|  |  |  |
| --- | --- | --- |
|  |  | Conception et Réalisation d'un systÈme DE MONITORING ET de détection dES anomalieS EN TEMPS RÉELS.  Sujet PFE |

Mars 2022

# PRISE DE NOTE DES RECHERCHES

To get started right away, just tap any placeholder text (such as this) and start typing.

Diagram

Description automatically generated

## INTRODUCTION

*Que sont les anomalies?*

Les anomalies sont des cas où les mesures de performance s'écartent du comportement normal et attendu. Cela peut se produire de plusieurs façons. Cependant, nous allons nous concentrer sur l'identification de ces anomalies à l'aide de seuils.

*Définition de l’AIOps*

L'intelligence artificielle pour les opérations informatiques n'est rien d'autre que la surveillance et l'analyse de plus grands volumes de données générés par les plateformes informatiques en utilisant l'intelligence artificielle et l'apprentissage automatique. Celles-ci aident les entreprises dans la corrélation des événements et l'analyse des causes profondes pour permettre une résolution plus rapide.

Autre définition: L’intelligence artificielle pour les opérations **AIOps** concerne l’application de modèles d’aprentissage automatique, de l’intelligence artificielle et d’analyses avancées aux données opérationnelles informatiques.

L’objectif est de donner aux professionnels de l’informatique et des opérations les données don’t ils ont besion pour prendre des decisions en fin de compte, résoudre et restaurer le service d’une application plus rapidement.

*Comment sont-elles signalées?*

Avec les systèmes de surveillance actuels, les anomalies sont signalées sur la base de seuils statiques. Il s'agit de valeurs constantes qui fournissent les limites supérieures d'un comportement normal. Par exemple, l'utilisation du CPU est considérée comme anormale lorsque la valeur est supérieure à 85 %. Lorsque des anomalies sont détectées, des alertes sont envoyées à l'équipe d'exploitation pour qu'elle les examine.

*Pourquoi est-ce important?*

La surveillance de la santé des serveurs est nécessaire pour assurer une allocation efficace des ressources. Des pics ou des baisses de performance inattendus, comme l'utilisation du CPU, peuvent être le signe d'une contrainte sur les ressources. Ces problèmes doivent être traités par l'équipe d'exploitation à temps, faute de quoi les applications associées aux serveurs risquent de tomber en panne.

*Alors, que sont les seuils, comment sont-ils significatifs?*

Les seuils sont les limites des performances acceptables. Toute valeur qui dépasse le seuil est indiquée sous forme d'alerte et fait donc l'objet d'une résolution prudente au plus tôt. Il convient de noter que les seuils ne sont définis qu'au niveau de l'outil, de sorte qu'en cas de dépassement, une alerte est générée. Ces seuils, s'ils sont manuels, peuvent être ajustés en fonction de la demande.

Il existe 2 types de seuils:

1. Les seuils de surveillance statiques: Ces seuils sont des valeurs fixes indiquant les limites des performances acceptables.
2. Les seuils de surveillance dynamiques: Ces seuils sont dynamiques par nature. C'est ce que fait un outil de surveillance informatique intelligent. Il apprend la plage normale d'un seuil haut et d'un seuil bas, à chaque moment de la journée, de la semaine, du mois, etc. Par exemple, un système dynamique saura qu'une utilisation élevée du CPU est normale pendant la sauvegarde, et qu'il en va de même pour les utilisations survenant les autres jours.

*Le mode d'identification des alertes par le seuil ne présente-t-il pas d’inconvénients?*

Ce n'est absolument pas le cas. Comme la plupart des choses dans la vie, elle comporte sa part de problèmes. En revenant de la philosophie à notre article, il y a des inconvénients dans la façon de faire du seuil statique, bien que ceux d'un seuil dynamique soient minimes. Nous devons également comprendre qu'avec la connaissance appropriée du domaine, il existe de nombreuses façons de surmonter ces inconvénients.

Considérez le scénario suivant: Imaginons un seuil de CPU fixé à 85 %. Nous savons que tout ce qui dépasse ce seuil est une anomalie générée sous forme d'alerte. Maintenant, considérez le même pourcentage de seuil comme un comportement normal dans une machine virtuelle (VM). Cette fois, l'outil de surveillance générera des alertes en continu jusqu'à ce qu'il atteigne une valeur inférieure au seuil. Si l'on ne s'en occupe pas, ce sera la pagaille, car il pourrait y avoir beaucoup de fausses alertes qui, à leur tour, pourraient empêcher l'équipe d'identifier le problème réel. Il y aura une chaîne de faux positifs. Cela peut perturber l'ensemble de la plateforme informatique et entraîner une charge de travail inutile pour l'équipe. Une fois qu'une plateforme informatique est en panne, cela entraîne des temps d'arrêt et des pertes pour nos clients.

Comme nous l'avons mentionné, il existe des moyens de surmonter ce problème grâce à la connaissance du domaine. Chaque organisation a ses propres secrets commerciaux pour éviter que cela ne se produise. Avec les bonnes connaissances, ce comportement peut être modifié et résolu rapidement.

*Que devons-nous faire maintenant? Les anomalies doivent-elles être résolues?*

#### Bien sûr, les anomalies doivent être résolues au plus tôt pour éviter que la plateforme ne soit mise en péril. Il existe de nombreuses méthodes et techniques d'apprentissage automatique pour y remédier. Avant d'entrer dans le vif du sujet, nous savons qu'il existe deux grandes techniques d'apprentissage automatique : l'apprentissage supervisé et l'apprentissage non supervisé. Il existe de nombreux articles sur Internet que l'on peut consulter pour se faire une idée de ces techniques. De même, il existe une variété de facteurs qui pourraient être classés dans ces catégories. Cependant, dans cet article, nous allons aborder une technique d'apprentissage non supervisé – **Isolation Forest** entre autres.

Elasticsearch

Comment sont choisies les données ? (Events, Metrics, Logs, Alerts)  
Dans la vraie vie cela dépend :

* L’architecture de notre application
* Le type de source de données applications à votre application
* Les exigences règlementaires en matière de données

Comment cela s’intègre dans notre modèle ?  
Il faudra réfléchir à trois étapes clés :

1. Surveiller et découvrir (Monitoring & Discovering)

A ce stade, les données sont ingérées par la plateforme AIOps et sont seuillées et créent des lignes de base pour votre application spécifique.

Ainsi nous allons réfléchir de la sorte :

* Qu’est-ce qui est normal pour mon application ?
* Quel est le taux d’ingestion de journaux ?
* Combien d’erreurs est acceptable en fonction de notre objectif de ralentissement ou de niveau de service ?

1. Engage Context (S’engager dans le contexte)

C’est donc là que l’AIOps brille vraiment : Il prend toutes les données ingérées et les présentent à un professionnel des opérations informatiques ou un ingénieur en fiabilité du site sous la forme d’une solution de la collaboration, également connue sous le nom d’opérations chat (chatOps)

Donc ici jusqu’à présent tout à été fait en background et dès que cet incident apparait dans notre application, il est apparu via des opérations de chat à notre ingénieur en fiabilité du site.

Moins c’est plus ici, ils ont le contexte sur l’emplacement de l’incident dans l’applications spécifiques recommandées pour résoudre ce problème et plus important encore comment ces actions sont-elles basées sur des incidents comme celui-ci qui se sont produits dans le passé.

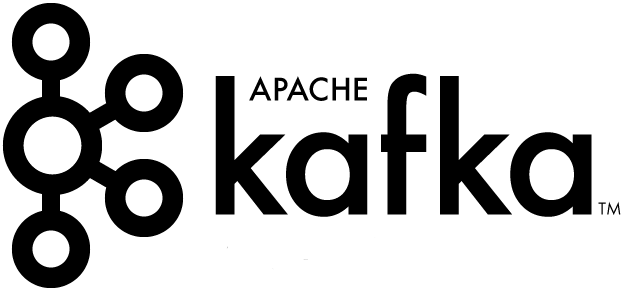
1. Agir et Automatiser (Act and Automate)

Les options suggérées disponibles via les opérations de chat permettent aux ITOps de sélectionner ce qui a fonctionné dans le passé et d’un **clic** d’activer un script ou un manuel d’exécution pour résoudre le problème dès qu’il est détecté.

Cela permet à notre application de se remettre en marche plus rapidement et de garantir que nos partenaires sont plus satisfaits de cette expérience.

*En résumé*, ce système permet à ses professionnels de résoudre les problèmes plus rapidement, de maintenir les applications opérationnelles et de protéger l’entreprise à long terme.

## APACHE KAFKA



Apache Kafka est une platefome de diffusion continue open source et distribué qui permet le développement d’applications événementielles en temps reel. Plus précisément il permet aux developpeurs de créer des applications qui **produisent** et **consomment** en permanence des flux d’enregistrements de données. Maintenant Kafka est *distribué*. Il fonctionne comme un cluster pouvant couvrir plusieurs serveurs ou même plusieurs centres de données. Les enregistrements produits sont répliqués et partitionnés de manière à permettre un grand nombre d’utilisateurs d’utiliser l’application simultanément sans aucun retard perceptible des performances. Donc avec ça, Apache Kafka est super **rapide**. Il maintient également un niveau de **précision très élevé** avec les enregistrements de données et Apache Kafka mainteint l’**ordre de leur occurrence**. Apache Kafka est également **résilient** et **tolérant** aux pannes. Ainsi ces caractériques réunies constituent une plateforme extrênement puissante………………………………………………………………………………...

*Diagram

Description automatically generated*

## ELASTICSEARCH

Elasticsearch fournit des recherches et des analyses en temps quasi réel pour tous les types de données. Qu'il s'agisse de texte structuré ou non, de données numériques ou de données géospatiales, Elasticsearch peut les stocker et les indexer efficacement de manière à permettre des recherches rapides. Vous pouvez aller bien au-delà de la simple extraction de données et agréger des informations pour découvrir des tendances et des modèles dans vos données. Et à mesure que le volume de vos données et de vos requêtes augmente, la nature distribuée d'Elasticsearch permet à votre déploiement de se développer en toute transparence.

Elasticsearch est le moteur de recherche et d'analyse distribué au cœur d'Elastic Stack. Logstash et Beats facilitent la collecte, l'agrégation et l'enrichissement de vos données et leur stockage dans Elasticsearch. Kibana vous permet d'explorer, de visualiser et de partager de manière interactive des informations sur vos données, ainsi que de gérer et de surveiller la pile. Elasticsearch est l'endroit où la magie de l'indexation, de la recherche et de l'analyse se produit.

Bien que tous les problèmes ne soient pas des problèmes de recherche, Elasticsearch offre la vitesse et la flexibilité nécessaires pour traiter les données dans une grande variété de cas d'utilisation :

* Ajouter un champ de recherche à une application ou un site web
* Stockez et analysez les journaux, les métriques et les données d'événements de sécurité.
* Utilisez l'apprentissage automatique pour modéliser automatiquement le comportement de vos données en temps réel.
* Automatisez les flux de travail en utilisant Elasticsearch comme moteur de stockage.
* Gérez, intégrez et analysez des informations spatiales en utilisant Elasticsearch comme système d'information géographique (SIG).
* Stockez et traitez des données génétiques en utilisant Elasticsearch comme outil de recherche bioinformatique.

Nous sommes continuellement étonnés par les nouvelles façons dont les gens utilisent la recherche. Mais que votre cas d'utilisation soit similaire à l'un de ceux-ci ou que vous utilisiez Elasticsearch pour vous attaquer à un nouveau problème, la façon dont vous travaillez avec vos données, vos documents et vos index dans Elasticsearch est la même.

## ELASTIC MACHINE LEARNING

Les fonctions d'apprentissage automatique analysent vos données et génèrent des modèles pour leurs modèles de comportement. Le type d'analyse que vous choisissez dépend des questions ou des problèmes que vous souhaitez aborder et du type de données dont vous disposez.

**Apprentissage automatique non supervisée**

Il existe deux types d'analyse qui permettent de déduire les modèles et les relations au sein de vos données sans formation ni intervention : la détection d'anomalies et la détection de valeurs aberrantes.

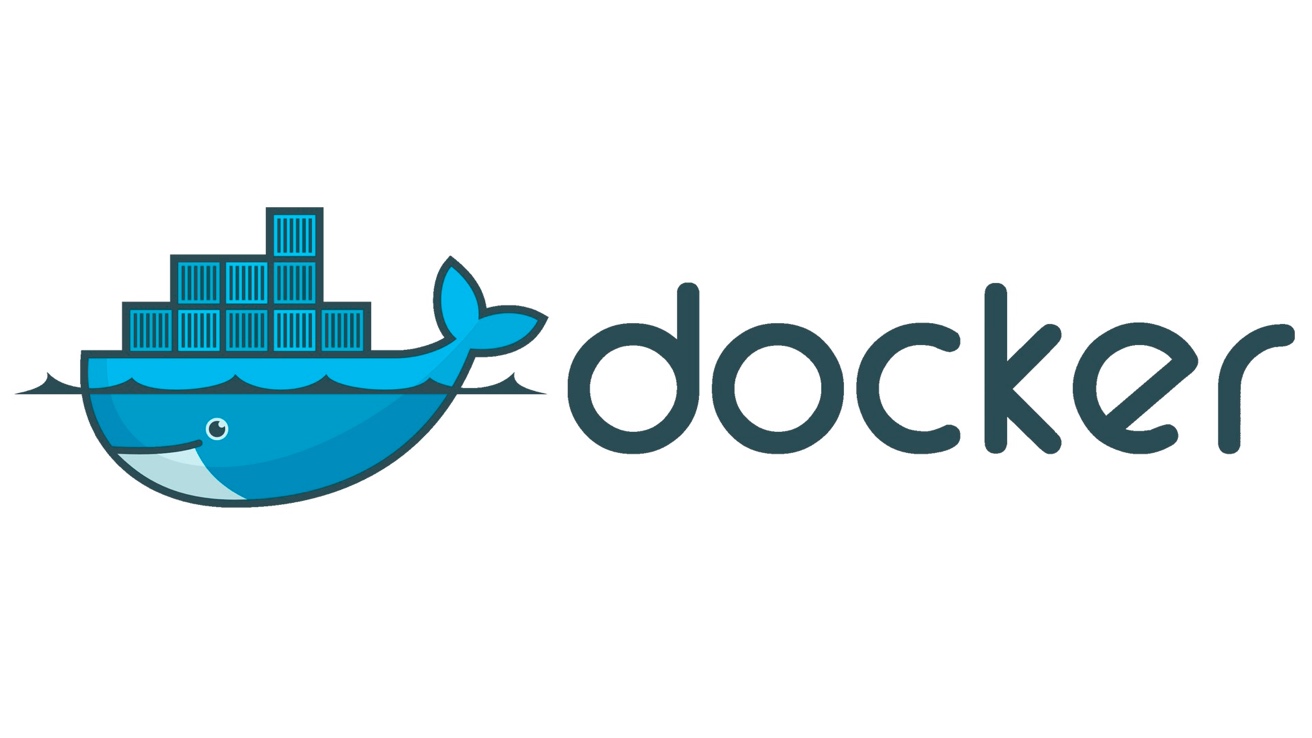
* La détection d'anomalies nécessite des données de séries chronologiques. Elle construit un modèle de probabilité et peut fonctionner en continu pour identifier les événements inhabituels au fur et à mesure qu'ils se produisent. Le modèle évolue dans le temps ; vous pouvez utiliser ses informations pour prévoir les comportements futurs.
* La détection de valeurs aberrantes ne nécessite pas de données de séries chronologiques. Il s'agit d'un type d'analyse de cadre de données qui identifie les points inhabituels dans un ensemble de données en analysant la proximité de chaque point de données par rapport aux autres et la densité du groupe de points qui l'entoure. Il ne fonctionne pas en continu ; il génère une copie de votre ensemble de données où chaque point de données est annoté d'un score aberrant. Ce score indique dans quelle mesure un point de données est aberrant par rapport aux autres points de données.

**Apprentissage automatique supervisée**

Il existe deux types d'analyse du cadre de données qui nécessitent des ensembles de données d'apprentissage : la classification et la régression.

Dans les deux cas, le résultat est une copie de votre ensemble de données où chaque point de données est annoté avec des prédictions et un modèle formé, que vous pouvez déployer pour faire des prédictions pour de nouvelles données.

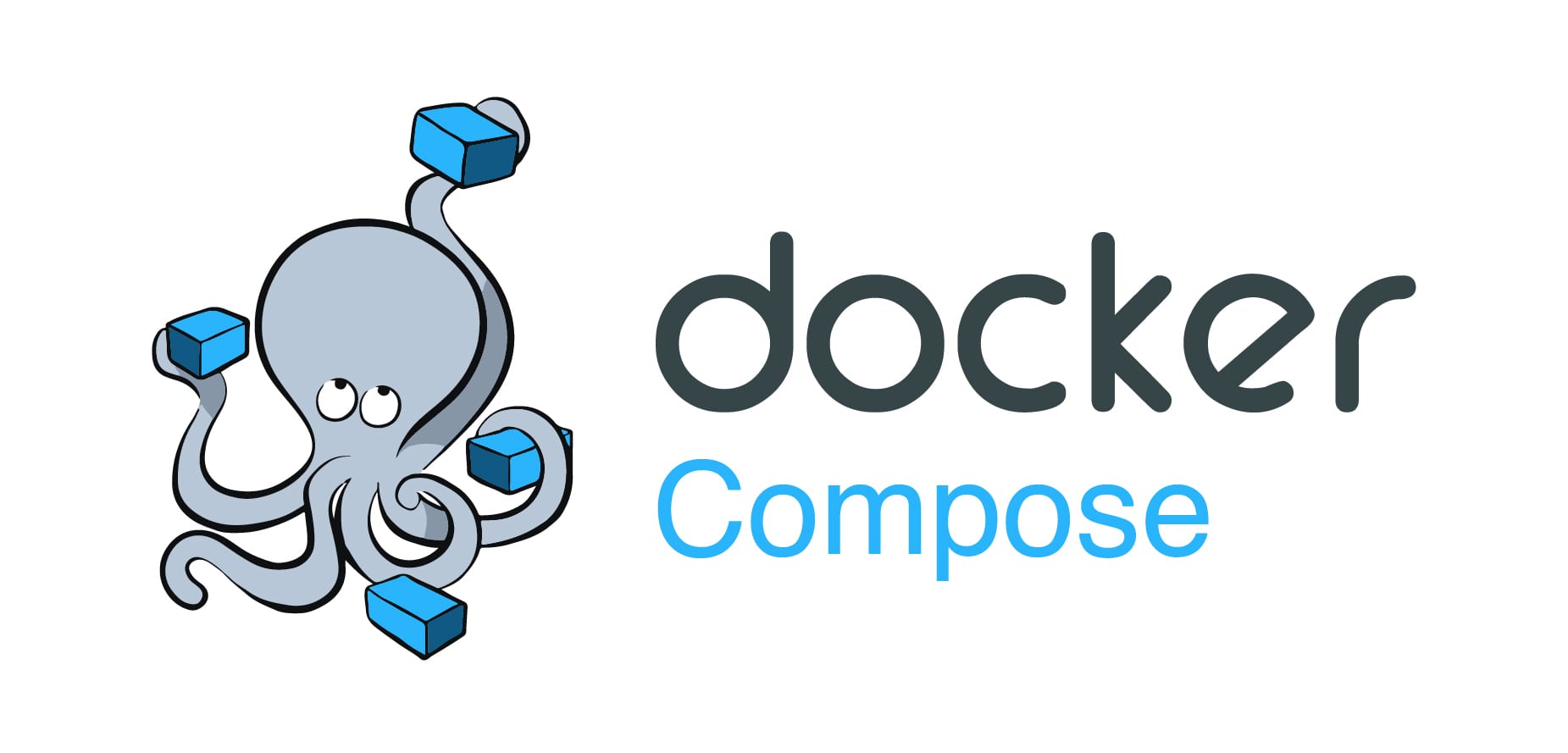
* La classification apprend les relations entre vos points de données afin de prédire des valeurs catégoriques discrètes, comme par exemple si une requête DNS provient d'un domaine malveillant ou bénin.
* La régression apprend les relations entre vos points de données afin de prédire des valeurs numériques continues, telles que le temps de réponse d'une requête Web.



Add some definition

**Images – Containers – Volumes** (to persist data) **- Networks**

Parler du fait que les données sont persistées ce qui facilitent les configs.



Specify also why we use docker composer pour executer plusieurs containers à la fois pour ELK stack. **Docker Compose runs containers on a single host machine.**



**Analyser le journal d'accès de Nginx avec GROK: Dataset**

**https://grokdebug.herokuapp.com**

Grok est un plugin de filtre qui analyse les données non formatées et plates des journaux et les transforme en champs interrogeables. Vous l'utiliserez très certainement pour analyser diverses données.

La définition du mot grok est "comprendre (quelque chose) intuitivement ou par empathie". Essentiellement, grok fait exactement cela en termes de texte - il utilise des expressions régulières pour analyser le texte et leur attribuer un identifiant en utilisant le format suivant : %{REGEX:IDENTIFIER}. La liste des motifs grok pour Logstash peut être trouvée ici. Il y a bien sûr certains caractères spéciaux qui peuvent exister dans les données brutes et qui entrent en conflit avec grok ; ils doivent être échappés avec un backslash. Il existe un outil utile en ligne pour déboguer et tester votre modèle grok.

Logstash dispose d'un ensemble de modèles prédéfinis pour analyser les formats de journaux d'applications courants, tels que Apache, MySQL, PostgreSQL, HAProxy, les journaux système et bien plus encore. Les utilisateurs sont libres d'écrire leurs propres modèles grok s'ils le souhaitent.

Dans le cas présent, nous travaillerons avec les données du journal de Nginx, nous configurerons Filebeat pour envoyer ces données à Logstash, pour lequel nous passerons la plupart du temps à la configuration.

**Données du journal - Journal de Nginx**

Les journaux de serveur Web (*Web server logs*) stockent des données d'utilisation précieuses - l'adresse IP du visiteur, l'activité de l'agent utilisateur, les urls du site visité, les méthodes HTTP utilisées, les octets transférés, divers paramètres de performance (c'est-à-dire le temps qu'il faut au serveur Web pour servir le contenu demandé), pour n'en citer que quelques-uns.

Le fichier journal plat (*flat log file*) contenant ces précieuses données est extrêmement difficile à lire pour la plupart des humains. Supposons qu'une entreprise souhaite savoir où se trouvent ses visiteurs les plus fidèles sur le globe et qu'elle vous a confié cette tâche. Comment aborderiez-vous ce défi ? Ou, par exemple, votre service d'assistance doit faire face à un afflux de clients signalant la lenteur frustrante du temps de réponse de votre service Web. Comment faire pour filtrer les serveurs qui sont hors norme par leur temps de réponse ?

Enfin, vous pouvez vous demander : "à quoi ressemble mon fichier journal Nginx ?" Si votre configuration Nginx est celle par défaut, alors le format du journal est défini dans votre fichier /etc/nginx/nginx.conf comme :

Text

Description automatically generated with low confidence

Ouvrons le fichier access.log et vérifions-le. Les données du journal généré avec la configuration par défaut ressemblent à ceci :

Text

Description automatically generated



Filebeat est un expéditeur léger qui permet de transmettre et de centraliser les données des journaux. Installé en tant qu'agent sur vos serveurs, Filebeat surveille les fichiers ou emplacements de journaux que vous spécifiez, collecte les événements de journaux et les transmet à Elasticsearch ou Logstash pour indexation.

## SOURCES:

#Comprehension du sujet

1. Anomaly Detection [En Ligne] Disponible : <https://www.gavstech.com/anomaly-detection-in-aiops/>
2. What is Elasticsearch? [En Ligne] Disponible: <https://www.elastic.co/guide/en/elasticsearch/reference/8.1/elasticsearch-intro.html>
3. Elastic Machine Learning [En Ligne] Disponible: <https://www.elastic.co/guide/en/machine-learning/current/machine-learning-intro.html#machine-learning-intro> (for elastic machine learning)
4. Nginx [En Ligne]. Disponible: <http://nginx.org/en/docs/http/ngx_http_log_module.html#log_format>
5. Mohiuddin Solaimani, “Anomaly Detection for Application Log Data” (2015). Thèse de doctorat, University of Texas at Dallas, Disponible: <https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1640&context=etd_projects>