Neodev

Fabrice Prigent

Université Toulouse 1 Capitole

Capitoul Jeudi 23 juin 2022

Pourquoi ?

Parce qu'en tant qu'informaticien j'ai 2 amies:

- La sauvegarde
- La journalisation

Mais la journalisation, elle cause, et pas qu'un peu !

Les logs

A l'UT1 c'est

- entre 200 et 800 Mo compressés par jour
- entre 6 millions et 25 millions de lignes de log par jour
- entre 50 et 21 000 lignes de log par seconde (sans compter le pare-feu externe, 300 millions de lignes par jour)

UNIVERSITÉ TOULOUSE 1 CAPITOLE

Les logs

Mais toutes les recommendations sécurité sont de

- lire ces logs !!!

Alors on fait comment ?

Idée 1: Deviator

- Repérer l'anormal
  - Outil deviator
  - Mise en place de regexp
  - Si cela ne matche pas, on envoie un mail en cumulant 5 minutes
  - L'administrateur se charge de faire une regexp si c'est du "normal".

Idée 1: Deviator : Résultat

Les résultats:

- Une regexp c'est aussi sympathique qu'une colonne de chars russes devant Marioupol
- Adhésion de l'équipe... mitigée. On se demande pourquoi !?!
  - Faire des regexp à n'en plus finir (c'est donc le psychopathe qui s'y colle).
  - Beaucoup trop de remontées (chaque 5 minutes, entre 5 et 50 lignes)
  - MAIS, début d'utilisation "après coup": Si on a un problème, on va voir les anciennes remontées dans les mails.

Objectifs : Amélioration

On veut améliorer les points suivants:

- Simplifier la gestion (arrêt des regexp pour l'utilisation)
- Réduire les remontées

Idée 2: Neodev

- On inverse le sens: on modifie quand c'est anormal
- Simplifier les lignes de logs (suite de caractères => C, suite de chiffres => D)
- Tout en conservant les informations utiles

Processus Neodev

- On passe tout en minuscule
- On garde le nom de la machine => en majuscule
- On garde le nom du processus => en majuscule
- Pour le reste
  - On repère les mots intéressants => en majuscule
  - On repère les IP intéressantes => IP_CATEGORIE
  - Toute suite de caractères alphabétiques => C
  - Toute suite de chiffres => D
- Si la suite résultante est inconnue, elle fera partie du rapport toutes les 5 minutes
- A l'administrateur de choisir si elle fait partie des choses "A notifier" (en l'ajoutant à un fichier des logs "notifiables").

UNIVERSITÉ TOULOUSE 1 CAPITOLE

Exemple Neodev

Lignes originelles:

```
Jun 17 17:04:42 fsiham2-as-db.ut-capitole.fr ansible-systemd: Invoked with no_block=False force=None
name=avagent daemon_reexec=False enabled=True daemon_reload=False state=started masked=None scope=None
user=None
Jun 21 10:12:24 fafnir.ut-capitole.fr kernel: FIREWALL_DENIED IN=eth1.11 OUT= MAC=56:59:ed:69:59:1d:00:16
SRC=10.8.52.43 DST=213.49.52.34 LEN=40 TOS=0x00 PREC=0x00 TTL=39 ID=8834 PROTO=ICMP TYPE=13 CODE=0
```

Lignes converties:

```
SIHAM!!ANSIBLE-SYSTEMD:!!C C C_BLOCK=C C=C C=C C_C=C C=C C_C=C C=C C=C C=C C=C
FAFNIR!!KERNEL:!!C_DENIED C=C.D C=C.D C=ETHERNET:ETHERNET:ETHERNET C=IP_C_EDUROAM C=IPV4 C=DD C=C
C=C C=D C=D C C=C C=D C=D C=D
```

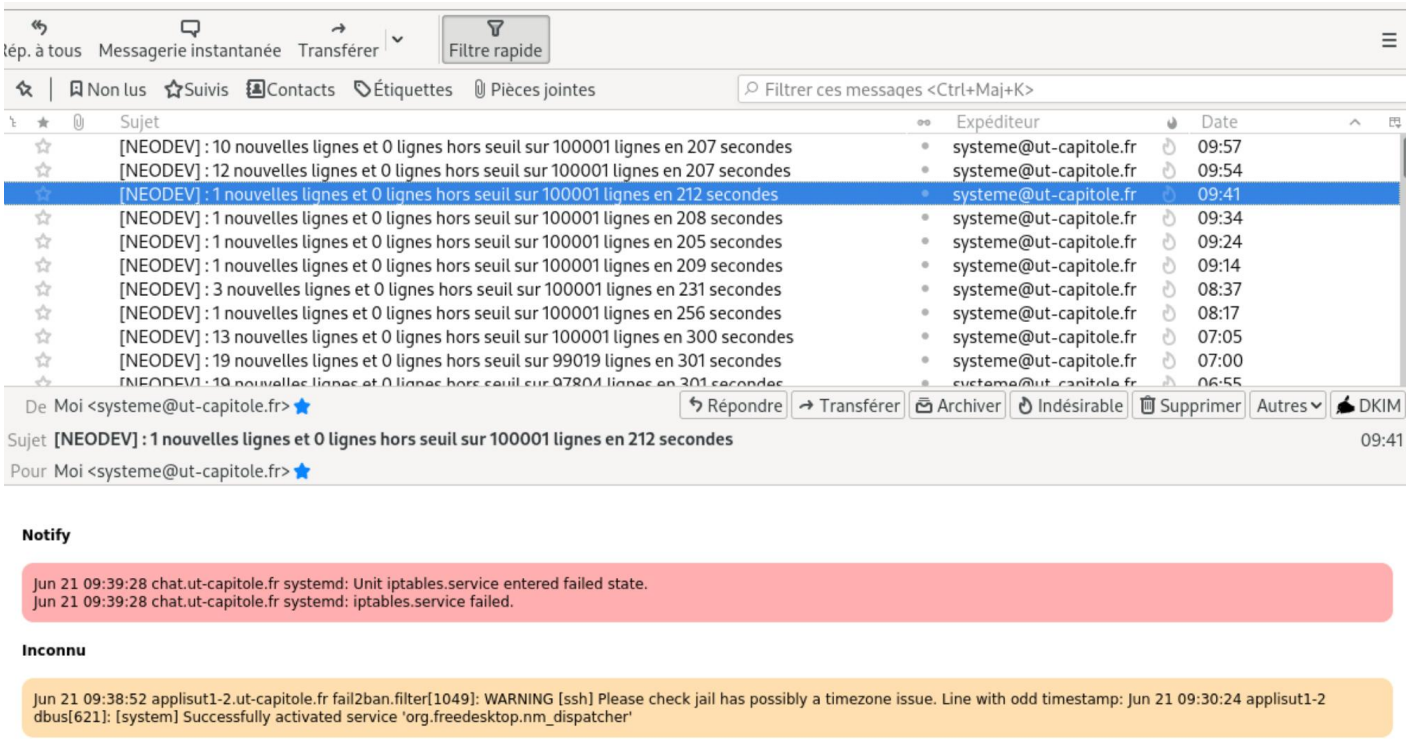
UNIVERSITÉ TOULOUSE 1 CAPITOLE

Résultats Neodev

- Beaucoup plus utilisé par l'équipe (et pas qu'en sécurité ! loin de là).
- Beaucoup de détections d'événements
- Beaucoup plus simple pour l'exploitation
- Beaucoup moins de remontées (même si cela a nécessité des modifications)
- Les IP/mots intéressants jouent beaucoup dans la lenteur de l'apprentissage, mais aussi dans la détection.

UNIVERSITÉ TOULOUSE 1 CAPITOLE

Résultats Neodev



Questions

Des questions ?

Suppléments

- Le script est en python
- Plus que largement améliorable
- 527 lignes pour le code
- 275 lignes pour le fichier de configuration (farci de regexp !!)
- 600 Mo de BD, mais on pourrait facilement se contenter de 100 Mo
- 15% de CPU, avec des pointes à 50% sur les 4 vCPU.
- 1 Mo de RAM.
- L'outil est un poil plus complexe que décrit.

UNIVERSITÉ TOULOUSE 1 CAPITOLE

Suppléments 2: les points compliqués

- Les logs Microsoft
- Les logs ruckus
- Les logs HaProxy

Suppléments 3: les pistes d'optimisation

Avec des risques de "perdre" de l'information

- regrouper les lettres et les chiffres.
- regrouper les "imprimables"
- limiter les catégories d'IP
- faire disparaitre les noms de machines.