Infra Chapril - Demande #5358

Entreposage des logs nginx bastion 2020 anonymisées

04/27/2021 02:56 AM - Christian P. Momon

Status:	Rejeté	**Start date:**	04/27/2021
Priority:	Normale	**Due date:**	
Assignee:	Christian P. Momon	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	Backlog		

Description

En 2020, expérimentalement, les logs nginx bastion ont été collectées.
Le but est de pouvoir générer des évaluations de l'activité du Chapril.

Cela a permis notamment de produire le graphique suivant dans le rapport morale sur l'année 2020 :

chapril-visites.png

Reste à traiter à deux questions :

1. où les entreposer ?
2. gérer l'anomymisation

History

#1 - 04/27/2021 02:58 AM - Christian P. Momon

- Status changed from Nouveau to En cours de traitement

- Assignee set to Christian P. Momon

Traitement des questions :

1) où les entreposer ?
Depuis début 2021, a été mis en place un archivage des logs nginx bastion avec le logiciel logar :
https://admin.chapril.org/doku.php?id=admin:machines_virtuelles:bastion#archivage_des_logs_nginx
Chaque début de mois, les logs du mois précédent sont extraites de /var/log/nginx/* vers /var/log/logar/*.
Du coup, cet emplacement semble idéal.

2) gérer l'anomymisation
C'est un engagement du Chapril que de ne pas conserver des données personnelles plus longtemps que nécessaire, a priori dans les 3 mois.
Se pose la question de comment anonymiser les fichiers de logs 2020.
Le logiciel logar a été étendu avec une commande anonymize et utilisé pour anonymiser les logs 2020.

#2 - 04/27/2021 02:59 AM - Christian P. Momon

- Status changed from En cours de traitement to Attente d'information

Les points sont traités. Avant de fermer, une période pour recenser des remarques, notamment sur l'anonymisation.

À vos avis \o/

#3 - 04/27/2021 09:38 AM - François Poulain

Coucou,

2 remarques :
- mapper chaque ip vers un identifiant tiré au hazard n'est pas de l'anonymisation mais de la pseudonymisation
- conserver les user agent permet le fingerprinting

Au delà de ça, j'adhère au principe de minimisation de la collecte. Donc perso, en dehors d'objectifs clairement définis je suis contre conserver les logs.

#4 - 04/27/2021 11:23 AM - François Poulain

Par contre on peut imaginer conserver des agrégats, je n'ai rien contre. Par ex. pour chaque url le nombre de visite et de visiteurs jours après jours. Un awstat peut faire ce job.

#5 - 04/27/2021 05:26 PM - Christian P. Momon

François Poulain a écrit :

> Par contre on peut imaginer conserver des agrégats, je n'ai rien contre. Par ex. pour chaque url le nombre de visite et de visiteurs jours après jours.

Le problème de l'agrégat, c'est qu'il ne permet pas un retraitement des logs pour par exemple avoir un nouveau point de mesure. Typiquement, en 2021 on travaille à définir de nouveaux points de mesure, c'est bien pratique de pouvoir les recalculer sur 2020.

> Un awstat peut faire ce job.

Awstats a deux défauts. D'abord, il n'est pas prévu pour qu'on en extrait des données. Ensuite, ses points de mesures sont restreints par rapport à ceux qui pourraient nous intéresser. En effet, il n'y a pas que des métriques http qui sont possibles. Chaque service a des métriques spécifiques dont certains sont calculables à partir des logs http.

> - mapper chaque ip vers un identifiant tiré au hazard n'est pas de l'anonymisation mais de la pseudonymisation
> - conserver les user agent permet le fingerprinting

La pseudonymisation implique la réversibilité. Dans notre cas, comment retrouves-tu le nom d'une personne à partir d'une ligne de nos logs ? D'un fingerprinting de nos logs ?

https://www.ccin.mc/fr/fiches-pratiques/anonymisation-ou-pseudonymisation

> Au delà de ça, j'adhère au principe de minimisation de la collecte. Donc perso, en dehors d'objectifs clairement définis je suis contre conserver les logs.

Si nous décidons que le niveau d'anonymisation présenté n'est pas suffisant alors je propose d'étendre l'expérimentation jusqu'à la fin de cette année, le temps de finaliser la mise en place en cours des métriques.

#6 - 05/05/2021 02:07 PM - Christian P. Momon

- Subject changed from Entreposage des log nginx bastion 2020 anonymisés to Entreposage des logs nginx bastion 2020 anonymisées

Suite aux remarques de François, j'ai procédé à une analyse des logs.

En 2020 :

```
176 106 124 lignes de logs
    341 333 ip différentes
    168 507 userAgent différents (dont 114 004 juste un nombre, 10 110 Mastodon, 5 233 Pleroma, 1 978 PeerTube
…)
     32 616 userAgent associables à des humains.
    430 856 couples ip-userAgent uniques
     18 501 userAgent présent dans un seul couple ip-userAgent
```

Conclusion : dans le cadre du Chapril, **57 % des userAgent d'humains sont associés à une seule ip** . Cela confirme la remarque de François, **le userAgent est un puissant fingerprint**.

#7 - 05/05/2021 07:11 PM - Christian P. Momon

- Project changed from Chapril to Infra Chapril

#8 - 09/24/2021 11:31 AM - Christian P. Momon

- Status changed from Attente d'information to Rejeté

En l'absence de résultat probant, abandon de l'expérimentation de stockage de logs anonymisées.