Déploiement d'un SIEM et enrichissement des données

Thibaud Badouard
GIP Renater
23 rue Daviel
75013 Paris

Résumé

Le déploiement d'une solution de gestion des événements de sécurité (SIEM) au sein du GIP s'est heurté à deux difficultés que nous avions sous-évaluées : la charge nécessaire à l'analyse des incidents et la complexité de la gestion des sources de données.

Dans notre poster, nous vous proposons de découvrir ce que nous avons mis en œuvre pour réduire ces difficultés. Nous y aborderons les thématiques suivantes :

- *l'enrichissement des événements remontés par le SIEM avec des sources d'informations :*
  - *externes (sites de réputation d'IP, blacklists publiques) pour augmenter le niveau de confiance dans les résultats fournis (cette machine fait-elle partie d'un botnet ? D'autres structures ont elles observé des évènements similaires ?) et faciliter le travail d'analyse ;*
  - *internes (référentiels, outils de gestion d'IP) pour adapter la réponse à apporter (est-ce que la source de l'évènement est interne au GIP ? Est-ce une machine de la communauté ?) et pouvoir automatiser une partie de la réponse à incident.*
- *l'amélioration de la gestion des sources de données en interfaçant le SIEM avec le puits de logs de l'infrastructure pour réduire le nombre d'évènements inutiles analysés et faciliter la récupération de nouveaux types d'événements.*

Mots-clefs

logs, sécurité, incidents, journalisation, SIEM, IOC

1 Pourquoi un SIEM ?

Un SIEM (*Security Information and Event Management*) est une solution visant à centraliser les logs de différentes sources et faciliter leur analyse en cas d'incident de sécurité. La plupart des solutions permettent également de définir des règles de corrélation afin de lever des alertes voire d'exécuter des scripts (du simple *whois* pour obtenir des informations complémentaires à la création de règles sur un pare-feu pour bloquer une attaque).

C'est un outil qui avant tout nous permet d'automatiser ce que nous faisions manuellement et de passer à l'échelle des volumes de logs générés sur nos SI. En effet, si *grep*, *sed*, *cut* ou *awk* restent des outils très pertinents pour l'analyse des fichiers de logs d'une machine précise, ils atteignent rapidement leur limite quand il s'agit d'analyser plusieurs téraoctets de logs structurés différemment.

Le pôle SSI du GIP RENATER a déployé en 2017[1] un SIEM sur un périmètre restreint à l'infrastructure de service. La démarche retenue consistait à débuter sur un nombre limité d'évènements à détecter et donc un nombre réduit de sources, puis à ajouter petit à petit d'autres scénarios et les sources d'évènements nécessaires à leur identification.

2 Pourquoi enrichir les données ?

Durant la première phase du test courant 2018, nous nous sommes aperçus que le temps nécessaire à l'analyse des résultats était nettement supérieur à ce que nous avions envisagé. La principale tâche coûteuse pour l'analyste est la distinction entre les faux positifs (les alertes remontées par le SIEM alors qu'il s'agit de trafic légitime) et les vrais positifs (les alertes correspondant à de réels incidents de sécurité). On a rapidement identifié que les « vrais positifs » sont souvent connus d'autres entités et partagés par divers moyens sur internet.

L'enrichissement des données vise à récupérer ces informations mises à disposition par la communauté et à enrichir les résultats du SIEM pour augmenter le niveau de confiance qu'on peut avoir dans nos alertes. Si le SIEM remonte une tentative d'attaque par une IP et que plusieurs entités tierces ont observé des comportements similaires de la même IP, on peut légitimement avoir plus confiance en cette alerte et automatiser une partie de la réponse à incident.

Cela permet de dégager du temps à l'analyste pour qu'il se concentre sur des tâches ayant une valeur ajoutée plus forte :

- la définition de nouvelles règles / nouveaux scénarios ;
- l'affinage des règles et seuils existants ;
- l'identification des faux négatifs (absence d'alerte alors qu'un incident de sécurité est en cours).

3 Architecture retenue

3.1 Choix d'architecture

La mise en place d'un SIEM s'inscrit dans une démarche plus large de gestion et de détection d'incidents. Pour qu'un projet de déploiement se déroule bien, les choix d'architecture doivent être adaptés à l'organisation de notre structure. Il convient pour cela de se poser les bonnes questions avant de se lancer. Les questions qui ont déterminé nos choix d'architecture étaient les suivantes :

- Veut-on utiliser le SIEM pour gérer uniquement des incidents de sécurité ou veut-on qu'il soit utilisable par les équipes projet pour accéder aux logs applicatifs ou récupérer des métriques métier ?

  → Certaines solutions sont très modulaires et permettent d'avoir une vision métier / exploitation / sécurité. Dans notre cas, nous disposions d'une autre solution de puits de logs qui répondait à ce besoin donc une approche purement sécurité était suffisante.

- Quel périmètre veut-on couvrir ?

  → L'infrastructure à mettre en œuvre sera radicalement différente si on veut couvrir le SI interne d'une structure, un datacenter ou un backbone. Dans notre cas, le périmètre visé était celui de l'infrastructure de services dans un premier temps puis du SI bureautique dans un second temps.

- Quels types d'évènements veut-on détecter ?

  → Cette question nous permettra notamment d'identifier les sources d'évènements qui devront être exploitées. Dans notre cas, nous souhaitions en premier lieu détecter des tentatives d'attaques depuis Internet vers notre infrastructure de Services.

- Avec quels composants de notre infrastructure le SIEM doit-il s'interfacer ?

  → La récupération d'informations tierces sur un annuaire (LDAP, Active Directory) peut par exemple être nécessaire ou au moins s'avérer utile pour identifier des accès illégitimes à des services ou informations. Dans notre cas, les évènements que nous souhaitions détecter ne nécessitaient pas d'interfacer le SIEM avec d'autres composants de notre infrastructure.

- Souhaite-t-on pouvoir personnaliser en détail la solution ou préfère-t-on une solution clé en main ?

  → De nombreuses solutions commerciales utilisent des briques open source (*logstash*, *elasticsearch*) qu'on pourra souvent personnaliser de façon approfondie. Dans notre cas, nous ne souhaitions pas partir sur une solution « boite noire » mais avoir la possibilité de modifier les différentes briques ou pouvoir basculer sur une solution entièrement open source.

- Est-on en mesure de faire une première étape de tri des évènements qui seront envoyés au SIEM pour limiter la quantité de données analysées ?

  → Le coût de licence de certaines solutions dépendant du nombre d'évènements analysés, l'incapacité à faire ce pré-processing peut rendre le

prix de ces licences exorbitant. Dans notre cas, nous n'étions pas en mesure de réaliser ce pré-processing et seule une solution dont le prix ne dépendait pas du nombre d'évènements analysés pouvait entrer dans notre budget.

Après étude « papier » des différentes solutions du marchés, deux répondaient aux attentes exposées ci-dessus et entraient dans notre budget : Keenai et Prelude. Dans notre contexte, les résultats obtenus durant les tests des deux outils se sont révélés assez proches, c'est donc le coût de la licence qui a été le facteur de choix déterminant et nous nous sommes tournés vers Keenai.

3.2 Infrastructure SIEM

La solution Keenai est composée de trois briques principales : le collecteur, la console et les nœuds *elasticsearch* (ES). Leurs rôles sont brièvement explicités dans le schéma suivant :

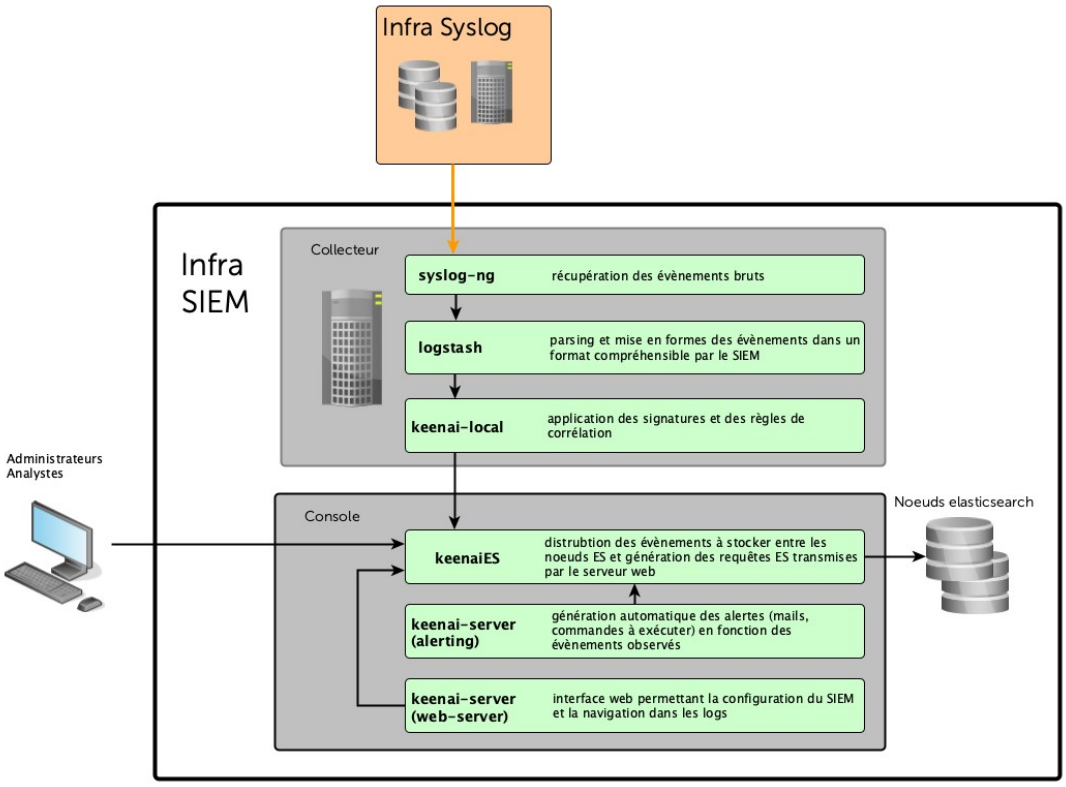

Figure 1: Composantes du SIEM déployé chez RENATER

Pour plus d’informations sur l’architecture d’une solution d’analyse de logs ou d’un SIEM, vous pourrez vous tourner vers la présentation de la solution Racdata faite par les équipes de l’académie de Nancy-Metz aux JRES 2017[2] ou vers le numéro 100 de MISC[3] dont le dossier principal porte sur les SIEM.

Bien que fonctionnelle, cette architecture pose plusieurs problèmes :

- le collecteur joue à la fois le rôle de *parser logstash* et de moteur de corrélation, deux rôles nécessitant des ressources CPU importantes ;
- certains évènements non pertinents sont reçus sur l'infrastructure de SIEM mais sont ignorés, utilisant là aussi des ressources CPU inutilement ;
- les logs sont dupliqués à plusieurs endroits de notre SI.

3.3 Amélioration de l'interface avec le puits de logs

L'équipe infrastructure du GIP RENATER a mis en place un puits de logs vRLI pour améliorer l'infrastructure de gestion de logs et fournir aux équipes métiers un outil d'accès aux logs des serveurs et équipements réseaux. Dans le cadre de ce projet, un framework a également été développé (vRLI Box) pour compléter l'API de vRLI et permettre aux utilisateurs de récupérer, enrichir et modeler les logs. Cette API présente plusieurs intérêts pour le SIEM :

- elle offre la possibilité de soulager le collecteur en ne récupérant que des logs qui seront réellement analysés par les différentes règles de corrélation et d'éviter ainsi une consommation inutile des ressources du collecteur ;
- elle renforce notre agilité en nous permettant d'augmenter ou de réduire le nombre de sources de données qui sont récupérées sans avoir à modifier l'infrastructure complète de gestion de logs ;
- à terme (en pointillés sur le schéma) :
  - elle pourrait nous permettre de nous passer entièrement de la brique *logstash* en récupérant directement les évènements au format JSON attendu par Keenai ;
  - on imagine pouvoir récupérer d'autres logs lorsqu'un évènement donné se produit sur le SIEM : par exemple, si une IP donnée répond aux règles de corrélation prédéfinies et lève une alerte, on récupérera auprès du puits de logs l'ensemble des évènements relatifs à cette IP.

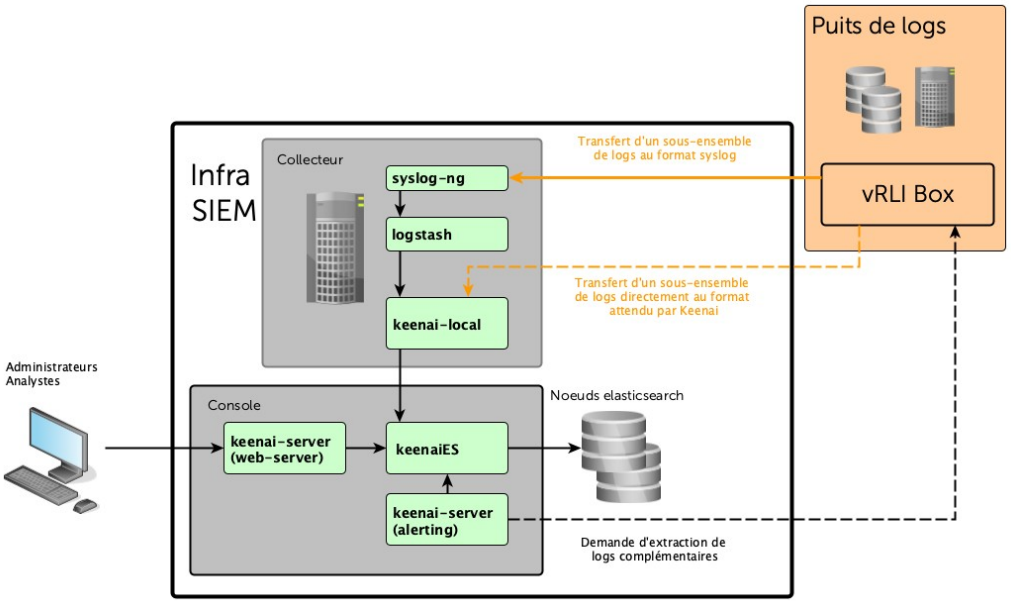

Figure 2: Intégration au puits de logs

3.4 Interfaces avec des sources internes

Afin d'enrichir les données remontées et obtenir des éléments de contexte sur les évènements identifiés, nous souhaitions tirer profit d'informations de diverses sources internes. Nous avons donc mis en place des interfaces entre notre infrastructure SIEM et ces sources.

En plus des logs de trafics générés par le firewall et parvenant au SIEM par l'infrastructure syslog, le firewall génère des rapports que l'on peut personnaliser et qui fournissent des indicateurs sur tous les flux transitant vers l'infrastructure de service (liste d'IP ayant le plus de sessions sur les 6 dernières heures, pays source / destination, etc.). L'équipement ne permet pas de transmettre ces indicateurs au travers de l'infrastructure de logs. Nous avons donc développé un script qui récupère ces rapports et un parser logstash spécifique pour les intégrer au SIEM en amont du moteur de corrélation.

En aval, des interfaces entre le SIEM et deux sources d'informations internes permettent d'adapter le traitement des évènements qui ont levé des alertes :

- avec l'IPAM pour vérifier si l'IP qui a levé l'évènement est une IP de la communauté ;
- avec le référentiel PASS pour obtenir des informations sur les contacts associés à cette IP lorsqu'il s'agit d'une IP de la communauté.

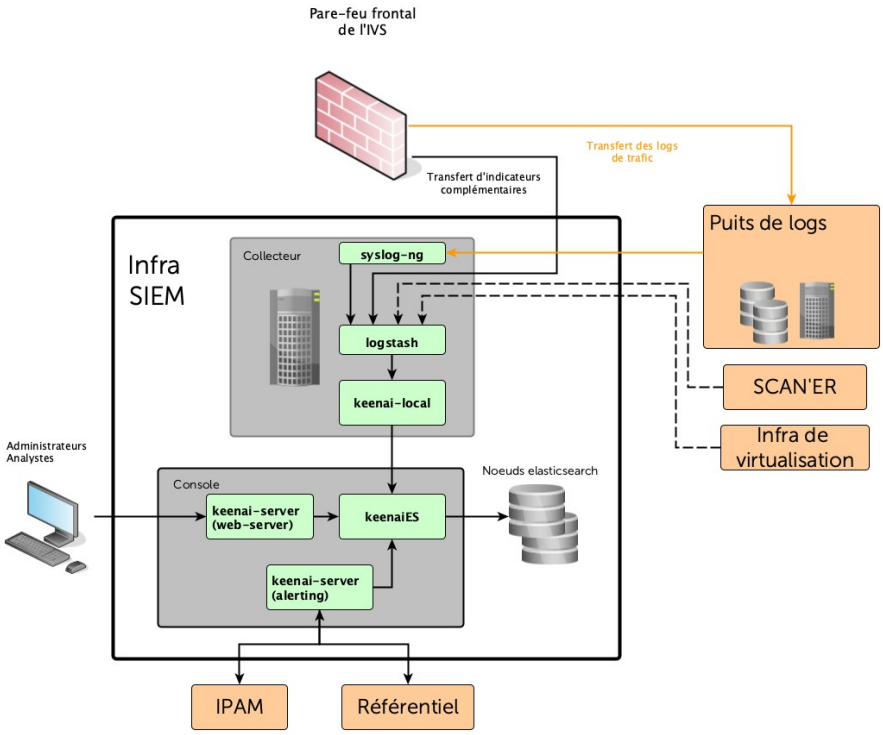

Figure 3: interfaces du SIEM avec diverses sources de données internes

L’ajout de deux autres sources d’informations est à l’étude (en pointillés sur le schéma) :

- SCAN’ER pour générer des alertes ou adapter en le niveau de sévérité en fonction de défauts ou vulnérabilités déjà identifiés sur les cibles ;
- l’infrastructure de virtualisation pour affiner les règles de corrélation en fonction des distributions et logiciels installés sur les machines et de leurs versions.

3.5 Interfaces avec des sources externe

Pour enrichir davantage les évènements traités, nous avons mis en place des interfaces entre le SIEM et des sources d'informations publiques.

Nous gérons ces interfaces avec l’outil open source Minemeld de PaloAlto qui récupère toutes ces données, les agrège et les remet à disposition sous forme de listes.

En plus de fournir des indicateurs en entrée du module de corrélation du SIEM, Minemeld est également utilisé pour générer et mettre à jour les listes d’IP qui seront bloquées par le pare-feu.

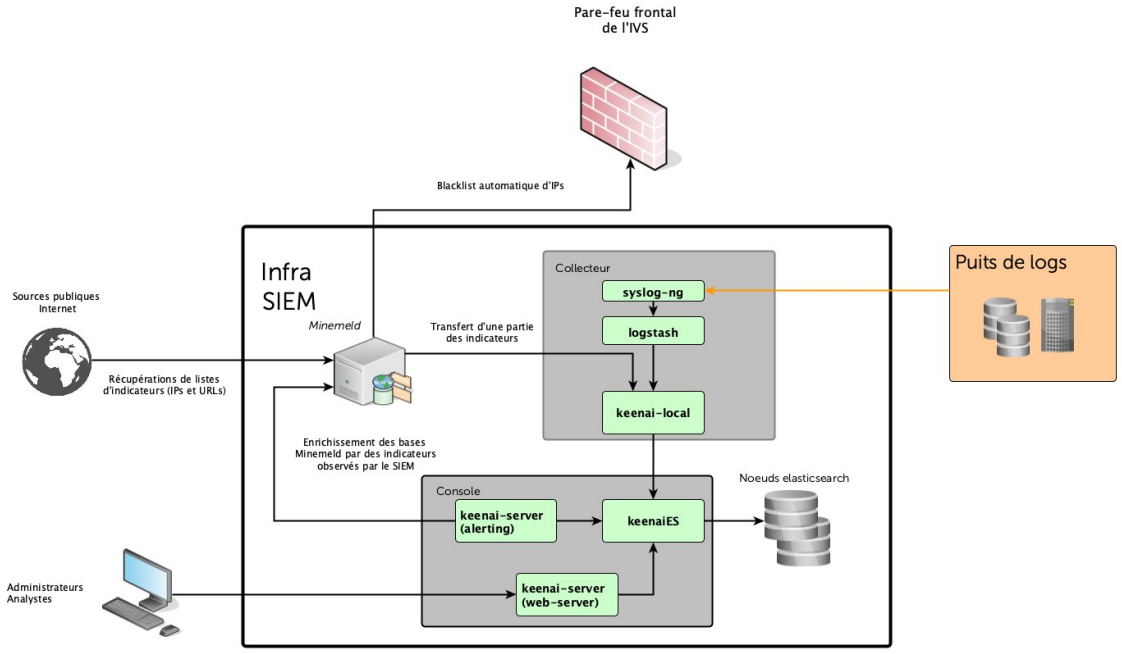

Figure 4: interface avec des sources externes et automatisation des réponses

Minemeld intègre plus de 200 sources fournies par des services de réputation d’IP (*blocklist.de*, *spamhaus*, *badips*, *greensnow*, etc.) mais également des listes fournies par des CERT ou des fournisseurs de services (plages d’IP de Google, d’AWS ou de Microsoft Office 365). Il est également possible de créer ses propres sources, soit

manuellement en ajoutant les IP ou les plages une à une, soit en récupérant des listes sur des URL données.

L'outil permet notamment d'attribuer un niveau de confiance (compris entre 1 et 100) aux sources et d'avoir en sortie des listes différentes en fonction de ce niveau de confiance. Cela nous permettra par exemple de bloquer automatiquement des IP en amont de l'IVS uniquement lorsque l'on a un niveau de confiance très élevé dans la source à l'origine de l'information.

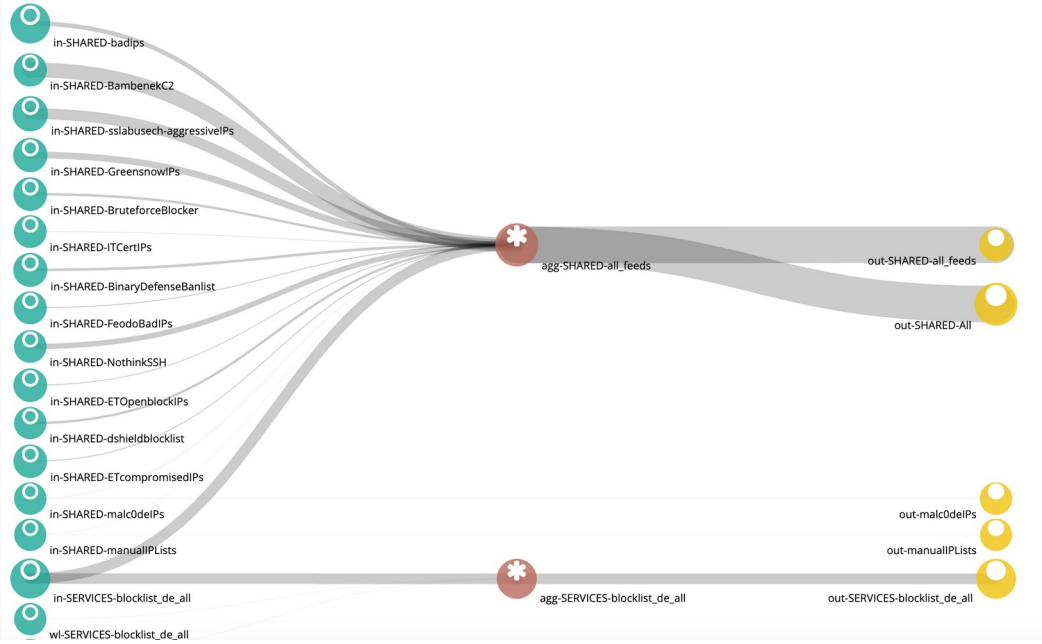

Figure 5: interface de configuration de Minemeld

3.6 Automatisation partielle de la réponse

Les informations récupérées dans ces différentes sources nous permettent d'adapter la réponse que l'on va apporter à un incident identifié par le SIEM :

- si une IP est présente sur une liste connue, on récupère la source et le niveau de confiance associé ;
- si la source qui a levé l'évènement est interne au GIP, une alerte est envoyée à l'équipe sécurité pour qu'une analyse manuelle soit réalisée rapidement ;
- s'il s'agit d'une IP de la communauté, les informations enrichies sont automatiquement transmises au CERT Renater qui sera en charge des investigations et de la communication auprès des contacts identifiés dans PASS ;
- si l'IP n'est pas de la communauté et qu'elle est présente dans une source connue ayant un niveau de confiance élevé, on l'ajoute à une liste qui la bloquera automatiquement lors de sa prochaine tentative de connexion ;
- dans les autres cas, une analyse manuelle devra être réalisée.

4 Feuille de route

Comme nous l'avons vu au fil de cet article, nous étudions actuellement la possibilité d'ajouter plusieurs interfaces pour compléter l'enrichissement des données (architecture de virtualisation, scanner de vulnérabilités, récupération de logs complémentaires).

Au-delà des problématiques d'enrichissements, plusieurs sujets techniques nous semblent intéressants pour améliorer la pertinence des résultats :

- la mise en place d'un module statistique qui nous permettrait de détecter des comportements anormaux à partir de motifs que nous déterminerions. Un tel module n'existe pour l'instant pas dans la solution que nous avons choisie, une demande d'intégration de cette fonctionnalité a donc été transmise au développeur ;
- des modules de « machine learning » qui visent à faciliter le travail des analystes (ou à permettre d'identifier l'intégralité des signaux faibles, si on en croit les plaquettes commerciales). Bien que les promesses de ces modules soient certainement exagérées, ils peuvent présenter un réel intérêt et nous chercherons à en tester pour nous faire notre idée sur la question.

Mais avant l'ajout ou l'essai de nouvelles solutions techniques, deux étapes nous semblent prioritaires pour que les résultats remontés par le SIEM apportent une vraie plus-value :

- l'ajout de scénarios métiers qui permettraient de couvrir des menaces propres à chaque service. En effet, au-delà des attaques applicatives génériques (bruteforce, tentatives d'injection), les analyses de risques mettent en exergue des scénarios dépendants des services (défauts d'autorisation, contournement des mesures en place) pour lesquels les capacités de corrélation du SIEM s'avéreraient très utiles ;
- l'élargissement du périmètre d'opération du SIEM au LAN bureautique et la définition des scénarios associés (compromission du poste d'un utilisateur, accès à des données internes, etc.).

5 Conclusion

Le gain en souplesse dans la gestion des sources est réel mais les performances sont pour l'instant insuffisantes pour nous passer du fonctionnement d'origine. Nous continuons donc de récupérer l'ensemble des évènements des équipements réseaux ce qui n'allège pas la charge du collecteur comme nous le souhaitions.

L'automatisation partielle des réponses et l'enrichissement facilite et allège le travail d'analyse des incidents mais la mise en place de ces interfaces est coûteuse et ne nous permet pas d'avoir plus de temps pour faire des analyses manuelles d'alertes intéressantes comme nous le pensions.

L'année qui vient nous permettra de savoir si le gain de temps induit par l'automatisation se rapproche davantage du modèle « théorique » ou du modèle « réel » proposés par xkcd (cf fig 6).

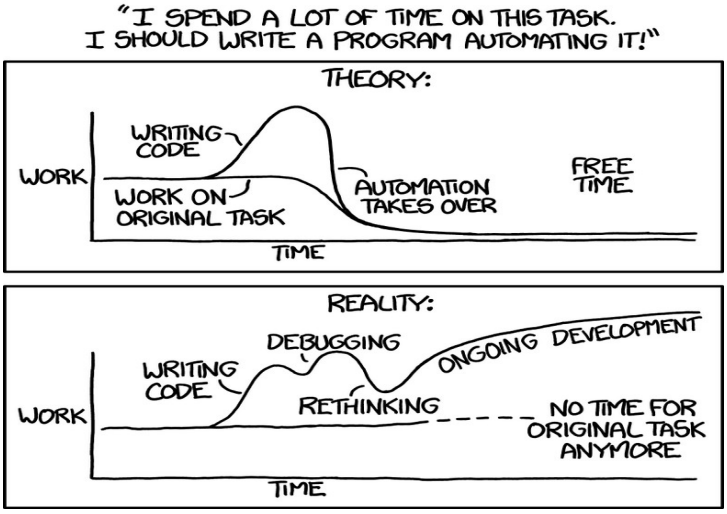

Figure 6: Automation - xkcd (https://xkcd.com/1319/)

Bibliographie

[1] Thibaud Badouard, Mise en place d’un SIEM – Premiers retours, JRSSI, 2018.

[2] Gwendan Colardelle, Michaël Coquard, Tiana Ralambondrainy, Stéphane Urbanovski, Racdata, une solution big data d’analyse de logs pour l’exploitation d’un datacenter, JRES, 2017.

[3] Émilien (gapz) Gaspar, Jean-Philip Guichard, Nicolas Hanteville, Eric Leblond, Laurent Oudot, Christian Perez, Bruno Wyttenbach, Supervision : retours d'expériences autour des SIEM, MISC, numéro 100, novembre / décembre 2018