networks [6]

Dans cet article, ils ont présenté les résultats du déploiement des Honeypot dans les réseaux à large bande. L'objectif est de capter et de caractériser les attaques ciblant les réseaux haut débit. Pour capturer ces attaques, ils ont aussi identifié six scénarios de déploiement Honeypot différents pour les réseaux à large bande. Ces scénarios de déploiement sont classés en fonction de leurs exigences réseau, de l'effet sur les réseaux sous-jacents et du type de données capturées. Pour démontrer l'efficacité du déploiement de Honeypot dans les réseaux à large bande, ils ont mis en œuvre l'un des scénarios les plus courants qui émule l'appareil IoT (routeur ADSL).

Scénario de déploiement 1 : (pot de miel actif derrière NAT) : ce scénario de déploiement ne nécessite aucune modification de la configuration du routeur ADSL. Il capture les attaques qui se propagent à l'aide de la technique de lecteur par téléchargement et ciblent les environnements SOHO.

Scénario de déploiement 2 (pot de miel actif + pot de miel passif derrière NAT): dans ce scénario, un environnement contrôlé est créé à l'aide d'une combinaison de haut et de bas interaction Pots de miel. L'objectif de ce scénario de déploiement est de capturer les attaques de type Drive by Pharming et d'observer le comportement post-infection.

Scénario de déploiement 3 (Honeypot passif) : Dans ce scénario, le routeur ADSL est configuré en mode DMZ. Ce scénario de déploiement capture un échantillon de logiciels malveillants, des analyses ciblant les sous-réseaux à large bande IP publics, des attaques se propageant dans les mêmes sous-réseaux par des systèmes infectés.

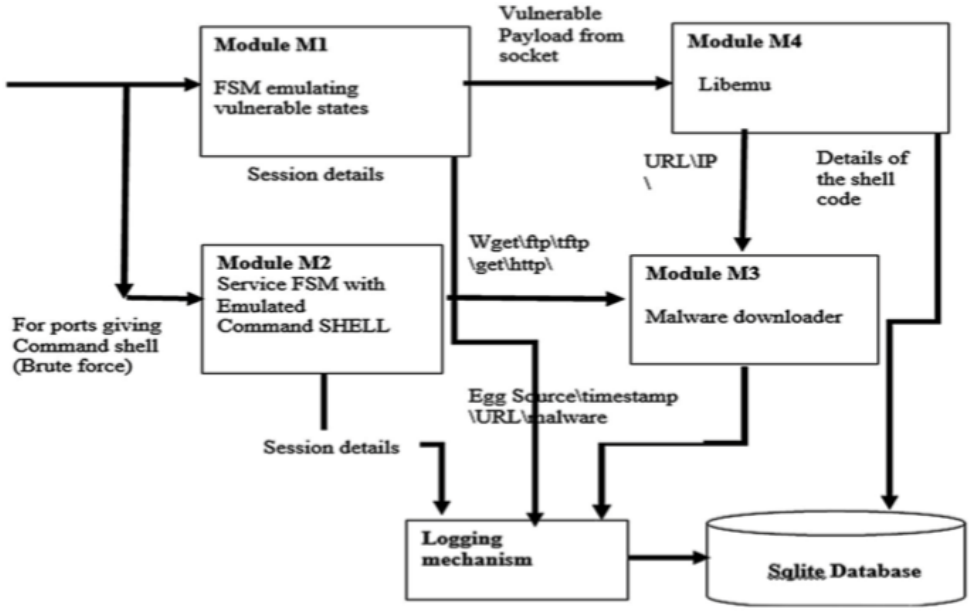
Scénario de déploiement 4 (Web Honeypot): ce type de déploiement émule l'environnement SOHO dans lequel une petite organisation déploie ses serveurs Web sur les réseaux à large bande. Ce scénario de déploiement capture les attaques d'applications Web, les analyses ciblant les sous-réseaux à large bande et les attaques ciblant l'interface Web WAN des appareils IoT

Scénario de déploiement 5 (pot de miel de routeur ADSL) : ces pots de miel émulent les services exécutés sur l'interface WAN du routeur ADSL (c'est-à-dire telnet, SSH, UPNP, etc.). Ce scénario capture les attaques ciblant les appareils IoT

Scénario de déploiement 6 (moniteurs Darknet) : Darknet est une plage d'adresses IP inutilisées qui est surveillée pour la collecte de données d'attaque. Ce scénario de déploiement capture les tendances d'attaque sur les réseaux larges.

Dans cet article ils se sont focalise sur le cinquième scenario vu qu'il y a une augmentation des attaques ciblant les appareils IoT.

Voici une figure qui montre le schéma fonctionnel du système Router Honeypot développé et ayant les modules suivants :



Module 1 : Le module M1 émule les vulnérabilités des services réseau

Module 2 : le module M2 est similaire au module M1, sauf qu'il émule les services ciblés par les attaques par force brute et que l'attaquant utilise pour obtenir le shell

Module 3 : est un module utilisé pour exécuter les tentatives de téléchargement de logiciels malveillants capturées par le module M1 et le module M2.

Module 4 : est le module de détection de shellcode. Il utilise la bibliothèque open source libemu qui utilise l'heuristique GetPC pour la détection de la charge utile du shellcode

Au cours de l'analyse dynamique d'échantillons de logiciels malveillants, il a été observé que ces logiciels malveillants ont la capacité de s'auto-répliquer et de se propager à l'aide d'une attaque par force brute telnet. De plus, ces logiciels malveillants ont des capacités de bot et ils communiquent avec leur serveur C&C en utilisant le protocole HTTP. Lorsqu'il est activé, le malware s'initialise en extrayant d'abord l'heure actuelle et le PID de son processus en cours d'exécution et l'utilise comme graine. Il vérifie ensuite la connectivité Internet en communiquant avec le serveur DNS de Google.

L'augmentation des attaques ciblant les appareils IoT est positivement corrélée à l'augmentation du nombre d'attaques DDoS lancées par des attaquants ciblant des infrastructures et des services critiques héberges sur Internet. En tant que cibles faciles, ces appareils IoT sont devenus un choix privilégié pour les attaquants à la recherche de compromis de masse pour la construction de botnets et offrant des services DDoS sur le dark web. En outre, cette tendance a entrainé l'émergence de nouvelles classes de logiciels malveillants ciblant les systèmes d'exploitation basés sur Linux et les architectures de processeur utilisées dans les appareils IoT. Une approche active est nécessaire pour résoudre ce problème car les mesures de sécurité conventionnelles échouent lamentablement. Il est également nécessaire de surveiller à grande échelle les incidents Internet pour garder un œil sur les tendances mondiales des attaques. Le scénario de déploiement que nous suggérons convient à la surveillance des environnements SoHo et peut compléter les efforts des dispositifs de sécurité conventionnels. La meilleure partie de ces scénarios de déploiement est qu'ils n'affectent pas l'utilisation normale d'Internet des utilisateurs dans l'environnement SoHo.

Chapitre III

Machine Learning and Cyber Security

Machine Learning

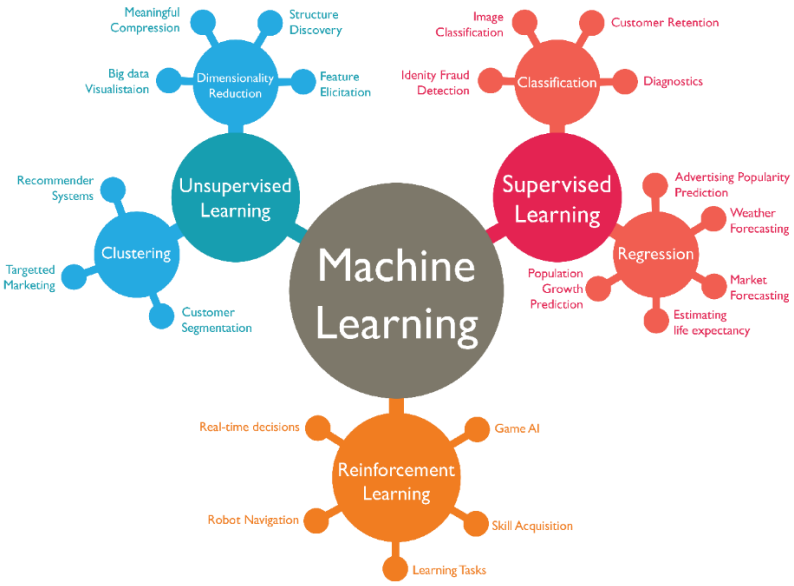
1. Définition Machine Learning

Bien que l'apprentissage automatique ne soit pas nouveau, sa définition précise confond encore beaucoup de gens. Plus précisément, il s'agit d'une science moderne consistant à découvrir des modèles et à faire des prédictions à partir de données basées sur des statistiques, l'exploration de données, la reconnaissance de modèles et l'analyse prédictive.

Le Machine Learning peut être défini comme une branche de l'intelligence artificielle englobant de nombreuses méthodes permettant de créer automatiquement des modèles à partir des données. Ces méthodes sont en fait des algorithmes. Un système Machine Learning ne suit pas d'instructions, mais apprend à partir de l'expérience. Par conséquent, ses performances s'améliorent au fil de son « entraînement » à mesure que l'algorithme est exposé à d'avantage de données.

Le Machine Learning est très efficace dans les situations où les insights doivent être découvertes à partir de larges ensembles de données diverses et changeantes, c'est à dire : le Big Data.

2. Les différents types d'algorithmes de ML



On distingue différents types d'algorithmes Machine Learning. Généralement, ils peuvent être répartis en trois catégories : l'apprentissage supervisés, non supervisés et par renforcement.

Dans le cas de **l'apprentissage supervisé**, les données utilisées pour l'entraînement sont déjà « étiquetées ». Par conséquent, le modèle de Machine Learning sait déjà ce qu'elle doit chercher

(motif, élément…) dans ces données. À la fin de l'apprentissage, le modèle ainsi entraîné sera capable de retrouver les mêmes éléments sur des données non étiquetées.

Parmi les algorithmes supervisés, on distingue :

- Les algorithmes de classification.
- Les algorithmes de régression.

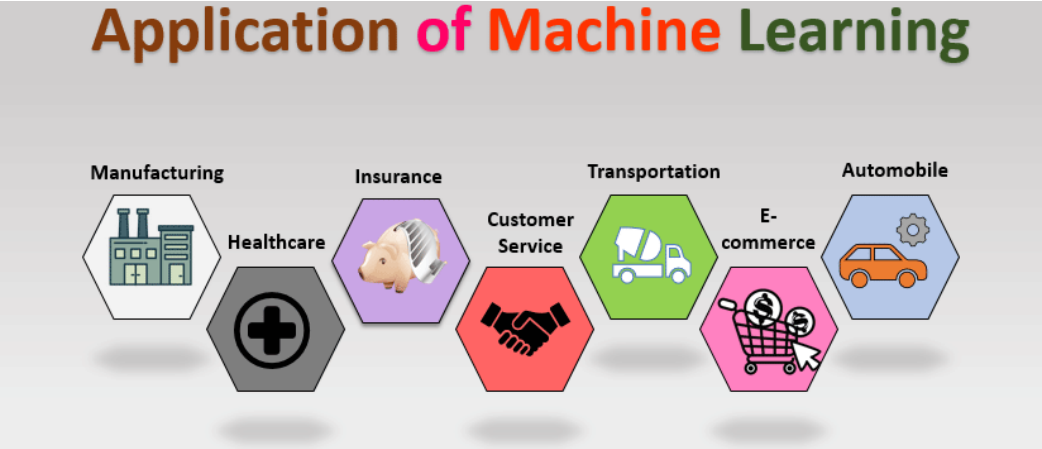
L'apprentissage non supervisé, au contraire, consiste à entraîner le modèle sur des données « sans étiquettes ». La machine parcourt les données sans aucun indice, et tente d'y découvrir des motifs ou des tendances récurrents. Cette approche est couramment utilisée dans certains domaines, comme la cybersécurité.

Parmi les modèles non-supervisés, on distingue :

- Les algorithmes de clustering.
- Les règles d'association.
- La réduction des dimensions.

Dans le troisième cas, **l'apprentissage par renforcement**. L'algorithme apprend en essayant encore et encore d'atteindre un objectif précis. Il pourra essayer toutes sortes de techniques pour y parvenir. Le modèle est récompensé s'il s'approche du but, ou pénalisé s'il échoue.

3. À quoi sert le Machine Learning ?



Le Machine Learning alimente de nombreux services modernes très populaires. On peut citer comme exemple les moteurs de recommandations utilisés par Netflix, YouTube, Amazon ou Spotify.

Il en va de même pour les moteurs de recherche web comme Google ou Baidu. Le fil d'actualité des réseaux sociaux tels que Facebook et Twitter reposent sur le Machine Learning, au même titre que les assistants vocaux tels que Siri et Alexa. Toutes ces plateformes collectent des données sur les utilisateurs, afin de mieux les comprendre et d'améliorer leurs performances. Les algorithmes ont besoin de savoir ce que regarde le spectateur, sur quoi clique l'internaute, et à quelles publications il réagit sur les réseaux. De cette manière, ils sont ensuite en mesure de proposer de meilleures recommandations, réponses ou résultats de recherche.

Cas d'usage et applications :

Quelques cas d'usages et applications de machine learning :

- Voitures autonomes.
- Le domaine des jeux.
- Images de radiographies médicales
- La traduction linguistique automatique.
- Conversion du discours oral à l'écran (speech-to-text).
- L'analyse de sentiment sur les réseaux sociaux.
- Etc…

4. Machine Learning et Big Data:

Les analyses prédictives donnent du sens au Big Data ?

Les analyses prédictives consistent à utiliser les données, les algorithmes statistiques et les techniques de Machine Learning pour prédire les probabilités de tendances et de résultats financiers des entreprises, en se basant sur le passé. Elles rassemblent plusieurs technologies et disciplines comme les analyses statistiques, le data mining, le modelling prédictif et le Machine Learning pour prédire le futur des entreprises. Par exemple, il est possible d'anticiper les conséquences d'une décision ou les réactions des consommateurs.

Les analyses prédictives permettent de produire des insights exploitables à partir de larges ensembles de données, pour permettre aux entreprises de décider quelle direction emprunter par la suite et offrir une meilleure expérience aux clients.

Cybersécurité

5. Définition Cybersécurité

La cybersécurité est la pratique consistant à protéger les ordinateurs, les serveurs, les appareils mobiles, les systèmes électroniques, les réseaux et les données contre les attaques malveillantes. Il est également connu sous le nom de sécurité des technologies de l'information ou de sécurité de l'information électronique. Le terme s'applique à une variété de contextes, de l'entreprise à l'informatique mobile.

Cybersécurité peut être divisé en quelques catégories courantes :

- **La sécurité du réseau** est la pratique consistant à sécuriser un réseau informatique contre les intrus, qu'il s'agisse d'attaquants ciblés ou de logiciels malveillants opportunistes.
- **La sécurité des applications** se concentre sur la protection des logiciels et des appareils contre les menaces. Une application compromise pourrait donner accès aux données qu'elle est censée protéger. Une sécurité réussie commence dès la phase de conception, bien avant le déploiement d'un programme ou d'un appareil.
- **La sécurité de l'information** protège l'intégrité et la confidentialité des données, tant en stockage qu'en transit.
- **La sécurité opérationnelle** comprend les processus et les décisions de gestion et de protection des actifs de données. Les autorisations dont disposent les utilisateurs lorsqu'ils accèdent à un réseau et les procédures qui déterminent comment et où les données peuvent être stockées ou partagées relèvent toutes de cette catégorie.

6. Types de cybermenaces

Les menaces contrées par la cybersécurité sont triples :

- **La cybercriminalité** comprend des acteurs ou des groupes individuels ciblant des systèmes à des fins financières ou pour provoquer des perturbations
- **Les cyberattaques** impliquent souvent la collecte d'informations à motivation politique.
- **Le cyberterrorisme** vise à saper les systèmes électroniques pour semer la panique ou la peur.

Alors, comment les acteurs malveillants prennent-ils le contrôle des systèmes informatiques ? Voici quelques méthodes courantes utilisées pour menacer la cybersécurité :

- **Logiciels malveillants** (**Malware**) : L'une des cybermenaces les plus courantes, les logiciels malveillants sont des logiciels créés par un cybercriminel ou un pirate informatique pour perturber ou endommager l'ordinateur d'un utilisateur légitime.
    - **Virus**
    - **Troie**
    - **Spyware**
    - **Ransomware**
    - **Adware**
    - **Botnets**

- **Injection SQL :** est un type de cyberattaque utilisée pour prendre le contrôle et voler des données d'une base de données.
- **Hameçonnage :** c'est quand les cybercriminels ciblent les victimes avec des e-mails qui semblent provenir d'une entreprise légitime demandant des informations sensibles.
- **Attaque de l'homme du milieu :** est un type de cybermenace où un cybercriminel intercepte la communication entre deux individus afin de voler des données.
- **Attaque par déni de service :** se produit lorsque les cybercriminels empêchent un système informatique de répondre à des demandes légitimes en submergeant les réseaux et les serveurs de trafic.

Chapitre IV

Réalisation et Mise en Œuvre

1. Définition du Projet

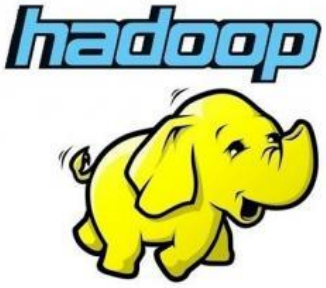
Dans les réseaux informatiques, une attaque est une tentative de vol, de mise hors service, de destruction, de modification, d'accès non autorisé ou d'utilisation non autorisée d'un actif. Les attaques de réseau peuvent entraîner le ralentissement des services du réseau, leur indisponibilité temporaire ou leur interruption pendant une longue période. Il est donc nécessaire pour les utilisateurs et l'administrateur réseau de détecter ces attaques avant qu'elles ne causent des dommages au système. Le problème actuel de la technologie de détection des intrusions dans le réseau est de parvenir à une détection des intrusions en temps réel et à grande vitesse.

C'est là où L'apprentissage automatique est devenu une technologie essentielle pour la cybersécurité. L'apprentissage automatique élimine de manière préventive les cybermenaces et renforce l'infrastructure de sécurité par la détection de modèles, la cartographie de la cybercriminalité en temps réel et des tests de pénétration approfondis.

À l'ère du Big Data et de l'Internet des objets, les volumes de collecte de données ne cessent de croître, ingérant parfois jusqu'à un pétaoctet d'événements de sécurité par jour, les taux d'ingestion ne faisant qu'augmenter de manière exponentielle au fil du temps. La connexion de milliards d'appareils à travers les réseaux et les Clouds a créé une vaste surface de points d'entrée à défendre contre les cyberattaques. À mesure que la sophistication, le volume et la variété des cyberattaques augmentent, le besoin d'une stratégie de défense de la cybersécurité robuste, axée sur les données et en temps réel, est d'autant plus pressant.

2. Technologies utilisées

2.1. Hadoop



Hadoop Ecosystem est une plate-forme ou une suite qui fournit divers services pour résoudre les problèmes de Big Data. Il comprend des projets Apache et divers outils et solutions commerciaux. Il existe quatre éléments principaux de Hadoop, à savoir HDFS, MapReduce, YARN et Hadoop Common. La plupart des outils ou solutions sont utilisés pour compléter ou soutenir ces éléments majeurs. Tous ces outils fonctionnent collectivement pour fournir des services tels que l'absorption, l'analyse, le stockage et la maintenance des données, etc.

Voici les composants qui forment collectivement un écosystème Hadoop :

- **HDFS :** système de fichiers distribué Hadoop.
- **YARN :** encore un autre négociateur de ressources.
- **MapReduce :** traitement de données basé sur la programmation.
- **Spark :** traitement des données en mémoire.
- **PIG, HIVE :** traitement des services de données basé sur des requêtes.
- **HBase :** base de données NoSQL.

- **Mahout, Spark MLLib :** bibliothèques d'algorithmes d'apprentissage automatique.
- **Solar, Lucene :** recherche et indexation.
- **Zookeeper :** Gestion du cluster.
- **Oozie :** Planification des tâches.

2.2. Elasticsearch

Elasticsearch est un moteur de recherche et d'analyse distribué gratuit et ouvert pour tout type de données, y compris les données textuelles, numériques, géospatiales, structurées et non structurées. Elasticsearch a été conçu à partir d'Apache Lucene et a été lancé en 2010 par Elasticsearch N. V. (maintenant appelé Elastic). Réputé pour ses API REST simples, sa nature distribuée, sa vitesse et sa scalabilité, Elasticsearch est le composant principal de la Suite Elastic, un ensemble d'outils gratuits et ouverts d'ingestion de données, d'enrichissement, de stockage, d'analyse et de visualisation. Couramment appelée la Suite ELK (pour Elasticsearch, Logstash et Kibana), la Suite Elastic comprend désormais une riche collection d'agents de transfert légers, appelés les agents Beats, pour envoyer des données à Elasticsearch.

2.3. Logstash

Logstash est une solution open source de traitement de données et de logs côté serveur.

Elle peut recueillir des données provenant d'une multitude de sources pour les analyser, les filtrer, les transformer, les enrichir, et enfin les transmettre à un autre système pour traitements additionnels ou stockage.

Ainsi, la solution Logstash est généralement utilisée comme un moteur de traitement polyvalent, permettant de s'intégrer à d'autres solutions et applications.

Étant l'une des composantes du stack ELK, Logstash transfert, la plupart du temps, les données collectées vers Elasticsearch.

Logstash collecte les logs provenant de différentes sources (logiciels, appareils électroniques, requêtes d'API, etc.), traite les données collectées et les transmet à une autre application pour traitement additionnels ou stockage ultérieur.

Logstash permet aussi le traitement et la transformation de données de natures différentes (structurées, semi-structurées ou non structurées).

2.4. Kibana

Kibana est une application frontend gratuite et ouverte qui s'appuie sur la Suite Elastic. Elle permet de rechercher et de visualiser les données indexées dans Elasticsearch. Si Kibana est connue pour être l'outil de représentation graphique de la Suite Elastic (précédemment appelée "la Suite ELK", acronyme d'Elasticsearch, Logstash et Kibana), elle sert aussi d'interface utilisateur pour le monitoring, la gestion et la sécurité des clusters de la Suite Elastic. Sans oublier qu'elle joue aussi le rôle de hub centralisé pour des solutions intégrées, développées sur la Suite Elastic. Créée en 2013 au sein de la communauté Elasticsearch, Kibana s'est développée pour offrir une vue à 360° sur la Suite Elastic, devenant ainsi un véritable portail pour les utilisateurs et les entreprises.

2.5. Beats

Beats est une plateforme gratuite et ouverte qui accueille des agents réservés au transfert de données. Que vous exploitiez des centaines ou des milliers de machines et de systèmes, les agents Beats se chargent de transférer vos données vers Logstash et Elasticsearch.

2.6. Suricata

Suricata est un logiciel open source de détection d'intrusion (IDS), de prévention d'intrusion (IPS), et de supervision de sécurité réseau (NSM). Il est développé par la fondation OISF (Open Information Security Foundation).
Suricata permet l'inspection des Paquets en Profondeur (DPI). De nombreux cas d'utilisations déontologiques peuvent être mis en place permettant notamment la remontée d'informations qualitatives et quantitatives.

2.7. Wazuh

Wazuh est une plateforme open source utilisée pour la prévention, la détection et la réponse aux menaces. Elle sécurise les environnements de travail sur site, virtualisés, conteneurisés et en cloud. Wazuh est largement utilisé par des milliers d'organisations à travers le monde, de la petite entreprise à la grande entreprise.

La solution Wazuh se compose de plusieurs agents de sécurité des terminaux, déployés sur les systèmes surveillés, et d'un serveur de gestion, qui collecte et analyse les données recueillies par les agents. En outre, Wazuh a été entièrement intégré à Elastic Stack, fournissant un moteur de recherche et un outil de visualisation des données qui permet aux utilisateurs de naviguer à travers leurs alertes de sécurité.

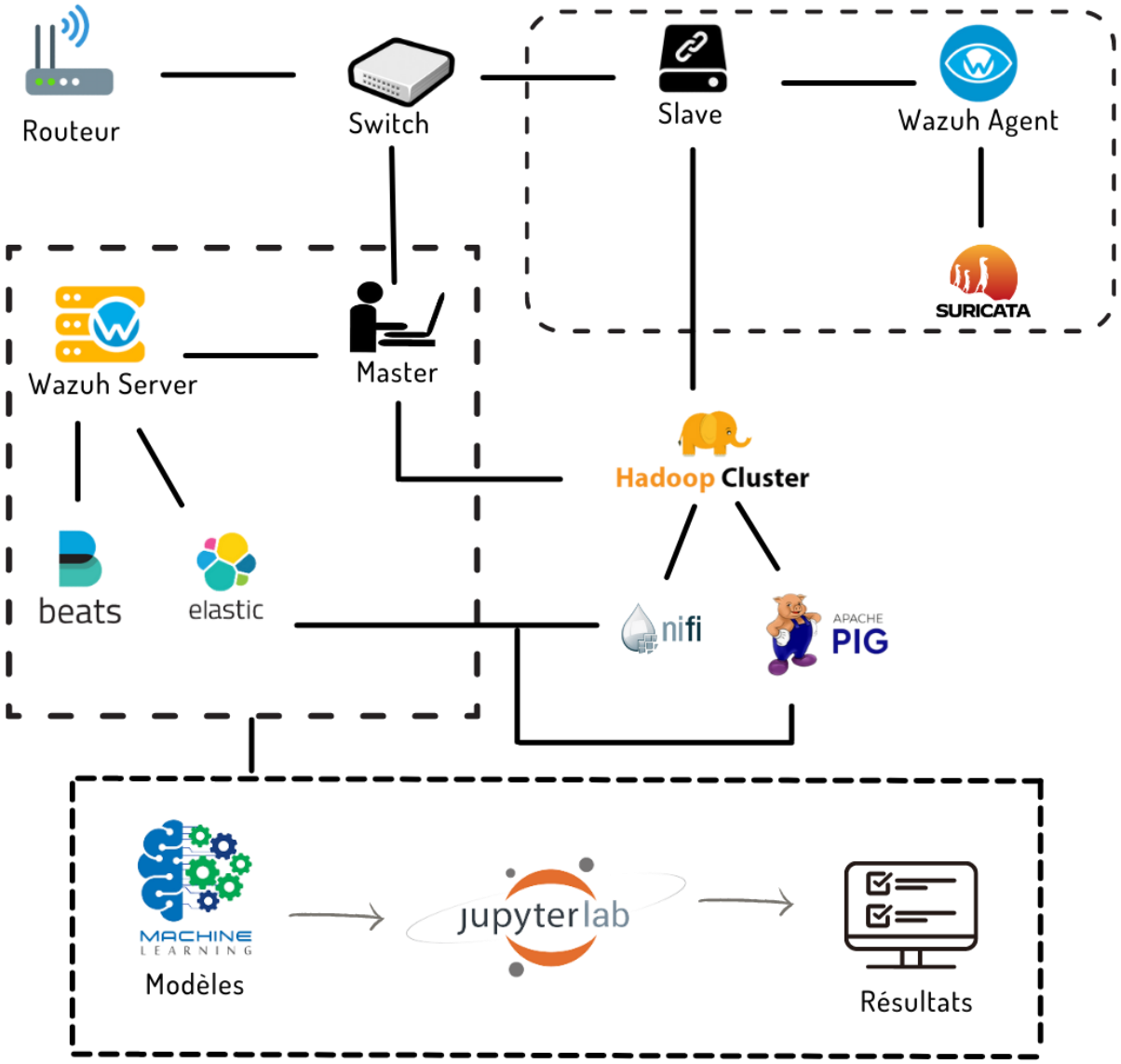
2.8. Pig

Pig3 est une plateforme haut niveau pour la création de programme MapReduce utilisé avec Hadoop. Le langage de cette plateforme est appelé le Pig Latin. Pig Latin s'abstrait du langage de programmation Java MapReduce et se place à un niveau d'abstraction supérieur, similaire à celle de SQL pour systèmes SGBDR. Pig Latin peut être étendue en utilisant UDF (User Defined Functions) que l'utilisateur peut écrire en Java, en Python, en JavaScript, en Ruby ou en Groovy et ensuite être utilisé directement au sein du langage.

2.9. Nifi

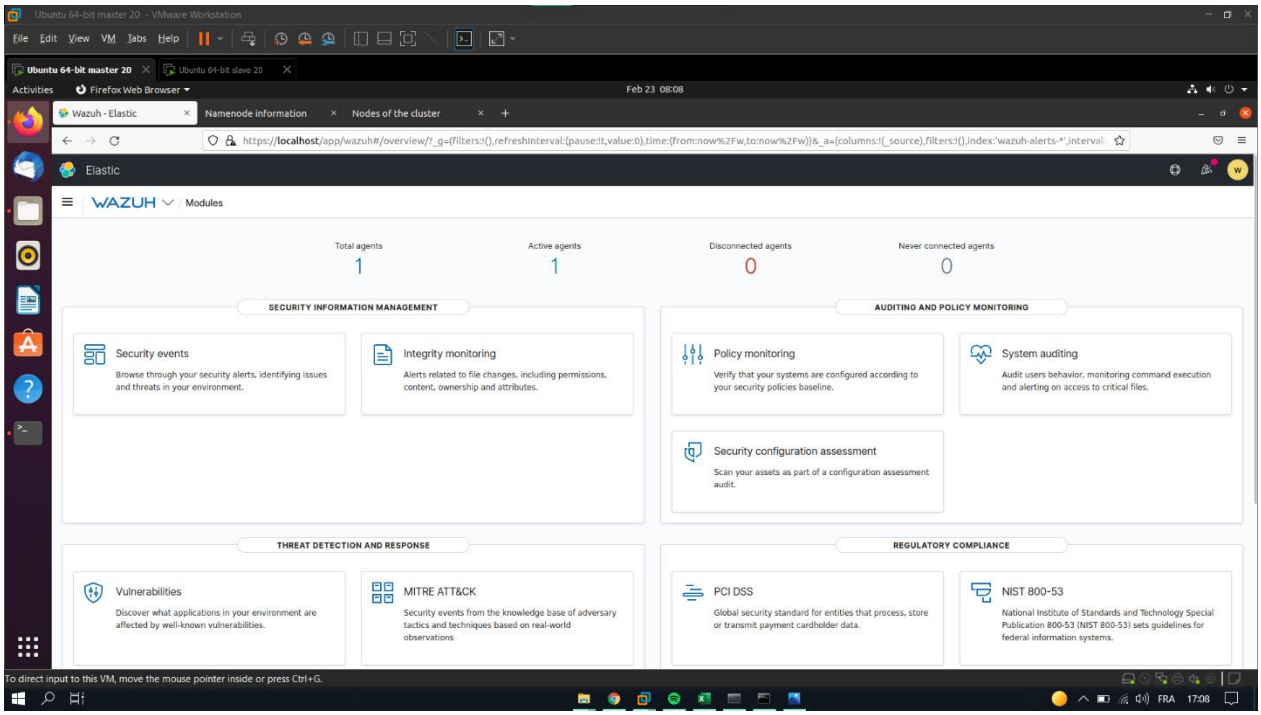
Apache NiFi est une plateforme intégrée de logistique de données qui sert à automatiser les mouvements de données entres des systèmes différents. Cette plateforme offre un pilotage en temps réel qui facilite la gestion des mouvements de données depuis n'importe quelle source et vers n'importe quelle destination. Indépendante, elle prend en charge des sources de données hétérogènes et distribuées ayant des formats, des schémas, des protocoles, des tailles et des débits différents, à l'instar des machines, appareils de géolocalisation, flux de clic, fichiers, réseaux sociaux, fichiers de journaux, vidéos, etc. Il s'agit de connexions configurables qui permettent le déplacement de données, au même titre que Fedex, UPS ou les autres services de livraison avec les colis. Et, tout comme ces services, Apache NiFi vous permet de suivre vos données en temps réel, comme pour une livraison.

3. Implémentation

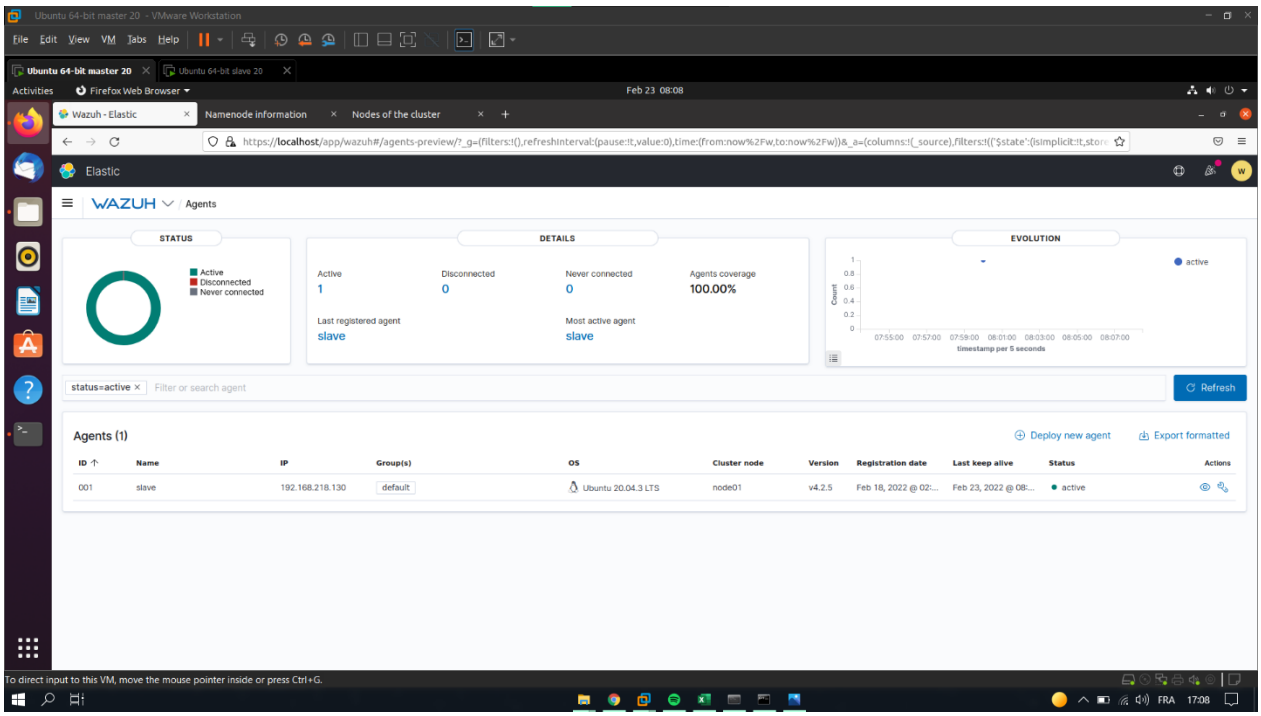


4. Réalisation

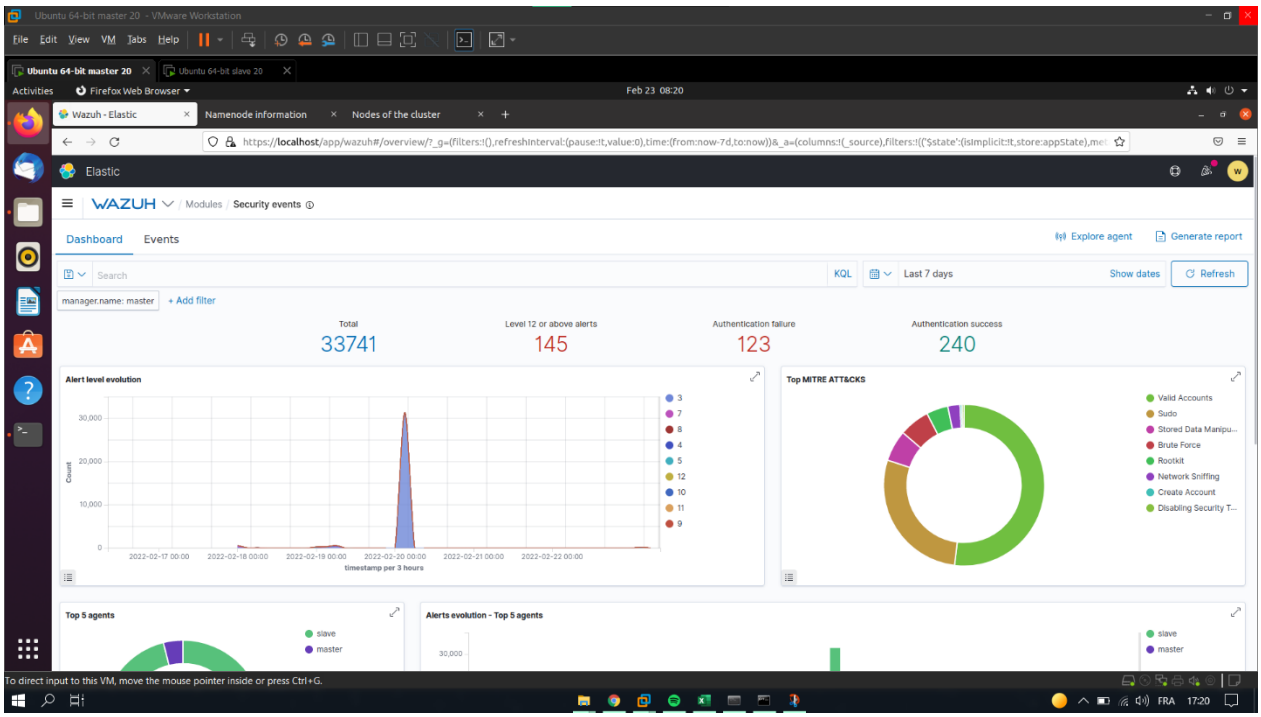
- Overview



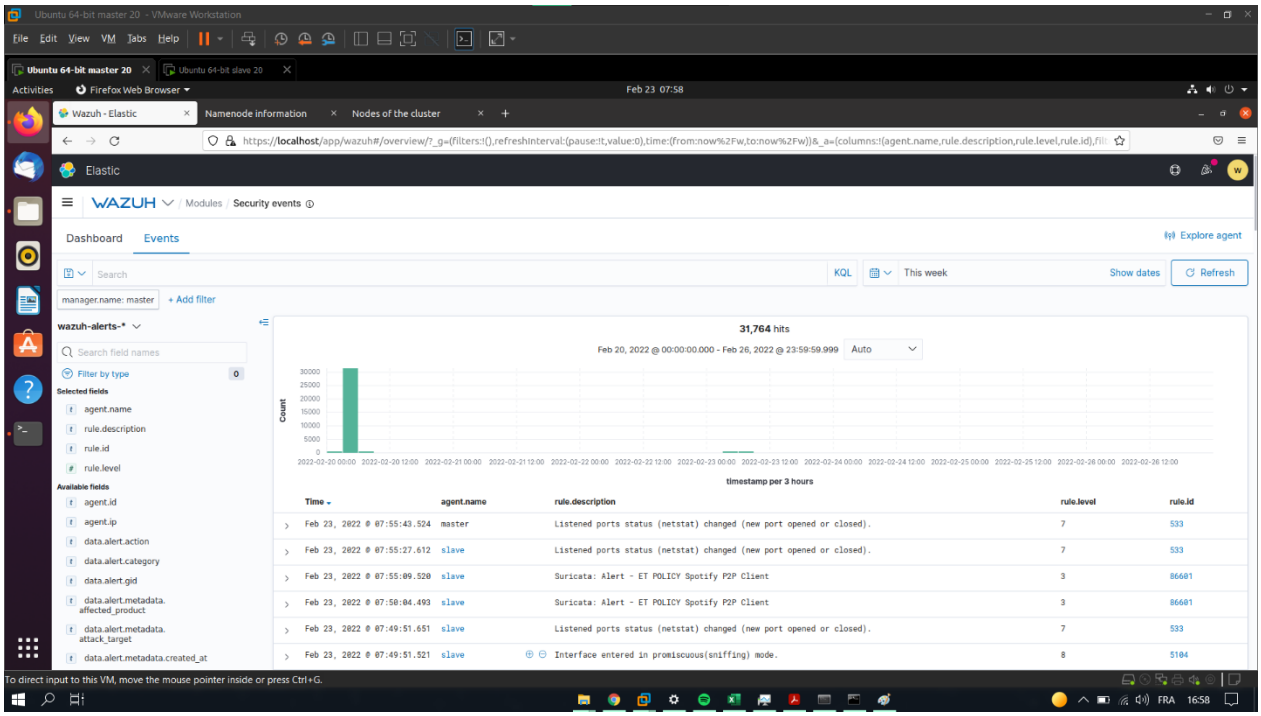
- Agents Preview



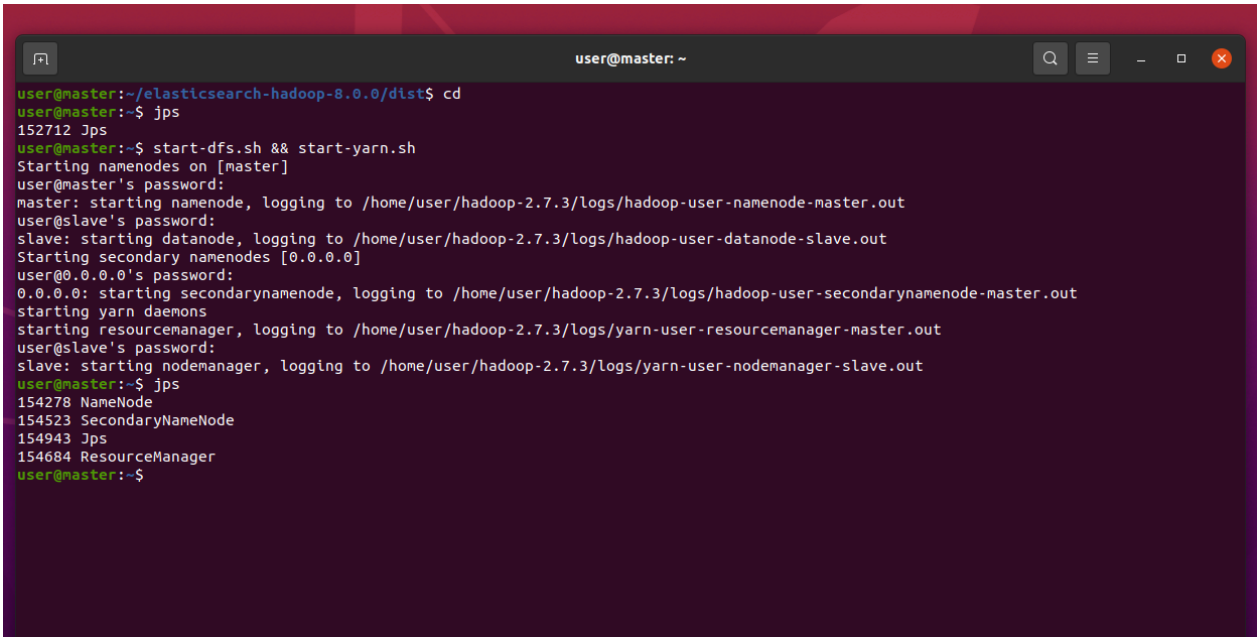
- Security Events



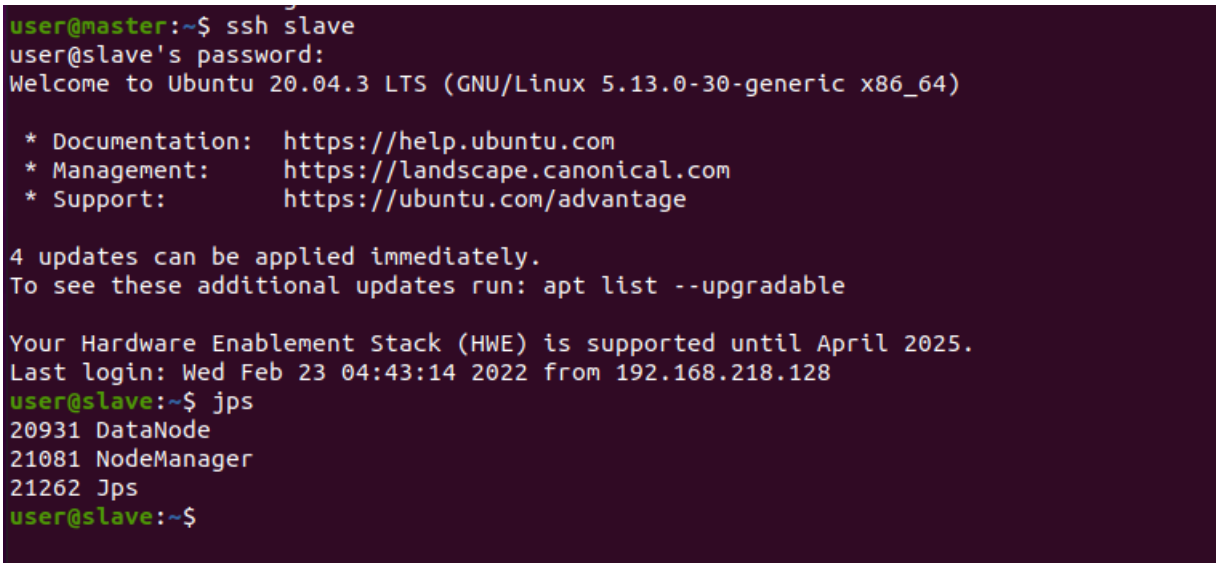
- Events : Filter by type



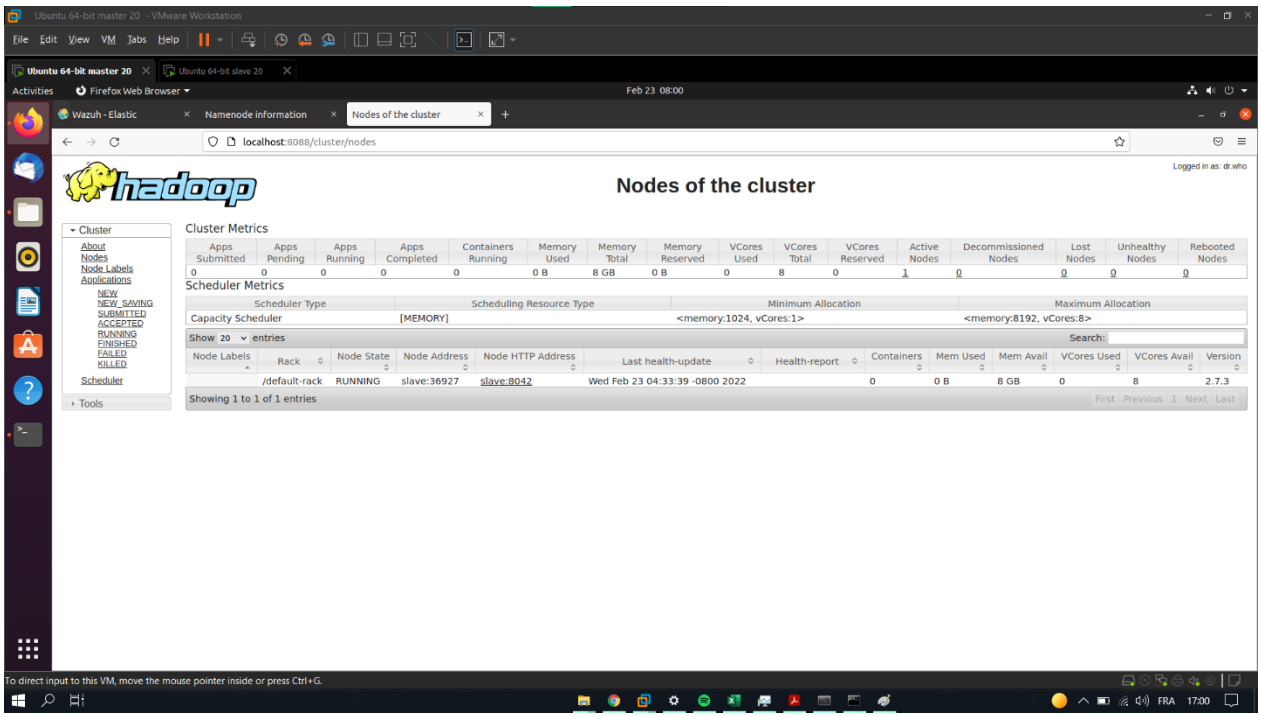
- JPS on Master node



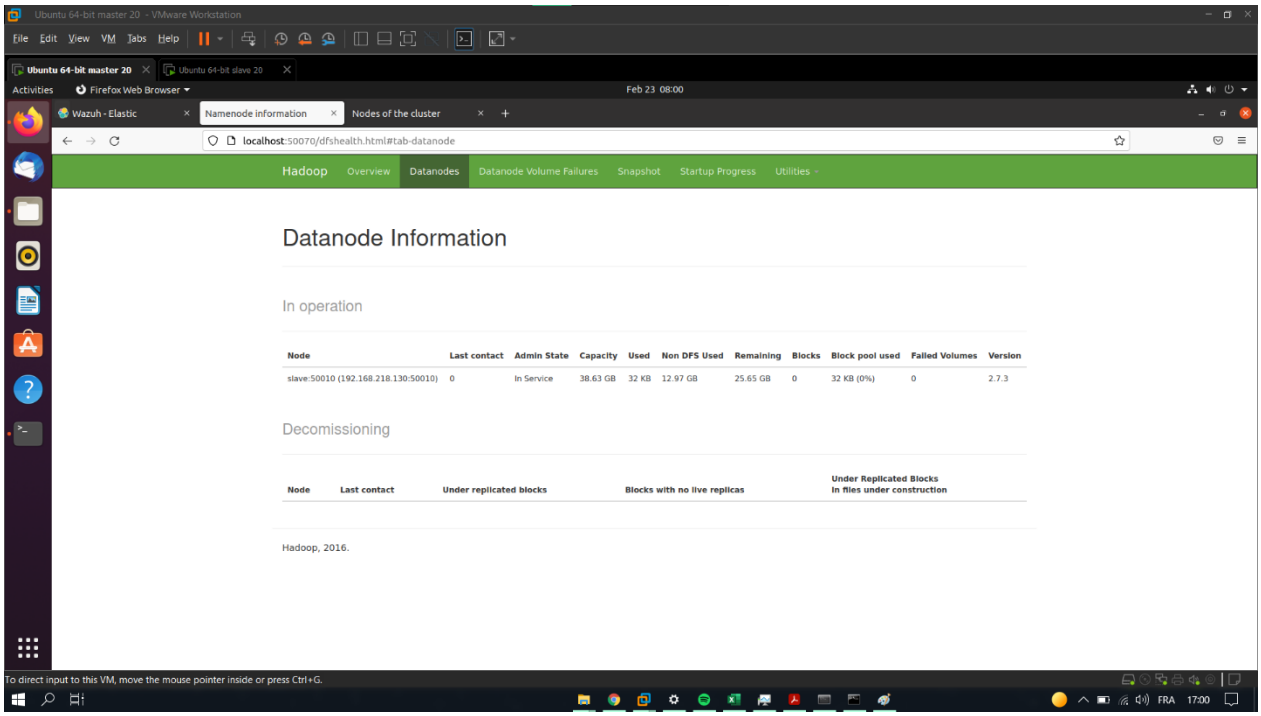
- JPS on Slave node



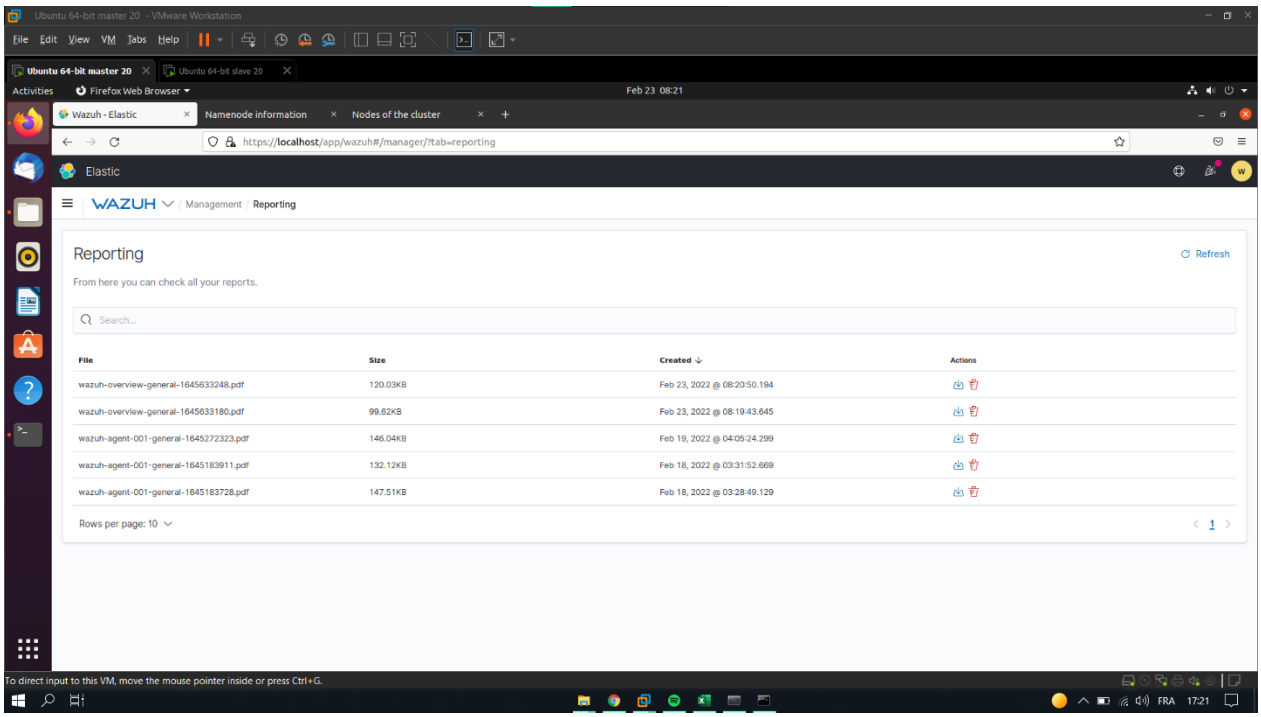
- Nodes Of The Cluster



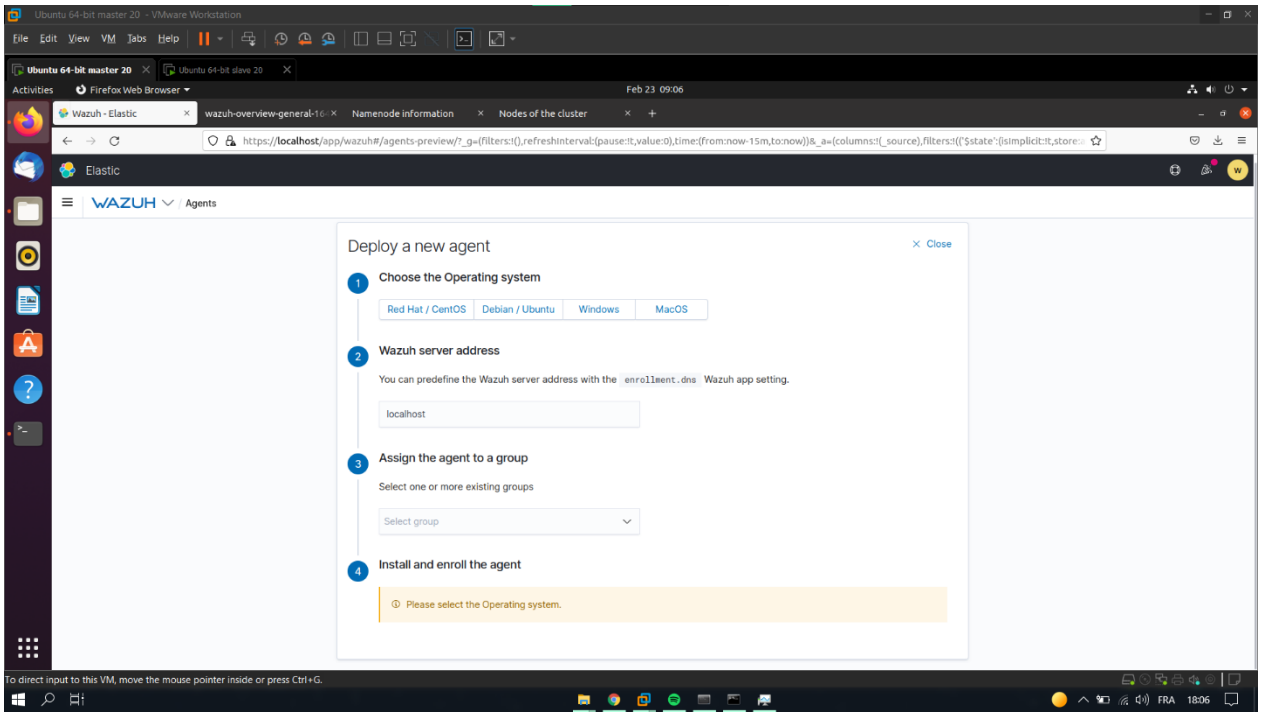
- Live node count and info about each live node



- Wazuh-Reporting



- Deploy a new agent



```
curl -so wazuh-agent-4.2.5.deb https://packages.wazuh.com/4.x/apt/pool/main/w/wazuh-agent/wazuh-agent_4.2.5-1_amd64.deb && sudo WAZUH_MANAGER='localhost' WAZUH_AGENT_GROUP='default' dpkg -i ./wazuh-agent-4.2.5.deb
```

Copy command

6 Start the agent

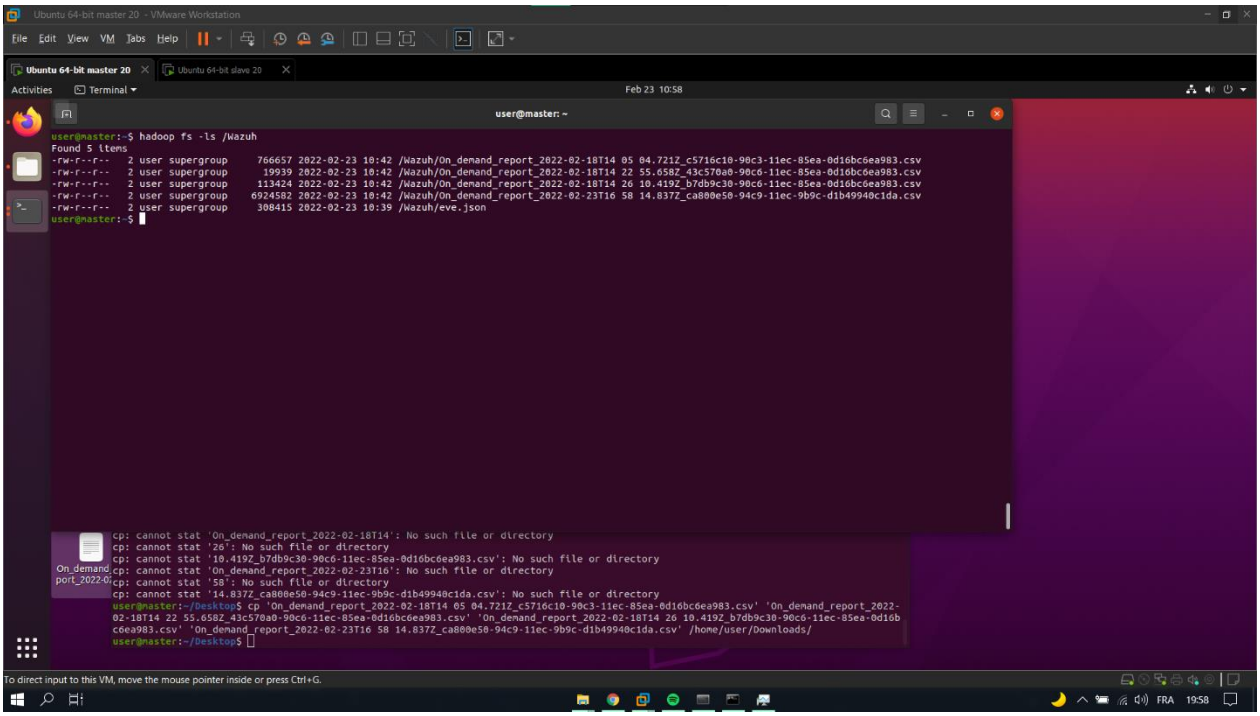
Systemd | SysV Init

```
sudo systemctl daemon-reload
sudo systemctl enable wazuh-agent
sudo systemctl start wazuh-agent
```

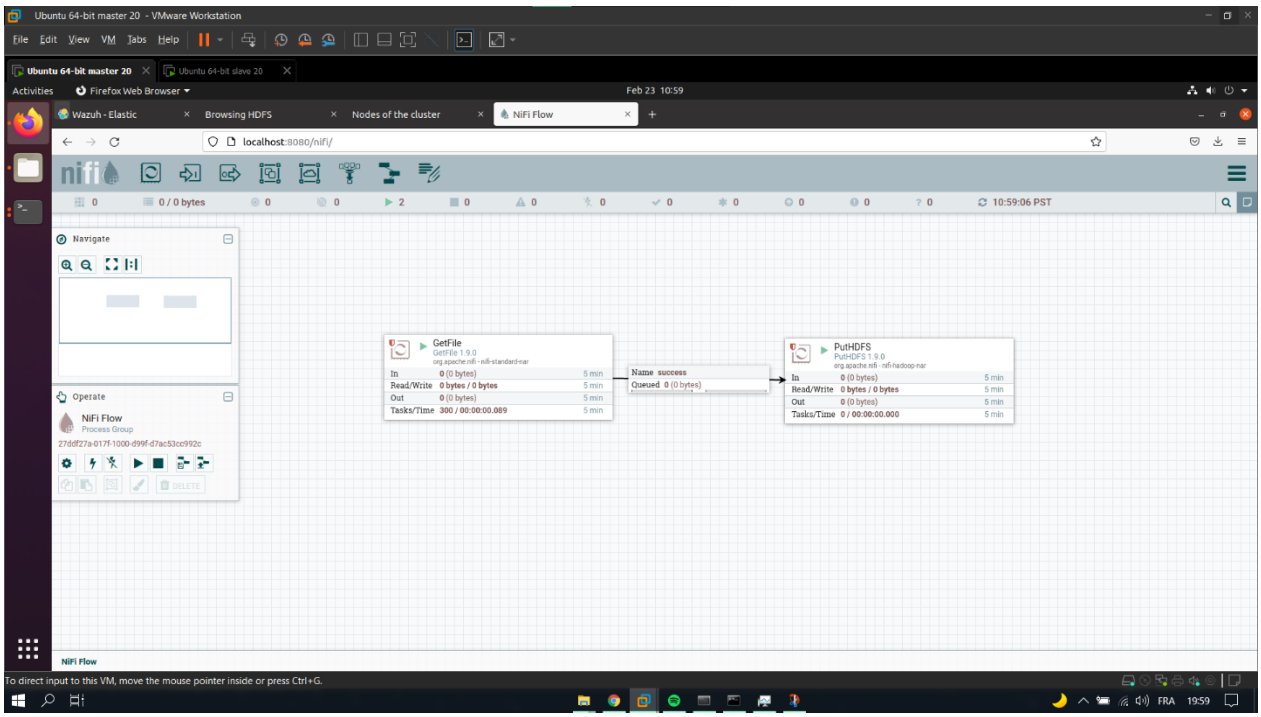
Copy command

To verify the connection with the Manager, please follow this document.

- Wazuh-Reporting save into Hadoop-Hdfs



- Send this Wazuh-Reporting to File Hadoop System



Conclusion générale

Tout au long de ce projet, nous avons appris de nouvelles technologies qui nous seront sans doute d'une grande utilité dans nos futures études et dans notre vie professionnelle. Cependant, avant d'arriver à la phase de réalisation, l'analyse et l'étude profonde du projet nous ont menées à en comprendre l'objectif, et donc de proposer une solution optimale.

Ce travail nous a donné la chance d'avoir un premier contact avec les systèmes de détection d'intrusion, ce qui a été une expérience très enrichissante et significative, dans le sens où on apprend que grâce à cette science on verra notre quotidien se révolutionner 100% sécurisé.

En addition, on a appris à travailler en groupe, à offrir le meilleur de nous dans une équipe afin d'avoir un résultat performant. Enfin, Ce projet nous a donné l'opportunité d'approfondir nos connaissances dans l'implémentation de problèmes d'apprentissage automatique et de mettre en pratique toutes les connaissances acquises grâce au module d'analyse prédictive, et surtout de nous familiariser avec les problèmes qui peuvent survenir à tout moment.

Bibliographie

[1] 2019 IJRAR November 2019, Volume 6, Issue 4
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3677845

[2] Citation Sara Abdalelah Abbas et Mahdi S. Almhanna 2021 J. Phys. : Conf. Ser. 1804 012136
https://iopscience.iop.org/article/10.1088/1742-6596/1804/1/012136

[3] 2009 Joint Conferences on Pervasive Computing (JCPC), Publisher: IEEE
https://ieeexplore.ieee.org/document/6996133

[4] Sensors 2021, Andrew Churcher, Rehmat Ullah, Jawad Ahmad, Sadaqat ur Rehman, Fawad Masood, Mandar Gogate, Fehaid Alqahtani, Boubakr Nour and William J. Buchanan
https://www.mdpi.com/1424-8220/21/2/446

[5] Ibrahim Yahya Mohammed AL-Mahbashi1, Prashant Chauhan, Shivi Shukla and M. B. Potdar - Vol-2 Issue-6 2016
http://ijariie.com/AdminUploadPdf/Review_on_Efficient_Log_Analysis_to_Evaluate_Multiple_Honeypots_using_ELK_ijariie3355.pdf

[6] Conference Paper in Lecture Notes in Computer Science · December 2016
https://www.researchgate.net/publication/310790755

Webographie

https://www.datto.com/blog/cybersecurity-101-intro-to-the-top-10-common-types-of-cybersecurity-attacks

https://www.kaspersky.com/resource-center/definitions/what-is-cyber-security

https://www.ibm.com/topics/cybersecurity#:~:text=Cybersecurity%20is%20the%20practice%20of,sensitive%20information%20from%20digital%20attacks.&text=Security%20system%20complexity%2C%20created%20by,expertise%2C%20can%20amplify%20these%20costs

https://www.omnisci.com/blog/the-rise-of-big-data-analytics-in-cyber-defense

https://www.elastic.co/fr/beats/

https://www.syloe.com/glossaire/logstash/

https://www.geeksforgeeks.org/hadoop-ecosystem/

https://www.elastic.co/fr/what-is/kibana