REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : RTIC24/M2/2019

Mémoire

présenté pour obtenir le diplôme de master académique en **Informatique**

Parcours : **RTIC**

Conception et réalisation d'un architecture de la détection d'intrusion sur les Big data

Par :

FIRANE MERIEM

Soutenu le 06 juillet 2019, devant le jury composé de :

Bouchana Belkacem	**MCA**	Président
Saouli Hamza	**MCB**	Rapporteur
Aloui Amel	**MCB**	Examinateur

Remerciements

Au terme de ce travail, je tiens à remercier Dieu le tout puissant de m'a donné la volonté et la patience pour terminer ce travail.

Que nos chers parents et familles, c'est la chance pour exprimer sur nos remerciements les plus profonds en reconnaissance de leurs sacrifices, aides.

J'ai l'honneur et la chance de présenter ma profonde mes remerciements à mon encadreur Dr. SAOULI Hamza, pour ses aides, ces orientations et le temps qu'il m'a donnée pour mon encadrement.

Je remercie prof tous les enseignants qui m'ont aidé pendant mon cursus.

Je remercie les jurys qui ont jugé notre travail.

Je remercie aussi tous les gens qui ont aidé de prêt ou de loin à la réalisation de mon mémoire.

Dédicace

Je dédie ce travail à :

A ma très chère mère Hamida, pour mon père Abdallah. Que dieu leur procure bonne santé et longue vie.

A mon frère Billal et ma sœur décédée Abir, sans oublié mes grands-parents, toute ma famille et mes amis.

Je vous dis merci.

Résumé

Le Big Data se définit par les technologies et méthodes utilisées stocke un grand volume de données issues de multiples ressources. Ces données peuvent être les informations que les utilisateurs laissent sur le Web ou les objets connectés, mais aussi les données internes à l'entreprise ou encore des informations générales, L'objectif du Big Data est de réussir à stocker ces données, en temps réel, pour échanger les données entre les entreprises. D'une part, c'est un atout important pour les entreprises pour la prise de décision d'autre part l'échange de ces données à travers le réseau ouvre la porte aux attaques et aux accès non autorisées des intrus ou d'un autre terme les intrusions.

Les entreprises utilisent des systèmes pur détecter les attaques et les intrusions des intrus à l'objectif de protéger les données .Mais il existe des cas d'attaques et d'intrusions ou les systèmes de détections ne peuvent pas protéger les données qui on des effets sur les données et sur les Big Data.

Ce travail vise à maintenir la détection d'intrusion sur les Big Data. Pour cela nous avons proposé une nouvelle architecture qui contient des divers composants, en prenant en compte les différents critères de détections d'intrusion et les caractéristiques des Big Data. Afin de montrer la faisabilité de l'architecture proposée, nous avons développé un prototype qui pourra résoudre les problèmes mentionnée.

Mots Clés : Détection d'intrusions, Big Data, Intrusion, attaque.

Abstract

Big Data is defined by the technologies and methods used to store alarge volume of data from multiple resources. This data can be the information that users leave on the Web or connected objects, but also the internal data to the company or general information, the purpose of Big Data is to successfully store these data, in real time, to exchange data between companies. On the one hand, it is an important asset for the professional organizations and the governments for the decision-making on the other hand the exchange of these data across the network open the door to attacks and unauthorized access of intruders or another term intrusions.

The companies use systems to detect intruder attacks and intrusions for the purpose of protecting data. But there are cases attacks and intrusions where detection systems cannot protect data that has effects on data and also on Big Data

This work aims to maintain intrusion detection on Big Data. For this we proposed a new architecture that contains various components, taking into account the different intrusion detection criteria and characteristics of Big Data. In order to show the feasibility of the proposed architecture, we have developed a prototype that can solve the mentioned problems

Keywords : Intrusion Detection , Big Data, Intrusion, Attack.

Table des matières

Table des figures

Table des figures

Liste des tableaux

ChapitreI:IDS et Big data

I.1. Introduction

La définition initiale de big data s'orientait d'abord vers la question technologique, avec la célèbre règle des 3V : un grand Volume de données, une importante Variété de ces mêmes données et une Vitesse de traitement s'apparentant parfois à du temps réel. Ces technologies étaient censées répondre à l'explosion des données dans le paysage numérique (le « data deluge »). Puis, ces qualificatifs ont évolué, avec une vision davantage économique portée par le 4ème V de la définition, celui de Valeur, et une notion qualitative véhiculée par le 5e V, celui de Véracité des données (disposer de données fiables pour le traitement). Ces cinq éléments ont servi pendant longtemps de boîte à outils pour comprendre les fondements du Big Data, à savoir l'apparition de technologies innovantes capables de traiter en un temps limité de grands volumes de données.

Dans ce chapitre on va présenter des notions générales et d'autres informations ayant rapport à la sécurité en général, au Big Data en tant que nouvelle technologie et enfin aux exigences sécuritaires imposées par cette dernière.

I.2. Big data

I.2.1. Émergence de Big data

Big data est considéré comme le dernier cri en matière de High Tech, il a résulté de la révolution de la technologie de l'information. Il a certainement une grande influence sur le présent et l'avenir à cause de son importance et du large spectre d'application dans le domaine de la communication et de traitement de données et la recherche scientifique.

La quantité d'information traitée par le micro-processeur se voit doublé chaque 18 mois environ selon la loi de *Moore.*

Il convient également de signaler que la capacité de stockage des données sur le disque dur évolue plus rapidement que la capacité de traitement de données sur le micro-processeur (loi de *Kryder).*

Le volume de données stocké est en croissance exponentielle. La capture, l'intégration et l'exposition de l'information était assez difficile. Les données non structurées prennent une importance considérable, cela est dû à l'évolution des réseaux sociaux et à leurs exigences.

Exemple : chaque appel téléphonique exige une masse de données ayant rapport à le lieu de l'abonné, son numéro, le temps de communication et le tarif unitaire. Un opérateur de télécommunications typique générera quelques téraoctets de données détaillées d'appels chaque mois. Nous constatons également une amélioration considérable dans le traitement du son (flac), de la vidéo (mkv) qui exigent plus de mémoire de stockage et de vitesse de traitement.

Cette masse énorme de données doit être stockée, et le cas échéant analysée et exploitée correctement pour répondre aux besoins des différents opérateurs industriels, économiques ou sociaux. En bref c'est l'ère de l'information.

Le Big data est une technologie révolutionnaire qui peut provoquer des perturbations aux niveaux économique, scientifique et culturelle. Cela est dû à l'importance des changements et des améliorations qu'il impose dans ces différents domaines et qui exigent une nouvelle réadaptation. [1]

I.2.2. Définition de Big data

Le terme «Big data» se réfère à des ensembles de données numériques structurées ou non structurées qui dépassent la capacité des outils traditionnels de traitement et d'analyse des données pour les manipuler.

Big data est défini par les trois V :

- ✓ Haut Volume.
- ✓ Haute Variété.
- ✓ Haute Vitesse.

Big data exigent des formes innovantes et rentables de traitement de l'information pour une meilleure compréhension et une prise de décision correcte. [6]

I.2.3. Modèle 5V

Big data peut être décrit en utilisant le modèle 5V illustré sur la figure 1. Ce modèle est une extension du modèle 3V précédemment cité, et comprend :

- ***Volume*** : il est d'une grandeur exceptionnelle par rapport aux normes connues. Les données produites sont de l'ordre de Zettabytes, et elles sont en croissance d'environ 40% chaque année.
- ***Vitesse*** **:** (the era of streaming data) la collecte et l'analyse des données doivent être rapides et en temps opportun afin de maximiser l'utilisation de la valeur commerciale du Big data.
- ***Variété*** : Les données sont sous plusieurs formats et types, elles comprennent des données semi-structurées et non structurées telles que l'audio, la vidéo, texte ainsi que les données structurées traditionnelles. La plupart des données existantes sont non-structurées ou semi-structurées.
- ***Valeur*** : les données sont devenues une marchandise qu'on peut vendre à des tiers pour une exploitation de nature commerciale, économique ou sociale. La maitrise et l'analyse correcte de ces données permettent une prise de décisions adéquates.
- ***Véracité*** : pour assurer l'exactitude et l'efficacité de cette masse de données, il est indispensable et même vital de procéder à son nettoyage de tout bruit qui peut générer des erreurs qui influent négativement sur les prises de décision. Des techniques sont utilisées pour atteindre cet objectif. [5]

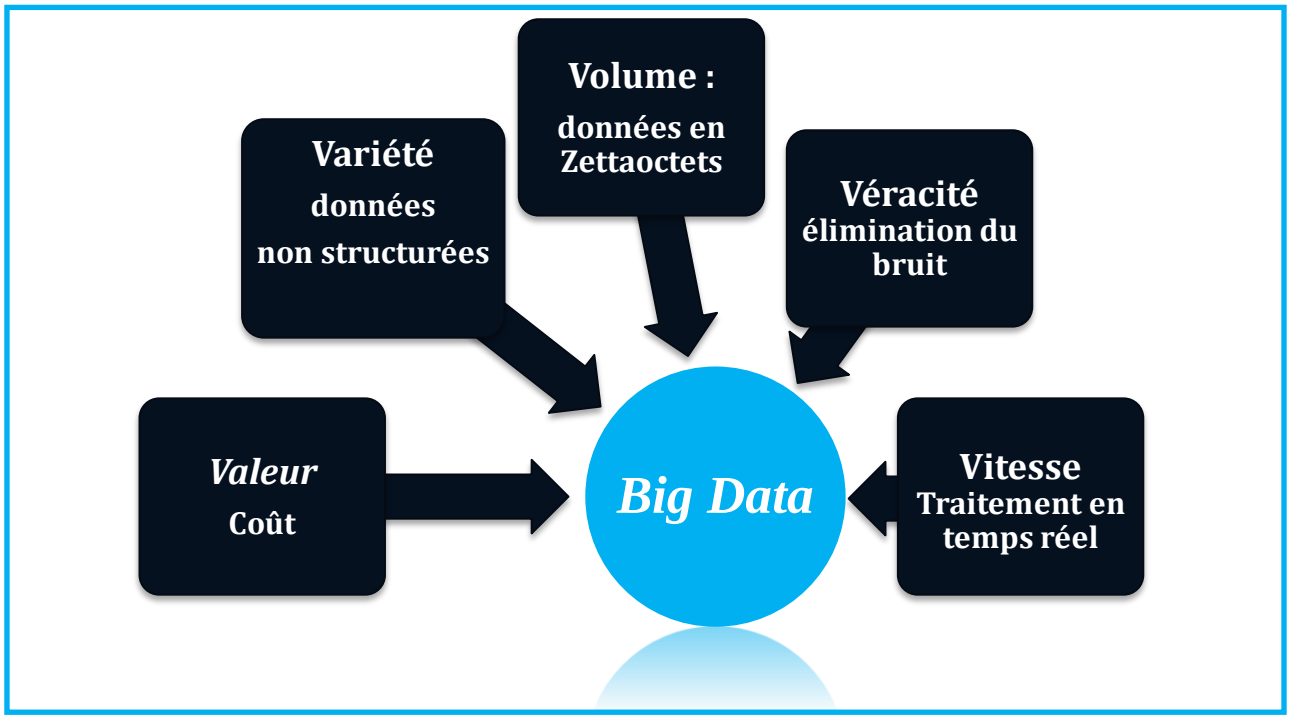

Figure I.1 : le modèle 5V qui définit Big data.

I.2.4. Concepts de Big data

Dans cette section, nous allons examiner de plus près les concepts techniques communs et modèles généralement utilisés dans la plupart des principaux outils et plateformes Big data

I.2.4.1. Cluster de Big data

Les concepts des Clusters Big data peuvent être divisés en deux grandes catégories :

1) Configuration et topologie des clusters.
2) Déploiement des clusters.

La première porte sur le modèle logique de la façon dont un cluster Big data est divisé selon les différents types de nœuds.

La seconde traite du déploiement réel de ces nœuds dans l'infrastructure matérielle physique.

I.2.4.1.1. Cluster Configuration et Topologie

Un cluster Big data est logiquement divisé en deux types de machines / nœuds, à savoir les nœuds de données et les nœuds de gestion.

Les nœuds de données servent deux objectifs fondamentaux d'une part, le stockage des données de manière distribuée et d'autre part le traitement secondaire pour la

transformation et l'accès. Les nœuds de gestion servent de façade pour les applications clientes pour l'exécution des cas d'utilisation.

I.2.4.1.2. Déploiements des Clusters

L'utilisation du Big data tourne en grande partie autour de l'accès / stockage d'un très grand volume de données. L'accès / stockage de données depuis / vers le disque est le processus le plus lent dans l'exécution d'une tâche dans une machine ou un cluster.

I.2.4.2. Concept de stockage de Big data

Le stockage de données constitue une importance capitale et se situe au cœur des outils et des plateformes Big data. Cette masse immense de données doit être mémorisée d'une manière appropriée pour permettre son analyse et son exploitation efficacement.
Le concept de stockage se résume comme suit :

I.2.4.2.1. Modèles de données

Il existe plusieurs types de modelés de données, tels que « le modèle relationnel, NoSQL

I.2.4.2.2. Partitionnement de données

Le partitionnement des données doit se faire sur plusieurs nœuds de données afin de pouvoir traiter ces données en simultanée par toutes ces machines.

I.2.4.2.3. La réplication de données

La réplication des données permet la protection de ces données en cas de défaillance d'un serveur et une plus large utilisation.

I.2.4.2.4. La compression de données

La compression nous permette de réduire l'espace de stockage et limite la bande passante nécessaire sur le réseau de transport. Toutefois elle présente un inconvénient en matière de temps de traitement de ces données suite à la compression et à la décompression qui s'imposent lors de l'écriture et de la lecture de ces données vers et à partir du disque de stockage.

I.2.4.2.5. Indexation de données

Les données sont reparties sur plusieurs blocs à travers différents nœuds de données du cluster Big data. L'indexation sert à identifier les enregistrements réalisés dans ces blocs et déterminera leurs positions.

I.2.4.3. Concept de récupération informatique du Big data

Il nous permet le traitement d'un grand nombre de données en temps réel ainsi que l'accès à ces données d'une manière aléatoire pour la lecture et l'écriture.

I.2.4.3.1. Moteur de traitement distribué

Les moteurs de traitement distribué répondent à la nécessité pour la gestion, le traitement, le filtrage, l'interrogation, la modélisation, l'exportation, ainsi que l'archivage de grand volume de données dans l'infrastructure Big data.

I.2.4.3.2. Sécurité des données

La sécurité présente une dimension vitale pour tout ensemble de données informatiques. Différentes méthodes sont appliquées pour réaliser cet objectif tel que le chiffrement, l'authentification et le cryptage.

I.2.4.4. La gestion des ressources

La gestion des ressources présente une grande importance pour technologie Big data. Ces ressources informatiques (CPU, Mémoires et Disques) de tous les nœuds et du réseau les reliant doivent être distribuées de manière appropriée. [3]

I.2.5. Résistance de Big data (Souplesse et maniabilité)

Le tremblement de terre qui frappa le Japon et qui fut suivi d'un Tsunami d'une ampleur sans précèdent puis de l'explosion nucléaire qui a eu lieu à Fukushima n ° 1 et des évènements qui suivirent (exode de la population, destruction d'une masse importante d'archives gouvernementale…etc.) a démontré l'utilité de la technologie Big data et le rôle que cette dernière pourrait jouer en sécurisant les données et en offrant des renseignements sur l'état des routes, l'intensité du trafic etc. Qui pourraient sauver des vies humaines et limiter relativement les dégâts. De nouvelles techniques permettant la préservation de données même en cas de catastrophe naturelles de grandes ampleurs sont apparues de nos jours tels que la technologie Cloud. [2]

I.2.6. Domaine d'application de Big data

Dans cette section nous présentons quelques principaux domaines d'applications du Big data :

I.2.6.1. Agriculture

D'ici 2050 on prévoie le dépassement de 9 milliards d'êtres humains sur le globe, ce qui rend l'agriculture un domaine prioritaire pour gérer les besoins alimentaires de la population mondiale. Le Big data représente un atout considérable pour l'organisation de l'agriculture à travers le monde, notamment pour la gestion de l'irrigation (l'eau potable étant une ressources de plus en plus rare), où nous avons besoin de gérer de gigantesque masses de données qui concernent les prédictions météorologique et la sécheresse du sol.

I.2.6.2. Assurance

L'assurance représente l'un des domaines directe d'application de Big data, vue qu'on est amené à effectuer des statistiques et des analyses sur les risques liés au comportement de millions d'individus.

La possibilité de récolter de gigantesques masses d'informations qui concernent la vie des individus permet de concevoir un modèle de vie pour chacun d'eux : hygiène de vie, conduite de voiture, amende, consommation électrique, relation professionnelle ….Etc. Ces modèles de vie permettent aux agences d'assurances d'améliorer leurs offres, d'optimiser leurs méthodes, et même de mener des enquêtes plus précises.

I.2.6.3. Marketing

Avec le marketing on est amené à gérer de gigantesques masses d'informations qui proviennent de divers sites et réseaux sociaux que des clients potentiels peuvent visiter. Mais ce qui révolutionne vraiment le marketing de nos jours c'est l'omniprésence de capteurs publics sur les centres commerciaux, métros, aéroports et universités, et qui sont destinés à capter le comportement des consommateurs, ce qu'il achète, ce à quoi ils s'intéressent, et les produits qu'ils ne trouvent pas aux marchés, ce qui permet d'analyser et d'étudier leurs besoins en temps réel afin de produire des solutions et des méthodes de marketing plus efficaces.

L'utilisation des capteurs permet de capter des données de diverses formes : images de visages pour analyse émotionnelle, vidéos pour description comportementale, données textuelles pour décrire la nature des produits achetés, données numériques et

statistiques. Cette diversité qui nécessite un traitement en temps réel ne peut être résolue qu'avec des méthodes de stockage et de traitement d'informations issues de Big data.

I.2.6.4. Au-delà du marketing

Le Big data a permis de refaçonner le monde du marketing en offrant les techniques et les stratégies nécessaires pour bénéficier des données que publient les consommateurs et fournisseurs en utilisant les : Réseau sociaux, applications mobiles, magasins, TV, catalogues, blog, presse, radios, etc. Sans la techniques Big data il sera tout simplement impossible de traiter les gigantesques masses d'informations que produisent ces moyens de publication. L'émergence de Big data a permis l'apparition de nouvelles notions telle que le pré-marketing et re-marketing qui représentant une nouvelle vision d'atteindre et de convaincre les consommateurs finaux.

I.2.6.5. Achat programmatique

L'achat programmatique est devenu la technique la plus utilisée pour l'achat/vente sur Internet, vue que cette technique permet d'utiliser un logiciel ou une plateforme intermédiaire entre les clients et les fournisseurs pour effectuer des opérations de : publicité, choix du meilleur prix, et payement électronique. L'achat programmatique permet d'alléger les tâches qui correspondent au processus d'achat/vente en s'occupant automatiquement du processus de négociation entre client et fournisseur ainsi que de toute opération manuelle traditionnellement demandée par le fournisseur. Cependant, l'achat programmatique impose la manipulation en temps réel de gigantesques masses d'informations qui sont échangées entre clients et fournisseurs en compétition pour trouver et acheter les meilleurs espaces publicitaires sur le Net. Les techniques de gestion des données issues du domaine Big data représentent un atout considérables et une alternative prometteuse pour la gestion des plateformes d'achat/vente intermédiaire.

I.2.6.6. Compétitivité et Innovation de produit

La possibilité de traiter de gigantesques masses d'informations en temps réel permet aux entreprises d'analyser les besoins de leurs clients afin de pouvoir optimiser et améliorer leurs propres produits et augmenter leur compétitivité sur le marché. C'est ainsi, que les services qu'offrent les fournisseurs de téléphonie mobile permettent aux touristiques de localiser, en temps réel, leurs clients habituels afin de leurs envoyer des offres d'excursions, les lieux et la nature des évènements touristiques, et les réductions hôtelières et les billets d'avion par exemple. Les techniques d'analyse en temps réel de gigantesques

masses d'informations, issues de Big data, permettent également aux entreprises de contrôler et d'être à jours par rapport aux produits des entreprises concurrentes ce qui garantit l'innovation et la compétitivité des produits.

I.2.6.7. Gestion de catastrophes naturelles

L'une des applications les plus intéressantes de Big data, est la possibilité d'analyser des données météorologiques en temps réel, ce traitement permet de suivre et de visualiser le déplacement des ouragans et de prédire les endroits géographiques où ces derniers vont frapper. C'est ainsi que les gouvernements locaux et les organisations internationales d'assistance humanitaire peuvent préparer les ressources nécessaires (couverture, alimentations, médicaments) ainsi que les moyens de transport et d'intervention rapide pour aider la population en détresse.

I.2.6.8. Contrôle d'épidémies

Le Big data peut contribuer à contrôler la propagation d'épidémies à travers le monde en surveillant par exemple la migration des insectes porteurs de maladies à travers le globe. Le big data est également utilisé pour traquer la population des rats dans les grandes villes telles que New-York ou Chicago où la police locale utilise un système Big data pour la surveillance visuelle et l'analyse des itinéraires des rats, afin de contrôler leurs croissances.

I.2.6.9. Prévention d'attaques cybernétiques

De nos jours, les techniques d'analyse de données qu'offre le Big data sont devenues incontournables pour pouvoir détecter les intrusions, les failles sécuritaires ainsi que les attaques cybernétiques, vue que le volume de données transportées sur le Net est devenu gigantesque, diversifier, et nécessitant un traitement en temps réel. Avec les techniques de traitement de données Big data on arrive à tracer le schéma relationnel entre les données et effectuer des calculs statistiques qui permettent de surveiller et d'intervenir, en temps réel, sur les menaces et les attaques cybernétiques à l'échelle mondiale. [9]

Assurance

Agriculture

Gestion de catastrophes naturelles

Marketing

Achat programming

Contrôle d'épidémies

Figure I.2 : Domaine d'application de big data.

I.2.7. Défis et enjeux

La grande progression de données constitue un énorme défi en matière d'acquisition, de stockage, de gestion et d'analyse. Les systèmes de gestion et d'analyse de données traditionnels relationnelles (SGBDR) utilisant un équipement onéreux et ne peuvent traiter des masses énormes de données hétérogènes. Cela à amener les chercheurs à proposer de nouvelles technologies telles que le Cloud Computing et les bases de données NoSQL.

Les principaux défis sont énumérés comme suit :

- ✓ *La réduction de la redondance et la compression des données :* réduire au maximum la redondance des données et procéder à leur compression pour limiter le coût de l'ensemble du système sans pour autant influer négativement sur la valeur de ces données.

- ✓ *La gestion du cycle de vie des données :* vu la masse énorme de données affluente, il convient de ne garder que les données utiles et mises à jour et supprimer tout ce qui est superflue afin d'éviter la saturation des systèmes.
- ✓ *Mécanisme analytique :* Traiter un gros volume de données hétérogène dans un temps limité.
- ✓ *La confidentialité des données : La* capacité limitée des fournisseurs ou propriétaires de ces volumes de données, ils n'ont pas les moyens de procéder à un traitement et à une analyse efficace. Ils ont recours à des professionnels ou à d'autres outils pour réaliser ces taches. Cela présente un risque de sécurité potentiel.
- ✓ *Gestion de l'énergie :* La consommation d'énergie au niveau des systèmes de stockage et d'analyse sont en nette progression. Il convient de contrôler cette consommation et l'optimiser dans la mesure du possible.
- ✓ É*volutivité :* le système d'analyse Big data doit prendre en charge les ensembles de données actuelles et futures. Les algorithmes doivent être en mesure de traiter des ensembles de données en expansion permanente. [4]

I.3. Système de détection d'intrusion (SDI)

I.3.1. Définition des SDIs

Le système de détection d'intrusion est une nouvelle technologie, ou le mot de détection d'intrusion est une technique de découverte d'accès non autorisé à un ordinateur, un système, ou un réseau informatique.

Cette technique découvrir des actions non autorisées dans les systèmes.

Elle est liée à la prévention d'intrusion qui donne la possibilité de filtrer rapidement le trafic réseau qui contient des intrusions ou des possibilités d'intrusions. [Joseph Migga Kizza]

I.3.2. Processus d'intrusion

Le processus d'intrusion fonctionne par l'application des étapes suivantes ;

1-identification et la connaissance du réseau ou système qui est le but de l'attaquant.

2-la reconnaissance des informations sur le réseau ou le système dans le cas de leurs présences.

3-La recherche des informations possibles sur le réseau ou le système et définir les point faibles pour les exploitent.

4-Le succès du l'accès dans le système.

5-l'utilisation du système facilement. [Joseph Migga Kizza]

Il existe une explication brève sur ces étapes qui sont présentés dans la figure I.3.

I.3.2.1.Reconnaissance

Est le processus qui fonctionne par la collection d'informations sur le système qui pose comme un but de l'intrusion d'une part.

D'une autre part, ce processus extraire les informations de leur fonctionnement et les points faibles qui se trouvent dans le système. [Joseph Migga Kizza]

I.3.2.2.intrusion physique :

Elle représente la phase qui utilise les informations collectées qui donne une chance pour l'attaquant qui entre dans le système d'une manière illégale. [Joseph Migga Kizza]

I.3.2.3.Déni du service:

Est une attaque qui fonctionne sur la surcharge du système ou l'intrus envoyer des requêtes depuis plusieurs endroits.

Le but qui donne ce d'attaque est obtenu l'accès pour enter dans une machine ou système mais ne collecter pas les informations. [Joseph Migga Kizza]

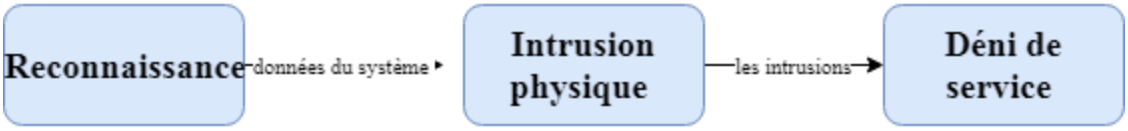

Figure I.3 : les étapes de processus d'intrusions

I.3.3. Danger d'intrusion :

L'intrusion dans le système laisse un ensemble des dangers qui sont :

I.3.3.1.perte de données personnelles :

Les données personnelles sont stockés dans un ordinateur personnel, l'attaquant entre dans le système et copier les données personnelles d'une manière facile. L'accès de l'intrus permet l'existence du danger de perte des informations personnelles. [Joseph Migga Kizza]

I.3.3.2.Confidentialité compromise :

les gens stockent leurs informations personnelles dans la ligne. D'une autre part, les autres informations stockées en ligne à travers les entreprises. Lorsqu'un système d'entreprise contient ces informations personnelles et on trouve des faiblesses dans le système, alors ces informations deviennent en danger. À conséquence la vie privée des gens est en danger. [Joseph Migga Kizza]

I.3.3.3.Responsabilité juridique : un réseau d'organisations possède des informations personnelles des clients. Si un pirate fait une attaque sue le réseau de l'organisation, le directeur de l'organisation est le responsable de la perte qui vient depuis l'intrus. En cas de l'attaque du pirate sur plusieurs niveaux sur le réseau, aussi le directeur est le responsable. [Joseph Migga Kizza]

I.3.4. Les Systèmes de détection d'intrusion

Un système de détection d'intrusion est un système qui permet de connaitre et découvrir les accès non autorisés dans le système de l'entreprise et les réseaux informatiques. La détection d'intrusion n'est pas une nouvelle technique, utilisée dans le passé pour les guerres. Avec l'évolution de la technologie, la détection d'intrusion devient un système pour protéger les systèmes des entreprises et les réseaux informatiques.

Il existe trois modèles de mécanismes de détection d'intrusion :

1-détection basée sur les anomalies.

2- détection basée sur la signature.

3-détection hybride. [Joseph Migga Kizza]

I.3.4.1-détection basée sur les anomalies :

Le système basé sur les anomalies est un système qui travaille avec le comportement du système, ont créé des profils sur l'état du système normal.

Ces profils sont les comportements normaux du système. Dans le cas de l'intrusion on fait une comparaison entre le comportement anormal et le comportement normal du système.et en faire une mise à jour sur les comportements normaux. [Joseph Migga Kizza]

I.3.4.2.détection basée sur la signature :

Ce type de détection porte le nom de la détection basée sur la mauvaise utilisation, il fonctionne avec une base de signatures des activités normales. Ce modèle doit créer une liste sur toutes les activités anormales ou les actions non autorisées, à cause de nombre limité des éléments de la liste. Pour la facilité de la gestion de la liste, on a trois classes d'activité :

1-l'accès non autorisé.

2-la modification non autorisée.

3-dénis de service. [Joseph Migga Kizza]

I.3.5. Type des SDIs

La classification des systèmes de détection d'intrusions se fait à travers un critère important qui est le champ de la surveillance. Il existe une surveillance sur une petite zone qui prend le nom de détection basée sur l'hôte et sur une zone vaste qui appelle détection d'intrusion basée sur le réseau. [Joseph Migga Kizza]

I.3.5.1. Systèmes de détection d'intrusion en réseau (NIDS) :

- Le NIDS est responsable de la surveillance du réseau par la détection des données non autorisées et anormales qui circulent dans le réseau surveillé par le NIDS, en conséquence la détection des intrusions dans les trafics réseaux.
- Il existe une différence entre le pare-feu et le NIDS, le pare-feu permet d'autoriser ou refuser l'accès à un réseau à base d'un ensemble des règles. Mais le NIDS capture tous les paquets arrivant au réseau et puis

  Cherche dans la liste des signatures, le résultat de la recherche peut être une intrusion ou non.
- Le NIDS s'exécute soit en tant qu'indépendante machine autonome qui surveille tout le trafic ou en tant qu'une machine cible qui fait la surveillance du son propre trafic. [Joseph Migga Kizza]

I.3.5.1.1.Architecture d'une détection d'intrusion basée sur le réseau :

Le système de détection d'intrusion compose des parties qui travaillent comme un ensemble pour détecter l'intrusion ces parties trouvent dans la figure I.4 :

I.3.5.1.1.1. Équilibreur de prise / charge réseau:

L'équilibreur de prise ou aussi porte le nom de la charge réseau est un composant important dans le système de détection d'intrusion, car il prend les données arrivant depuis le réseau extérieur au réseau local puis la diffuse a tous les capteurs réseaux. Aussi l'équilibreur peut-être un logiciel s'exécute dans un élément réseau, ainsi il est responsable de tout le trafic réseau et l'évitement de la perte des paquets.

I.3.5.1.1.2.Capteur de réseau / surveillance :

Le capteur réseau est un programme informatique qui est responsable de la réception du trafic arrivé à partir de l'équilibreur dans le réseau avec l'équilibreur, sinon dans le cas contraire le capteur reçoit directement le trafic réseau et sépare les différents trafics. Ce programme exécute dans des machines dans des zones critiques.

I.3.5.1.1.3. Analyseur :

L'analyseur donne le niveau de la menace à partir de la nature et le danger de virus trouvé dans le trafic. Depuis la collection des données, plusieurs couches de la surveillance travaillent comme un groupe pour la connaissance du danger de la menace, la portée, la fréquence de la menace.

I.3.5.1.1.4. Notificateur d'alerte:

Le Notificateur d'alerte contacte l'agent responsable de la gestion des incidents dans le cas de la détection de la menace, depuis l'application de la politique sécurité de l'organisation.

I.3.5.1.1.5. Console de commande / gestionnaire:

La console de commande possède le rôle d'agir comme une commande centrale pour le contrôle du système, aussi fait la gestion de la menace. On peut avoir un contrôle à distance pour la gestion de n'importe quel ordinateur à distance. Dans la console, il existe un gestionnaire d'évaluation et un gestionnaire de la cible, un gestionnaire d'alertes.

I.3.5.1.6. Sous-système de réponse:

Le sous-système de réponse avoir des capacités pour construire des mesures sur système. Les réponses sont initialisées d'une façon automatique sinon on les génère.

I.3.5.1.1.7. Base de données:

La base de données et le moyen de stockage de toutes les informations observées par le système de détection d'intrusion. Ces informations incluent des statistiques comportementales et d'utilisation des signatures, les statistiques sont importantes pour la modélisation de l'historique. [Joseph Migga Kizza]

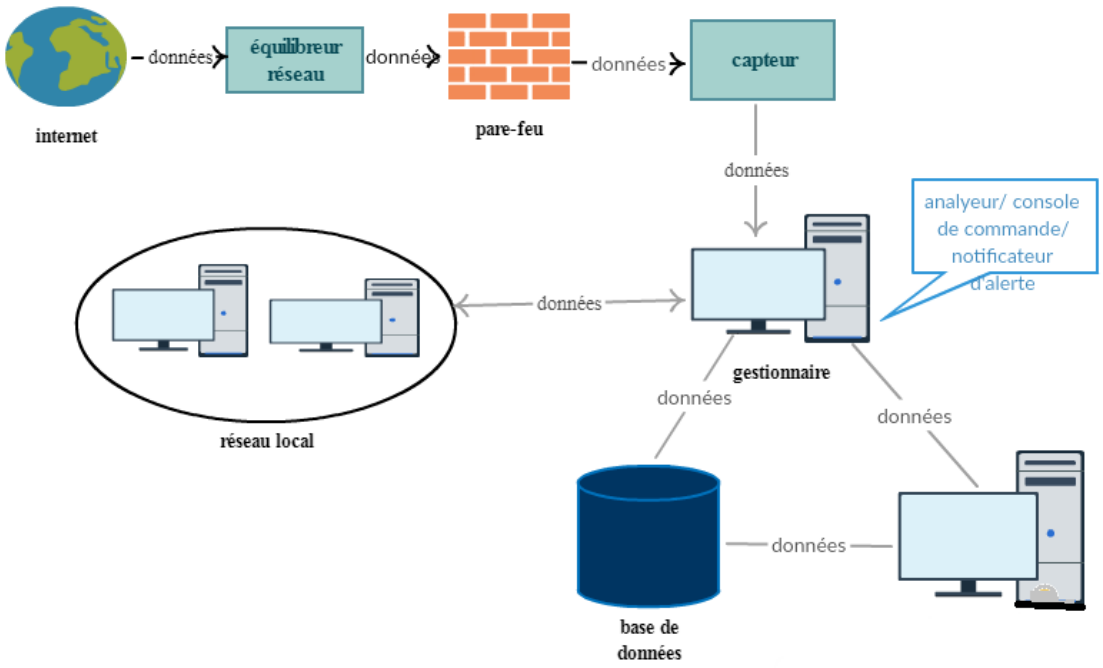

Figure I.4 : Architecture d'une détection d'intrusion basée sur le réseau

I.3.5.1.1.8. Emplacement des capteurs IDS

La position des capteurs IDS dépend sur plusieurs critères comme la topologie du réseau, le type de sécurité …etc., on trouve 4 domaines où on placer les capteurs est :

1-À l'intérieur de la zone démilitarisée : le DMZ protège le réseau

Contre les attaques entrantes, on place les capteurs au-dehors du premier pare-feu dans le réseau à l'intérieur de la zone démilitarisée. Aussi un autre emplacement c'est dans des zones définies dans la zone démilitarisée. Le

dernier emplacement c'est à l'intérieur de chaque pare-feu. Sans la zone démilitarisée, l'emplacement sera les points d'entrée / sortie.

2-entre le pare-feu et l'internet : cet emplacement permet au système de détection d'intrusions de connaitre le trafic internet déroulant dans le réseau mais il nécessite des capteurs supportent le volume élevé du trafic.

3-derrière le pare-feu avant le réseau : c'est la position idéale car le pare-feu arrête les trafics anormaux et les capteurs fait la gestion de les mauvais trafics.

4-à l'intérieur du réseau : on place les capteurs dans des points stratégiques pour la possibilité d'avoir le trafic réseau. [Joseph Migga Kizza]

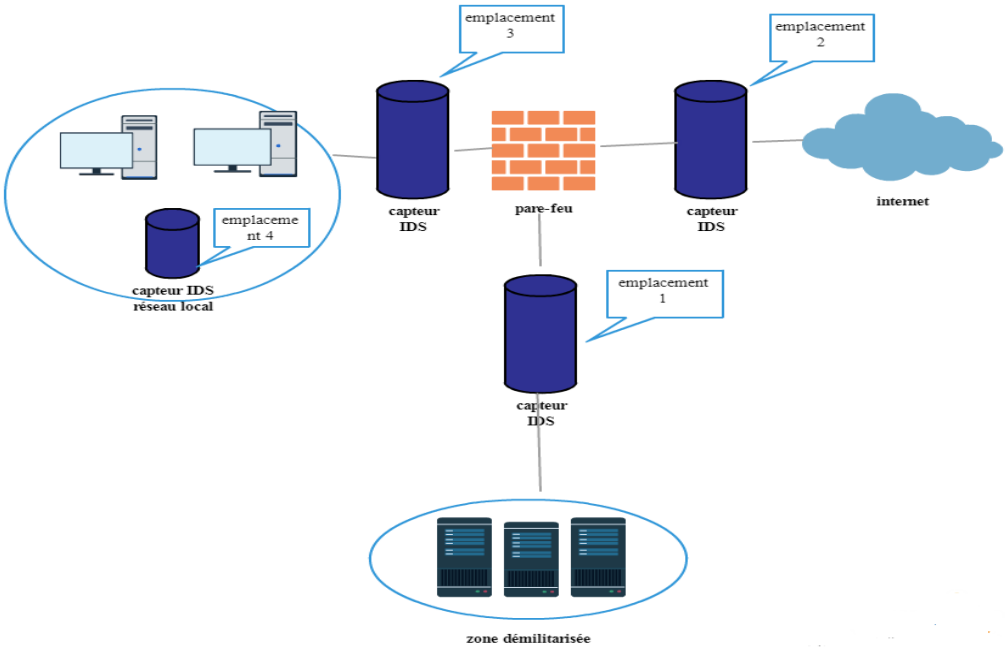

Figure I.5 : Les différents endroits où placer les capteurs IDS

I.3.5.1.2. Avantages des systèmes de détection d'intrusion sur réseau :

1- la capacité de détecter les attaques d'un système qui utilise une machine qui est perdue parce que le NIDS contrôle le trafic réseau.

2- Difficulté à éliminer les preuves : le NIDS situé dans des machines fortement protégées, donc il est difficile pour la suppression des preuves. Dans une autre part le NIDS contrôle et capture le trafic réseau même dans le cas de l'attaque, pour cette raison l'intrus n'est pas la chance d'éliminer les preuves.

3- Détection et réponse en temps réel : l'emplacement des capteurs réseaux dans les endroits stratégiques permet de donner la réponse en temps réel, car les capteurs détectent les attaques puis il envoie un message d'alerte à l'administrateur réseau.

4- Capacité à détecter les attaques infructueuses et les intentions malveillantes : l'utilisation de NIDS avec le DMZ permet de détecter ce type d'attaque par le pare-feu extérieur puis e les rejeter par le pare-feu intérieur. Dans ce cas, le NIDS mémorise la fréquence de l'attaque. [Joseph Migga Kizza]

I.3.5.1.3. Inconvénients de NIDS:

1- Angles morts : le NIDS contient des angles morts où les capteurs ne sont pas capables de contrôler le trafic déroulant dans ces endroits, ni détecter les attaques qui se produisent dans ces place

2- Données cryptées : le NIDS ne contient pas un algorithme de déchiffrer les données cryptées, car il connaîtra la partie non cryptée du paquet tel que l'entête qui ne donne pas des informations suffisantes sur les paquets el les attaques. [Joseph Migga Kizza]

I.3.5.2. Systèmes de détection d'intrusion basée sur l'hôte (HIDS) :

- Le HIDS et la solution de problème des intrus dans les organisations, car il dépend sur un système de vérification de réseau de l'organisation.
- Le HIDS est un système qui fonctionne sur un seul machine.
- C'est une technique de détections des activités malveillante qui l'implémente sur un seul hôte.
- Il travaille avec un logiciel de surveillance le système d'exploitation avec des journaux spécifiques tel que les journaux systèmes…etc
- On peut installer le HIDS dans une machine à distance ou sur un ensemble des machines sur le réseau. [Joseph Migga Kizza]

I.3.5.2.1. Avantages du HIDS :

Possibilité de vérifier rapidement le succès ou l'échec d'une attaque :

Le HIDS capture des événements en temps réel, car on obtient des informations plus précises et correctes.

- surveillance de bas niveau : le HIDS surveille sur une machine, donc il capture les évènements bas niveau. les rapports envoyés dans le temps réel et envoie une alerte à l'administrateur réseau.
- Détection et réponse en temps quasi réel : la détection des intrusions par le HIDS se fait rapidement avec un temps proche du temps réel.
- Capacité à gérer des environnements cryptés et commutés : à contraire du NIDS, le HIDS peut traiter les informations cryptées. le HIDS avait une visibilité sur les réseaux commutés.
- Rentabilité : l'absence des matériels secondaires pour implémenter le HIDS, donc à minimisé le cout de l'entreprise. Mais avec les grands réseaux découpant en segments, le nombre de NIDS devient important. Alors le cout s'augmente. [Joseph Migga Kizza]

I.3.5.2.2. Inconvénients du HIDS :

- Point de vue limité : Depuis qu'ils sont déployés chez un hôte, ils ont un très limité du vue sur le réseau.
- Comme ils sont proches des utilisateurs : ils sont plus susceptibles d'être altérés illégalement. [Joseph Migga Kizza]

I.3.5.3. Le système hybride de détection d'intrusion :

Le système hybride est la combinaison entre les NIDS et le HIDS. Le système hybride prend les avantages de chaque système avec le complètement des faiblesses de chaque système pour l'augmentation de la sécurité du réseau. Le système hybride offre une grande flexibilité dans les options de déploiement. [Joseph Migga Kizza]

I.3.6. la nature des outils de SDI

- Les nouveaux outils d'IDS concentrent sur les attaques des personnes qui ont des informations sur le réseau.
- les nouveaux outils d'IDS en cours de développement pour arrêter les intrusions.
- Des nouveaux modèles en train de développement pour la connaissance de comportement de l'être humain.
- D'après ces changements, les systèmes d'identité changent aussi. [Joseph Migga Kizza]

I.3.7. Types des SDI plus avancé

Il existe d'autres types de SDI, ces types sont les plus utilisées qui sont :

I.3.7.1. Intégrité du système Vérificateurs (SIV) :

Les vérificateurs d'intégrité du système qui permettent la surveillance et le contrôle des fichiers critiques dans le système.

I.3.7.2. Moniteurs de fichier journal (LFM) : qui permet de :

1- la création d'un enregistrement des fichiers journaux qui les généré par les services réseaux.

2-la surveillance d'un enregistrement.

3-la recherche des tendances systèmes et les modèles dans les fichiers journaux qui donnent une suggestion d'un attaquant dans le cas d'attaque.

I.3.7.3. Pots de miel :

- Le pot de miel est un système développé pour connaitre les attaquants et leurs outils d'attaque.
- Le pot de miel est un outil et n'est pas des systèmes de détection d'intrusion.
- Est un bon système pour la sécurité du système pour l'étude des intrus.
- La position du pot de miel est le DMZ pour atteindre l'objectif pour les réseaux avec DMZ et derrière le pare-feu pour les réseaux sans DMZ, la position du pot de miel est présentée dans la figure I.5.

La position du pare-feu est idéale pour ces raisons :

1-La plupart des pare-feu enregistrent tout le trafic le traversant; par conséquent, cela devient un bon moyen de suivre toutes les activités des intrus.

2-La plupart des pare-feu ont une capacité d'alerte, ce qui signifie qu'avec quelques ajouts à la base de règles du pare-feu.

3-Le pare-feu peut contrôler le trafic entrant et sortant. Cela signifie que les intrus peuvent trouver, sonder et exploiter notre pot de miel, mais ils ne peuvent pas compromettre d'autres systèmes. [Joseph Migga Kizza]

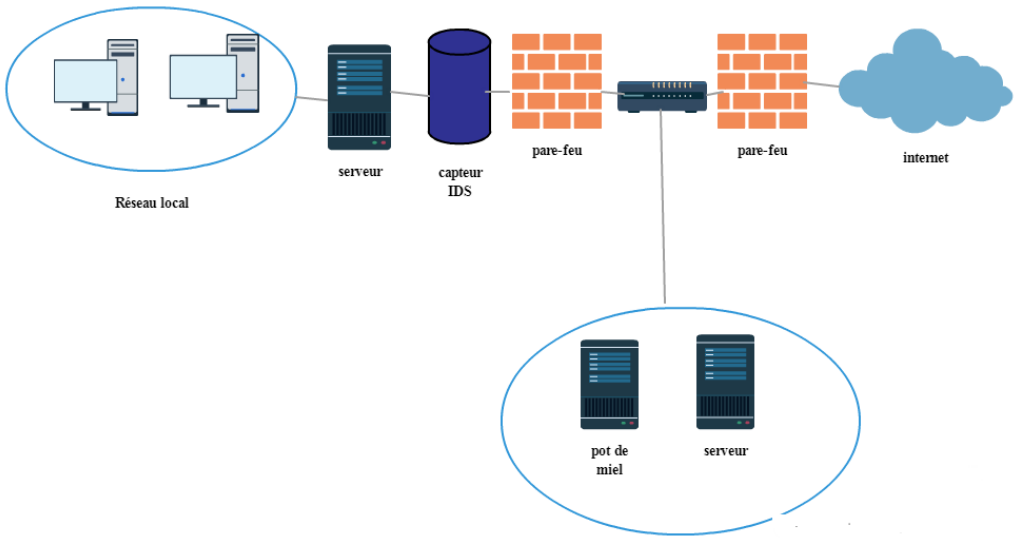

Figure I.6 : Le positionnement d'un pot de miel

I.3.8. Réponses à la détection d'intrusion :

- La bonne réponse se produit à partir d'un système de détection d'intrusions efficace qui donne une alerte précise.
- le type de réponse dépend sur le type d'attaque.
- la réponse efficace contient des contre-mesures prédéfinies.
- La collecte des informations utilisées dans le futur. [Joseph Migga Kizza]

I.3.8.1. Équipe d'intervention en cas d'incident :

Une équipe de réponse aux incidents (IRT) est un ensemble des personnes qui le contacter dans le cas d'incident, qui assure les responsabilités suivantes :

1-découvrait les nouvelles menaces et l'incident.

2-un point de communication principale pour l'avertissement des incidents.

3-évaluation des degrés et les impacts des incidents

4-découvrir comment éviter les attaques.

5- se remettre de l'incident

6- Lors du traitement d'un incident : qui contient plusieurs étapes :

La priorité pour les actions trouvées dans la politique de sécurité de l'organisation qui prendre l'ordre suivant :

1-la sécurité des personnes et la vie personnelle.

2- les données et les fichiers couteux.

3-préventions des attaques aux systèmes.

4-minimiser la destruction des systèmes.

7- évaluation des conséquences produites par l'incident.

8- évaluation et le signal des personnes concernées.

9-récupérations après l'incident : se forme d'un rapport post mortem qui contient les étapes qui suivent pour l'intervention dans le futur. [Joseph Migga Kizza]

I.3.8.2. Journaux IDS en tant que preuves :

- Les journaux IDS sont stockés pour la protection de l'organisation, en cas des procédures légales.
- Quelques personnes ont vu l'IDS comme un système d'écoute. [Joseph Migga Kizza]

I.3.9. Défis des SDIs

La technologie IDS avoir un long chemin et un futur en développement, le IDS fait face un ensemble des défis, parmi les défis le plus danger est le développement d'IDS dans des environnements commutés.

I.3.9.1. Déploiement du SDI dans des environnements commutés :

- Le déploiement des IDS pose un grand problème dans les réseaux.
- Les capteurs IDS voit le trafic réseau, mais avec les réseaux commutés, la vue du trafic en utilisant des composantes spéciales parce que le trafic est protégé.
- La solution du problème est la liaison de capteur réseau par un port miroir dans le commutateur. Donc il existe le problème la surcharge du port
- Les solutions sont :

1-Ecoute : cette solution permet de créer une ligne pour l'écoute des connexions Ethernet, et puis envoyer des copies avec le capteur correspondant à un autre commutateur pour minimiser la surcharge.

2- en utilisant les listes de contrôle d'accès Cisco standard : pour le stockage des images pour la vérification.

Des autres problèmes diminuaient l'efficacité d'IDS :

1-les fausses alarmes : les IDSs peuvent produire des alertes qui ne prendre pas en compte depuis le système.

2- la non-défense de l'IDS contre les attaques des grands réseaux : les capteurs réseaux dans les grands réseaux ne permettent pas de voir tous le trafic réseau. Avec cette raison, la technologie IDS ne permet pas de détecter les attaques dans les zones hors surveillance. [Joseph Migga Kizza]

I.3.10. Comment implémenter un SDI ?!

Un IDS est efficace à l'aide des autres systèmes qui sont :

1-Systèmes d'exploitation : un système d'exploitation qui caractérise par des fonctionnalités qui permet le contrôle et la protection des ressources critiques.

2-Le pare-feu ; il faut avoir les capacités pour la détection des attaques et les intrusions à travers des fonctionnalités.

3- Plate-forme de gestion de réseau : contient des services réseaux qui utilisent des outils responsables de la configuration des alertes. [Joseph Migga Kizza]

I.3.11. Les Systèmes de Prévention d'Intrusion

L'IDS est un système de détection d'intrusions sans faire aucune action, l'action se fait par un système de protection contre les intrusions. Qui s'appelle un système de prévention d'intrusion. Il existe deux types :

I.3.11.1. Systèmes de prévention des intrusions sur le réseau (NIPS) :

- Le NIDS détecte les attaques seulement sur le réseau sans faisant une action.
- Les fonctions de l'IDS complètent les fonctions du pare-feu par la détection des attaques puis on utilise IPS pour l'action de protection.
- IPS modifie les règles d'accès du contrôle du pare-feu par le blocage de l'attaquant.
- IPS est une nouvelle technologie dans la sécurité réseau.
- Il contient normalisateur de trafic, scanner d'entretien du système, moteur de détection et trafi shaper.

I.3.11.1. 1. Normalisateur de trafic :

- Le normalisateur dans le chemin réseau pour capturer le trafic, puis tester s'est-il existé des anomalies avant l'envoie.
- Après la normalisation du trafic, il existe la phase de suppression des paquets qui ne respectent pas la politique de sécurité.

I.3.11.1. 2. Moteur de détection :

- La gestion des modelés qui n'est pas avoir une gestion par le normalisateur.
- Ces modèles ne sont pas en fonction des états des protocoles.

I.3.11.1. 3. Traffic Shaper :

- le trafic se classifie par un formateur dépend sur le protocole trafic.
- la classification se fait avant la sortie. [Joseph Migga Kizza]

I.3.11.1.4. Avantages du NIPS :

1- Prévention zéro latence : la notification se fait dans circuit câblé,

qui permet la minimisation de la latence.

2-Hygiène efficace du réseau : le NIPS efface les paquets traités par le NIDS dans les attaques connues.

3-Gestion simplifiée : la présence de NIDS et le pare-feu comme une seule entité permet à minimisation dans l'espace du stockage. [Joseph Migga Kizza]

I.3.11.2. Systèmes de prévention des intrusions sur l'hôte (HIPS):

- Le HIPS travaille avec bac a sable qui permet la limitation de définir des règles de comportements trouvés dans le HIPS.
- Le HIPS fonctionne la prévention par un agent reliant à l'hôte.
- L'agent capture les appels et les messages système par le changement des bibliothèques ont des relations avec le (dll).
- Le changement se réalise par l'addition des dll système existantes avec le DLL stub fournisseur
- Avec cette opération les mauvais appels sont traités. [Joseph Migga Kizza]

I.3.11.2.1. Avantages de HIPS :

1- prévention efficace basée sur le contexte : l'utilisation des agents HIPS permet d'avoir des informations sur l'état complet de l'environnement, donne la possibilité de défense contre les attaques.

2-Efficace contre les attaques de jour zéro : l'utilisation de la méthode sandbox permet de faire des modifications dans les paramètres des applications avec une façon acceptable, donc l'agent doit être capable d'avoir tous les paquets. [Joseph Migga Kizza]

I.3.12. Quelques outils d'SDI

Les outils d'SDI travaille encore mieux après l'analyse des vulnérabilités fonctionne d'une façon efficace. Le tableau suivant représente les outils d'identifications utilisées [Joseph Migga Kizza]

		SDI basée sur la connaissance			SDI basée sur la comportement	
origine	Nom de SDI	SA	STA	STATE	HB	NB
	AIDE					
PCI	OSSEC				X	
	Osiris					X
	Samhain					
AlientVault	OSSIM				X	X
Source Forge	Sguil				X	X
ArcSight	ESM			X	X	X
Symtrex, Inc	SNARE	X			X	X
US Air Force	ASIM	X			X	X
	Snare BackLog for Solaris					
SECSI Project, France	ORCHIDS			X	X	X
Software. Informer	Net Ranger				X	X
SRI International	Emerald	X		X		X
Uc Davis	GrIDS	X		X		X
WindowsItPro	Kane Security Monitor				X	X

Tableau des outils IDS

I.4. Conclusion

Ce chapitre présente les notions principales sur le BigData et le système de détection d'intrusions. D'une part, nous présentons des généralités sur le BigData qui contient les définitions et les domaines d'application du BigData avec les défis.

D'une autre part, nous parlent sur le système de détection d'intrusions avec ces approches et ces types. Dans cette partie, nous présentons le système de prévention d'intrusions.

Dans le chapitre prochain, nous parlent sur les travaux connexes qui ont travaillait sur le sujet du système de détection d'intrusions dans BigData.

Chapitre II : Travaux connexes et étude comparative

II .1.Introduction :

L'augmentation du volume des données à travers l'évolution des technologies informatiques cause la naissance du BigData, ces données volumineuses circulent dans le réseau, elles sont sensibles aux attaques.

L'une de ces attaques est les intrusions ou l'intrusion est un accès non autorisée au système ou un réseau, les entreprises protègent leur système et BigData contre l'intrusion à l'aide d'une technique qui s'appelle la détection d'intrusion. La détection d'intrusion dans BigData est développée au cours du temps. Elle permet de définir et détecter les intrusions, plusieurs chercheurs proposent des approches pour détecter les intrusions dans BigData présentées dans ce chapitre.

Il contient une partie qui présente les approches avec les points forts et faibles de chaque approche, puis nous construisent un tableau comparatif entre les différentes approches définies dans la partie précédente.

II .2.les travaux connexes :

II .2.1.Approche basée Arbre de décision

Les auteurs [Mariem Ajabi, Imen Boukhris and Zied Elouedi ,2016], proposent une approche basée sur l'arbre de décision fonctionne parallèlement par l'utilisation Hadoop et MapReduce

Elle est une méthode de classification des données massives(BigData) dans des environnements incertains.

L'arbre de décision permet de résoudre le problème de l'ambiguïté des données massives par la classification des données en utilisant la théorie de fonction de croyance.

II.2.1.1.Les éléments de la solution proposée :

II.2.1.1.1.Hadoop :

Hadoop donne un ensemble des clusters en même temps où chaque cluster compose d'un ensemble des nœuds.

Hadoop contient deux composants principaux :

II.2.1.1.1.1.MapReduce : c'est un modèle de programmation utilise sa grande puissance pour l'exécution des données volumineuses pour notre solution.

II.2.1.1.1.2.Système de fichiers distribués Hadoop (HDFS) : c'est un système qui permet le stockage des données d'une manière distribuée qui caractérise par la non-perte des données.

II.2.1.1.2.Arbre de décision de croyance :

II.2.1.1.2.1.Théorie de fonction de la croyance : qui représente les connaissances incertitudes par la définition de l'assignation de croyance de bas par la formule mathématique.

II.2.1.1.2.2.arbre de décision de croyance : c'est un arbre de décision qui fait la classification des données dans un environnement incertain.

II.2.1.2.la solution proposée : la solution proposée basée sur la création de l'arbre de décision puis l'adaptation de MapReduce, la création se fait à l'aide de l'approche de moyenne par créer l'arbre puis assurer la classification.la solution passe par deux étapes :

II.2.1.2.1.Traitement de la structure de données dans la méthode proposée

Les données traitées par cette solution définie par catégories ou attributs numériques. L'ensemble de données est généralement composé d'attributs en colonnes, d'instances en lignes dénotées par un identifiant unique. L'utilisation de MapReduce permet l'échange des données sans perte d'information à cause du clonage de l'information pour chaque nœud. Les structures des données sont : la table d'attributs, la table de comptage, la table d'hash.

II.2.1.2.2.Adaptation du MapReduce : Le processus de l'adaptation de MapReduce contient quatre étapes, avec l'utilisation des fonctions Map et Reduce qui sont :

1- préparation des données.2-la sélection des attributs.

3-la mise-à-jour des tables.4- la croissance des arbres. [Tarek Gaber · Aboul EllaHassanien ,Nashwa El-Bendary · Nilanjan Dey]

II.2.1.3.Les inconvénients :

1 – la maintenance des tables de longues tailles prend une longue période.

2-une approche basée sur un modèle mathématique.

II .2.2.Méthode orienté Cloud

Les auteurs [Halim Gorkem Gulmez, Emrah Tuncel, and Pelin Angin], ont proposé une approche analytique du BigData pour détection d'intrusion dans les systèmes Cloud, elle utilise des réseaux de neurones récurrents qui fonctionnent sur l'intégration du comportement temporel du système dans la tâche de détection d'anomalies avec un temps quai réel.

Cette approche collecte les mesures du système depuis les plateformes Cloud, puis le traitement de ces mesures sous forme des flux par le moteur de traitement de flux. Enfin, l'envoie de ces mesures au réseau de neurone récurrent qui signale une alarme aux administrateurs du système lorsqu'on détecte une anomalie.

II.2.2.1.la solution proposée : cette solution fonctionne avec les réseaux de neurones récurrents sur les systèmes Cloud, elle compose de 3 étapes principales :

II.2.2.1.1.Collection des métriques :

La collection des métriques du système se fait par des agents appelés les agents des machines virtuelles invités. Les agents contiennent deux composants : le producteur collecte les mesures du système et d'application des différentes interfaces, et le consommateur rassemble les métriques à partir le producteur et transmettre à la phase de traitement.

II.2.2.1.2.traitement des métriques :

L'étape de traitement est la deuxième étape pour la détection des intrusions dans Cloud, il existe le traitement par lots ou par flux. On choisit le traitement des données par flux car la détection est en temps réel et le traitement en flux se fait sur les données lorsqu'on arrive dans le même temps. Au contraire le traitement en lot prendre un temps pour la création des lots depuis les données puis le traitement des lots des données.

II.2.2.1.3.Apprentissage basé sur RNN pour la détection d'anomalies :

Cloud est un environnement ouvert. Il peut exister des attaques connues ou non connues. On utilise Les méthodes d'apprentissage sans surveillance qui détectent les changements dans le système. On utilise cette méthode d'apprentissage dans le réseau de neurone récurrent qui est des machines modèles d'apprentissages composants des nœuds

reliant entre eux, ces nœuds mémorisent et traitent les informations en séquence. Le réseau de neurone récurrent utilise les informations du passé pour les événements de futur pour les mêmes sujets. Le réseau prend l'entrée x, puis applique un traitement sur x. ensuite, le résultat est la sortie h qui est utilisée dans l'étape suivante.

On utilise LSTM-RNN, ou LSTM (Long Short Tem Memory) ajoute des portes supplémentaires pour contrôler l'état de la cellule. LSTM-RNN fonctionne les étapes suivantes :

1-la première couche obtient l'entrée et la sortie actuelle de la série chronologique passée.

2- autre couche sigmoïde, qui décide des valeurs à mettre à jour.

3- une couche tangente hyperbolique crée des valeurs candidates pouvant éventuellement être incluses à l'état de cellule.

4- les résultats de toutes les étapes précédentes sont combinés afin de créer une mise à jour de l'état de la cellule

5- enfin, la sortie est décidée. [Min Luo • Liang-Jie Zhang]

II.2.2.12. Les Inconvénients :

1-La sécurité est le grand défi pour le Cloud et les réseaux de neurones récurrents.

II.2.3.SDI pour réseau a haute vitesse

Les auteurs **[M. Mazhar Rathore · Awais Ahmad ·Anand Paul , 2016],** ont proposé un système qui présente un système de détection d'intrusion ultra-rapide qui permet de détecter les intrusions dans le réseau avec plus d'efficacité et d'exactitude dans un temps réel.

II.2.3.1.La solution proposée : la solution proposée est un système qui détecte les anomalies en temps réel avec une vitesse ultra rapide, on présente l'architecture et l'algorithme qui sont utilisées pour ce système.

II.2.3.1.1.L'architecture proposée : l'architecture est implémentée sur n'importe quel périphérique réseau haute vitesse, il compose de quatre couches principales sont :

II.2.3.1.1.1.Couche de capture du trafic ; cette couche est responsable de capter le trafic en utilisant des périphériques avec une vitesse rapide et un débit élève, puis il envoie le trafic à la couche suivante. Tous les paquets sont captés sans perte.

II.2.3.1.1.2.Serveur de filtrage et d'équilibrage de charge : cette couche a deux fonctions principales :

1- cette couche fait un filtrage sur le trafic qui n'est pas décidé comme une intrusion **o**u un trafic normal pour la recherche.

2- il envoie les flux de trafic non identifiés et un paquet contient les informations de l'entête au maitre de couche Hadoop.

3- il équilibre la charge en faire une décision de choix des maitres d' Hadoop qui font recevoir les flux en fonction des adresses IP.

II.2.3.1.1.3.couche Hadoop : Le maître prend le trafic et génère le fichier séquence pour chaque flux, le nœud maître extrait les informations nécessaires de chaque paquet et stocke ces informations dans le fichier de séquence. Ensuite, le fichier de séquence est envoyé aux nœuds de données qui sont équipés d'un algorithme de calcul de valeur de caractéristique implémenté dans MapReduce. Les valeurs envoyant à la couche suivante.

II.2.3.1.1.4.couche serveur de décision : le serveur de décision classifie les flux en utilisant des algorithmes de classification, et décide les flux normaux ou des flux d'attaque. Les décisions de chaque flux sont stockés dans la base de données des intrus on mémoire pour être utilisés par la couche de filtrage

II.2.3.1.2.l'algorithme proposé : un algorithme proposé pour toutes les couches afin d'identifier l'intrus. Les flux sont distincts par quatre champs, à savoir IP source, IP de destination, port source et destination

1- chaque paquet est capturé, puis on filtre chez FLBS

2-FLBS transmettent les paquets appartenant aux flux qui ne sont pas identifiés comme intrus ou flux normaux pour les traiter

3-un nœud maître vérifie l'absence du paquet dans le flux, puis crée comme un nouveau flux.

4-le paquet appartient au flux enregistré, les informations de paquets sont alors venues d'entrer dans le fichier de séquence correspondante.

5-Le nœud maître continu à copier les informations de paquets dans le fichier de séquence jusque le seuil de durée s'écarte, la séquence du fichier est envoyée à l'un des nœuds de données pour le calcul des paramètres de flux.

6-Le nœud de données utilise les fonctions Map and Reduce équipées de paramètres. Code de calcul pour mesurer les valeurs finales de chacune des neuf fonctions d'intrusion détection. [M. Mazhar Rathore , Awais Ahmad ,Anand Paul]

II.2.3.2.les inconvénients :

1-L'augmentation de la taille de fichier de séquence.

2-Le problème de sécurité.

II.2.4.SDI multicouches

Les auteurs [Mostafa Doroudian, Narges Arastouie, Mohammad Talebi, Ali Reza Ghanbarian, 2015], ont proposé une méthode pour la détection d'intrusions dans base de données, le système présenté permet de détecter les tâches d'utilisateurs anormaux dans les bases de données. Le système passe par deux phases principales pour la détection des intrusions dans le système de gestion de la base de données. La première phase est l'apprentissage permet d'identifier les règles de dépendance et de séquence entre les échanges existant dans les travaux d'utilisateur, puis une phase de détection qui fonctionne sur la connaissance des tâches normales et anormales d'utilisateur en utilisant les règles définies dans la phase précédente pour chaque nouvelle transaction.

II.2.4.1.la solution proposée : la solution proposée donne un système de détection d'intrusion dans les bases des données, ce système compose de deux étapes principales :

II.2.4.1.1.Une phase d'apprentissage : cette étape contient deux modules :

1-module de spécification génération : il définit les échanges trouvant dans l'entreprise, puis on donne une spécification des échanges par le module de créateur de spécification.

2-Module de génération des règles : dans ce module, le préprocesseur de la règle rassemble les échanges existant dans le journal de formation de la base de données dans les tâches d'utilisateurs. Ensuite, depuis ces échanges regroupés le générateur de la règle crée des règles de dépendance et de séquence.

II.2.4.1.2.Une phase de détection : cette phase contient trois composants :

1-Préprocesseur de détection : il prend les informations importantes depuis les nouveaux échanges et qui ont stocké dans le journal de la base de données, puis il envoie ses informations au détecteur de transaction malveillante et le détecteur des tâches utilisateurs malveillants.

2-Détecteur de transaction malveillante : il compare et relie entre les nouveaux échanges et les spécifications définies pour chaque échange à travers les règles définies.

3-Détecteur des taches utilisateurs malveillantes : il identifie les relations correspondantes les échanges dans les tâches utilisateur, puis il connaitre la tâche utilisateur normal ou anormale en utilisant les échanges trouvant dans une nouvelle tâche utilisateur et les relations correspondantes. [Mostafa Doroudian, Narges Arastouie, Mohammad Talebi, Ali Reza Ghanbarian]

II.2.4.2.Les inconvénients :

1-La maintenance de règles existantes prend beaucoup de temps.

2-le choix des informations optimales influence la phase de détection.

II.2.5.Détection par estimation statistique

Les auteurs [Nour Moustafa, Gideon Creech, and Jill Slay, 2017], ont proposé une solution qui représente une Framework qui permet du développement de l'approche de détection d'anomalies (l'approche comportementale) pour un système de détection d'intrusions qui est efficace pour connaitre les anomalies dans les systèmes réseaux. Un modèle statistique est développé pour être utilisé dans la solution proposée qui donne un niveau plus efficace au niveau de détermination des anomalies dans BigData.

II.2.5.1.La solution proposée : les auteurs donnent une solution pour la détection d'intrusion par le développement de la méthode détection d'anomalies en utilisant un modèle statique qui est modèle de mélange de Dirichlet. La solution permet la détection d'intrusion avec plus d'efficacité.

II.2.5.1.1.Le modèle de mélange de Dirichlet : est un modèle statistique qui dépend de la probabilité de l'obtention d'un comportement normal ou un comportement anormal qui signifie une attaque basée sur la fonction de densité probabiliste, avec ce modèle qui donne une estimation statistique pour la détection d'intrusion. Il compose de deux étapes principales :

II.2.5.1.1.1.Phase de formation des instances normales : cette phase permet la construction des profils des utilisateurs ou réseau par l'utilisation des informations normales (instances) en utilisant la fonction de densité probabiliste. On applique

l'algorithme suivant pour chaque instance qui prend comme entrée les instances normales(les informations) et donne comme sortie le profile normal pour un ensemble des instances, l'algorithme contient les étapes suivantes :

1-pour chaque enregistrement i parmi l'ensemble des instances, on calcule les paramètres (π_i, α_i, Z_i) par des équations mathématiques utilisant dans la fonction de densité probabiliste.

2-calcule la fonction de densité avec les paramètres calculés dans l'étape 1 pour chaque enregistrement.

3-on calcule les deux résultats lower et upper qui représentent le quartile de la première fonction de densité probabiliste et le quartile de la troisième fonction de densité probabiliste respectivement.

4-le calcule d'IQR qui représente la substruction de deux résultats upper et lower.

5-le définition du profile normale qui contient les paramètres (π_i, α_i, Z_i) et les résultats lower et upper et IQR.

II.2.5.1.1.2.La phase de test et méthode décisionnelle : cette phase permet de tester et décider l'état des enregistrements observés soit sont des enregistrements normaux, soit une intrusion (une attaque). L'algorithme pour cette phase prend en entrée les instances observées et le profile normal qui est le résultat de la phase précédente, qui contient les étapes suivantes :

1-Le calcule de la fonction de densité probabiliste pour chaque enregistrement observée avec le mêmes paramètres de la profile normale.

2-La comparaison entre la valeur de chaque fonction et les valeurs suivantes : (lower-w*(IQR)) et (upper+w*(IQR)), qui donne deux cas :

1-si (la valeur de fonction < (lower- w*(IQR)))|| (la valeur de fonction > (upper+ w *(IQR))) alors il existe une attaque

2-dans le cas contraire, l'enregistrement observé est normal

II.2.5.1.2.Les modules de la solution proposée : la solution proposée contiennent trois modules :

1-module de capture et de journalisation : ce module permet de capturer les données réseau et les informations circulent dans le réseau en utilisant des outils de capture, ce module avoir une autre fonctionnalité qui est le stockage de données captées pour le traitement dans le module moteur de décision, les données captées transmettent vers le module de prétraitement pour appliquer un prétraitement sur les données.

2-module de prétraitement : ce module prend les données stockées dans le module de capture, il applique un prétraitement sur les données pour faire un filtrage et définition sur les données réseau, ce module passe par 4 étapes :

1- création d'entités : cette étape permet de créer les entités/ des attributs à partir des données de réseau en utilisant des différents outils.
2- Conversion d'entités : elle permet de transformer les attributs symboliques à des attributs numériques car le traitement dans le moteur de décision se fait à l'aide des fonctions mathématiques qui utilisent des attributs numériques.
3- Réduction d'entités : elle permet de réduire le nombre des attributs en évitant les attributs non nécessaires l'utilisation de la technique PCA parce qu'il est le meilleur technique avec un petit espace de stockage et un temps de transfert et traitement est court.
4- Normalisation d'entité : il permet de classer la valeur de chaque attribut dans un intervalle défini afin d'éviter le bruit qui peut détruire la valeur de chaque attribut en utilisant une fonction mathématique.

3-module de moteur de décision : il donne une décision sur les entités sortant du module de prétraitement en utilisant le modèle de mélange de Dirichlet, il prend une décision entre l'attaque et les données normales. On applique le modèle de mélange sur les entités qui sont les résultats du module de prétraitement, qui permet la détection des intrusions et les attaques. [Nour Moustafa, Gideon Creech, and Jill Slay]

II.2.5.2.Les inconvénients :

1-cette solution ne permet pas de détecter les attaques telles que DOS et backdoors.

2-la solution basée sur une estimation statistique qui ne peut être pas donné les bons résultats pour la détection d'intrusions.

II.2.6.Exploitation de Kmeans pour la DI

L'auteur [**Kai Peng, 2017**], propose une approche qui fonctionne sur la classification et le regroupement en utilisant la fonction de petit lot Kmeans qui est capable de la détection d'intrusions sur le BigData. L'approche utilise la fonction Kmeans avec des petits lots et la méthode d'analyse des composants principaux, on définit ces deux concepts.

II.2.6.1.la fonction de Mini lots Kmeans : c'est l'extension de la fonction Kmeans originale qui utilise l'ensemble des données pour être classifiée dépend sur le centre de cluster le plus proche pour chaque enregistrement en utilisant la fonction f(C, x) puis on fait la sélection d'un nombre fini k des enregistrements d'une manière aléatoire avec Kmeans++. Lorsque le volume de données devient très grand, on utilise la fonction Mini lots Kmeans pour réduire le temps de regroupement car on choisit un sous-ensemble des enregistrements d'une manière aléatoire depuis l'ensemble des données

II.2.6.2.la méthode d'analyse des composants principaux : cette méthode repose sur l'idée de trouver un vecteur de direction et le projeter sur l'ensemble des données pour un but de diminuer le volume de l'ensemble de données. La méthode passe par les étapes suivantes :

Chaque donnée appartient a l'ensemble de données représenté par une matrice $T_{m*n.}$ avant tout, on calcule le moyen pour chaque colonne de la matrice.

1-le calcul de la soustraction de la moyenne des colonnes et les donnés qui ont traité, donc le résultat est une autre matrice $TAdjust_{(m*n).}$

2-on calcule la matrice de covariance de $TAdjust_{(m*n)}$ par la formule : $c_{ij}=Cov(x_i, x_j)= E\{[x_i - E(x_i)][x_j - E(x_j)]\}$, on obtient la matrice C.

3- le calcul des valeurs propres et des vecteurs propres de la matrice de covariance C_{n*n}

4-la sélection d'un nombre d fini des valeurs propres les plus grandes sont les valeurs propres sont classifiées d'un ordre décroissant, puis on prend les d-vecteurs propres correspondantes aux les valeurs propres sélectionnées en formant une matrice de ces vecteurs qui appelle la matrice des vecteurs propres $EigenVectors_{n*d}$.

5-la projection de la matrice de covariance sur la matrice des vecteurs propres sélectionnés, on obtient la matrice des données finales $Finaldata_{m*d}$ en utilisant la formule : $FinalData_{m*d}=TAdjust_{m*n}*EigenVectors_{n*d.}$

II.2.6.3.la solution proposée : la méthode proposée est la méthode détection d'intrusions en utilisant une méthode de regroupement fonctionne pour BigData. La méthode compose de deux étapes :

II.2.6.3.1.le prétraitement des données : cette étape permet d'applique un prétraitement sur les données qui contiennent un processus de chaine et de normalisation ou le processus de chaine permet la transformation des caractères trouvés dans les données à des nombres en utilisant une fonction de remplacer car le traitement se fait avec les nombres ,puis on applique le processus de normalisation qui permet de normaliser les nombres existant dans les ensembles de données qui ne sont pas dans une forme définie par la fonction de normalisation. Donc on obtient le résultat qui est un ensemble de données numériques et normalisées qui sont les entrées dans l'étape suivante.

II.2.6.3.2.Mini lots Kmeans avec l'analyse de composant principal pour système de détection d'intrusion : cette étape prend les données sortant depuis l'étape précédente et appliquer les étapes suivantes pour faire la classification et la détection d'intrusions :

1-on réduire la dimension des données en fait appel de la fonction d'analyse des composants principaux par le suivre des étapes définie dans l'élément II.2.6.2 dans la partie précédente.

2-on obtient un nombre définit k des centres du cluster en appliquant la fonction Kmeans++ qui permet la sélection d'un K centres de cluster par l'utilisation d'une fonction qui calcule la distance entre les enregistrements et le centre de cluster sélectionné qui représente un enregistrement sélectionné aléatoirement qui donne un deuxième centre de cluster puis on appelle une fonction de probabilité pour déterminer les autres centres de cluster qui répète jusqu'à définir les K centres.

3-la sélection d'une manière aléatoire des lots avec des tailles b, puis on forme une collection de ces lots sélectionnés qui appelle M.

4-la calcul de la distance entre chaque lot et les centres, ensuite associer chaque enregistrement au centre du cluster le plus proche.

5-on recalculer le centre de chaque cluster.

6-finalement, les K clusters sont formés par l'association des enregistrements proches à ces k centres et une valeur CH est calculée en appliquant une fonction de score ou CH est une valeur pour la l'évaluation des résultats de classification. [Kai Peng]

II.2.6.4.les inconvénients :

La complexité de l'algorithme proposé est très élevée.

II.2.7.Apprentissage basé DI

Les auteurs [Rachana Sharma & Priyanka Sharma, Preeti Mishra & Emmanuel S. Pilli], proposent une solution pour la détection d'intrusion en utilisant l'approche de détection d'anomalies basée sur les algorithmes de Machine Learning dans l'environnement BigData. Cette solution utilise MapReduce comme un modèle de programmation qui offre une détection des anomalies d'une manière rapide et évolutive

II.2.7.1.la solution proposée : cette approche utilise les algorithmes de machine Learning pour la classification des intrusions dans BigData, elle basée sur MapReduce comme un modèle de programmation. Cette approche utilise deux algorithmes de classification utilisée pour la Machine Learning

II.2.7.1.1.Machine Learning et la classification utilisant MapReduce

II.2.7.1.1.1.MapReduce basée sur l'algorithme de naïve Bayes : l'algorithme basé sur le théorème de Bayes qui permet de calculer la probabilité conditionnelle d'un événement basée sur des événements qui font des hypothèses entre les prédicteurs. L'algorithme qui utilise MapReduce passe par deux étapes :

1-La phase Map qui fait le calcul de la somme de chaque prédicteur de l'objet pour chaque classe, la valeur de somme est la valeur entrée pour l'étape suivante.

2-La phase Reduce qui prend la valeur de somme comme entrée depuis l'étape précédente puis faite le calcul la probabilité des deux classes sélectionnées pour le test ou la classe qui a une probabilité élevée c'est la classe sélectionnée pour le test.

II.2.7.1.1.2.MapReduce basée sur l'algorithme K plus proche voisin : l'algorithme utilise pour la classification, il permet de trouver les objets de nombre k définit le plus proche pour construire de la classe pour le test. L'algorithme compose de deux étapes :

1-La phase Map qui fait le calcul la distance euclidienne pour chaque instance de test .la distance calculée est une entrée pour la phase suivante.

2-La phase Reduce permet de trier les distances puis choisit les objets les plus proches qui ont utilisé pour trouver la classe correspondante pour la classification. [Rachana Sharma & Priyanka Sharma, Preeti Mishra & Emmanuel S. Pilli]

II.2.7.2.les inconvénients :

La complexité des algorithmes utilisés est élevée.

II .3.Tableau comparative :

Approches et critères	évolutivité	efficacité	rapidité	Les algorithmes de classification	L'utilisation de hadoop Et MapReduce	La Détection d'intrusions
Approche basée sur l'arbre de décision	dépend sur le temps d'exécution	dépend sur le temps d'exécution	N'est pas mentionné	oui	oui	X
Méthode orienté Cloud	X	X	En temps Quai réel	X	X	moyenne
SDI pour réseau a haute vitesse	X	X	Ultra rapide	X	oui	élevée
SDI multicouches	X	X	X	X	X	moyenne
Détection par estimation statistique	oui	X	X	X	oui	moyenne
Exploitation de Kmeans pour la DI	X	élevée	faible	oui	oui	moyenne
Apprentissage basé DI	oui	X	élevée	oui	oui	moyenne

Tableau comparative

II .4.Synthèse et discussion :

L'étude des approches définies dans ce chapitre avec la définition du fonctionnement et les points forts et faibles de chaque approche, ensuite la représentation du tableau comparatif pour ces approches. Ce tableau permet de définir les limites suivantes :

1- La plupart des approches ont un niveau de détection d'intrusion moyenne.

2- La plupart des approches ne prend pas en compte l'efficacité du système.
3- Les approches étudiées ne donnent pas la réponse en temps réel

II .5.Conclusion :

Ce chapitre présenté les différentes approches et méthodes pour la création d'un système de détection d'intrusions dans les environnements BigData.

On fait une étude comparative entre les travaux mentionnés par la définition des avantages et les limites pour chaque travail.

La chapitre suivant contient la définition de notre architecture d'un système de détection d'intrusions dans BigData, on prend en considération les inconvénients de ces travaux.

Chapitre III : Conception du systéme

III.1.Introduction :

La sécurité est un grand défi qui pose des problèmes dans les BigData tel que les attaques et les intrusions. La résolution du problème des intrusions est le développement d'un système de détection d'intrusions dans BigData.

On voit dans le chapitre précédant les divers travaux pour la construction d'un système de détection d'intrusion qui sont proposées par plusieurs chercheurs.

Dans ce chapitre, on présente l'architecture globale de notre système puis la modélisation du système avec les diagrammes UML qui permet la détermination de la structure du système.

III.2.Architecture globale

Dans cette section, on présente l'architecture globale pour la construction d'un système de détection d'intrusions dans les environnements BigData.

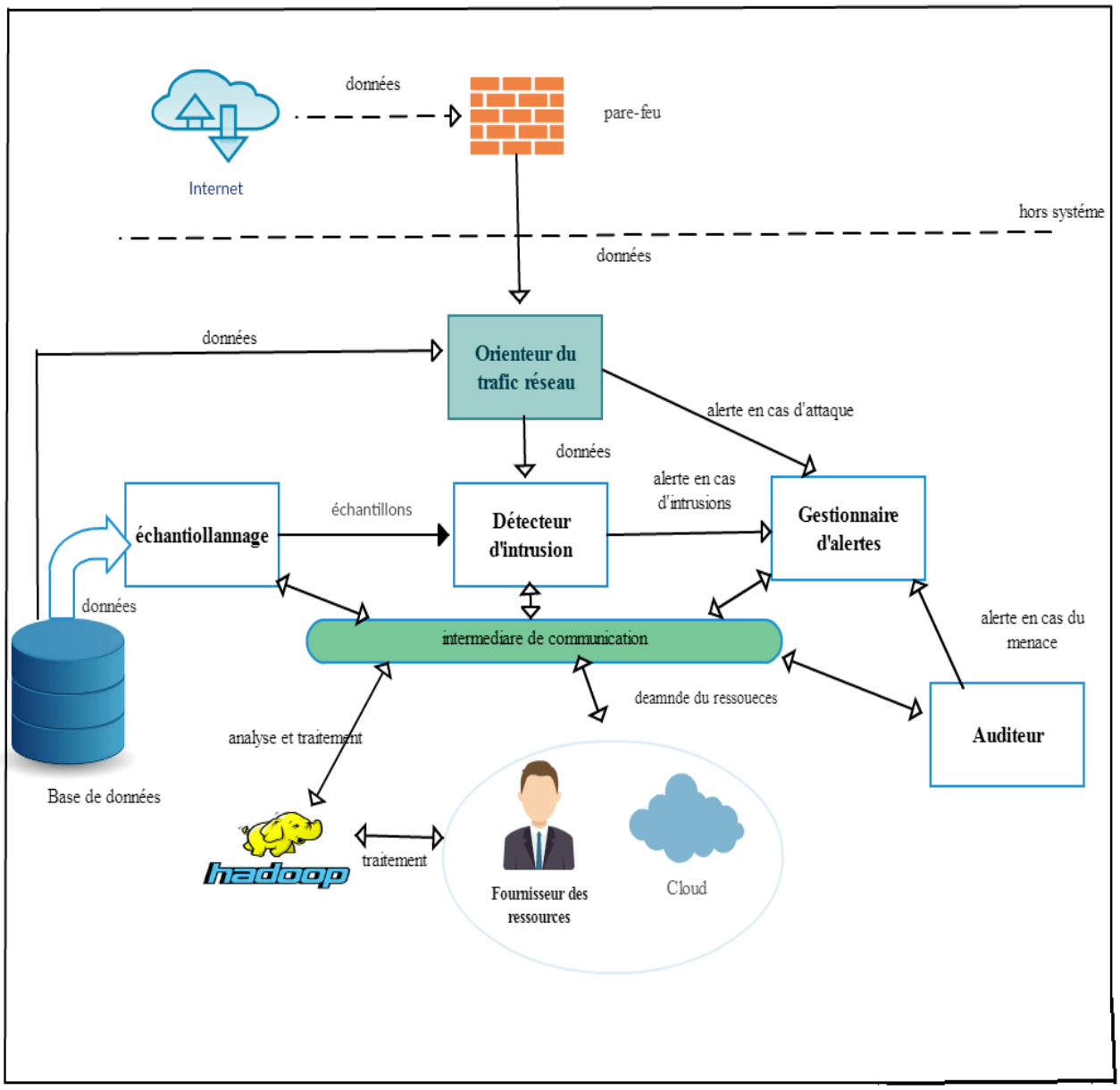

Figure III.1: l'architecture générale

III.2.1.La solution proposée : l'architecture de notre système contient plusieurs composants qui travaillent en collaboration qui permet la capacité de détecter les intrusions dans les environnements BigData. Ces composants sont :

1-Orienteur du trafic réseau. 2-Fournisseurs des ressources.

3-Auditeur. 4- échantillonnage. 5- détecteur d'intrusions.

III.2.2.Cas d'utilisations des composants du système :

III.2.2.1.Diagramme de cas d'utilisation du composant orienteur du trafic réseau :

Ce composant fonctionne sur la minimisation d'impact de l'attaque Dos par l'analyse du nombre des requêtes et données dans le système.

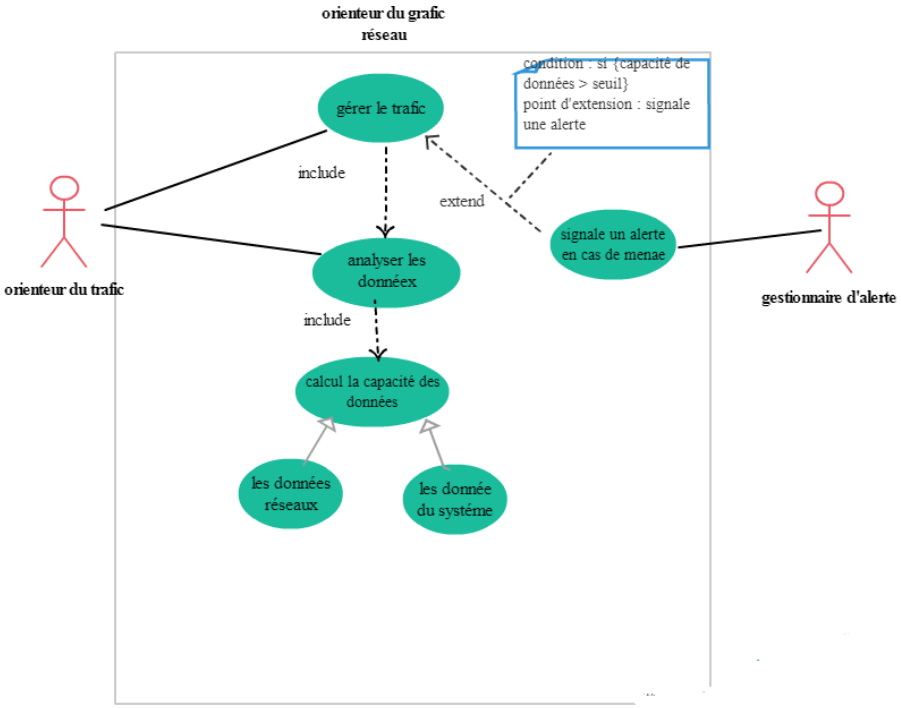

Figure III.2 : diagramme cas d'utilisation du l'orienteur du trafic réseau

III.2.2.2.Diagramme de cas d'utilisation du composant fournisseur des ressources :

Ce composant permet l'allocation des ressources pour les autres composants du système.

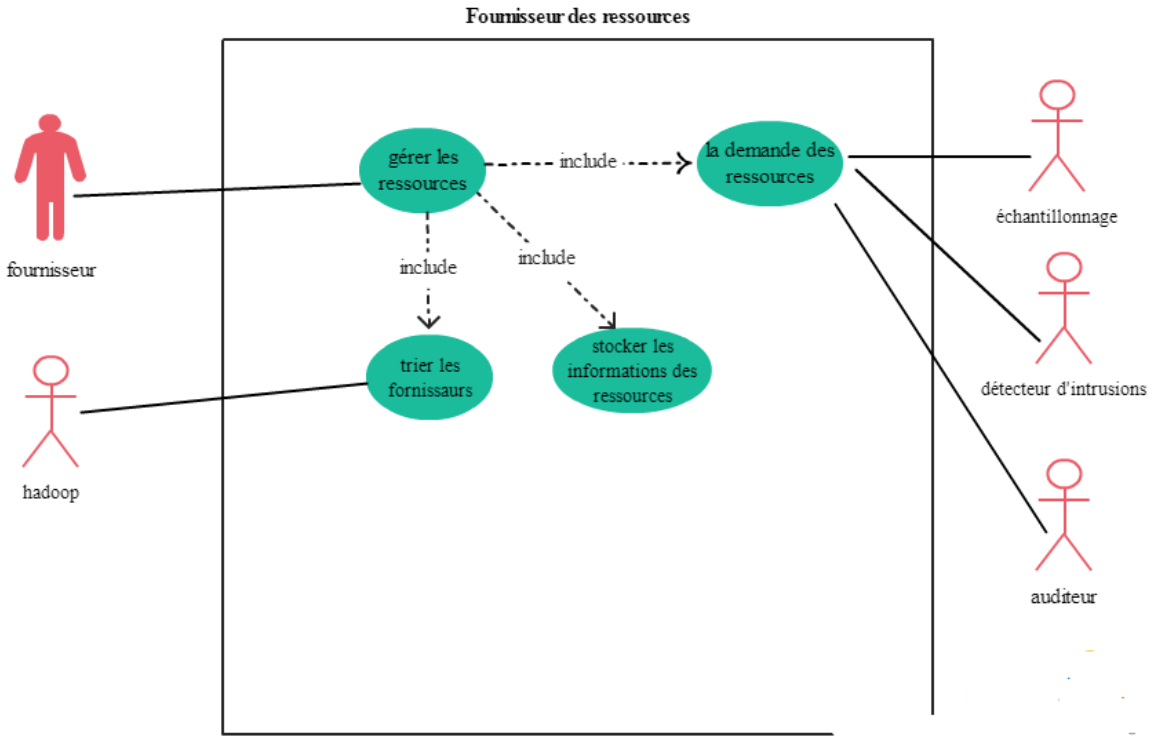

Figure III.3: diagramme de cas d'utilisation du fournisseur des ressources.

III.2.2.3.Diagramme de cas d'utilisation du composant échantillonnage :

Le composant d'échantillonnage fonctionne sur la construction des échantillons a partir des données de système, puis sélectionne aléatoirement les échantillons pour le composant détecteur d'intrusions.

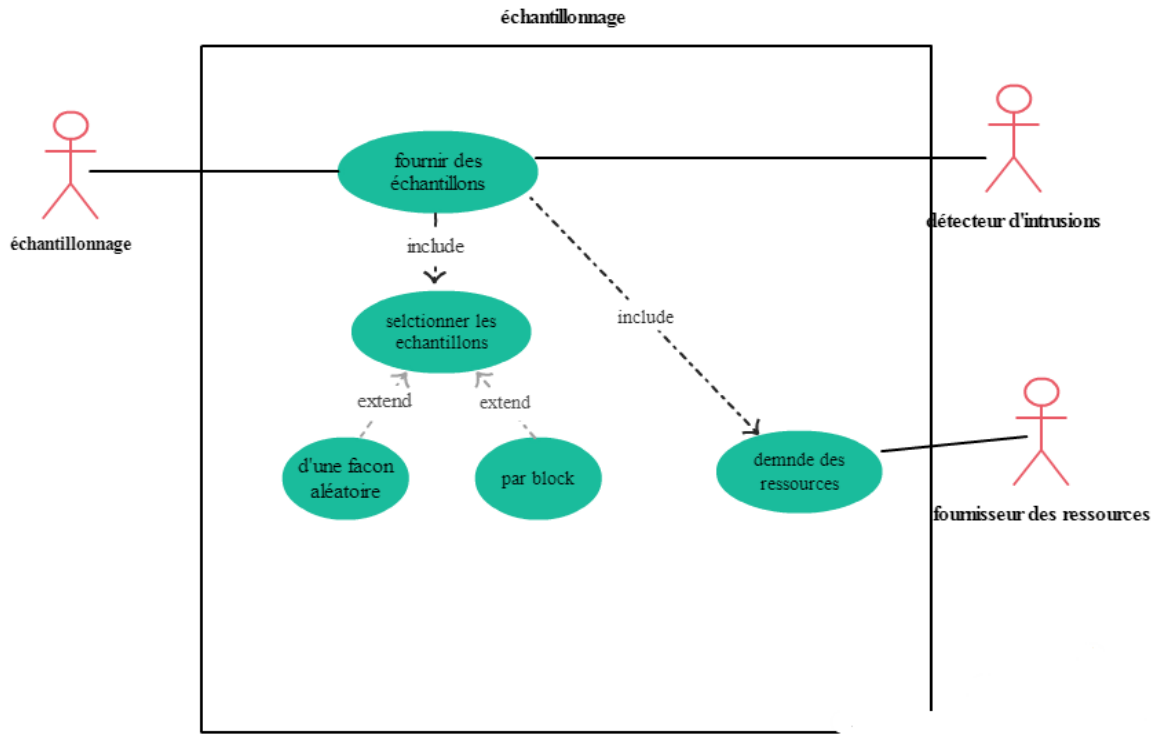

Figure III.4: diagramme de cas d'utilisation du l'échantillonnage

III.2.2.4.Diagramme de cas d'utilisation du composant détecteur d'intrusions:

Le composant de détecteur d'intrusions permet la détection d'intrusions par la comparaison entre les échantillons et les données circulent dans le système puis le signale à une alerte en cas de menace.

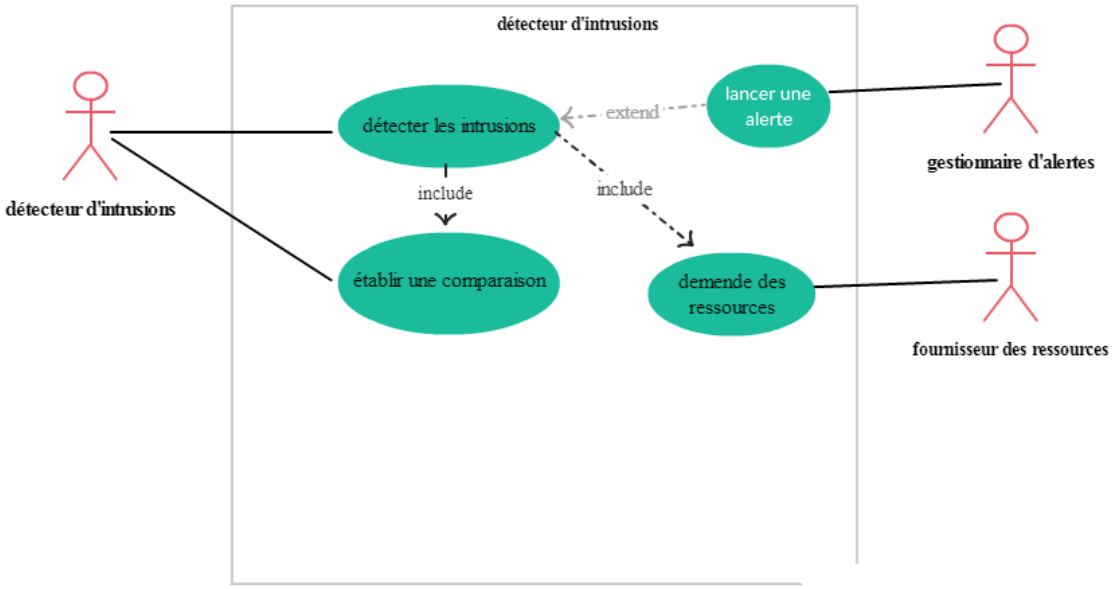

Figure III.5: diagramme de cas d'utilisation du détecteur d'intrusions

III.3.2.5.Diagramme de cas d'utilisation du composant auditeur:

Ce composant permet d'analyser les comportements normaux d'utilisateurs puis selon cette analyse, il bloquer l'accès pour les utilisateurs anormaux et lance une alerte.

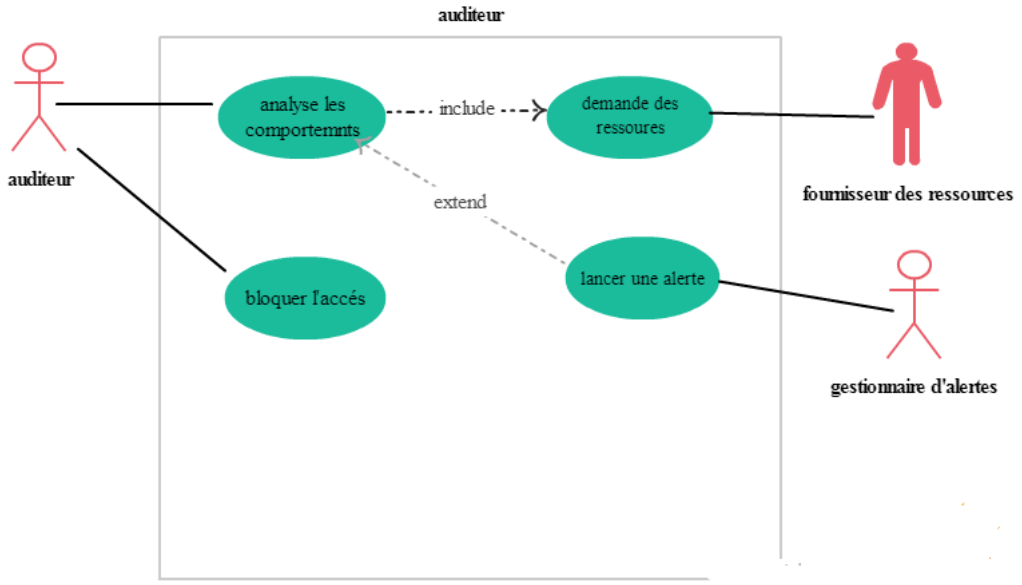

Figure III.6: diagramme de cas d'utilisation d'auditeur

III.3.3.Activités et tâches des composants :

III.3.3.1.diagramme d'activité du composant orienteur du trafic :

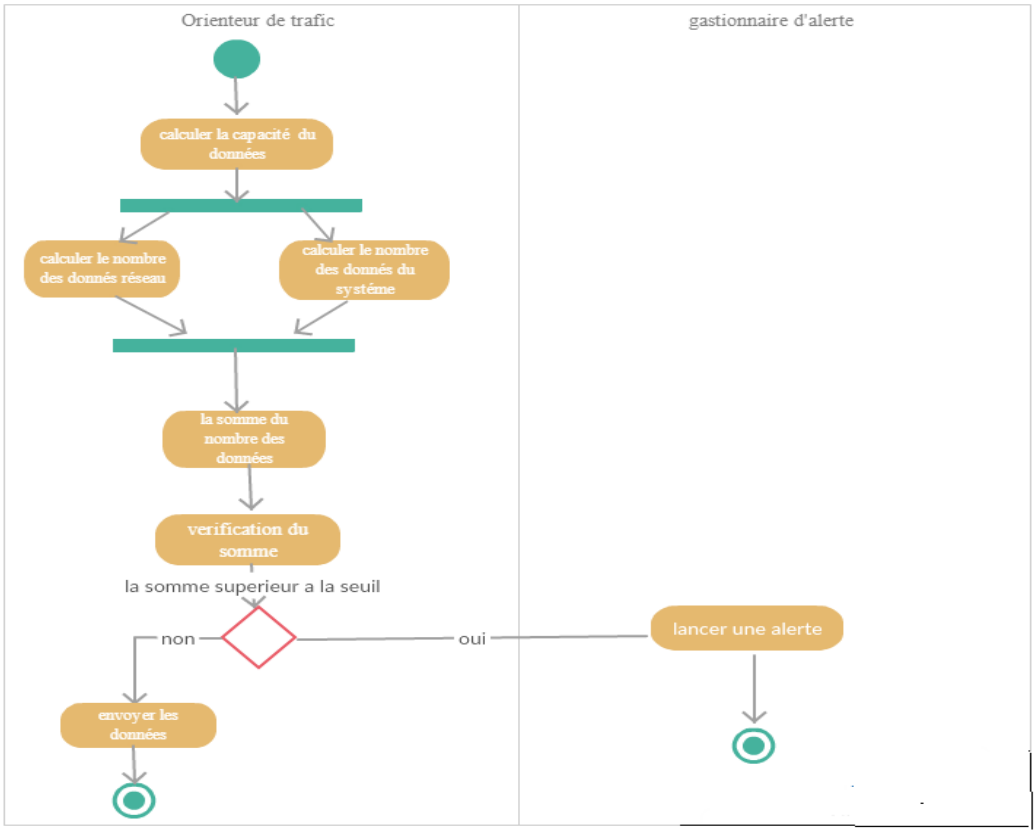

Figure III.7: diagramme d'activité du composant orienteur du trafic réseau

III .3.3.2.diagramme d'activité du composant fournisseur des ressources :

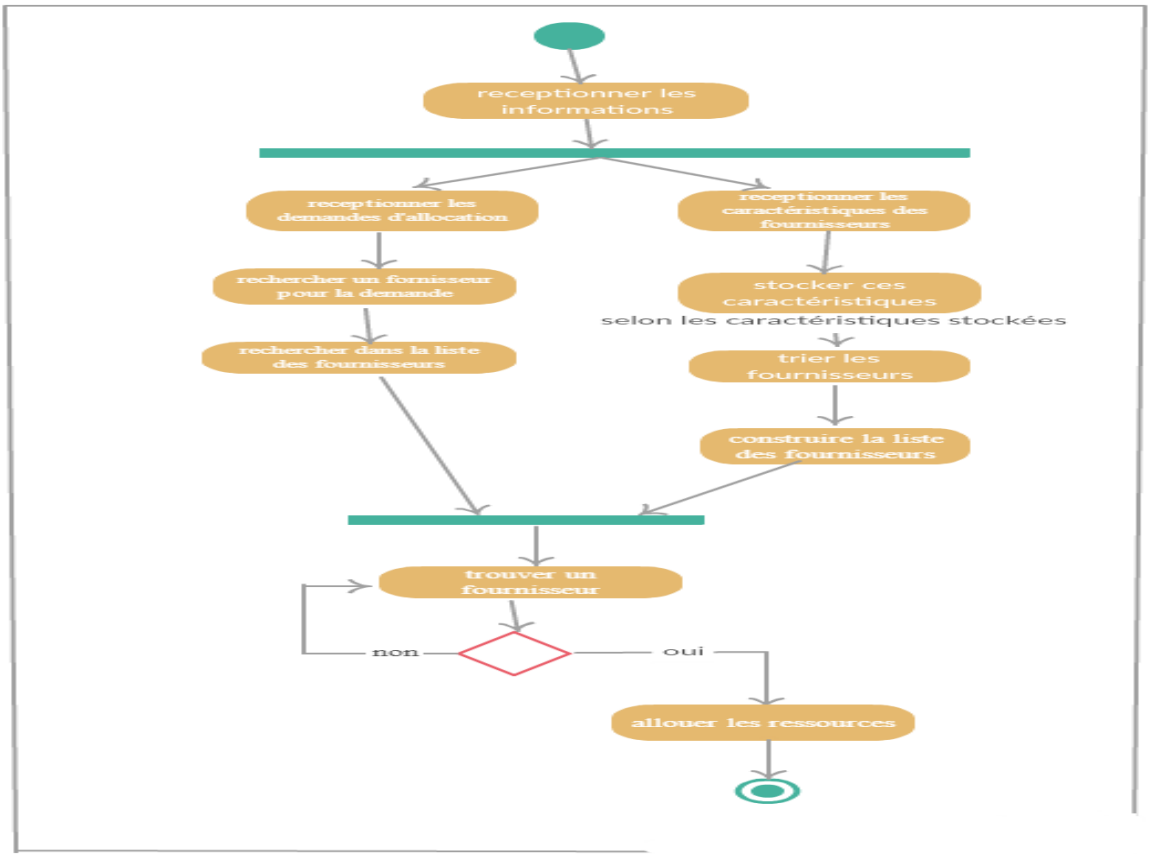

Figure III.8: diagramme d'activité du composant fournisseur des ressources

III.3.3.3.diagramme d'activité du composant échantillonnage :

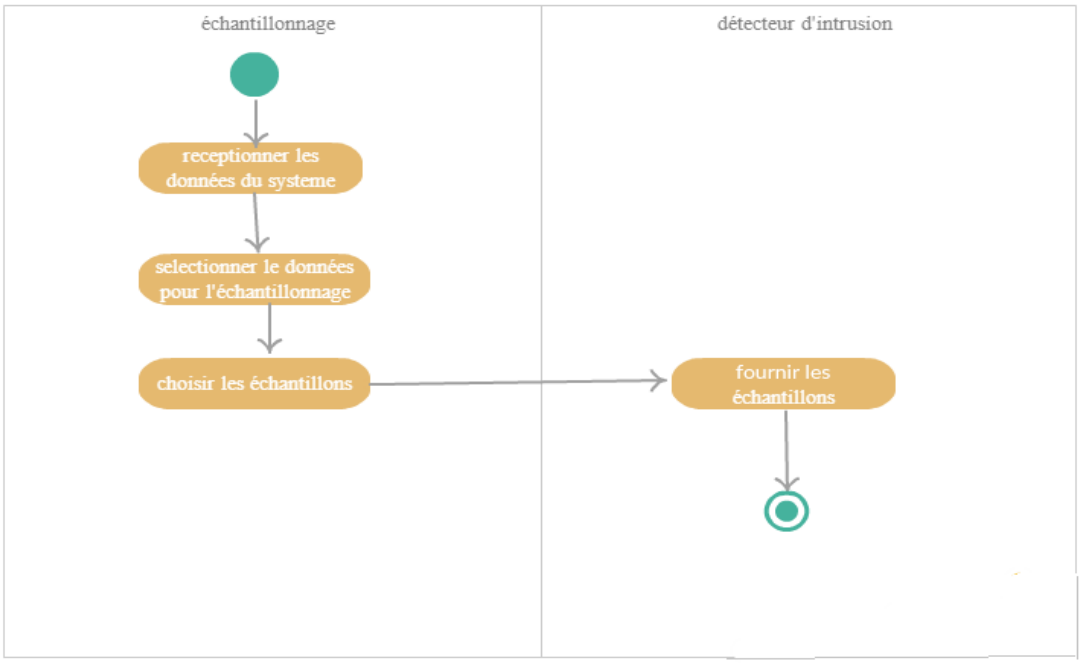

Figure III.9: diagramme d'activité du composant échantillonnage

III.3.3.4.diagramme d'activité du composant détecteur d'intrusions :

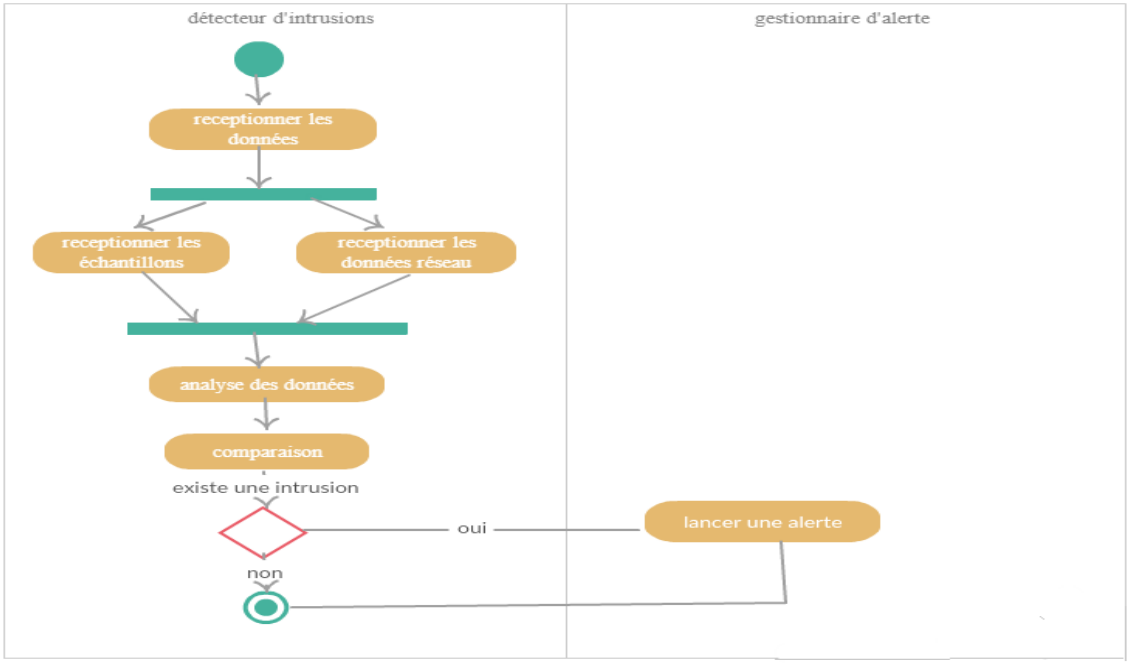

Figure III.10: diagramme d'activité du composant détecteur d'intrusions

III.3.3.5.diagramme d'activité du composant auditeur :

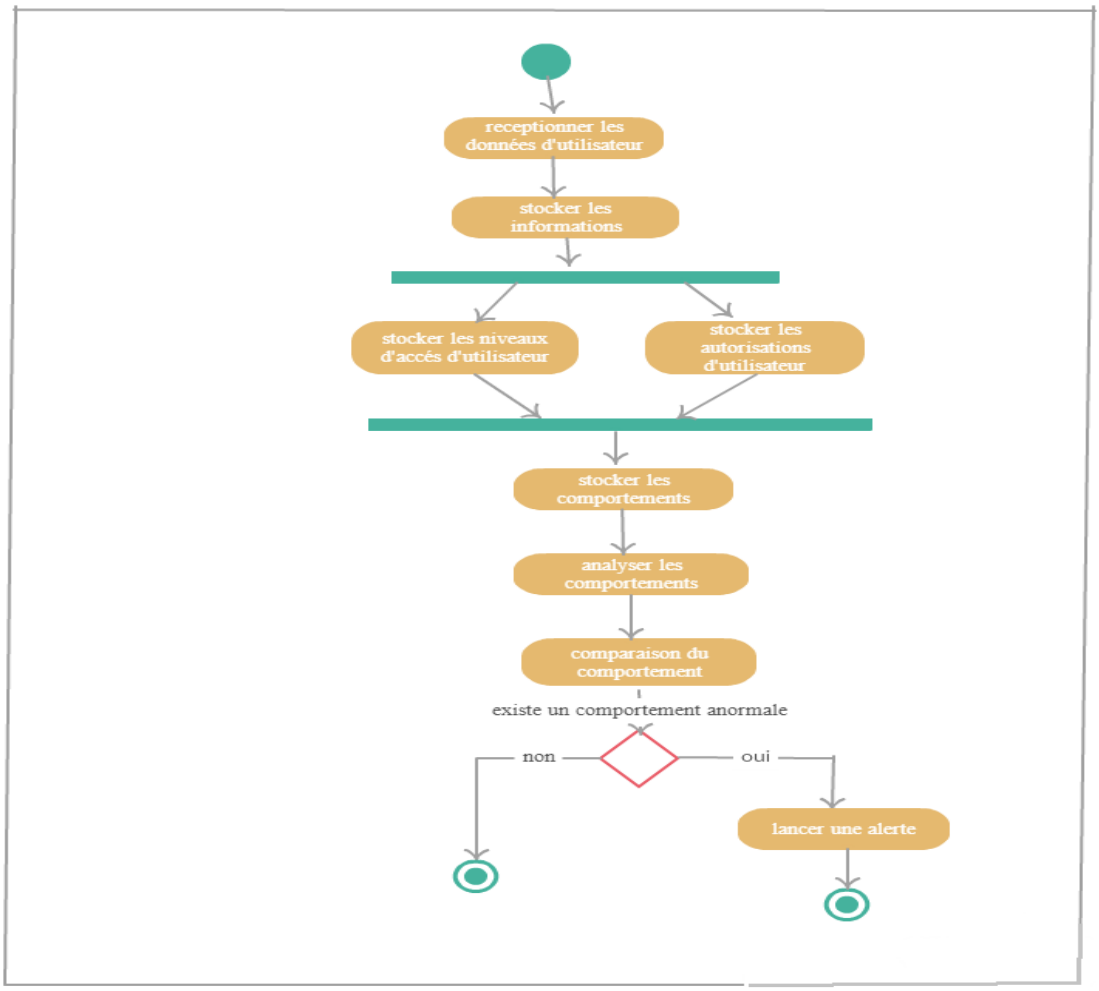

Figure III.11: diagramme d'activité du composant auditeur

II.3.4.Scénario temporelle d'exécution globale :

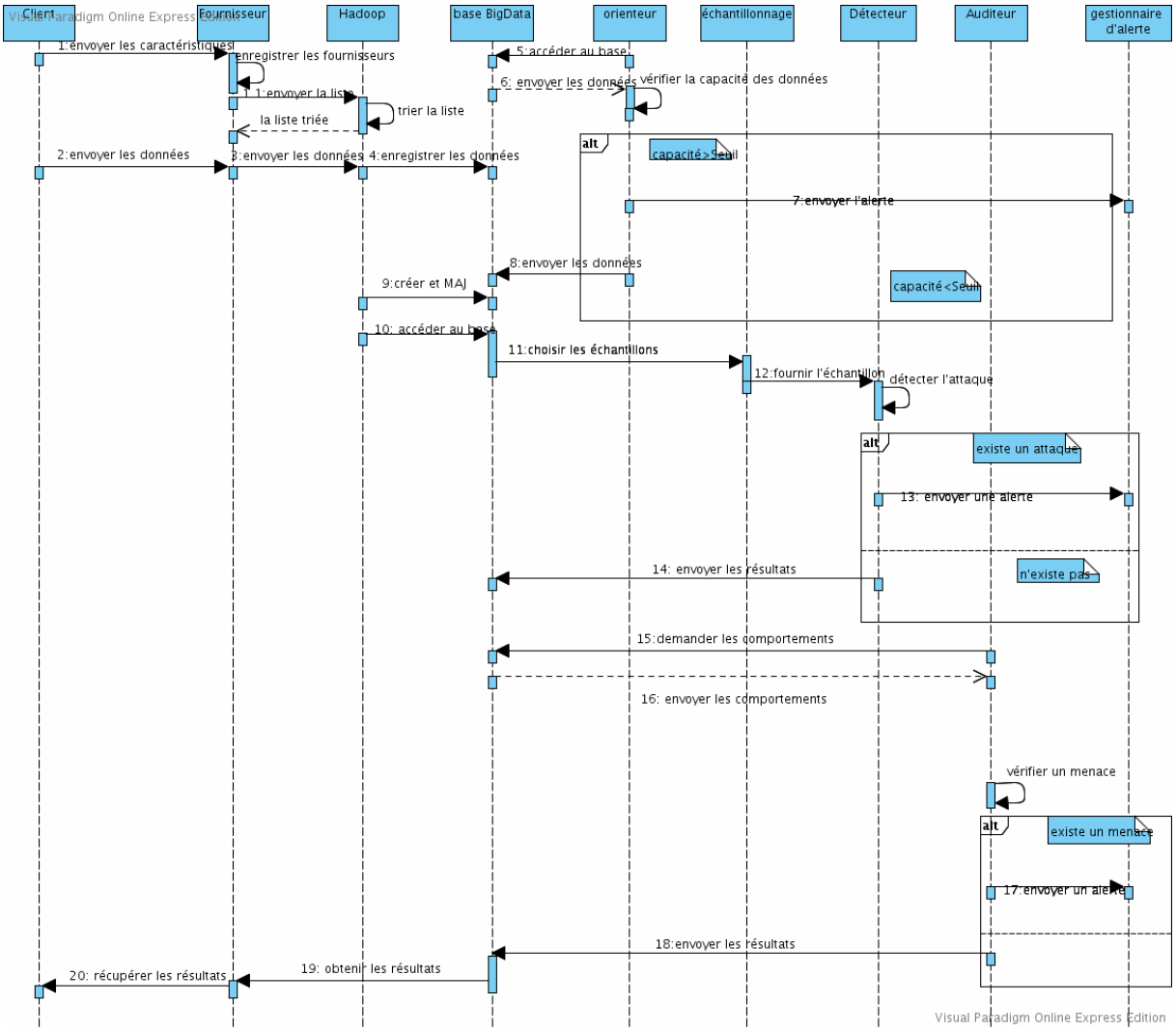

Figure III.12: diagramme de séquence du système

III.3.5.Communication entre les composants du système :

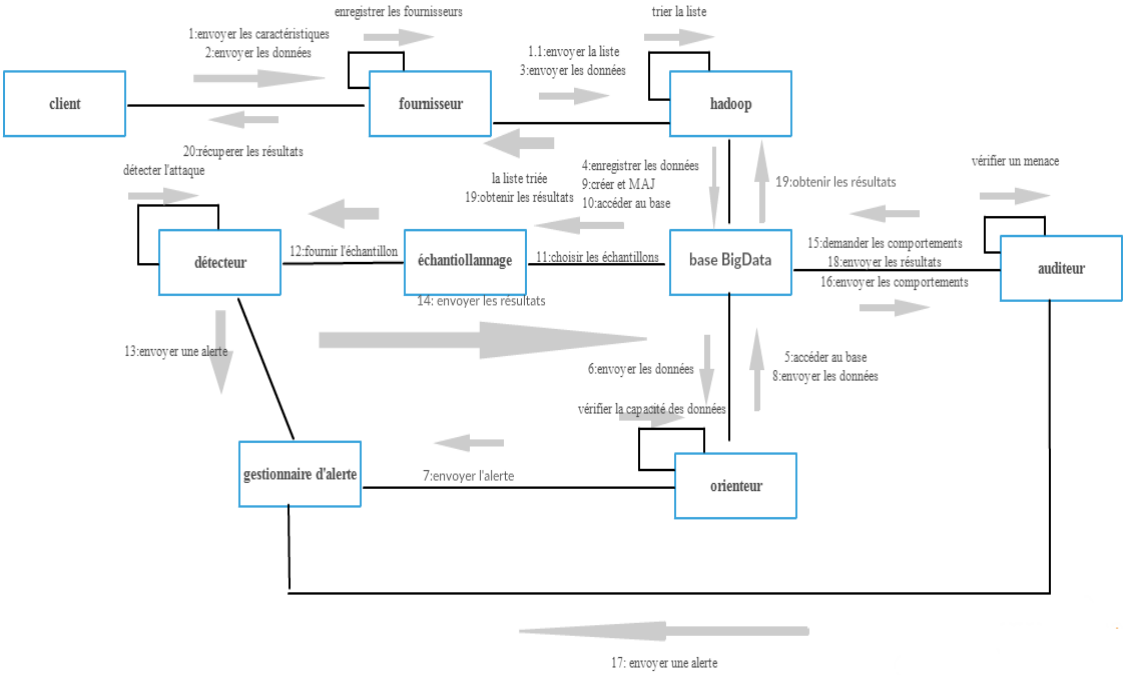

Figure III.13: diagramme de communication du système

III.4.Conclusion :

Dans ce chapitre, nous présentons le système de la détection d'intrusion dans BigData par l'assurance de détecter les intrusions

Cette étude qui contient les détails conceptuels de l'architecture générale de notre système.

Le prochain chapitre présente l'implémentation de notre système avec les techniques nécessaires pour l'application de notre système.

Chapitre IV: Implémentation et Réalisation

IV.1. Introduction

Après avoir présenté en détails notre système dans le chapitre précédent, ce chapitre sera consacré à la phase d'implémentation. Nous aborderons l'aspect pratique de notre application, il s'agit ici d'expliquer l'environnement matériel sur lequel notre système a été développé, les langages de programmation et les outils/technologies utilisés. Pour terminer, nous allons présenter les interfaces graphiques en décrivant les différentes fonctionnalités de notre application et nous présenterons aussi un exemple réel sur des données de la base Big Data qui représente les données des paquets du trafic

IV.2. Outils d'Implémentation

Avant de commencer l'implémentation de notre application, nous allons tout d'abord spécifier les langages de programmation et les outils utilisés qui nous ont semblé être un bon choix vu les avantages qu'ils offrent.

IV.2.1. MySQL & phpMyAdmin :

- **MySQL** : MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il fait partie des logiciels de gestion de base de données les plus utilisés au monde. MySQL fait référence au Structured Query Language, le langage de requête utilisé.
- **phpMyAdmin** : PhpMyAdmin est une interface d'administration pour le SGBD MySQL. Il écrit en langage PHP et s'appuie sur le serveur HTTP Apache.

IV.2.2. Eclipse Java

Eclipse est une plate-forme de développement Java gratuite, connue pour ses plug-ins, qui permettent aux développeurs de développer et de tester du code écrit dans d'autres langages de programmation. Un produit basé sur Eclipse est structuré comme une collection de plug-ins. Chaque plug-in contient le code qui fournit certaines des fonctionnalités du produit.

Le code et les autres fichiers d'un plug-in sont installés sur l'ordinateur local et sont activés automatiquement, le cas échéant. Les plug-ins d'un produit sont regroupés dans des fonctionnalités. Une fonctionnalité est une unité de fonctionnalité téléchargeable et installable séparément.

. La nature fondamentalement modulaire de la plate-forme Eclipse facilite l'installation de fonctionnalités et de plug-ins supplémentaires dans un produit basé sur Eclipse, ainsi que la mise à jour des fonctionnalités et des plug-ins existants du produit.

. Vous pouvez le faire en utilisant le support d'installation et de mise à jour de la plate-forme Eclipse disponible dans le menu Aide. Eclipse vous permet de découvrir, de télécharger et d'installer des fonctionnalités et des plug-ins à partir de sites Web spécialisés du logiciel Eclipse.

IV.2.3. MongoDB

MongoDB est un système de gestion de base de données open source (SGBD) qui utilise un modèle de base de données orienté document qui prend en charge diverses formes de données.

Il est destiné à être utilisées dans des applications Big Data et d'autres tâches de traitement impliquant des données mal adaptées à un modèle relationnel rigide.

Au lieu d'utiliser des tables et des lignes comme dans les bases de données relationnelles, l'architecture MongoDB est composée de collections et de documents

IV.3. Diagramme de classe de l'application

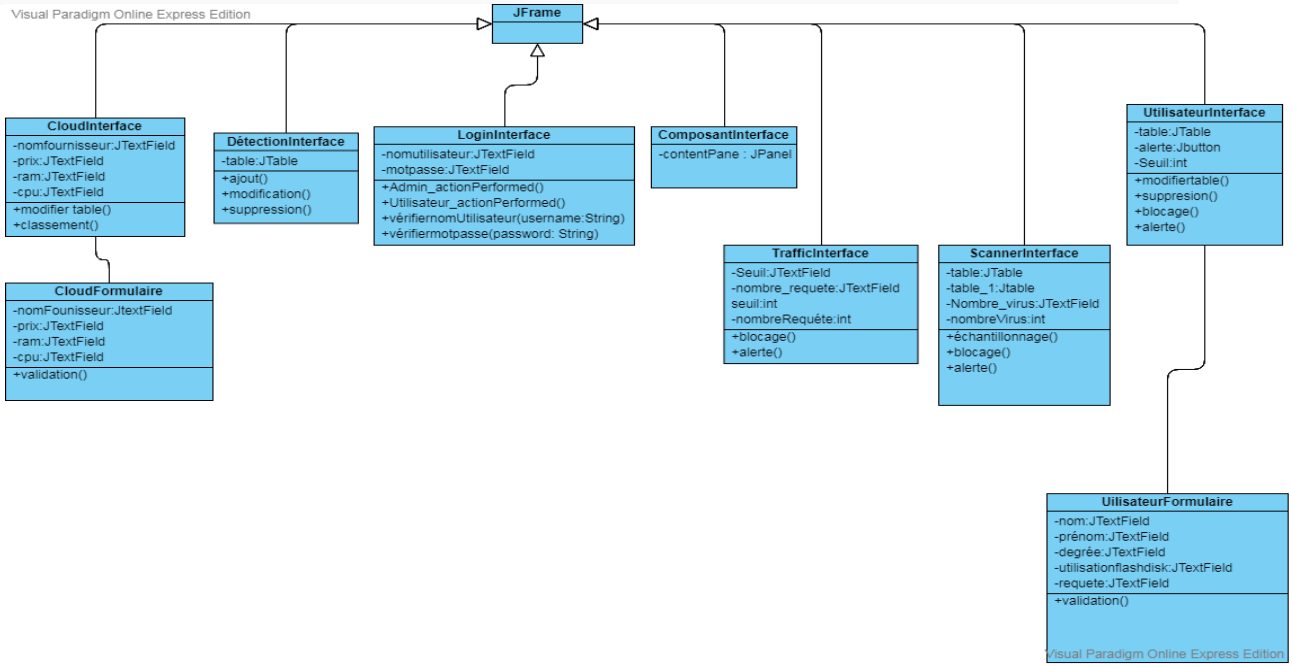

Figure IV.1 : Diagramme de classe du système

IV.4. Base de données du système proposé

IV.4.1. Schéma générale de la base de données

login(id:entier , nomutilisateur:chaine , motdepasse: chaine)

servicecloud(id:entier , prix:entier , RAM:entier , CPU:entier)

utilisateur(id:entier , nom_utilisateur :chaine , prénom_utilisateur :chaine, degrée :chaine, utilisation_flashdisk :entier, requete_non_autorisée :entier statut :chaine)

IV.4.2. Principaux table de la base de données

IV.4.2.1. table login : cette table représente les listes des utilisateurs qui utilisant l'application, .elle est représentée dans la figure suivante :

id	nomutilisateur	motdepasse
1	meriem	26101995

Figure IV.2 : figure de la table login

IV.4.2.2. table servicecloud : cette table représente la liste des fournisseurs Cloud, elle est représenter dans la figure suivante :

	id	nom	prix	RAM	CPU
ıer	1	Amazon	3	100	100
ıer	2	Google	3	10	10
ıer	3	IBM	11	50	13
ıer	5	Microsoft	3	10	10

Figure IV.3 : figure de la table service cloud

IV.4.2.3. table utilisateur : cette représente la liste les comportements des utilisateurs normaux et anormaux, elle est représentée dans la figure suivante :

id	nom_utilisateur	prénom_utilisateur	degrée	utilisation_flashdisk	requete_non_autorisée	statut
1	firane	meriem	A	1	1	normal
2	leksouri	maria	B	1	0	normal

Figure IV.4 : figure de la table utilisateur

IV.4.3. Base de Big Data

La base Big Data représente les informations sur les paquets de trafic réseau qui sont des données de test pour notre application, elle est stockée dans Mongodb sous des Documents ,ou un ensemble des documents forment un collection , aussi un ensemble de collections forment un base de données dans Mongodb. Elle est représentée dans la figure suivante :

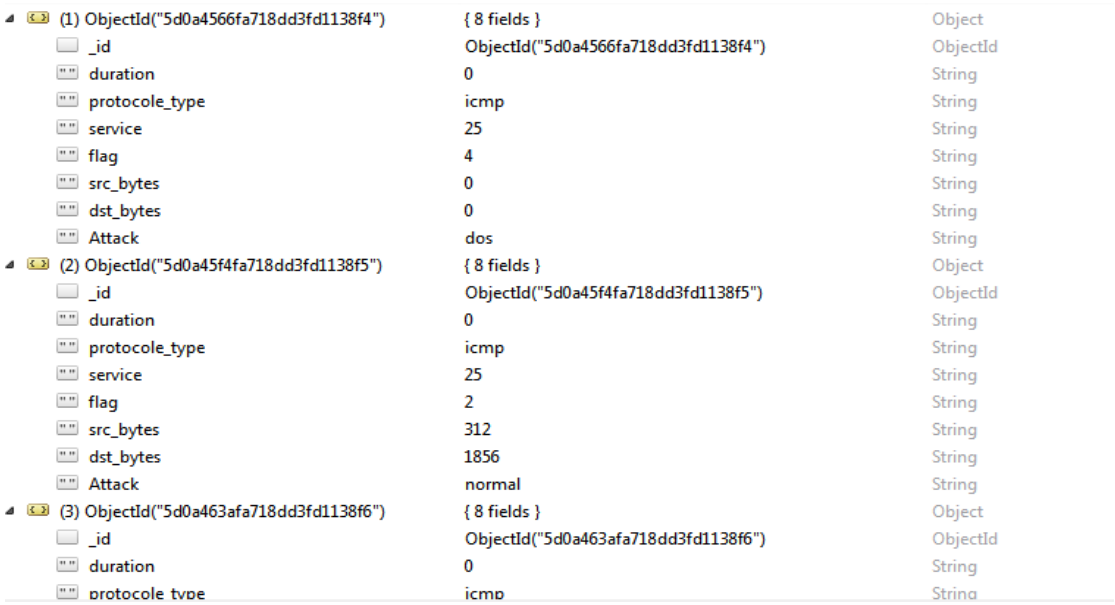

Figure IV.5: figure du Base Big Data

IV.5. Les principaux algorithmes du système proposé :

IV.5.1. Pseudo code du composant Détection

Algorithm IV.1. :MAJ Détection .

Entrées: mongoClient:MongoClient, db : DB , packetInforamtion :DBCollection,port : interger;

Sorties: table : JTable ;

```
        Port=27017 ;
 //connexion to mongodb server
         mongoClient<- new MongoClient(localhost,port) ;
//accès à la base de données
       db<- mongoClient.get(«nom base de données ») ;
//accées à la collection
            DBCollection coll<- db.getCollection(« nom de collection ») ;
// ajout un documet
   Insert(« nom de document a inséré »)
//supprimer un document
  Remove(« nom de document a supprimer »)
//modifier d'un document
Update(« nom de document a modifier »)
// afficher les documents
 //pour trouver les documents de la collection
  Cursor<-coll.Find()
//afficher chaque document dans un ligne de model
Tantque(il existe des documents)
    //Ajouter le document dans un ligne de model
```

```
    AddRows(ligne a ajouter)
 FinTantque
// afficher dans un table
  table.SetModel(model)
```

IV.5.2. Pseudo code du FounisseurCloud

Algorithm IV.2. : Cloud.

Entrées: conn : Connection , cloud : CloudFourmulaire, Modification : JButton, Ajoute : JButton, Suppression : JButton ;

Sorties: table : JTable ;

```
 //connexion a la base de données
Conn<-Connexion.getConnection() ;
// ajouter un fournisseur en utilisant un button ajoute
      Ajoute.addActionListener() ;
   // utiliser une formulaire pour ajouter un fournisseur
           Cloud<- new CloudFormulaire() ;
   // cette formulaire permet l'ajout d'un fournisseur avec interaction avec les requêtes sql
  //modifier un fournisseur
   // utilise le bouton modification avec leur méthode addActionListener et actionPerformed
        Modification.addActionListener() ;
//supprimer un fournisseur en utilisant un button suppression
      Suppression.addActionListener() ;
//pour afficher la liste des fournisseurs
   ModifierTable()
// pour classer les fournisseurs
  Classement()
```

IV.5.3. Pseudo code de l'Orienteur traffic réseau

Algorithm IV.3. : TrafficRéseau.

Entrées: Blocage :JButton, Alerte : JButton , Seuil : entier,nombre Requête :entier;
Sorties: message de blocage, message d'alerte ;

```
//fixer le seuil      Seuil<-1000000 ;
// calculer un nombre aléatoire a chaque fois
  Nombre Requéte<-random.nextInt(un nombre max)
  //le fonction de button Blocage
   Si(nombreRequéte=Seuil)alors
      Afficher un message de blocage
   Sinon
      Ne faire rien
   FinSi
//le fonction de button alerte
   Si(nombreRequéte=Seuil)alors
      Afficher un message de blocage
   Sinon
      Ne faire rien
   FinSi
```

IV.5.4. Pseudo code de Scanner

Algorithm IV.4. : Scanner.

Entrées: Blocage :JButton, Alerte : JButton ,échantillonnage : JButton ,mongoClient:MongoClient, db : DB , packetInforamtion :DBCollection,port : entier, nombre_virus :entier, limit :entier,count : entier,skip : entier,offset :entier;
Sorties: message de blocage, message d'alerte ;

```
Port=27017 ;
 //connexion to mongodb server
        mongoClient<- new MongoClient(localhost,port) ;
//accès à la base de données
      db<- mongoClient.get(«nom base de données ») ;
//accées à la collection
          DBCollection coll<- db.getCollection(« nom de collection ») ;
// fonction  de button échantillonnage
   //fixer le nombre seuil    limit<-5 ;
//compter les nombre le nombre de documents dans la collection
     Count<-coll.Count() ;
Si(count<=limit)alors
    //appel de la fonction find
      Cursor<-coll.Find() ;
   //afficher les documents dans table
      Tantque(il existe des documents)
        // ajouter document dans un liste
```

```
    Add(document) ;
        Pour(chaque élément du liste)
            //ajouter l'élément dans un ligne d'un default table model
              addRow (nouvel objet représente l'élément) ;
         FinPour
       FinTantque
Else
    Compter skip par une formule mathématique

 //appel de la fonction find
      Cusor<-coll.Find() ;
   Tantque(la longueur de la liste < limit)
      Compter offset par une formule mathématique
       Choisir le document avec l'utilisation du offset
       cursor.skip(offset).next();
       Ajouter le documnent choisi au liste
         Add(document) ;
       Tantque(il existe un document dans la liste)
          Ajouter le document comme un ligne dans la table
            addRow(document) ;
         FinTantque
          Afficher les resultat dans table
```

FinTantque

FinSi

La fonction de button blocage

Si(nombre_virus>0)alors

Afficher un messsage de blocage

Sinon

Ne fait rien

FinSi

La Fonction du button Alerte

Si(nombre_virus>0)alors

Afficher un messsage d'alerte

Sinon

Ne fait rien

FinSi

IV.5.5. Pseudo code de gestion utilisateur

Entrées: Blocage : JButton, Alerte : JButton, degrée : chainede caractére,seuil :entier ;

Sorties: message de blocage, message d'alerte ;

La fonction du button blocage

Si (il existe un utilisateur avec statut='suspect')alors

Afficher un message de blocage

Sinon

Ne fait rien

FinSi

La fonction du button alerte

Si (il existe un utilisateur avec statut=’suspect’)alors

Afficher un message d’alerte

Sinon

Ne fait rien

FinSi

Si(degrée=’A’)alors

Il avait l’autorisation

FinSi

Si((degrée=’B’)and (flashsisk_utilisation>=seuil)and(requete_non_autorisée>0))alors

Afficher un message d’alerte

FinSi

Si((degrée=’C’)and(falsh_utilisation>=seuil))alors

Afficher un message d’alerte

FinSi

IV.6. Manuel d'utilisation de l'application

IV.6.1. Interface Login :

Cette interface est responsable de l'authentification des utilisateurs. Elle est représentée dans la figure suivante :

Figure IV.6: figure du l'interface de connexion

Dans l'interface, il existe deux buttons :

Le bouton choix : utilisé pour choisir que soit un administrateur ou un utilisateur.

Le bouton connexion : utilisé pour l'authentification et passé par lui a la deuxième interface.

IV.6.2. Interface Composants du Système :

Cette présente les différents composants de notre système, elle est présentée dans la figure suivante :

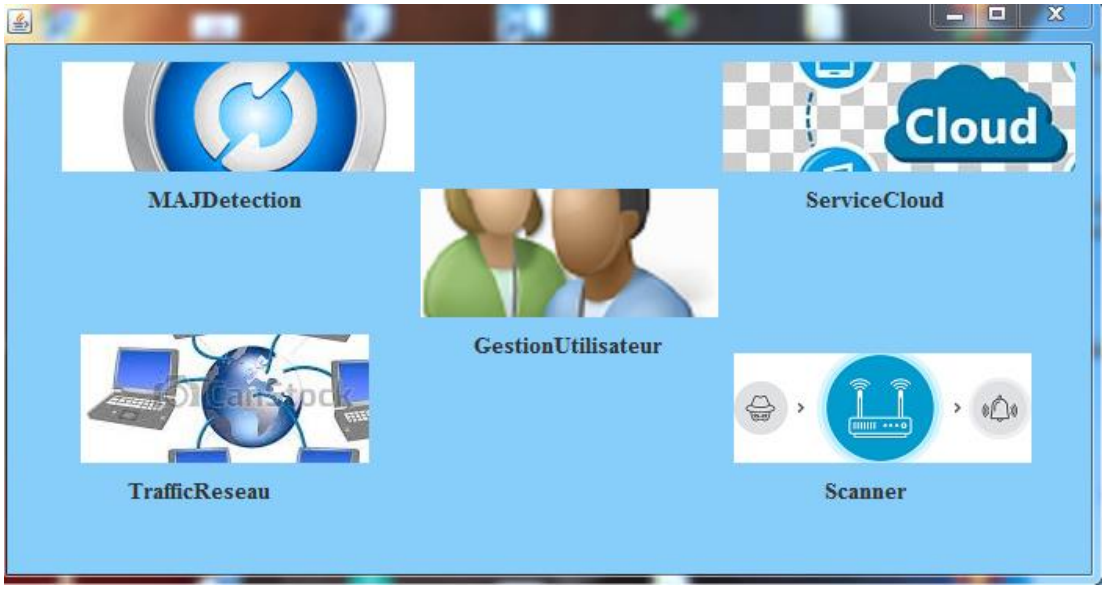

Figure IV.7: figure du l'interface des composants

L'interface contient 5 composants :

1-MAJDetection : qui présente un lien vers l'interface de détection.

2-Servicecloud: qui présente un lien vers l'interface Cloud.

3-TrafficRéseau : qui présente un lien vers l'interface traffic.

2-Scanner : qui présente un lien vers l'interface Scanner.

2-Gestion Utilisateur: qui présente un lien vers l'interface utilisateur.

IV.6.3. Interface Détection:

Cette interface fonctionne sur la manipulation de la base des signatures d'attaques stockées dans Mongodb, elle est présentée dans la figure suivante :

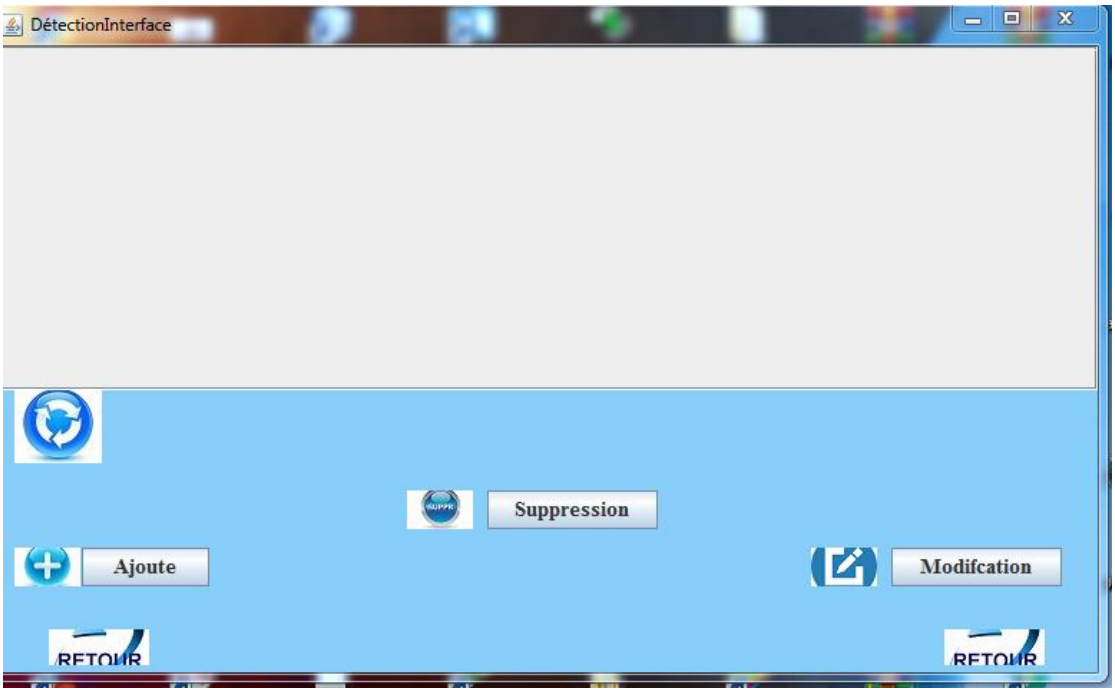

Figure IV.8: figure du l'interface de détection

Cette interface contient 3 boutons sont :

1-bouton d'ajout : fonctionne sur l'ajout une nouvelle signature d'attaque.

2-bouton de suppression : fonctionne sur la suppression d'une signature d'attaque.

4- bouton de modification : fonctionne sur la modification d'une signature d'attaque.

IV.6.4. Interface Cloud:

Cette interface fonctionne sur une base de données des fournisseurs Cloud, elle utilise une autre interface pour appliquer les opérations ajout, modification, suppression, elle est présentée dans la figure suivante :

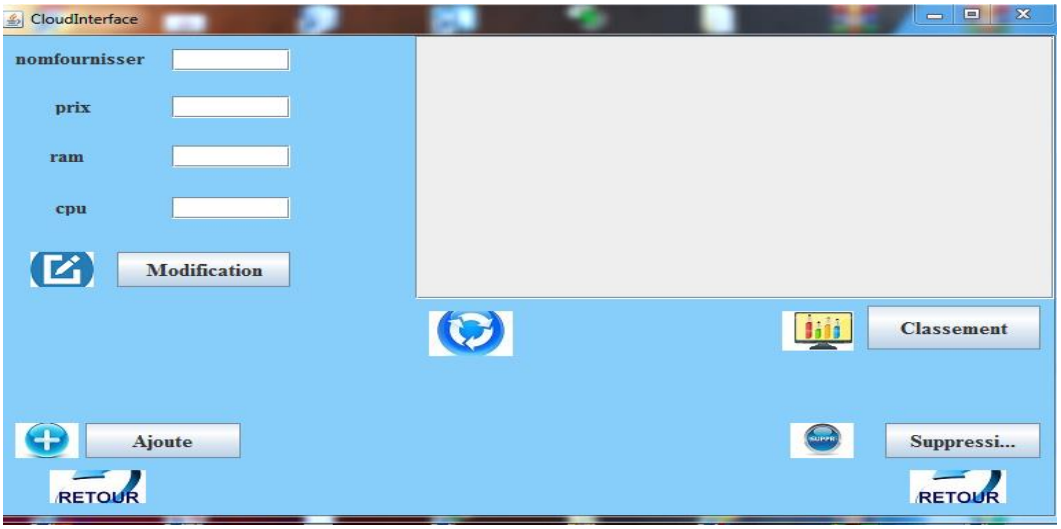

Figure IV.9: figure du l'interface Cloud

Cette interface compose de 4 boutons :

1- bouton de classement : pour classer les fournisseurs par rapport le prix puis la processeur puis la mémoire.
2- bouton ajout : fonctionne sur l'ajout d'un nouveau fournisseur.
3- bouton modification : fonctionne sur la modification d'un fournisseur.
4- bouton suppression : fonctionne sur la suppression d'un fournisseur

Les opérations d'ajout et suppression nécessite une autre formulaire pou applique ces derniers, elle est présentée dans la figure suivante :

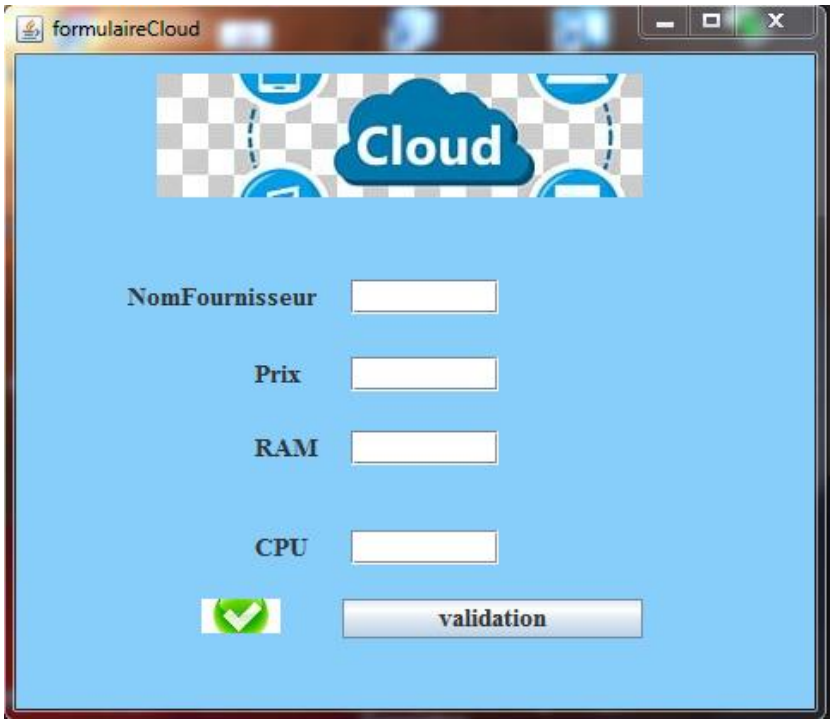

Figure IV.10: figure du le formulaire Cloud

IV.6.5. Interface de Trafic Réseau:

Cette Interface fonctionne sur la diminution de l'impact des attaques DDOS et protéger le système d'être bloqué, elle représentée dans la figure suivante :

Figure IV.11: figure du l'interface traffic réseau

Elle compose de deux boutons :

1- bouton stop : fonctionne sur la comparaison entre la valeur de nombre de requête par seconde et la valeur de seuil ou il lance un message de blocage lorsque les des valeurs sont égaux.
2- bouton alerte : fonctionne sur la comparaison entre la valeur de nombre de requête par seconde et la valeur de seuil ou il lance un message d'alerte lorsque les des valeurs sont égaux.

IV.6.6. Interface de Scanner:

Cette interface fonctionne sur deux concepts le premier est l'échantillonnage de la base Big Data, la deuxième est la comparaison entre la base Big Data et la base des signatures d'attaques pour être capable de détecter les données contiennent des intrusions, elle est présentée dans la figure suivante :

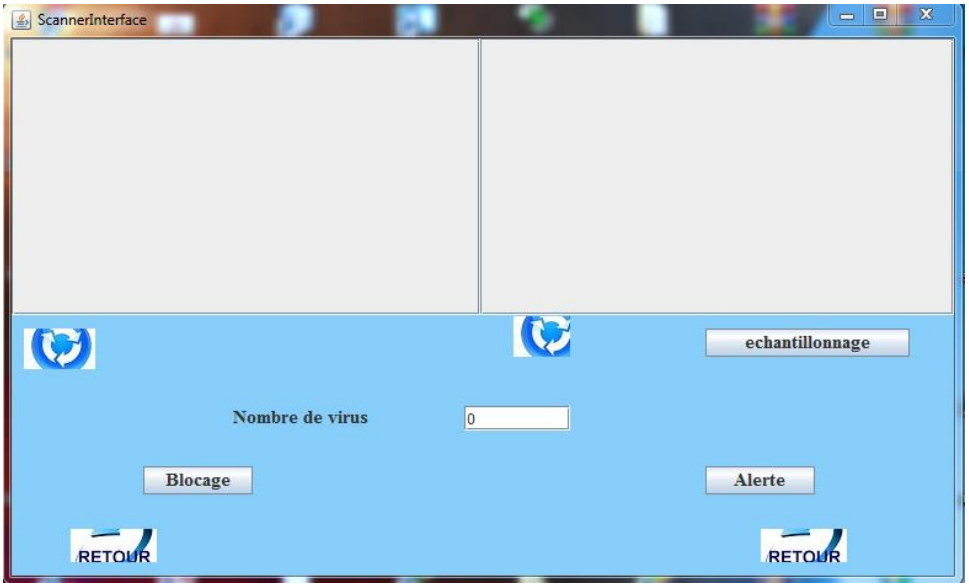

Figure IV.12: figure de l'interface Scanner

Cette interface contient 3 boutons :

1- bouton d'échantillonnage : fonctionne sur la sélection aléatoire d'ensemble de la base Big Data.
2- bouton blocage : fonctionne sur la valeur de nombre de virus, lorsque cette valeur être différent de 0, il affiche un message de blocage.
3- bouton alerte : fonctionne sur la valeur de nombre de virus, lorsque cette valeur être différent de 0, il affiche un message d'alerte.

IV.6.7. Interface d'Utilisateur :

Cette interface fonctionne sur le comportement des utilisateurs, cette interface définit les utilisateurs normaux et les utilisateurs suspects. Elle est présentée dans la figure suivante :

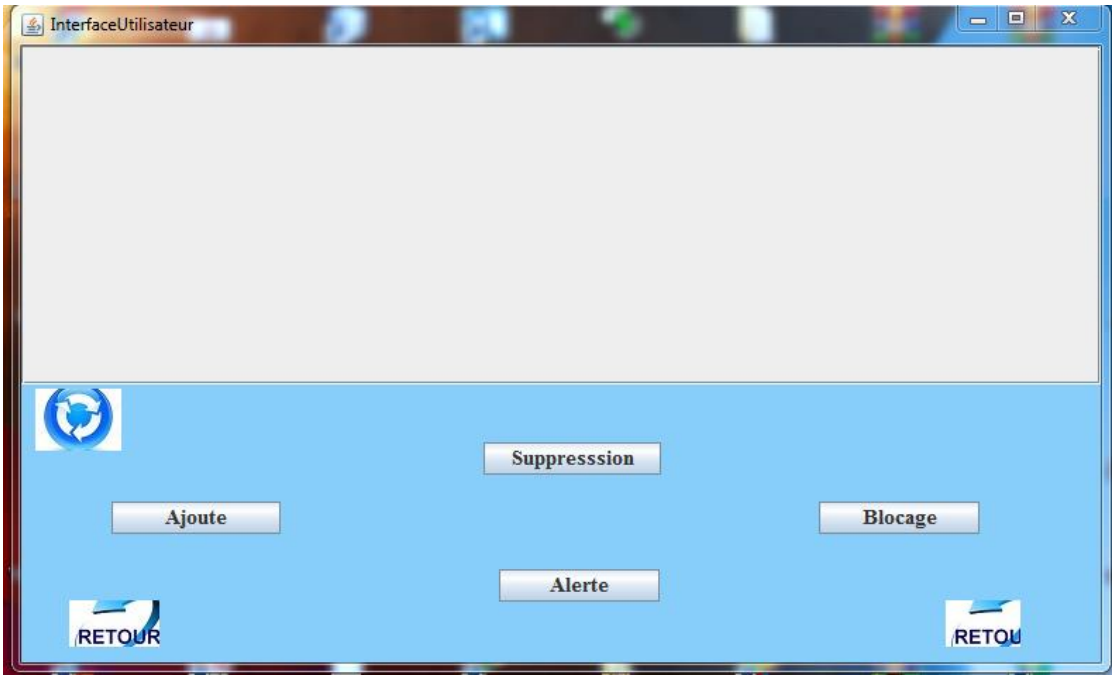

Figure IV.13: figure de l'interface utilisateur

Cette interface compose de 4 boutons :

1-bouton ajout : fonctionne sur l'ajout d'un nouvel utilisateur

2-bouton suppression : fonctionne sur la suppression d'un utilisateur.

3-bouton blocage : fonctionne sur le blocage des utilisateurs suspects.

4-bouton alerte : fonctionne sur l'affichage d'une alerte lorsqu'il existe un utilisateur suspect.

Pour les opérations de l'ajout et la suppression, il existe un formulaire pour appliquer ces opérations, elle est représentée dans la figure suivante :

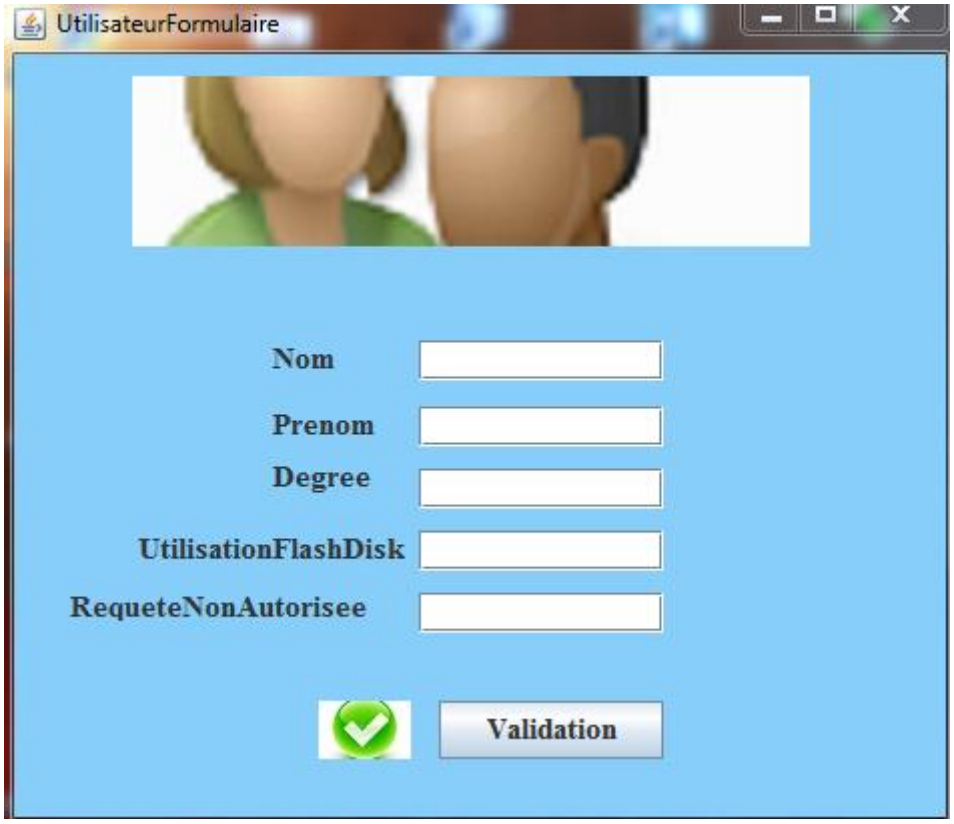

Figure IV.14: figure du formulaire utilisateur

IV.7. Conclusion

Dans ce dernier chapitre nous sommes proposés un nouveau système pour résoudre les problèmes de détection d'intrusions, nous avons montré l'implémentation de notre système, et décrit les outils utilisés pour cette implémentation. Nous avons illustré le diagramme de classe pour notre système, puis le schéma de la base de données et la base Big Data. Nous avons illustré les interface graphique avec un description textuelle, ensuite nous avons fait des tests sur le système proposé.

Chapitre V: Conclusion et Perspectives

V.1.Conclusion :

Grâce à des nouvelles technologies de stockage et surtout d'analyse Big Data permet de collecter, de stocker, et d'analyser toutes ces données à des coûts raisonnables.

Les Big Data devraient désormais ouvrir de nouvelles opportunités de revenu pour les entreprises et, en même temps, faciliter la vie de tous les jours.

Mais cette opportunité ne pourra pas être saisie sauf si la protection des données contre les attaques et les intrusions est assuré.

V.2.Perspectives:

Comme perspectives, nous pouvons envisager les points suivants:

- Ajouter un composant pour la gestion d'alerte dans le but d'organiser les alertes
- Ajouter d'autres fonctionnalités et lancer la version commerciale.

Chapitre VI : Bibliographie

Bibliographie

[1] [Evan Stibbs] , "Big Data,Big Innovation", 2014, pp.

[2] [Koichiro. Hayashi], "Social Issues of Big Data and Cloud: Privacy, Confidentiality, and Public Utility", International Conference on Availability, Reliability and Security, 2013.

[3] [Shui yu . Song Guo], "Big Data Concepts, Theories, and Applications", pp 31_50.

[4] [Min Chen · Shiwen Mao] · Yunhao Liu, "Big Data: A Survey", 2014, pp 175_176.

[5] [Lisbeth.R , Aaron. C , Jose. Luis, Jair. C, Jorge luis. G, Giner. A], "A general perspective of Big Data: application, tools, challenges and trends", New York 2015.

[6] wikipedia.

[9] [Saouli.H, Kazar.O, Kassimi.D], "Applications et enjeux des Big Data dans le contexte des défis mondiaux", Laboratoire LINFI.

[Joseph Migga Kizza], "Guide to Computer Network Security", pp 271_292.

[Tarek Gaber · Aboul EllaHassanien ,Nashwa El-Bendary · Nilanjan Dey],

" Advances in Intelligent Systems and Computing'', 2015, pp 369.

[Min Luo • Liang-Jie Zhang], " Cloud Computing –CLOUD 2018", 2018, pp 377

[M. Mazhar Rathore , Awais Ahmad ,Anand Paul], " Real time intrusion detection system for ultra-high-speed big data environments", New York 2016.

[Mostafa Doroudian, Narges Arastouie, Mohammad Talebi, Ali Reza Ghanbarian], "Multilayered Database Intrusion Detection System for Detecting Malicious Behaviors in Big Data Transaction ", 2015.

[Nour Moustafa, Gideon Creech, and Jill Slay], "Big Data Analytics for Intrusion DetectionSystem: Statistical Decision-Making Using Finite Dirichlet Mixture Models ",2017

[Kai Peng], "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data", 2017

[Rachana Sharma & Priyanka Sharma, Preeti Mishra & Emmanuel S. Pilli], "Towards MapReduce Based Classification approaches for Intrusion Detection", 2016