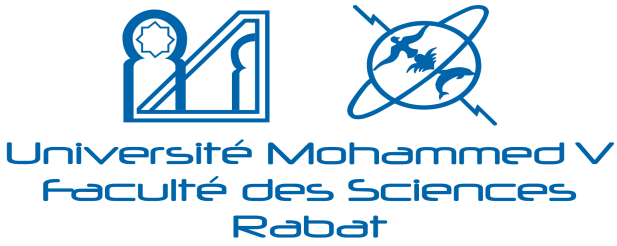

Master de recherche informatique et télécommunication MIT

Détection d'intrusions par l'analyse Big Data des fichiers logs

Projet tutoré présenté par :

AMMOURI Yassire
ALLAL Yahia
AIT ICHOU Mustapha

Sous la direction de

Prof. Yassine Benjelloune Touimi

Devant le jury :

Prof. Yassine Benjelloune Touimi Professeur à la Faculté des Sciences de Rabat

Année Universitaire : 2023-2024

Table des matières

Remerciements **V**

Introduction générale **1**

1 Système de Détection d'Intrusion **3**
1.1 Introduction . . . 3
1.2 Notion de base . . . 4
1.2.1 Sécurité d'un système (informatique ou d'information) . . . 4
1.2.2 Service de sécurité . . . 4
1.3 Fichiers Logs et Intrusions . . . 5
1.3.1 Définition . . . 5
1.3.2 Type des fichiers Logs) . . . 6
1.3.3 Qu'est-ce qu'une intrusion . . . 7
1.4 Qu'est ce qu'un système de détection d'intrusion IDS . . . 8
1.4.1 Définition . . . 8
1.4.2 Type d'IDS . . . 8
1.5 Méthode de détection des intrusions . . . 9
1.5.1 Détection par signature . . . 9
1.5.2 Détection par anomalie . . . 10
1.5.3 Détection par hybride . . . 10
1.6 Conclusion . . . 11

2 Big Data **12**
2.1 Introduction . . . 12
2.2 Définition : . . . 12
2.3 Architecture Big Data . . . 13
2.4 Big data et la sécurité informatique (Cyber sécurité) . . . 15
2.4.1 La sécurité des Big Data . . . 16
2.4.2 Solution de sécurité pour les environnements Big Data . . . 16
2.4.3 Le Big Data au service de la sécurité . . . 17
2.5 Conclusion . . . 18

3 Intelligence Artificiel (IA) **19**
3.1 Introduction . . . 19
3.2 Définition . . . 19
3.3 Machine Learning . . . 20
3.3.1 Definition . . . 20

3.3.2 Supervised Machine Learning . . . 20
3.3.3 Unsupervised Machine Learning . . . 23
3.4 Deep Learning . . . 25
3.4.1 Introduction . . . 25
3.4.2 Définition . . . 26
3.4.3 Artifitial Neural Network (ANN) . . . 26
3.4.4 Composantes d'un Réseau de Neurone . . . 26
3.4.5 Les Réseaux de Neurones Convolutifs (CNN) . . . 27
3.4.6 Comment fonctionnent les réseaux neuronaux convolutifs ? . . . 28
3.4.7 Conclusion . . . 28

4 Implémentation et Réalisation **29**
4.1 Les ressources materielles et logicielles . . . 29
4.1.1 Matériels utilisés . . . 29
4.1.2 Logicielles utilisés . . . 29
4.2 Architecture & Overview du Système . . . 32
4.2.1 Architecture générale du système . . . 32
4.2.2 Comment fonctionne Kafka . . . 34
4.2.3 DataSet utilisé : KDD Cup 99 . . . 35
4.2.4 Démarche suivie pour la conception du modèle de machine learning . . . 35
4.2.5 Résultats et mise en œuvre réelle . . . 37
4.3 Conclusion générale . . . 39

Table des figures

1.1 Services principaux de la sécurité informatique [2]. 4
1.2 Schéma montrant une intrusion causée par un intrus au système. 7

2.1 Schéma de l'architecture Big Data. 14

3.1 Decision Tree 21
3.2 Random Forest 22
3.3 Régression Linéaire 23
3.4 K-means clustring 25
3.5 La Structure d'un Reseau de Neurone 27
3.6 Comment fonctionnent les CNN 28

4.1 Architecture générale du système 32
4.2 les deffirent partition de Kafka 34
4.3 Exploration de l'ensemble de données 35
4.4 Nombre de Protocoles et Types d'Attaques 36
4.5 étiquettes prédites 36
4.6 étiquettes prédites 37
4.7 Capture d'un système en temps réel 37

Remerciements

Nous tenons à exprimer notre sincère gratitude au Professeur Yassine Benjelloune Touimi pour ses conseils et ses idées précieuses, ainsi que pour son examen attentif de ce travail et l'intérêt qu'elle porte à notre projet. Nous souhaitons également exprimer notre reconnaissance envers le laboratoire de recherche Informatique et Télécommunications pour les privilèges qui nous ont été accordés, en particulier au Professeur MINAOUI Khalid pour son soutien continu.

Enfin, nous tenons à remercier chaleureusement nos collègues pour leur soutien et leur aide précieuse dans la correction et le développement de ce rapport. Nous sommes reconnaissants envers toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail, et nous leur sommes très reconnaissants pour leur soutien et leurs encouragements tout au long du processus.

Introduction générale

Ce rapport aborde trois axes majeurs : les systèmes de détection d'incréments, le Enormous Information et l'intelligence artificielle appliquée en sécurité informatique.

La première partie se concentre sur les systèmes de détection d'intrusion. Nous aborderons les ideas de base, la sécurité des systèmes d'information, les sorts de fichiers log, et les différentes méthodes de détection des interruptions telles que la détection standard signature, standard anomalie et standard méthode hybride.

La deuxième partie du rapport explore le monde du Big Data, allant de ce qu'est le Big Data et pourquoi il est devenu l'un des secteurs majeurs en informatique aujourd'hui, jusqu'à expliquer sa structure et technologie, et finir par l'importance du Big Data dans le secteur de la cybersécurité (sécurité informatique) et les solutions qu'il propose afin de se préserver contre les attaques des hackers.

La troisième partie looks at l'intelligence artificielle. We open with a general presentation of AI before delving into the procedures for supervised and unsupervised learning, ANN, and CNN. Explain how they work and what their role has been in security.

En dernier lieu, nous allons décrire tout d'abord l'implémentation et la réalisation des concepts présentés ci-dessus. Nous allons décrire les res-

sources matérielles et logicielles utilisées ; l'architecture proposée ; les données du dataset KDDCup99 qui nous ont permis de réaliser le modèle de machine learning ; ainsi que les résultats obtenus dans cette étude.

L'objectif de ce compatibility est de fournir une compréhension approfondie des advances actuelles de sécurité des systèmes d'information, du traitement des données à grande échelle et de l'intelligence artificielle, ainsi que de leur application pratique dans la détection et la prévention des intrusions.

Chapitre 1

Système de Détection d'Intrusion

1.1 Introduction

Les systèmes informatiques occupent une place prédominante dans les entreprises, les administrations et dans le quotidien des particuliers, c'est pour ça la sécurité informatique est devenue primordiale, et d'une priorité absolue pour la bonne exploitation de ces systèmes. Et pour assurer cette sécurité il faut connaître tous les obstacles que nous pouvons rencontrons.

Parmi ces obstacles nous avons les logiciels malveillants(Malwares), les attaques informatiques qui représentent une réelle menace pour la sécurité des systèmes informatiques. Dans cette partie nous nous sommes intéressées au Malware qui constitue une menace à croissance rapide dans le monde informatique moderne. La production des logiciels malveillants est devenue une industrie de plusieurs milliards de dollars.

Dans ce chapitre, nous commençons par présenter les différentes propriétés relatives à la sécurité informatique, une classification des malwares informatiques selon différents aspects, ensuite nous abordons les différents moyens et techniques de protection contre les malwares informatiques.

1.2 Notion de base

1.2.1 Sécurité d'un système (informatique ou d'information)

C'est un ensemble de moyens techniques, organisationnels, juridiques et humains nécessaires et mis en place pour réduire la vulnérabilité d'un système contre les menaces accidentels ou intentionnels an d'assurer les services de sécurité [1].

1.2.2 Service de sécurité

FIGURE 1.1 – Services principaux de la sécurité informatique [2].

— **Confidentialité** La confidentialité est à peu près équivalente à la vie privée. Les mesures de confidentialité sont conçues pour empêcher les tentatives d'accès non autorisées aux informations sensibles. Il est courant que les données soient classées en fonction de la quantité et du type de dommages qui pourraient être causés si elles tombaient entre de mauvaises mains. Des mesures plus ou moins contraignantes peuvent alors être mises en place selon ces catégories.

— **Intégrité** implique de maintenir la cohérence, l'exactitude et la fiabilité des données tout au long de leur cycle de vie. Les données ne doivent pas être modifiées en transit et des mesures doivent être prises pour s'assurer que les données ne peuvent pas être modifiées par des personnes non autorisées (par exemple, en cas de violation de la confidentialité).

— **Disponibilité** La disponibilité signifie que les informations doivent être constamment et facilement accessibles pour les parties autorisées. Cela implique de maintenir correctement l'infrastructure matérielle et technique et les systèmes qui contiennent et affichent les informations.

1.3 Fichiers Logs et Intrusions

1.3.1 Définition

À chaque utilisation, différentes commandes et opérations sont exécutées sur un ordinateur, un serveur ou un routeur. Le scénario idéal verrait chaque système fonctionner sans aucun accroc. Toutefois, si des erreurs et des problèmes se produisent ou si le besoin d'optimiser certains aspects se présente, il est nécessaire de vérifier et d'analyser les opérations qui se sont déroulées. Pour cette raison, la plupart des systèmes courants créent des fichiers .log. Ces derniers tiennent lieu de journal de bord ou de journal du système, dans lequel tous les événements peuvent être consignés et analysés ultérieurement. Un fichier .log est généralement un document dans lequel les différents événements relatifs au système sont enregistrés en texte clair.

Cela permet aux utilisateurs de vérifier par la suite comment certaines erreurs ont pu se produire ou de constater si des actions indésirables ont été effectuées dans un système. Certains fichiers .log vous permettent de récupérer des données supprimées ou perdues. Il est important de noter ici que la taille des fichiers .log est souvent limitée. Lorsque cette valeur

est atteinte, les anciennes entrées sont automatiquement supprimées à la faveur des lignes les plus récentes.

1.3.2 Type des fichiers Logs)

Les fichiers logs se présentent sous différents types, chacun étant généré pour répondre à des besoins spécifiques. Parmi les plus courants, on trouve les suivants :

Log système : Ils enregistrent les événements généraux du système, y compris les messages du noyau et les événements du système d'exploitation.

Log de sécurité et d'audit : Suivi des événements liés à la sécurité, tels que les tentatives de connexion, les succès et les échecs d'authentification et les incidents de sécurité.

Log d'application : Capture les événements spécifiques aux applications et aux services, tels que les erreurs, les avertissements et les informations opérationnelles.

Log de réseau : Ils enregistrent les activités et les erreurs liées aux services de réseau, telles que les demandes d'accès et les incidents de serveur.

Log de maintenance et de sauvegarde : Documentent les activités de maintenance, y compris l'exécution des tâches planifiées et les opérations de sauvegarde.

Log de dépannage : Ils contiennent des informations utiles au diagnostic et à la résolution des problèmes du système, notamment les messages générés lors du démarrage et les détails des pannes.

Log de configuration : Ils enregistrent les actions liées à la gestion et à la configuration des logiciels, telles que les installations, les mises à jour et les suppressions de paquets.

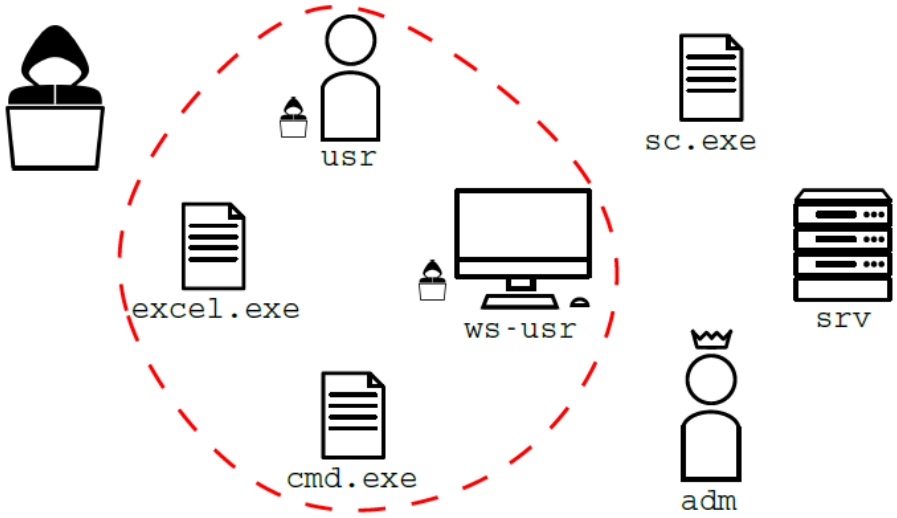

FIGURE 1.2 – Schéma montrant une intrusion causée par un intrus au système.

1.3.3 Qu'est-ce qu'une intrusion

Une intrusion peut être définie comme l'entrée non autorisée d'une personne dans un périmètre donné. Dans le contexte de la sécurité de l'information, les attaques sophistiquées sont menées par des menaces persistantes avancées (APT) qui sont difficiles à détecter par les systèmes de détection d'intrusion traditionnels. Ces attaques visent à voler des informations, mais peuvent également inclure des actions de sabotage ou d'extorsion.

Entre l'intrusion initiale et l'impact final, l'attaquant mène plusieurs actions, offrant ainsi des opportunités de détection. La chaîne de la mort cybernétique de Lockheed Martin définit sept étapes d'une intrusion avancée : reconnaissance, armement, livraison, exploitation, installation, commande et contrôle, et action finale. Pendant ces étapes, l'attaquant collecte des informations sur le réseau ciblé, conçoit un logiciel malveillant, le livre au réseau, l'exploite pour obtenir un accès initial, installe une porte dérobée persistante, établit une communication avec l'infrastructure de l'attaquant, puis atteint son objectif final.

1.4 Qu'est ce qu'un système de détection d'intrusion IDS

1.4.1 Définition

Les systèmes de détection des intrusions sont l'un des moyens possibles de sécuriser les données ou les informations. Il existe d'autres moyens, comme la cryptographie ou la sténographie, qui permettent à l'utilisateur de partager des informations sans les compromettre. Mais ces approches sont préventives et impliquent de cacher les données ou de les crypter.

Un intrus est un utilisateur légitime qui utilise la vulnérabilité du système pour pénétrer dans les zones du système qui lui sont réservées et accéder ainsi à l'information. Il est donc nécessaire de disposer d'un outil permettant d'identifier ces activités. Les systèmes de détection d'intrusion nous fournissent donc des outils permettant de détecter ces activités. Une fois l'activité détectée, le travail du système de prévention consiste à faire en sorte que la même vulnérabilité ne soit pas utilisée à nouveau pour une attaque d'intrusion.

1.4.2 Type d'IDS

Selon l'integration du system de detection d'intrusion, on destingue deux type d'IDS :

Network based intrusion detection system

Les systèmes de détection d'intrusion sont classés en deux grandes catégories : les systèmes de détection d'intrusion basés sur l'hôte et les systèmes de détection d'intrusion basés sur le réseau. Les NIDS surveillent le réseau en temps réel, identifiant ainsi les paquets malveillants et détectant les attaques telles que le déni de service, le dépassement de mémoire tampon, l'analyse de protocole, les attaques CGI.

Host based intrusion detection system

Comme son nom l'indique, le système IDS basé sur l'hôte prend en compte les données d'un système unique, telles que les mémoires tampons, les journaux du système, le système de fichiers et divers événements. Il dépend principalement des données de la piste d'audit et des journaux d'appels système pour détecter les comportements anormaux. Il existe deux approches pour détecter les intrusions : la détection des abus et la détection des intrusions basée sur les anomalies. Détection des abus.

1.5 Méthode de détection des intrusions

Les méthodes de détection d'intrusions IDS peuvent être classées en trois grandes catégories : la détection par anomalie, la détection par signature et la détection hybride. Voici un aperçu de chacune de ces méthodes :

1.5.1 Détection par signature

La détection basée sur les signatures analyse les paquets réseau pour détecter les caractéristiques ou comportements d'attaque propres à une menace spécifique. Une séquence de code apparaissant dans une variante particulière de logiciel malveillant est un exemple de signature d'attaque.

Un IDS basé sur les signatures maintient une base de données de signatures d'attaque avec laquelle il compare les paquets réseau. Si un paquet déclenche une correspondance avec l'une des signatures, l'IDS le signale. Pour être efficaces, les bases de données de signatures doivent être régulièrement mises à jour avec de nouveaux renseignements sur les menaces à mesure que de nouvelles cyberattaques émergent et que les attaques existantes évoluent. Les toutes nouvelles attaques qui n'ont pas encore

0. IDS : Intrusion Detection System

été analysées pour les signatures peuvent échapper à un IDS basé sur les signatures.

1.5.2 Détection par anomalie

Les méthodes de détection basées sur les anomalies font appel à l'apprentissage automatique pour créer et affiner en permanence un modèle de référence de l'activité réseau normale. Elles comparent ensuite l'activité réseau au modèle et signalent les écarts, comme un processus qui utilise plus de bande passante que d'habitude ou un appareil qui ouvre un port habituellement fermé.

Étant donné qu'ils signalent tout comportement anormal, les IDS basés sur les anomalies peuvent souvent détecter de nouvelles cyberattaques susceptibles d'échapper à la détection basée sur les signatures. Par exemple, les IDS basés sur les anomalies peuvent détecter des attaques "*zero-day*", c'est-à-dire des attaques qui exploitent les vulnérabilités des logiciels avant que le développeur ne s'en aperçoive ou n'ait eu le temps d'appliquer un correctif.

Toutefois, les IDS basés sur les anomalies peuvent également être plus sujets aux faux positifs. Même une activité bénigne, comme un utilisateur autorisé accédant à une ressource réseau sensible pour la première fois, peut déclencher un IDS basé sur les anomalies.

1.5.3 Détection par hybride

La méthode de détection des intrusions hybride combine les techniques de détection par signature et les techniques de détection par anomalie pour améliorer l'efficacité et la précision de la détection des intrusions. Cette approche cherche à tirer parti des avantages des deux méthodes tout en minimisant leurs inconvénients.

1.6 Conclusion

Dans ce chapitre, nous avons présenté de manière globale les différents concepts relatifs au domaine de la sécurité informatique, en mettant un accent particulier sur les différentes méthodes de détection des intrusions. Nous avons discuté des notions de base, y compris la sécurité des systèmes informatiques et des informations, ainsi que des services de sécurité. Nous avons également abordé les fichiers de log et les intrusions, en définissant ce qu'est une intrusion et en explorant les types de fichiers de log utilisés pour la détection des intrusions.

Ensuite, nous avons expliqué ce qu'est un système de détection d'intrusion (IDS), ses définitions et ses différents types. Enfin, nous avons détaillé les méthodes de détection des intrusions, y compris la détection par signature, par anomalie et par approche hybride, en soulignant les avantages et les inconvénients de chaque méthode.

Dans le chapitre suivant, nous nous intéresserons au domaine du big data avec ses différentes méthodes, telles que les méthodes ensemblistes, et les mesures d'évaluation.

Chapitre 2

Big Data

2.1 Introduction

Depuis la révolution numérique, la quantité de données produites chaque jour dans ou en dehors d'un Système d'Information a pris de telles proportions qu'il est difficile de continuer à utiliser les outils traditionnels pour les manipuler de façon performante. Il est devenu nécessaire de développer de nouvelles méthodes pour gérer et analyser cette énorme quantité d'informations. Ainsi, est né le Big Data ; un concept qui porte sur la recherche, l'analyse, la capture, le stockage, le partage et la présentation de ces données.

2.2 Définition :

Littéralement, le terme Big Data signifie mégadonnées, grosses données ou encore données massives. Il désigne un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut traiter. En effet, on procrée environ 2,5 trillions d'octets de données tous les jours. Ce sont les informations provenant de partout : messages que nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore.

Cependant, aucune définition précise ou universelle ne peut être donnée au Big Data. Étant un objet complexe, sa définition varie selon les com-

munautés qui s'y intéressent en tant qu'usager ou fournisseur de services. Parmi les autres définitions, nous citons :

1. **Selon Gartner** : les Big Data sont des ressources d'information volumineuses, à grande vélocité et à grande variété qui exigent des formes innovantes et rentables de traitement de l'information pour améliorer la compréhension et la prise de décision.
2. **Selon Lisa Diforti** : le Big Data est un ensemble de technologies, d'architecture, d'outils et de procédures permettant à une organisation de très rapidement capter, traiter et analyser de larges quantités et contenus hétérogènes et changeants, et d'en extraire les informations pertinentes à un coût accessible.

2.3 Architecture Big Data

L'architecture Big Data est une infrastructure essentielle pour la gestion et l'analyse efficaces de vastes ensembles de données provenant de diverses sources. Elle fournit un cadre robuste pour collecter, stocker, traiter et analyser ces données afin de générer des insights significatifs pour les organisations. Cette architecture est conçue pour faire face aux défis uniques posés par le volume massif, la variété des formats, la vélocité de génération et la véracité des données. Dans cette section, nous explorerons les différents composants et étapes clés de l'architecture Big Data, ainsi que leur rôle dans la gestion du cycle de vie des données.

La Figure 2.1 montre le schéma de l'architecture Big Data, qui illustre le flux des données depuis leur source jusqu'à la visualisation et le reporting, en passant par les différentes étapes de traitement et de stockage.

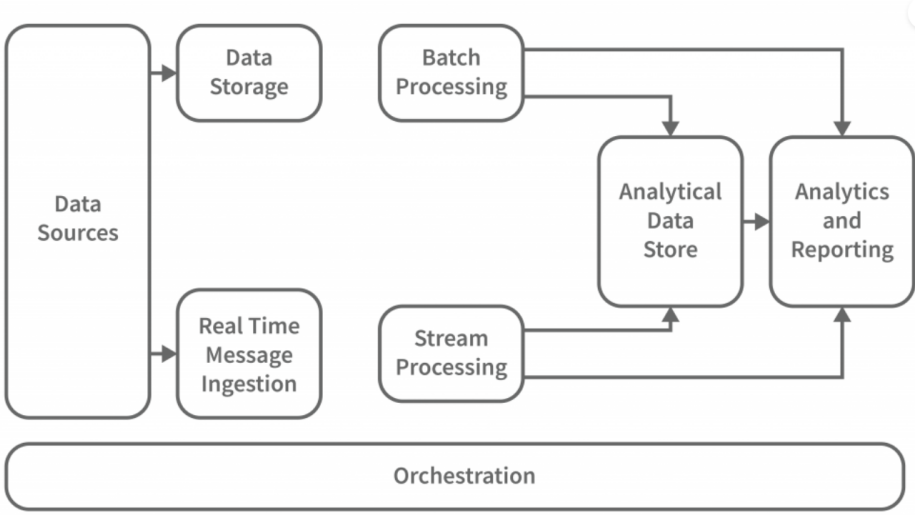

FIGURE 2.1 – Schéma de l'architecture Big Data.

— **Sources de Données** :
  — Données Structurées : Provenant de bases de données relationnelles, fichiers CSV, etc.
  — Données Semi-Structurées : Provenant de fichiers XML, JSON, etc.
  — Données Non-Structurées : Provenant des médias sociaux, emails, vidéos, etc.
  — Données en Temps Réel : Provenant de capteurs, journaux d'événements, etc.
— **Ingestion des Données** :
  — Batch Processing : Ingestion de grandes quantités de données à intervalles réguliers.
  — Stream Processing : Ingestion de flux de données en temps réel.
— **Stockage des Données** :
  — Data Lake : Stockage de données brutes et non structurées.
  — Data Warehouse : Stockage de données structurées et prêtes pour l'analyse.
— **Traitement des Données** :

  - Traitement Batch : Utilise des frameworks comme Hadoop pour traiter de grandes quantités de données en lots.
  - Traitement en Temps Réel : Utilise des outils comme Apache Kafka et Apache Spark pour traiter des flux de données en temps réel.
- **Analyse des Données** :
  - Analyse Descriptive : Analyse des données historiques pour comprendre ce qui s'est passé.
  - Analyse Prédictive : Utilisation de techniques de machine learning pour prédire des tendances futures.
  - Analyse Prescriptive : Recommandations basées sur l'analyse des données pour optimiser les décisions.
- **Visualisation et Reporting** :
  - Outils de Visualisation : Utilisation de tableaux de bord interactifs et de graphiques pour représenter les données de manière compréhensible.
  - Rapports : Génération de rapports détaillés pour la prise de décision.

Chaque étape de cette architecture est conçue pour gérer efficacement le volume, la variété, la vélocité et la véracité des données, permettant ainsi aux organisations de tirer des informations précieuses et de prendre des décisions éclairées basées sur leurs données.

2.4 Big data et la sécurité informatique (Cyber sécurité)

Le Big Data suscite énormément de débats dans le domaine de sécurité, mais de quoi discutons-nous vraiment ? En termes de cyber sécurité, le Big Data représente à la fois une opportunité et une menace pour les

entreprises. Alors deux problèmes différents se posent : d'une part la sécurité des informations de l'entreprise et de ses clients dans un contexte de Big Data, d'autre part l'utilisation des techniques du Big Data pour analyser, ou prévoir, les incidents de sécurité.

2.4.1 La sécurité des Big Data

Un grand nombre d'entreprises utilisent le Big Data pour le marketing et les recherches, mais ne maîtrisent pas forcément les concepts de base, en particulier la sécurité. Ces entreprises utilisent cette technologie pour stocker et analyser des pétaoctets de données, notamment les journaux Web, les données sur le parcours de navigation et le contenu des réseaux sociaux, et ce, dans le but de mieux connaître leurs clients et leurs activités. Cependant, les cybercriminels peuvent eux aussi profiter de ces opportunités pour accéder à des quantités massives d'informations sensibles en utilisant les technologies les plus avancées. Par conséquent, la classification des informations devient encore plus critique et il convient de déterminer la propriété des informations pour permettre une classification acceptable. Pour cela, les entreprises doivent faire recours à des techniques telles que le chiffrement pour protéger les données sensibles et appliquer les contrôles d'accès et les outils analytiques du Big Data pour détecter et prévenir les cyber attaques.

2.4.2 Solution de sécurité pour les environnements Big Data

Comme nous avons déjà mentionné les environnements Big Data doivent être dotés d'une protection des données sensibles. À cet effet une solution de IBM® Security Guardium® assure les environnements Big Data en :

— Surveillant étroitement l'activité des bases de données Hadoop et NoSQL par les applications et les utilisateurs en temps réel ; déclenchant des alertes en cas de violation de règles ; en faisant le suivi des tentatives d'accès et d'utilisation des données afin de déceler

tout comportement inhabituel parmi les utilisateurs privilégiés et externes ; et en notifiant les tableaux de bord SIEM (Security Information and Event Management) pour engager les mesures correctives adéquates (alerte, blocage, résiliation de connexion).

— Implémentant des contrôles automatisés et centralisés au sein de l'entreprise (bases de données, applications, fichiers, Big Data, etc.).

— Protégeant les données sensibles au moyen de techniques de chiffrement, de masquage et d'occultation.

— Évaluant et résolvant les faiblesses de l'environnement de manière à sécuriser l'ensemble du Big Data.

En résumé, l'objectif de la solution Guardium est d'améliorer la sécurité des données et les décisions en matière de sécurité en se fondant sur des informations exploitables et priorisées, issues du contrôle et de la surveillance de l'ensemble de l'environnement.

2.4.3 Le Big Data au service de la sécurité

À l'heure des Big Data, ces nouvelles ressources doivent permettre aux entreprises de dépasser les niveaux de sécurité classiques. Désormais les Big Data vont permettre aux entreprises d'accéder à un troisième niveau de protection contre les cyber attaques.

En ce qui concerne les paliers de sécurité informatique, on peut distinguer trois niveaux :

1. Le premier niveau, qui consiste à sécuriser l'entreprise des attaques provenant de l'extérieur.
2. Le deuxième niveau est cependant plus complexe, puisqu'il s'agit de protéger l'entreprise de l'intrus, vis-à-vis de ses propres utilisateurs qui peuvent être une source de vulnérabilités au sein du système d'information.

3. Le troisième niveau de sécurité consiste à mesurer l'impact que la menace détectée a eu sur l'infrastructure, ce qui suppose de pouvoir remonter dans le temps, donc il s'agit d'engager le trafic qui transite par les réseaux de l'entreprise afin de traquer le chemin emprunté par le logiciel malveillant dès qu'il a ete repere et de prendre des mesures e caces en consequence.

2.5 Conclusion

Dans ce deuxième chapitre, nous avons abordé la technologie Big Data en détaillant ses définitions, ses caractéristiques, ainsi que les différents domaines dans lesquels elle est utilisée. Nous avons présenté les fondements du Big Data, soulignant son importance croissante dans divers secteurs grâce à sa capacité à gérer et analyser des volumes de données massifs.

En fin de chapitre, nous avons exploré la relation bidirectionnelle entre le Big Data et la sécurité informatique. La cybersécurité est devenue un enjeu majeur de Big Data, car la taille et la complexité des données de sécurité sont désormais trop vastes pour être gérées et analysées par les outils de sécurité traditionnels. Le Big Data offre la possibilité d'améliorer la connaissance de la situation et la sécurité de l'information en détectant et en prévenant les attaques en temps réel. De plus, l'analyse des données fournie par le Big Data permet de prédire les attaques et les menaces dans un contexte de sécurité. En conclusion, nous avons montré comment le Big Data représente une avancée révolutionnaire dans divers domaines, et en particulier dans la cybersécurité, qui constitue le sujet principal de notre étude.

Chapitre 3

Intelligence Artificiel (IA)

3.1 Introduction

L'IA est devenue un terme fourre-tout pour désigner toute application qui exécute des tâches complexes qui nécessitaient auparavant une intervention humaine, comme communiquer avec des clients en ligne ou jouer aux échecs. Le terme est souvent utilisé de manière interchangeable avec les domaines qui composent l'IA, tels que l'apprentissage automatique (ML) et l'apprentissage profond (deep learning).

3.2 Définition

L'intelligence artificielle (IA) est une branche de l'informatique qui se consacre à la création de systèmes capables d'accomplir des tâches qui requièrent généralement l'intelligence humaine. Ces tâches comprennent l'apprentissage, le raisonnement, la résolution de problèmes, la perception, la compréhension du langage et même la créativité. L'IA est devenue une partie intégrante de la technologie moderne, influençant divers domaines et industries.

3.3 Machine Learning

3.3.1 Definition

machine learning est une forme d'intelligence artificielle (IA) qui vise 'a cr eer des syst'emes qui apprennent ou am eliorent leurs performances en fonction des donn ees qu'ils traitent. Les algorithmes sont les moteurs de Machine Learning. En g en eral, deux grands types d'algorithmes d'apprentissage automatique sont utilis es aujourd'hui : Supervised Machine Learning (l'apprentissage supervis e) et Unsupervised Machine Learning L'apprentissage supervisé vs l'apprentissage non supervisé. La différence entre les deux est définie par la méthode utilisée pour traiter les données afin de faire des prédictions.

3.3.2 Supervised Machine Learning

apprentissage Les algorithmes d'apprentissage automatique supervisé sont les plus couramment utilisés. Dans ce modèle, un scientifique des données sert de guide et enseigne à l'algorithme les conclusions qu'il doit tirer. Tout comme un enfant apprend à identifier les fruits en les mémorisant dans un livre d'images, dans l'apprentissage supervisé, l'algorithme apprend à partir d'un ensemble de données déjà étiquetées et dont le résultat est prédéfini. les algorithmes d'apprentissage supervisée sont classés en : - les algorithmes de classification. - les algorithmes de régression.

Algorithmes de classification

Les algorithmes de classification sont utilisés pour attribuer des étiquettes de classe aux données en fonction des caractéristiques fournies.

— **Arbres de Décision** : Un arbre de décision est un algorithme d'apprentissage supervisé non paramétrique, utilisé pour les tâches de classification et de régression. Il présente une structure arborescente

hiérarchique, composée d'un nœud racine, de branches, de nœuds internes et de nœuds feuilles.
Comme le montre le diagramme ci-dessus, un arbre décisionnel commence par un nœud racine, qui ne comporte aucune branche entrante. Les branches sortantes du nœud racine alimentent ensuite les nœuds internes, également appelés nœuds de décision. Sur la base des caractéristiques disponibles, les deux types de nœuds effectuent des évaluations pour former des sous-ensembles homogènes, désignés par les nœuds feuilles ou les nœuds terminaux. Les nœuds feuilles représentent tous les résultats possibles au sein de l'ensemble de données.

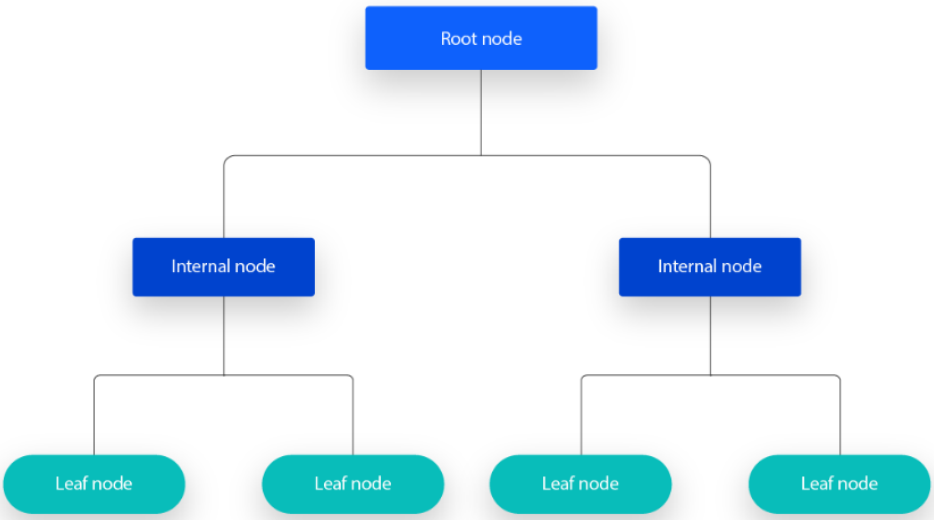

FIGURE 3.1 – Decision Tree

— **Random Forest** :
La forêt aléatoire est un algorithme d'apprentissage automatique couramment utilisé, dont la marque a été déposée par Leo Breiman et Adele Cutler, qui combine les résultats de plusieurs arbres de décision pour obtenir un résultat unique. Sa facilité d'utilisation et sa flexibilité ont favorisé son adoption, car il traite à la fois les problèmes de classification et de régression.

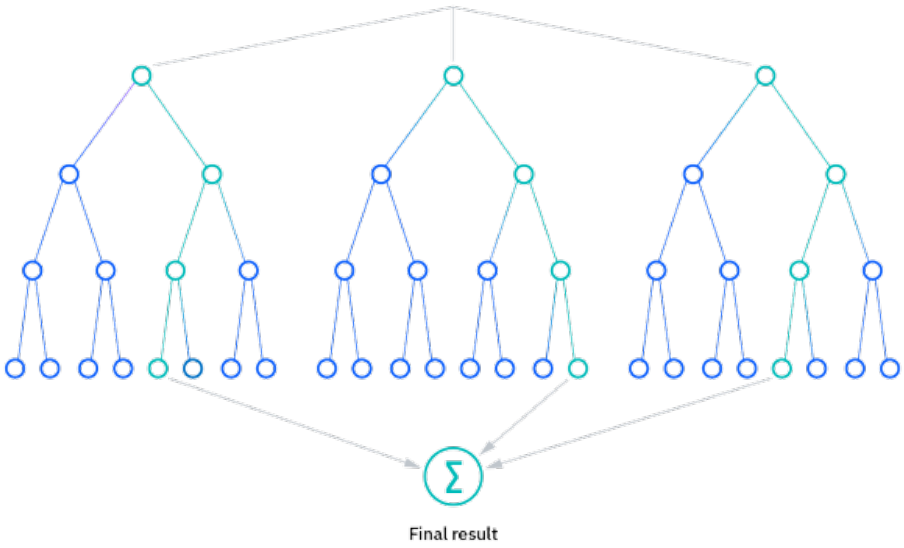

FIGURE 3.2 – Random Forest

Algorithmes de régression.

Les algorithmes de régression sont utilisés pour prédire des valeurs continues en fonction des entrées. Voici quelques-uns des algorithmes de régression les plus couramment utilisés :

— **Régression Linéaire** :
La régression linéaire est une technique d'analyse de données qui prédit la valeur de données inconnues en utilisant une autre valeur de données apparentée et connue. Il modélise mathématiquement la variable inconnue ou dépendante et la variable connue ou indépendante sous forme d'équation linéaire. Supposons, par exemple, que vous disposiez de données sur vos dépenses et vos revenus de l'année dernière. Les techniques de régression linéaire analysent ces données et déterminent que vos dépenses représentent la moitié de vos revenus. Ils calculent ensuite une dépense future inconnue en réduisant de moitié un revenu futur connu.

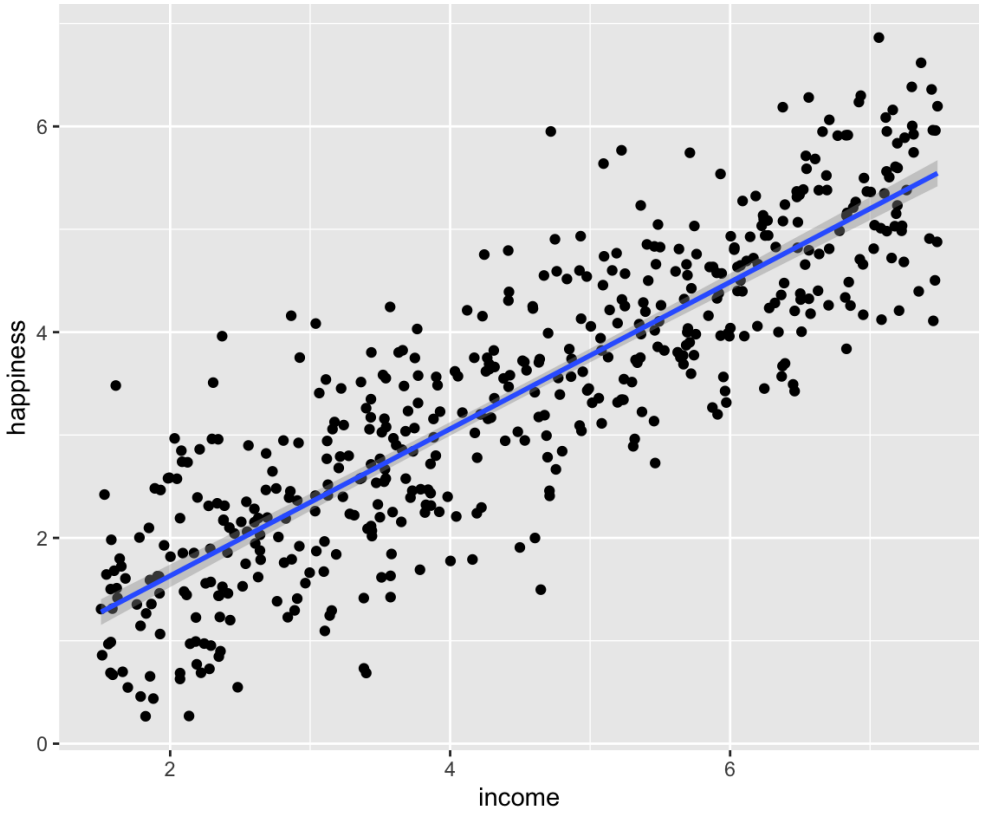

FIGURE 3.3 – Régression Linéaire

3.3.3 Unsupervised Machine Learning

Unsupervised machine learning utilise une approche plus indépendante dans laquelle un ordinateur apprend à identifier des processus et des modèles complexes sans orientation humaine cohérente et rigoureuse. L'apprentissage automatique non supervisé implique une formation sur des données sans étiquette ni résultat spécifique défini. On distingue trois types de l'apprentissage non supervisé :
- Les algorithmes de clustering.
- Les algorithmes d'association.
- Les algorithmes de réduction dimensionnelle.

Algorithmes de clustering

Pour trouver des groupes d'objets similaires . On demande à la machine de grouper des objets dans des ensembles de données les plus homogènes

possible.
Cette technique peut sembler proche de celle de la classication dans l'apprentissage supervis é, mais à la diérence de cette dernière, les classes ne sont pas pré-remplies par un humain, c'est la machine qui invente ses propres classes, à un niveau de nesse pas toujours évident pour un humain.

Algorithmes d'association

Pour trouver des liens entre des objets. est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données.

Algorithmes de réduction dimensionnelle

a réduction de la dimensionnalité consiste à passer d'un espace d'apprentissage à haute dimension à un espace de calcul plus petit. En d'autres termes, il s'agit de réduire le nombre de variables ou de caractéristiques utilisées pour entraîner le modèle d'IA. Si les données sont représentées dans un tableau, la réduction de la dimensionnalité consistera à réduire le nombre de colonnes. Quant à un modèle tridimensionnel tel qu'un cube ou une sphère, il peut être réduit à un seul plan, un carré ou un cercle. Et parmi les algorithmes les plus populaire d'apprentissage non supervisée on trouve :

— **K-means clustring** :
Le regroupement par k-moyennes est une méthode de quantification vectorielle, issue du traitement du signal, qui vise à répartir n observations en k grappes dans lesquelles chaque observation appartient à la grappe dont la moyenne est la plus proche (centre de la grappe ou centroïde de la grappe), qui sert de prototype de la grappe. Il en résulte un partitionnement de l'espace de données en cellules de Voronoï. Le regroupement par k-moyennes minimise les variances à

l'intérieur des grappes (distances euclidiennes au carré), mais pas les distances euclidiennes régulières, ce qui serait le problème de Weber le plus difficile : la moyenne optimise les erreurs au carré, alors que seule la médiane géométrique minimise les distances euclidiennes.

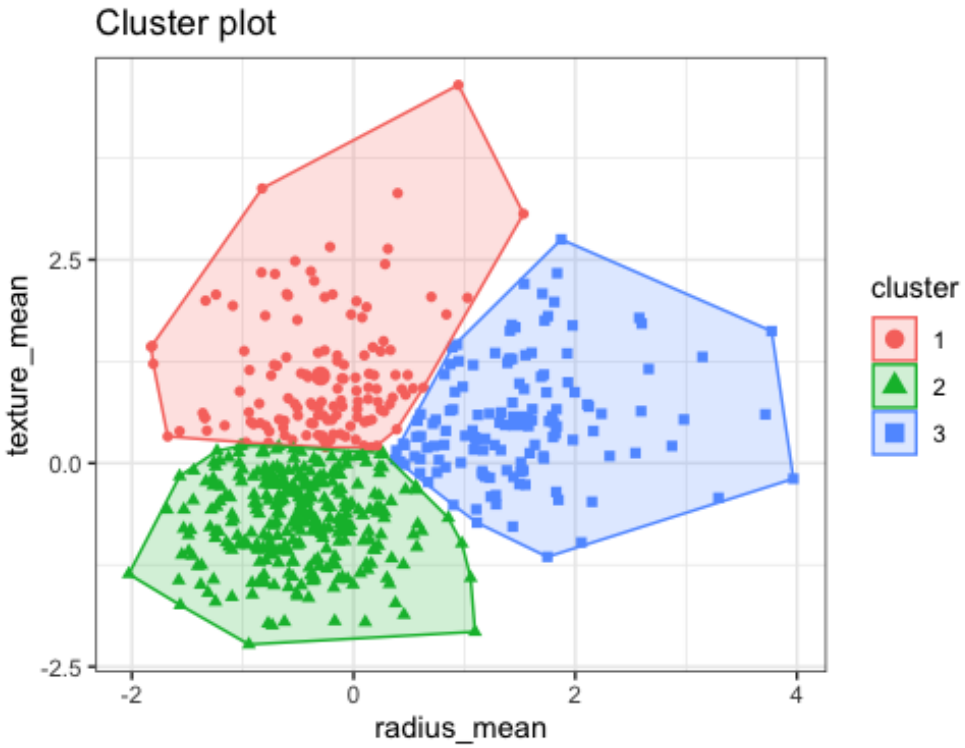

FIGURE 3.4 – K-means clustring

3.4 Deep Learning

3.4.1 Introduction

Le Deep Learning est un nouveau domaine de recherche du Machine Learning, qui a été introduit dans le but de rapprocher le ML de son objectif principal L'intelligence artificielle. Il concerne les algorithmes inspirés par la structure et le fonctionnement du cerveau. Ils peuvent apprendre plusieurs niveaux de représentation dans le but de modéliser des relations complexes entre les données.

3.4.2 Définition

Le Deep Learning est un ensemble d'algorithmes d'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux, correspondant à diférents niveaux d'abstraction. Il a la capacité d'extraire des caractéristiques à partir des données brutes grâce aux multiples couches de traitement composé de multiples transformations linéaires et non linéaires et apprendre sur ces caractéristiques petites à petit à travers chaque couche avec une intervention humaine minime.[7]

3.4.3 Artifitial Neural Network (ANN)

Un réseau de neurones articiels est un système dont la conception est à l'origine inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques. Les réseaux de neurones articiels sont des réseaux fortement connectés par des processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire (neurone articiel) calcule une sortie unique sur la base des informations qu'ils reçoivent.

3.4.4 Composantes d'un Réseau de Neurone

Un réseau de neurones est composé de :

Neurones : ensemble de fonction

Ils prennent une donnée d'entrée et produisent une donnée de sortie. Un certain nombre de neurones sont groupés en couches .

Couches : groupement de neurones

Les couches contiennent des neurones et aident à faire circuler l'information. Il existe au moins deux couches dans un réseau de neurones : la couche d'entrée et la couche de sortie. Les couches, autres que les couches d'entrée et de sortie, sont appelées les couches cachées .

Poids et Biais : valeurs numériques Les poids et biais sont des variables du modèle qui sont mises à jour pour améliorer la pré- cision du réseau.

Un poids est appliqué à l'entrée de chacun des neurones pour calculer une donnée de sortie.Les réseaux de neurones mettent à jour ces poids de manière continue. Il existe donc une boucle de rétro-action mise en ÷uvre dans la plupart des réseaux de neurones. Les biais sont également des valeurs numériques qui sont ajoutées une fois que les poids sont appliqués aux valeurs d'entrée.

Fonction d'activation : Algorithmes mathématiques appliqués aux valeurs de sortie. Les fonctions d'activation lissent ou normalisent la donnée de sortie avant qu'elle ne soit transmise aux neurones suivants. Ces fonctions aident les réseaux de neurones à apprendre et à s'améliorer.

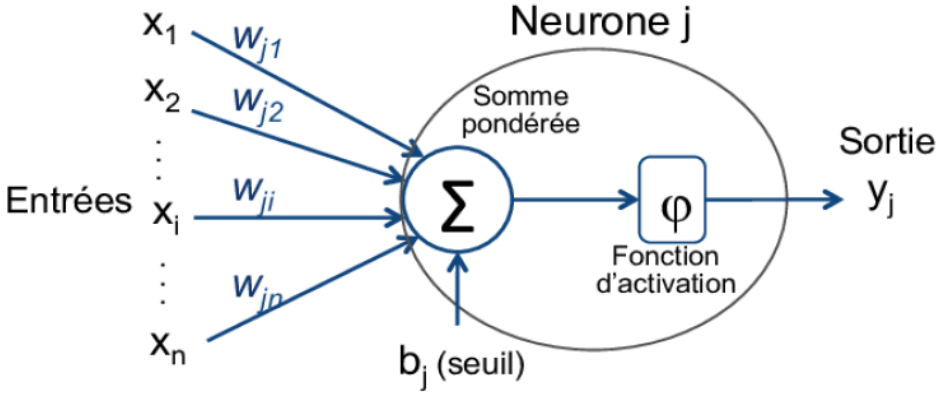

FIGURE 3.5 – La Structure d'un Reseau de Neurone

3.4.5 Les Réseaux de Neurones Convolutifs (CNN)

Les réseaux neuronaux sont un sous-ensemble de l'apprentissage automatique et sont au cœur des algorithmes d'apprentissage profond. Ils sont constitués de couches de nœuds, contenant une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque nœud se connecte à un autre et est associé à un poids et à un seuil. Si la sortie d'un nœud individuel est supérieure à la valeur seuil spécifiée, ce nœud est activé et envoie des données à la couche suivante du réseau. Dans le cas contraire, aucune donnée n'est transmise à la couche suivante du réseau.

3.4.6 Comment fonctionnent les réseaux neuronaux convolutifs ?

Les réseaux neuronaux convolutifs se distinguent des autres réseaux neuronaux par leurs performances supérieures avec les signaux d'image, de parole ou audio. Ils comportent trois principaux types de couches, à savoir

— la couche convolutive

— Couche de mise en commun

— Couche entièrement connectée (FC)

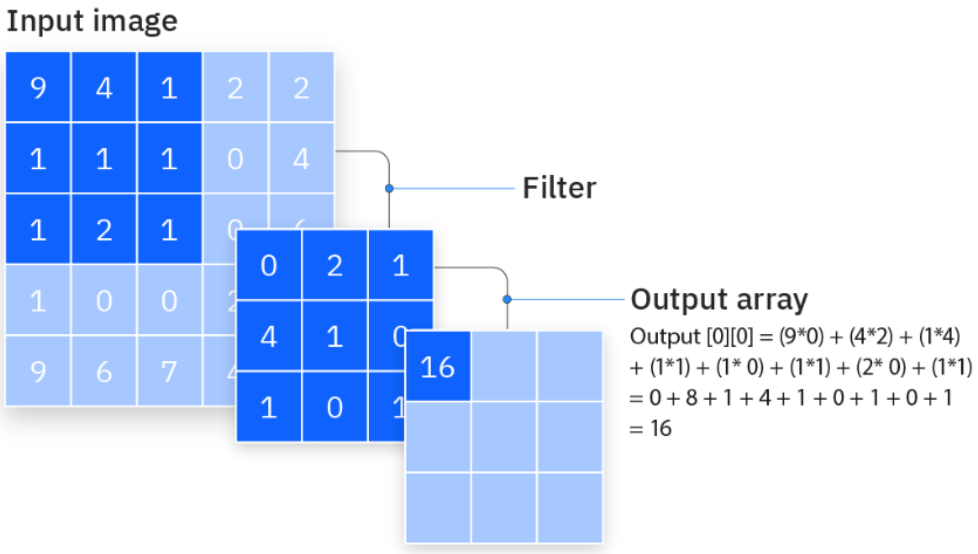

FIGURE 3.6 – Comment fonctionnent les CNN

3.4.7 Conclusion

Ce chapitre a offert une vue d'ensemble des concepts fondamentaux de l'intelligence artificielle (IA) et de ses sous-domaines. Nous avons exploré l'apprentissage automatique, distinguant l'apprentissage supervisé et non supervisé, et avons détaillé les algorithmes couramment utilisés.

Dans la section sur l'apprentissage profond, nous avons expliqué les réseaux de neurones artificiels (ANN), leurs composants, et le rôle des fonctions d'activation. Nous avons également abordé les réseaux de neurones convolutifs (CNN) et leur capacité à traiter les données visuelles.

Chapitre 4

Implémentation et Réalisation

4.1 Les ressources materielles et logicielles

Dans cette section, nous vous presentons les ressources materielles et logicielles que nous avons utilisees pour mettre en uvre le cas dutilisation propose :

4.1.1 Matériels utilisés

L'implémentation de notre système a été réalisée sur une machine virtuelle possédant les caractéristiques suivantes :

— Processeur : 4 core.

— Mémoire : 4 Go de RAM.

— Disque dur : 100 Go.

4.1.2 Logicielles utilisés

a) Système d'exploitation :

Ubuntu 16.02 : Ubuntu est un système d'exploitation Linux complet, il est adapté pour être utilisé comme poste de travail ou serveur.

b) Outils de développement :

Python3 : Python est un langage de programmation interprété à usage

général, interactif, orienté objet et de haut niveau. Python combine une puissance remarquable avec une syntaxe claire. Il comporte des modules, des classes, des exceptions, des types de données dynamiques de très haut niveau et un typage dynamique. Il existe des interfaces pour de nombreux appels système et bibliothèques, ainsi que pour divers systèmes de fenêtrage.

Bibliothèque Python3 Scikit-Learn :

Scikit-learn est une bibliothèque libre développée en Python destinée à l'apprentissage automatique (machine learning), elle propose plusieurs types d'algorithmes de classification, régression et regroupement et peut être utilisée comme middleware, notamment pour des tâches de prédiction.

Snort :

Snort est un système de prévention et de détection d'intrusions open-source qui analyse le trafic réseau en temps réel.

Zeek :

Zeek est un cadre de surveillance du réseau de haute performance, flexible et open-source.

TensorFlow :

TensorFlow est une bibliothèque open-source développée par Google pour le calcul numérique et l'apprentissage automatique.

Apache Spark :

Apache Spark est un moteur de traitement de données rapide dédié au Big Data. Il permet d'effectuer un traitement de larges volumes de données de manière distribuée. Ses principaux avantages sont sa vitesse, sa simplicité d'usage et sa polyvalence.

Kafka :

Apache Kafka est une plateforme distribuée de diffusion de données en continu, capable de publier, stocker, traiter et souscrire à des flux d'enregistrement en temps réel. Elle est conçue pour gérer des flux de données provenant de plusieurs sources et les fournir à plusieurs utilisateurs. En bref, elle ne se contente pas de déplacer un volume colossal de données d'un point A à un point B : elle peut le faire depuis n'importe quels points vers n'importe quels autres points, selon vos besoins et même simultanément.

4.2 Architecture & Overview du Système

4.2.1 Architecture générale du système

Le schéma suivant représente l'architecture générale de notre système :

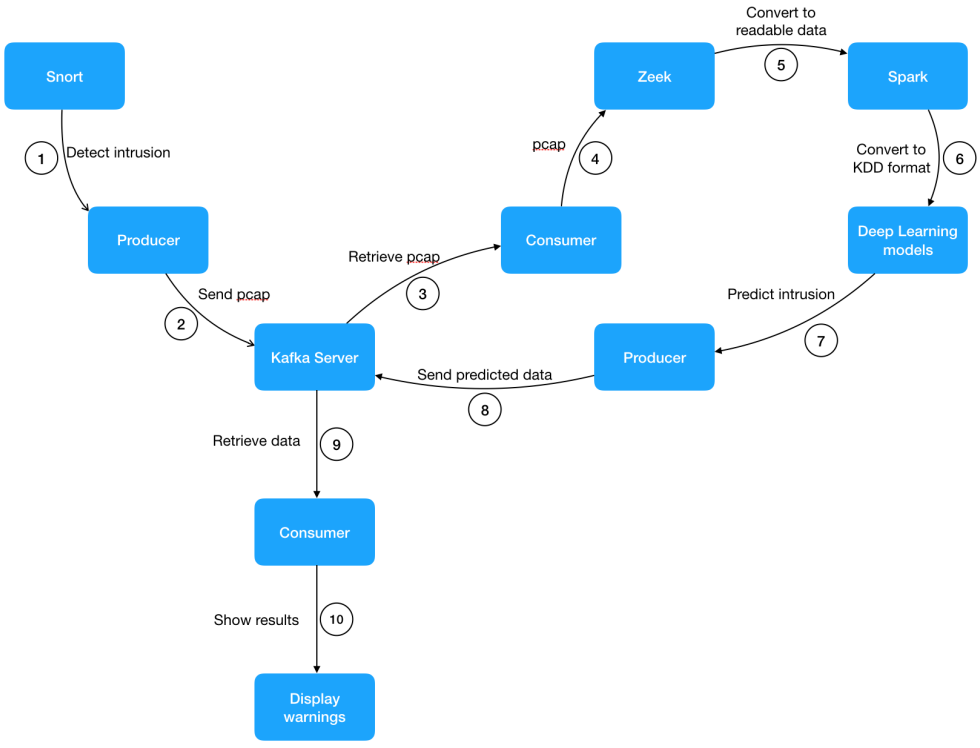

FIGURE 4.1 – Architecture générale du système

Dans notre système, nous utilisons Snort comme IDS pour renifler et capturer les paquets suspects. Kafka joue un rôle rôle de plateforme de

traitement en continu. Dans notre architecture, le concept de base de Kafka est le suivant d'avoir des producteurs et des consommateurs qui communiquent avec le serveur Kafka. Les producteurs publient des données sur Les producteurs publient des données sur le serveur et les consommateurs récupèrent les données de ce serveur pour les traiter.

Nous utilisons Zeek comme langage de programmation optimisé pour le réseau afin d'analyser les logs pcap et de les transformer en données lisibles. Zeek (anciennement Bro) est la première plateforme mondiale de surveillance de la sécurité des réseaux. Spark est utilisé pour convertir les données au format KDDCup99 que nous avons déjà présenté dans le rapport précédent. Ensuite, un modèle d'IA prédit à quel type d'attaque les données appartiennent, ou s'il s'agit simplement de données normales. données.

La figure 4.1 présente une vue d'ensemble du système. Comme nous pouvons le voir, ce processus comporte 10 étapes.. Nous expliquons brièvement chaque brièvement chaque étape comme suit.

1. Snort capture le trafic suspect sur la base de règles prédéfinies.
2. Un producteur envoie le trafic suspect au serveur Kafka.
3. Un consommateur récupère les données envoyées à l'étape 2.
4. Le consommateur attend un lot de N (argument prédéfini) échantillons de données, puis applique Zeek.
5. Zeek transforme les données au format pcap en données lisibles. Ces données lisibles ont également leur propre format pour permettre le calcul des données KDDCup99.
6. Spark convertit les données au format KDDCup99.
7. Le modèle d'IA prédit le type d'attaques pour les données.
8. Un producteur renvoie les données prédites au serveur Kafka.
9. Un consommateur récupère les données prédites sur le serveur Kafka.
10. Afficher les résultats et avertir les utilisateurs.

4.2.2 Comment fonctionne Kafka

Apache Kafka fonctionne comme un cluster qui stocke des enregistrements(record), appelés événements, dans des catégories appelées sujets(topics). Chaque enregistrement dans Kafka se compose d'une clé, d'une valeur et d'un horodatage. Les sujets sont essentiellement des journaux qui contiennent de nombreux enregistrements et sont divisés en partitions. Ces partitions sont des séquences ordonnées et immuables d'enregistrements et peuvent être répliquées sur plusieurs brokers. Un brokers est un serveur Kafka chargé de stocker les données et de répondre aux demandes des clients. En répartissant les partitions entre les courtiers, Kafka garantit la tolérance aux pannes et l'évolutivité. Cette architecture permet à Kafka de traiter efficacement de grands volumes de données, ce qui en fait un choix idéal pour les pipelines de données en temps réel et les applications de streaming.

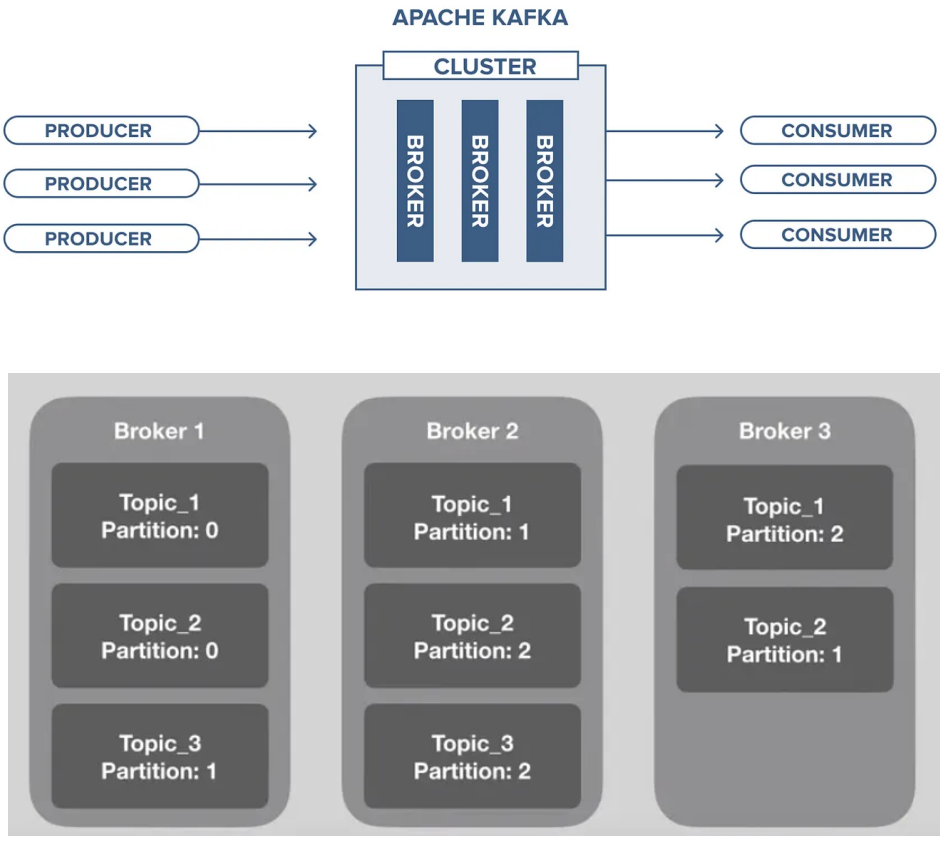

FIGURE 4.2 – les deffirent partition de Kafka

4.2.3 DataSet utilisé : KDD Cup 99

KDD Cup 99 est un ensemble de données largement utilisé dans la recherche en sécurité informatique, en particulier pour les systèmes de détection d'intrusion (IDS). Cet ensemble de données a été créé pour la compétition KDD Cup de 1999, organisée dans le cadre de la Conférence Internationale sur la Découverte de Connaissances et l'Extraction de Données (KDD) [8].

— **Nombre d'Instances :** Environ 4,9 millions d'instances de connexions réseau.

— **Nombre de features :** 42 features

— **Types d'Attaques :** 22 types d'attaques réparties en quatre catégories principales :

  — **Denial of Service (DoS), Remote to Local (R2L),User to Root (U2R), Probing**

4.2.4 Démarche suivie pour la conception du modèle de machine learning

Dans cette partie, nous abordons l'aspect technique de la solution proposée, en expliquant les algorithmes de machine learning utilisés pour prédire les attaques, le traitement préalable à la prédiction, ainsi que d'autres éléments liés à notre solution et Préparation des Données.

Exploration de l'ensemble de données

Exploring the dataset

```
[ ]  # Read Train and Test dataset
     data_train = pd.read_csv("/content/sample_data/KDDTrain+.txt")

[ ]  # Check data
     data_train.head()
```

	0	tcp	ftp_data	SF	491	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.18	2	2.1	0.00	0.00.1	0.00.2	0.00.3	1.00	0.00.4	0.00.5
0	0	udp	other	SF	146	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	1	0.0	0.0	0.0	0.0	0.08	0.15	0.00
1	0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	123	6	1.0	1.0	0.0	0.0	0.05	0.07	0.00
2	0	tcp	http	SF	232	8153	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	5	5	0.2	0.2	0.0	0.0	1.00	0.00	0.00
3	0	tcp	http	SF	199	420	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	30	32	0.0	0.0	0.0	0.0	1.00	0.00	0.09
4	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	121	19	0.0	0.0	1.0	1.0	0.16	0.06	0.00

FIGURE 4.3 – Exploration de l'ensemble de données

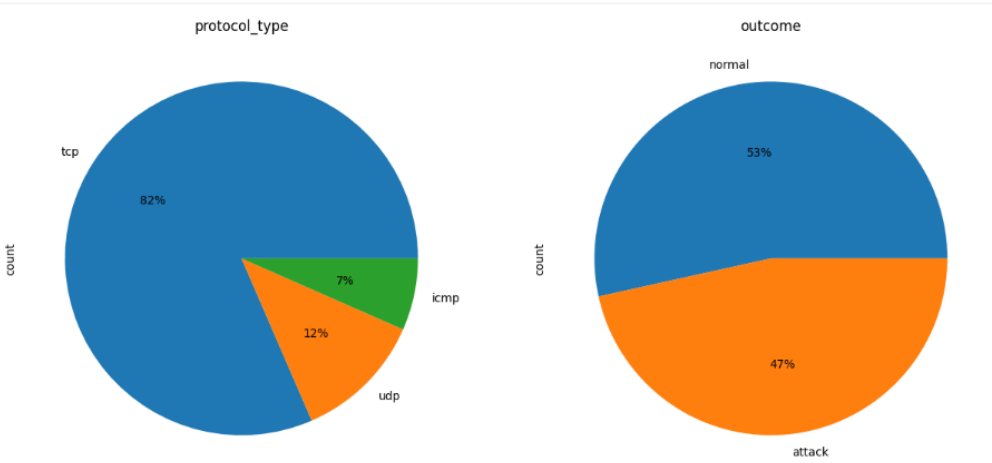

FIGURE 4.4 – Nombre de Protocoles et Types d'Attaques

Dans l'image ci-dessus, nous présentons le nombre de protocoles et les types d'attaques, ainsi que le pourcentage de chaque type.

Modèle Random Forest : étiquettes prédites

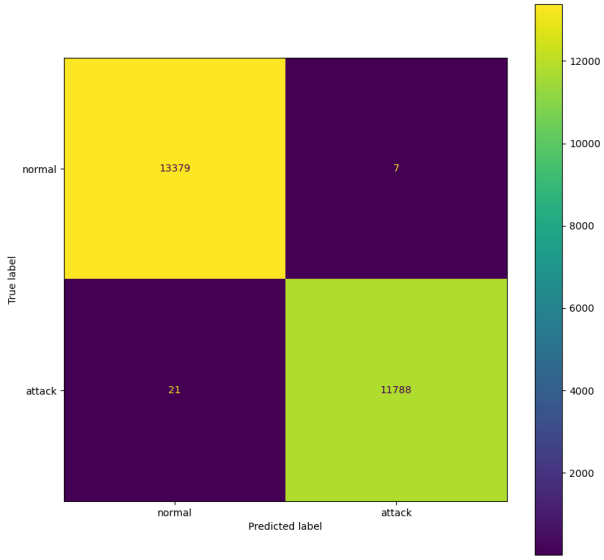

FIGURE 4.5 – étiquettes prédites

Modèle Artificial Neural Network (ANN)

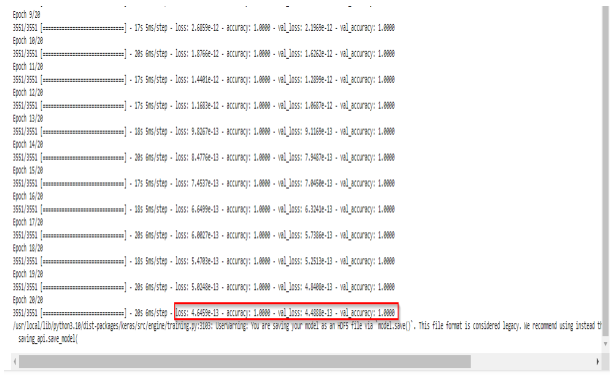

FIGURE 4.6 – étiquettes prédites

4.2.5 Résultats et mise en œuvre réelle

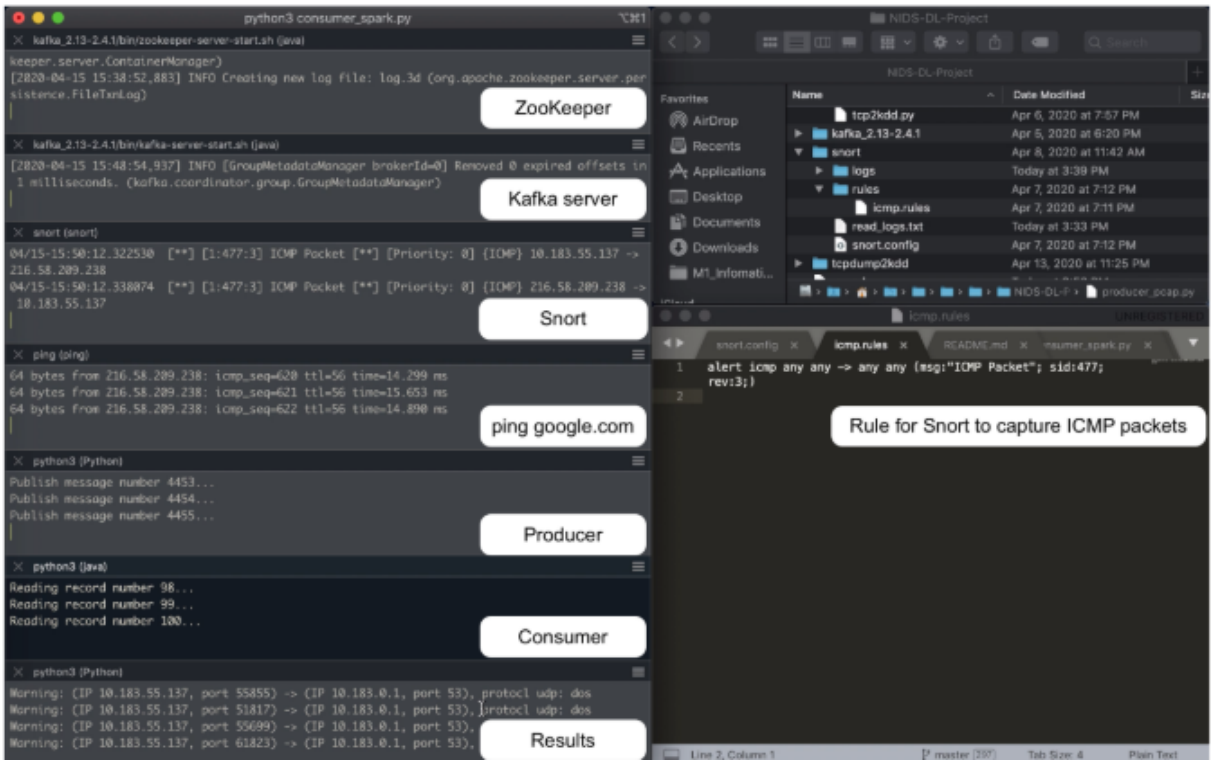

FIGURE 4.7 – Capture d'un système en temps réel

Dans la démo, nous définissons une règle pour capturer tous les paquets ICMP qui passent par notre ordinateur. Nous envoyons un ping à www.google.com pour prouver que Snort fonctionne bien. Le producteur est activé pour envoyer les paquets capturés par Snort au serveur Kafka qui a déjà été lancé. Un consommateur récupère les journaux pcap du

serveur Kafka, puis les traite avec Zeek, Spark et un modèle d'intelligence artificielle. Après avoir obtenu Après avoir obtenu les prédictions du modèle d'IA, un autre producteur renvoie les résultats au serveur Kafka. Un consommateur récupère enfin les résultats et les affiche aux utilisateurs. Si nous portons notre attention sur les résultats, ils doivent être améliorés à l'avenir. Un grand nombre de paquets ICMP dans la démo sont prédits comme des attaques "dos".

4.3 Conclusion générale

L'évolution des réseaux informatiques, notamment Internet, au cours de ces dernières années a entraîné une augmentation considérable du nombre d'utilisateurs. Cette croissance est principalement due à la facilité d'accès et à la diversité des services utiles offerts. Cependant, l'ouverture de ces réseaux, bien qu'elle rende l'accès à l'information plus simple et rapide, les rend également plus vulnérables et exposés aux menaces. Ainsi, la mise en place d'une politique de sécurité permettant de garantir la protection de ces réseaux contre les risques est indispensable.

En réponse à ce défi, une nouvelle génération de solutions d'analyse de sécurité a émergé ces dernières années, capable de surveiller la performance des systèmes et de détecter les problèmes de sécurité, notamment grâce à l'analyse des fichiers logs utilisant big data. Ce projet visait à développer un système de détection d'intrusion en temps réel capable de capturer les intrusions réseau et de prédire leur type d'attaque à l'aide d'un modèle d'intelligence artificielle.

Nous avons utilisé Snort pour capturer les intrusions réseau, fournissant ainsi une première ligne de défense efficace. Kafka a été utilisé comme serveur de streaming pour coordonner les données entre les différentes composantes du système. Zeek a été employé pour analyser le trafic capturé et transformer les données en formats lisibles. Un modèle de machine learning a été développé pour prédire les types d'attaques avec une précision significative. Les différentes composantes ont été intégrées pour fonctionner ensemble en temps réel, permettant une détection et une réponse rapides aux menaces.

Ce projet a démontré la viabilité et l'efficacité de l'utilisation des technologies de Big Data et de l'intelligence artificielle dans la détection des intrusions réseau. .

Bibliographie

[1] AMAD Mourad.Sécurité des Systémes Informatiques et de Communications. Principes et Concepts Fondamentaux de la Sécurité des Systèmes Informatiques.2017, pages 3-12-13 ,Université Bouira.

[2] https://becht.com/becht-blog/entry/cost-schedule-quality-pick-two-isn-t-always-true/

[3] https://encyclopedia.kaspersky.fr/knowledge/history-of-malicious-programs/

[4] https://www.avast.com/fr-fr/c-malware

[5] https://www.fortinet.com/resources/cyberglossary/ddos-attack

[6] MohamedBELAOUED.ApprochesCollectivesetCoopÃl'rativesenSÃl'cu PhDthesis,Skikda,2016

[7] https://learn.microsoft.com/en-us/troubleshoot/windows-client/deployment/dynamic-link-library

[8] https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html