Malware Traffic Analysis with ELK stack
Project : Manipulating ELK Stack

Maxime Geay

June 2020

Formation : Mastère Spécialisé Big Data : gestion et analyse de données massives
Auteur : Maxime Geay
Année : 2020-2021
Encadrant : Julien Dreano

Contents

1 Contexte et rappel des objectifs

Ce projet est proposé dans le cadre du cours Cybersécurité du Mastère Spécialisé Big Data dispensé à Télécom Paris.

L'objectif étant de construire la stack ELK et de la manipuler a des fin d'analyse de la donnée. Cette donnée est en fait des fichiers pcap.

Un fichier pcap (« packet capture ») est une interface de programmation permettant de capturer un trafic réseau. Il s'agit plus particulièrement d'un type de fichier généré lors de la capture du trafic réseau contenant des informations sur des paquets de réseau. (Wireshark est un logiciel permettant une telle capture.)

Etapes du projet :

- Construire le framework (Docker)
- Collecter les données (Wireshark + Logstash)
- Indexation des données (Logstash)
- Dashboard (Kibana)

2 Data

Pour ce projet, on se servira de l'application Netresec NetworkMiner (https://www.netresec.com/?page=pcapfiles) qui est une application classée comme un programme d'analyse légale de réseau.
NetworkMiner permet aux utilisateurs de déterminer des informations relatives au réseau comme les `types de système` et `les versions des ordinateurs` qui sont connectés au réseau ou qui essaient d'accéder aux ressources du réseau, ainsi que les `adresses IP` de ces machines, les `ports ouverts`, les `noms d'hôtes` et `les journaux de session active` entre autres.

Point de vue métier :

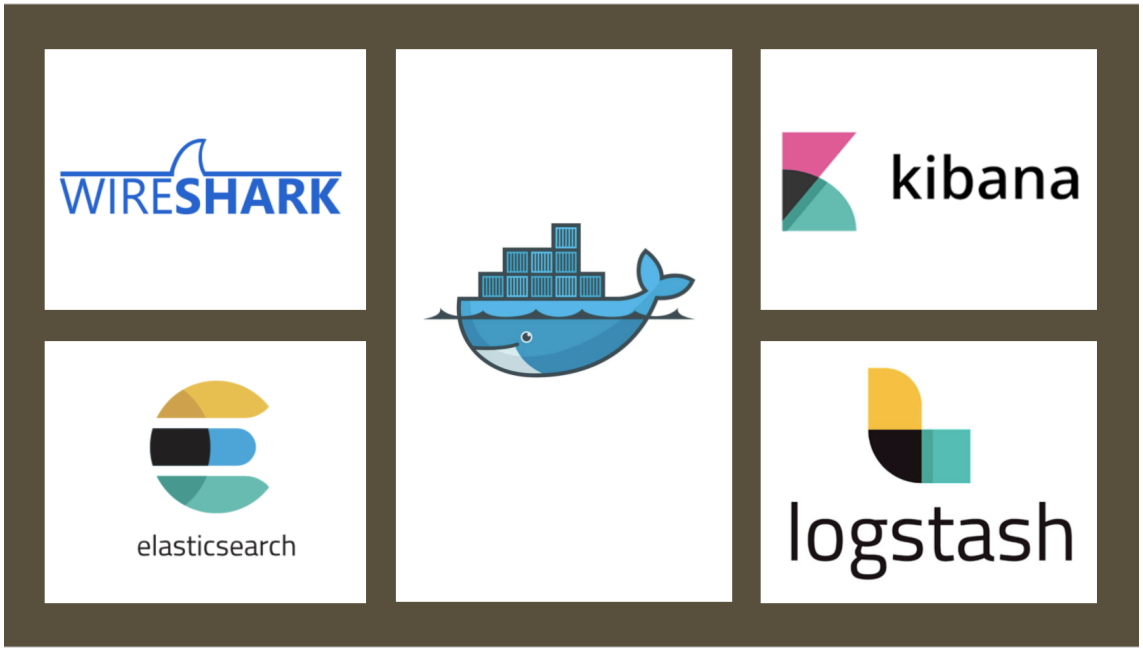
D'autres fonctionnalités de capture de paquets et de reniflage passif sont également intégrées dans NetworkMiner de Netresec. Elles peuvent être utilisées par les administrateurs système et le personnel informatique pour sécuriser le réseau et optimiser l'efficacité des ressources liées à la protection et à la confidentialité qui sont disponibles pour tous les ordinateurs et appareils auxquels l'accès au réseau a été accordé.

NB: NetworkMiner de Netresec a été conçu pour recueillir des détails criminalistiques sur le réseau et non des données sur le trafic réseau.

Nous allons utiliser les données téléchargeables ici :
https://www.malware-traffic-analysis.net/2020/12/31/index.html il s'agit de donnée d'analyse du trafic des malwares par des logiciels malveillants sous Windows. (size : environ 20MB)

3 Framework

La suite ELK est disponible gratuitement sur le web. Il s'agit de 3 projets open source : **E**lasticSearch pour la recherche et l'analyse de la donnée, **L**ogstash pour le traitement de la donnée coté serveur et **K**ibana coté client pour visualiser la donnée.



elk veut faciliter et accélérer la recherche et l'analyse de grands ensembles de données.

- Elasticsearch va permettre d'extraire les données, Logstash normalise toutes sortes de données temporelles et Kibana apporte un insight.

Bien qu'Elasticsearch, Logstash et Kibana aient été conçus pour fonctionner ensemble, chacun d'entre eux est un outil distinct.

- ElasticSearch est un moteur de recherche et d'analyse qui utilise le format JSON. Son objectif est d'extraire efficacement les données à partir de sources de données structurées ou non structurées en temps réel. Elasticsearch utilise Lucene pour fournir les capacités de recherche en texte intégral les plus puissantes disponibles dans n'importe quel produit open-source.
- Logstash est un outil pour la saisie, le traitement et la sortie des données logs. Sa fonction est d'analyser, filtrer et découper les logs pour les transformer en documents formatés à destination d'Elasticsearch. Kibana est un tableau de bord interactif et paramétrable qui permet de visualiser les données stockées dans ElasticSearch.
- Kibana apporte un insight sur les tendances et les modèles sous toutes formes de diagrammes et courbes. Ce dashboard peut être partagé et associé à des visualisations de données pour une communication rapide et intelligente.

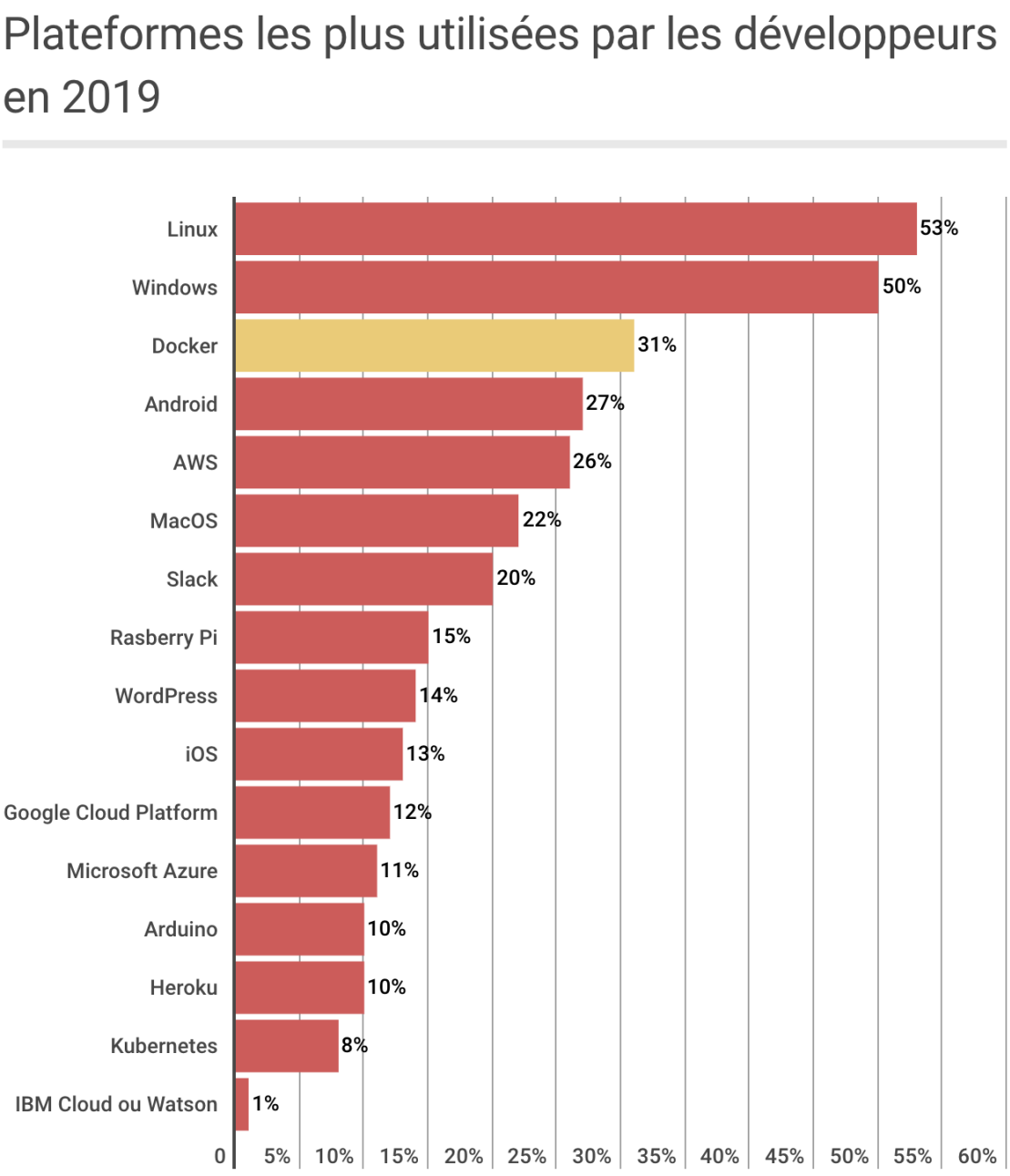
Afin d'utiliser ces trois couches, Docker est un candidat idéal puisqu'il s'agit d'un conteneur qui contient déja la stack ELK.

Qu'est-ce que Docker ?

Docker permet d'embarquer une application (fichiers source,environnement d'exécution, librairies, outils et fichiers) dans un ou plusieurs containers logiciels qui pourra s'exécuter sur n'importe quel serveur machine, qu'il soit physique ou virtuel. Docker fonctionne sous Linux comme Windows Server. C'est une technologie qui a pour but de faciliter les déploiements d'application, et la gestion du dimensionnement de l'infrastructure sous-jacente. Elle est proposée par la société Docker, en partie en open source (sous licence Apache 2.0).

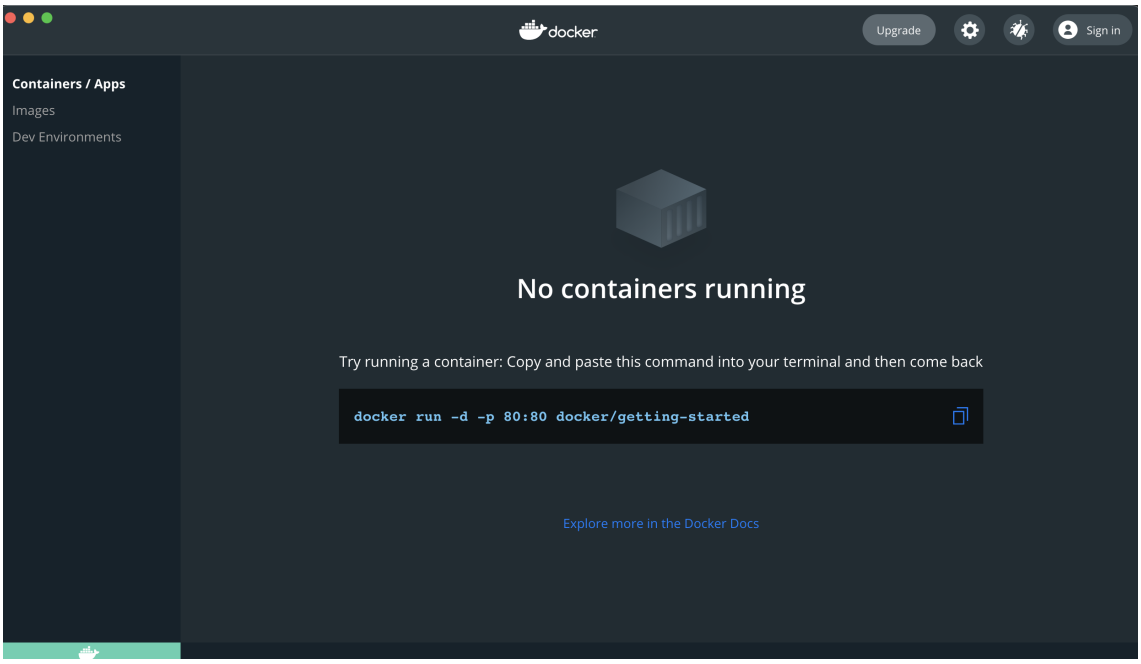
Quels sont les avantages de Docker ?

J'ai décidé d'utiliser Docker car à la différence d'une VM, il est plus léger et il se lance rapidement : Contrairement à la virtualisation de serveurs et à une machine virtuelle, le conteneur n'intègre pas de noyau, il s'appuie directement sur le noyau de l'ordinateur sur lequel il est déployé.



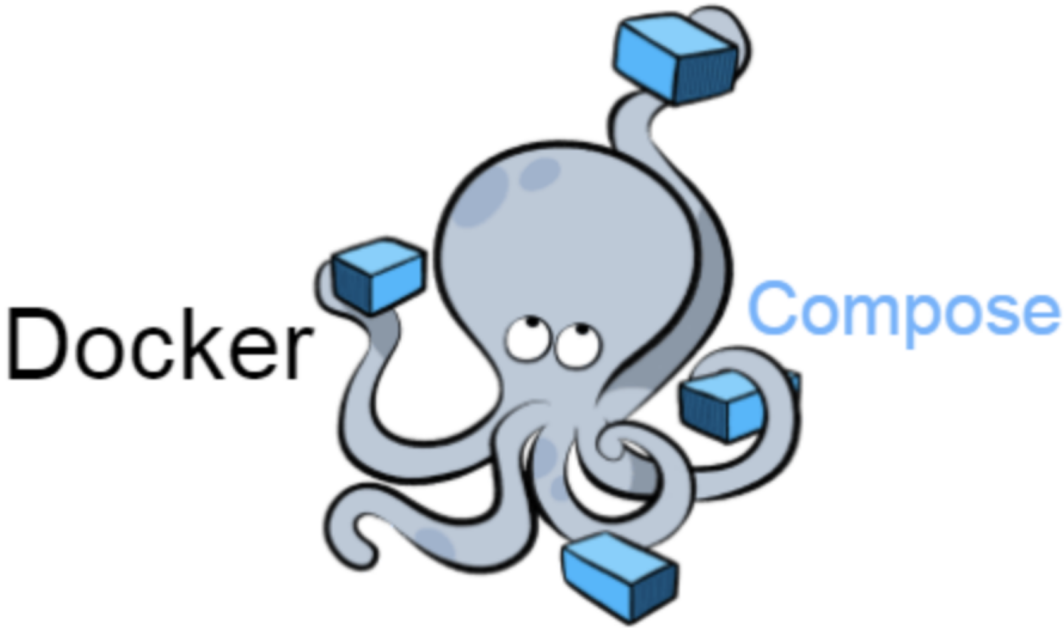
4 Installation

Il est aisé d'installer docker gratuitement sur le site officiel : https://docs.docker.com/get-docker/



J'utilise personnellement Mac OS, ce qui induit que j'ai directement Docker Compose avec Docker Engine, cependant si vous ne possédez pas de Mac il vous faudra télécharger Docker Compose.

Qu'est ce que Docker Compose ?



Docker Compose est un outil qui permet de décrire (dans un fichier YAML) et gérer (en ligne de commande) plusieurs conteneurs comme un ensemble de services inter-connectés. Si je travaille sur une application ELK + Wireshark, je vais ici décrire un ensemble composé de 4 conteneurs :

- ElasticSearch
- Logstash
- Kibana
- Tshark

On peut donc à l'aide d'une seule commande `Docker-compose up` lancer les 4 conteneurs ensemble.
Un avantage certain est celui du travail collaboratif puisqu'ici il n'y a pas besoin d'instructions très détaillées afin de lancer les applications. Cela se fait de manière automatique
Dans chaque conteneurs se trouve un fichier docker-compose.yml, qui décrit un ensemble de paramètres qui correspondent aux options disponibles lors d'un docker run : l'image à utiliser, les volumes à monter, les ports à ouvrir, etc. Mais on peut également y décrire des éléments supplémentaires, comme la possibilité de « construire » (docker build) une image à la volée avant d'en lancer le conteneur.
Le dossier Logstash possède également un fichier conf contenant la pipeline sur le traitement des données : input, filtres, cleaning et output vers Elasticsearch.
Le dossier Wireshark contient lui un script bash supplémentaire permettant de convertir les fichiers pcap en fichiers json à l'aide de l'utilitaire TShark.

Qu'est ce que Wireshark ?

Wireshark est une application qui intègre des fonctionnalités permettant de surveiller l'activité pertinente du réseau et de stocker ces éléments de données dans des journaux et des bases de données en vue d'une analyse ultérieure.
Différents protocoles de réseau peuvent être surveillés, suivis et analysés par Wireshark, ce qui fournit aux administrateurs de systèmes de réseau et au personnel informatique un moyen rapide et facile d'améliorer l'efficacité des ports de réseau et des transmissions de données vers et depuis les ordinateurs de réseau, les autres dispositifs connectés au réseau et les serveurs Internet.

Une fois installé on peut cloner un github qui contient les images ELK Docker, prenons par exemple ce github : **Github à cloner**

Une fois cloné il suffit de suivre les étapes de set up uniquement car nous n'utiliserons pas des fichiers de log comme présenté dans ce github mais des fichiers pcap comme mentionné précédemment.

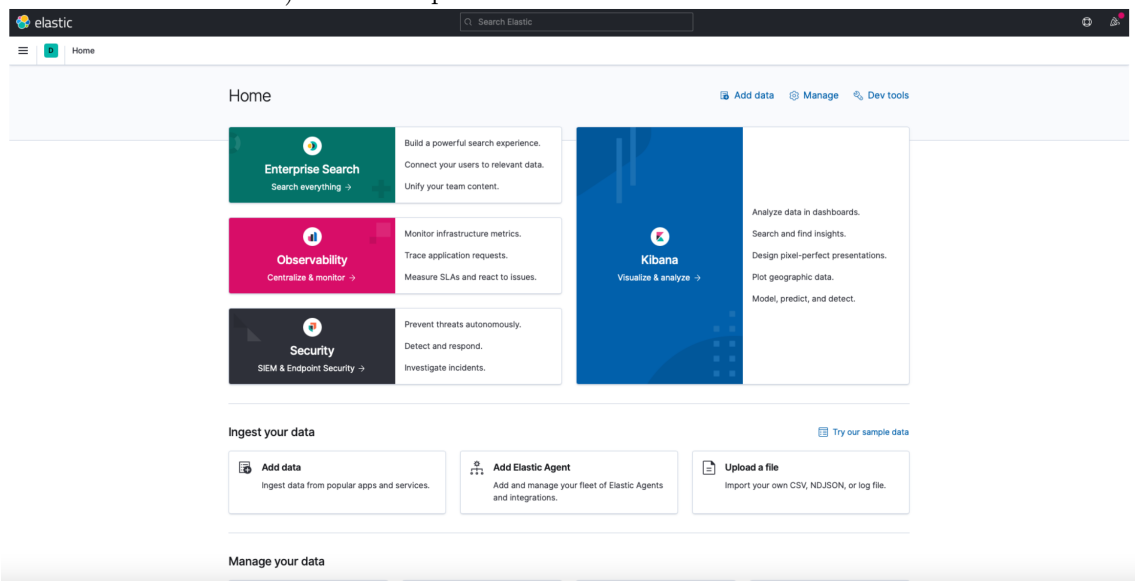
On peut ensuite ajouter le conteneur wireshark en se plaçant dans le dossier nouvellement crée dans le repertoire cloné via la commande suivante : `docker pull nomdeimage`

Vous trouverez via ce lien l'image wireshark que vous souhaitez : **Image Wireshark**

Une fois fait, il suffit de lancer avec la commande `docker-compose up` la stack. Vous devriez voir apparaître dans docker les images :
Une fois la stack lancée il suffit d'attendre quelques secondes pour voir apparaître les informations sur le cluster **Elastic**:

```
{
  "name" : "025f2c6d7e17",
  "cluster_name" : "docker-cluster",
  "cluster_uuid" : "JbSyYuEsRzqjqidtKP1dqA",
  "version" : {
    "number" : "7.13.0",
    "build_flavor" : "default",
    "build_type" : "docker",
    "build_hash" : "5ca8591c6fcdb1260ce95b08a8e023559635c6f3",
    "build_date" : "2021-05-19T22:22:26.081971330Z",
    "build_snapshot" : false,
    "lucene_version" : "8.8.2",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

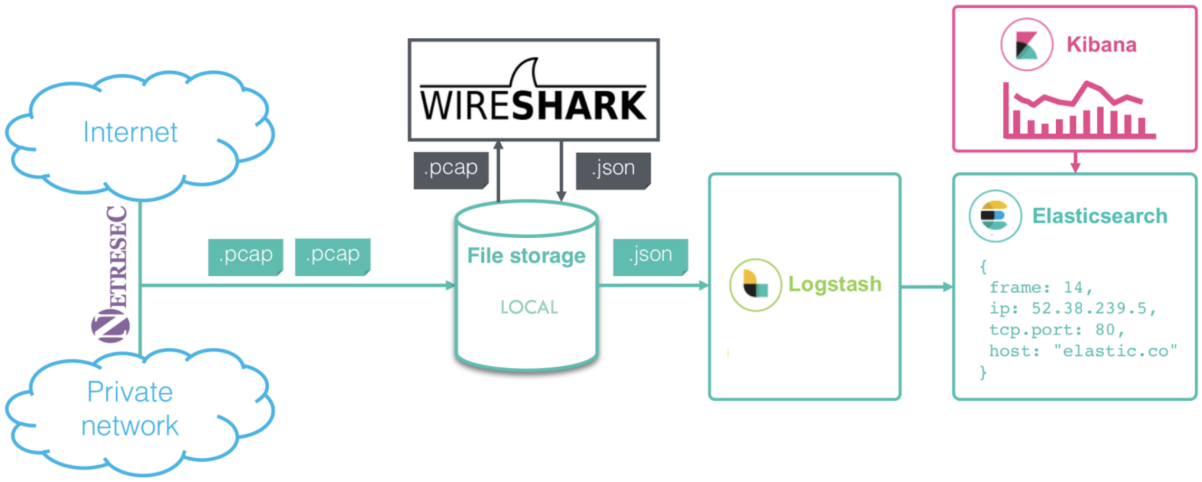

puis après quelques minutes (le temps que wireshark récupère les données, que logstash les traites et les transfère à Elasticsearch) on devrait pouvoir accèder à l'interface de **Kibana**



Pour stopper l'environnement il suffit de lancer la commande bash : `docker-compose down -v`
passons maintenant à l'architecture :

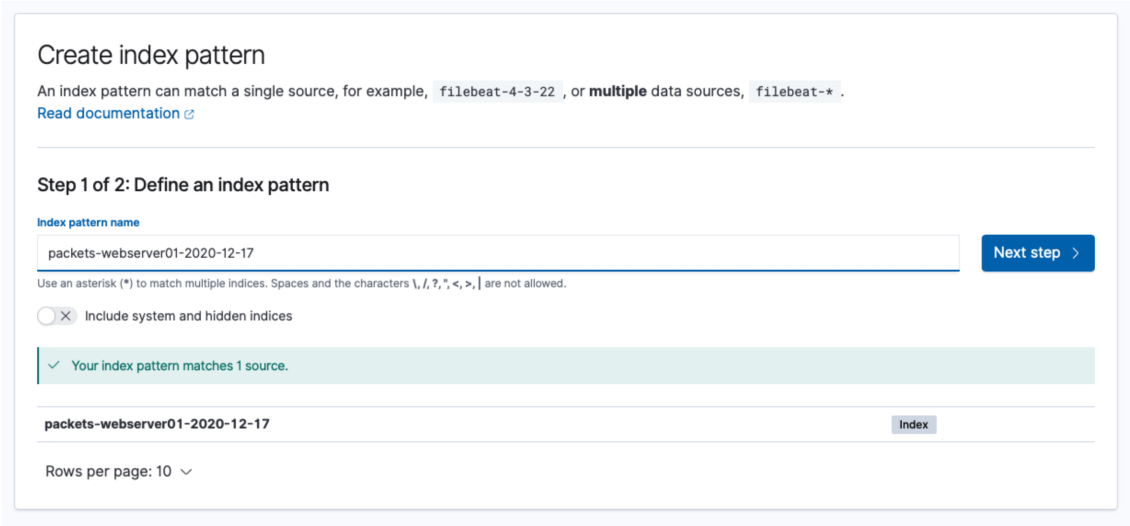
5 Architecture

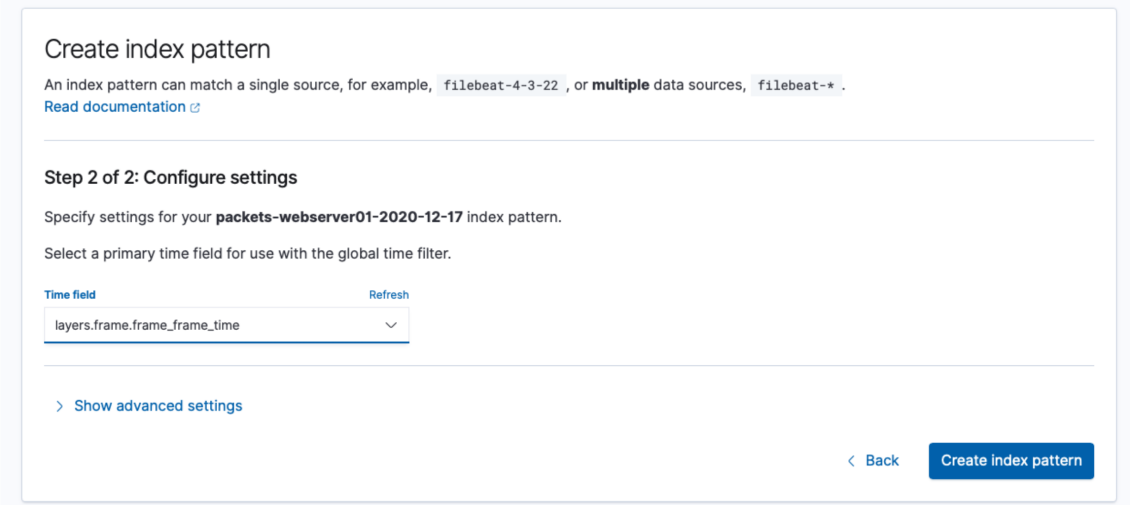
On récupère les données sur le site Netressec et on les place en local dans le dossier data qui convient, puis wireshark se charge de convertir les données pcap en json pour les transmettre à logstash qui via sa pipeline de traitement des données (cf fichier `logstash.conf` qui notamment parse convenablement les données en octroyant le bon timestamp c'est à dire celui d'origine et non celui du traitement par wireshark) va finalement envoyer les données à Elasticsearch. Elles seront visible dans kibana après avoir crée un index.



6 Indexation des données

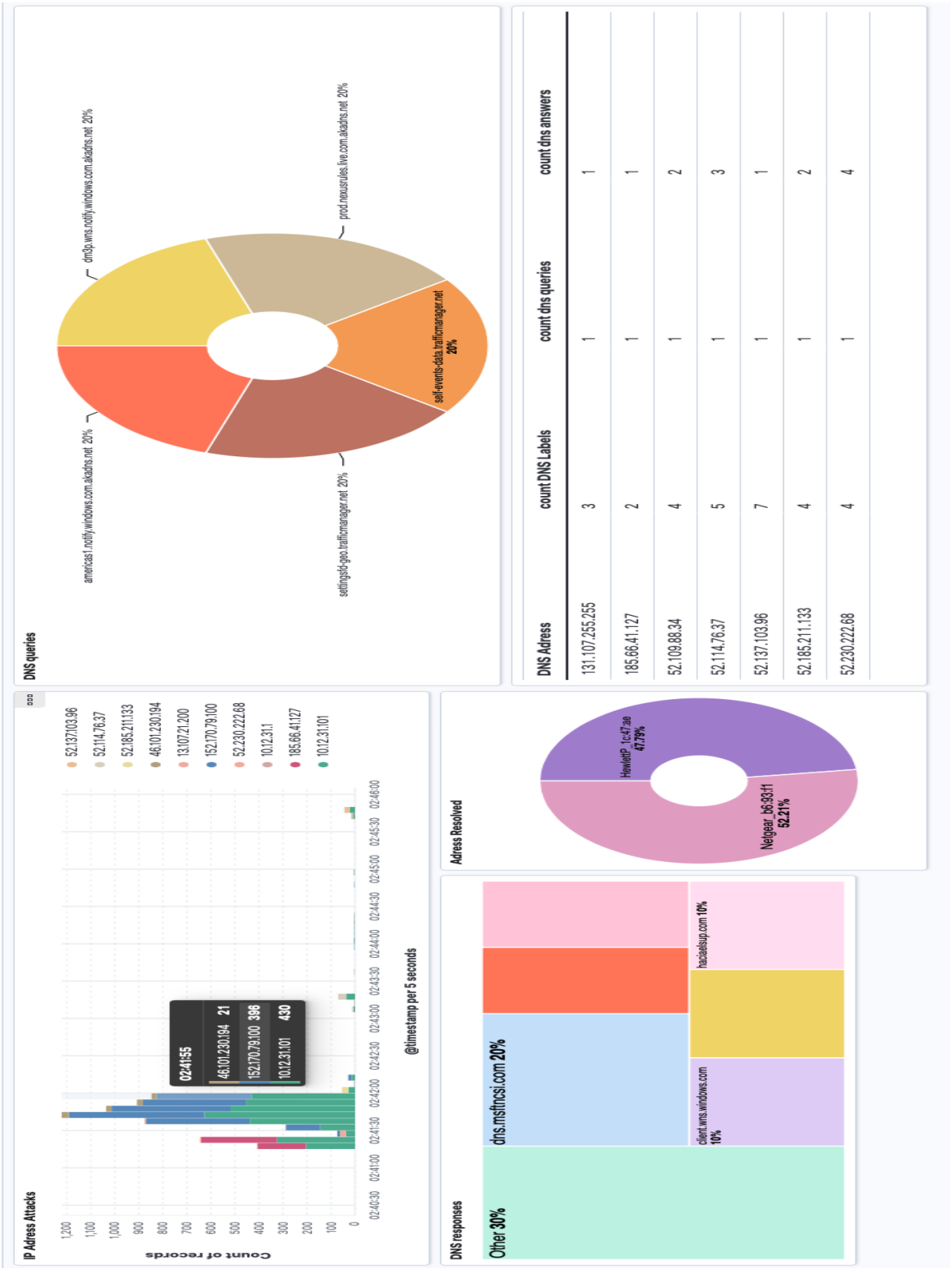
Dans l'onglet discover de Kibana vous pouvez créer un index si les données ont bien été transmises à Elasticsearch





Il est maintenant possible de créer des dashboards intéractifs

7 Dashboard Kibana



Dans l'onglet Dashboard, il est possible de créer des graphs et tableaux et d'en faire un dashboard, voici un exemple de Dashboard sur les données de Malware.

NB : Pensez à bien parametrer le time range

8 Références

Théorie :

https://www.elastic.co/fr/what-is/elk-stack
https://docs.docker.com/compose/
https://web.leikir.io/docker-compose-un-outil-desormais-indispensable/
https://www.reviversoft.com/fr/file-extensions/pcap
https://www.journaldunet.fr/web-tech/guide-de-l-entreprise-digitale/1146290-docker-definition-docker-compose-docker-hub-docker-swarm-160919/

Pratique :

https://github.com/deviantony/docker-elk
https://www.elastic.co/fr/blog/analyzing-network-packets-with-wireshark-elasticsearch-and-kibana

Table of contents