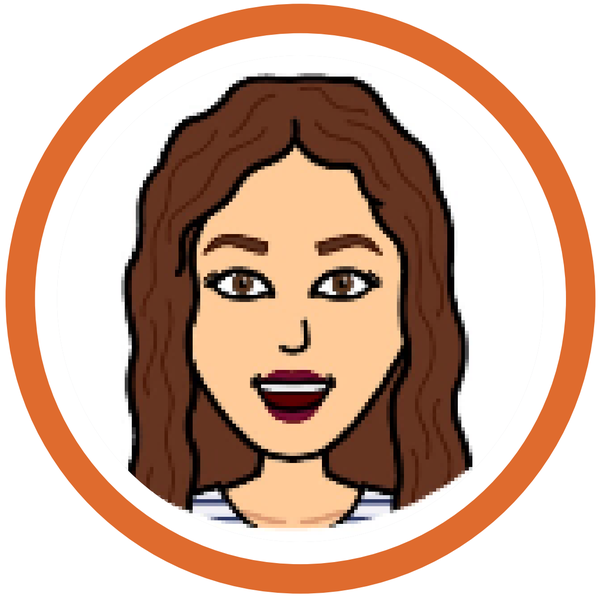
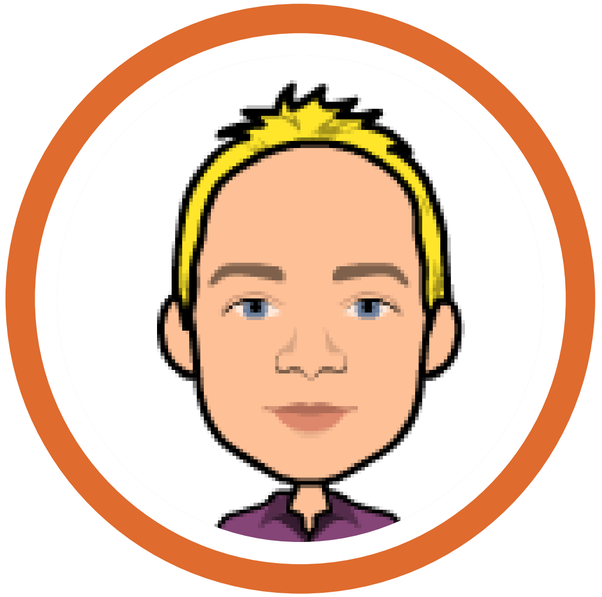
Exploration de données opendata avec la suite Elastic






```
{
  "prenom" : "Marine",
  "nom" : "MESNAGE",
  "age" : 23,
  "tags" : ["big data", "elastic", "hadoop"],
  "email" : "mmesnage@aneo.fr",
  "geek" : true
}
```

```
{
  "prenom" : "Cyrille",
  "nom" : "ADAM",
  "age" : 37,
  "tags" : ["big data", "elastic", "grid"],
  "email" : "cadam@aneo.fr",
  "geek" : true
}
```

Déroulement de la présentation

- ✓ Présentation des données open source
- ✓ Présentation de la suite Elastic
- ✓ Logstash & l’intégration de données
- ✓ Elasticsearch & le stockage de données
- ✓ Kibana pour en tirer de la valeur
- ✓ Un mot sur ANEO

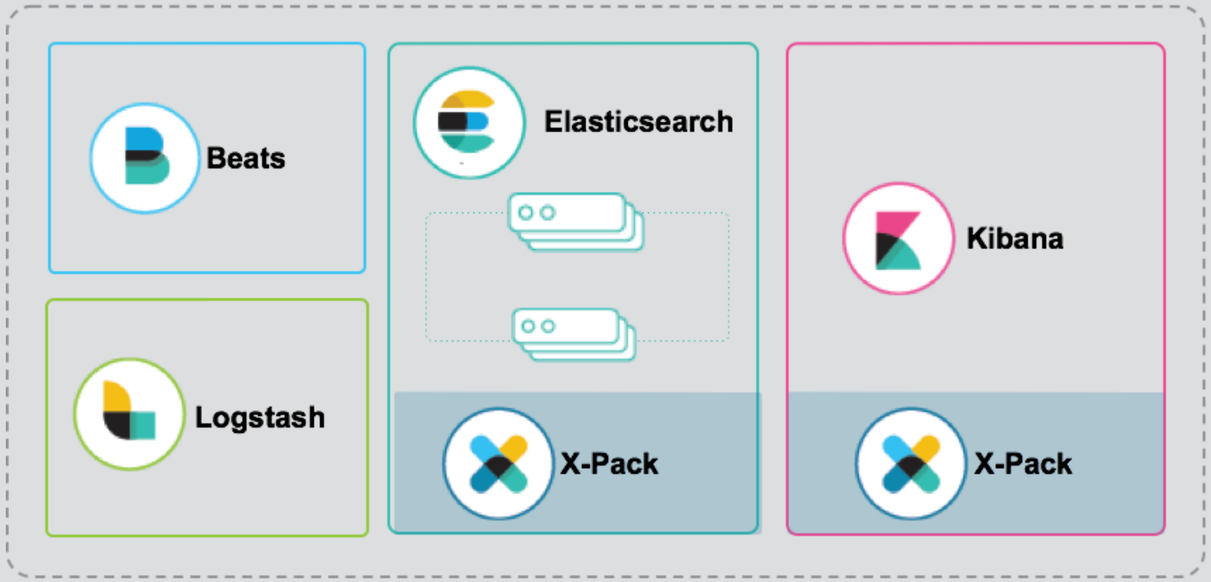
A la fin de cette présentation vous :

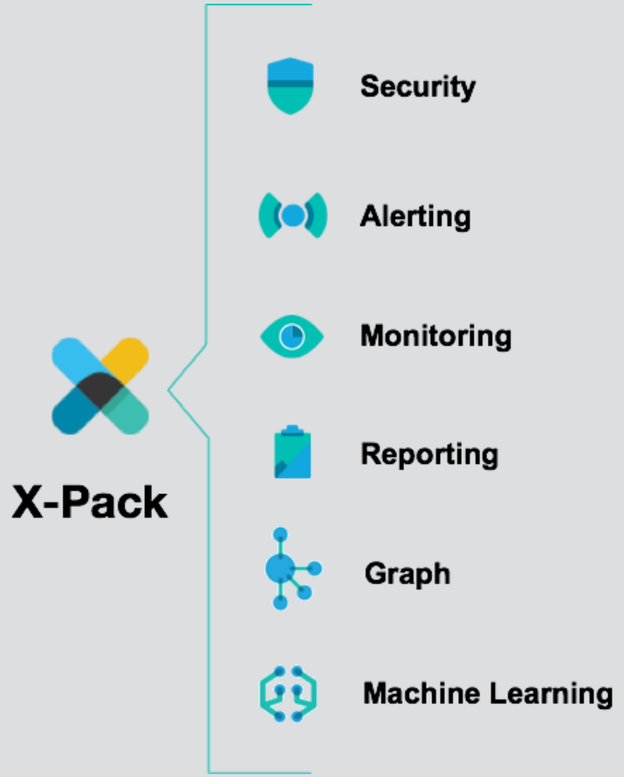
- Connaitrez la suite Elastic : ses composants et leur fonctionnement
- Saurez refaire les démos proposées dans cette présentation
- Pourrez explorer vos propres données

Sources de données explorées

http://opendata.paris.fr

http://www.enedis.fr/open-data





La suite Elastic : bien plus qu'une simple solution de gestion de logs

Installer la stack

Récupérer l’ensemble des binaires disponibles et les mettre en place

Ou

Utiliser les scripts fournis dans notre package
(powershell sous Windows, à passer en bash pour Linux/Max) :

- **_InstallElastic.ps1**
- **_StartElasticsearch.ps1**
- **_StartKibana.ps1**
- **_StartCerebro.ps1**

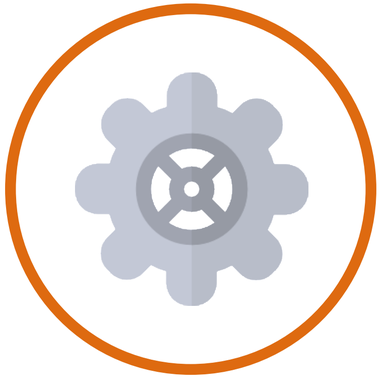
https://github.com/cyadam/explore-opendata-elastic

logstash

Collecte de données

input { }

Analyse et transformation



filter { }

Transport de données

output { }

input {

beats { … }
elasticsearch { … }
exec { … }
file { … }
jdbc { … }
github { … }
http { … }
kafka { … }
log4j { … }
meetup { … }

}

filter {

aggregate { … }
anonymize { … }
csv { … }
date { … }
dns { … }
fingerprint { … }
geoip { … }
grok { … }
mutate { … }
xml { … }

}

output {

csv { … }
elasticsearch { … }
email { … }
exec { … }
file { … }
google_cloud_storage { … }
http { … }
kafka { … }
mongodb { … }
s3 { … }

}

1er exemple : les bureaux de votes

La source de données est disponible ici :
https://opendata.paris.fr/explore/dataset/bureaux-de-votes

Ce jeu de données contient la localisation de tous les bureaux de votes de Paris.

Pour le charger avec Logstash, utiliser le script _StartLogstash_bureaux_de_votes.ps1 :

- Télécharge le fichier
- Exécute logstash pour charger les données

```
input {
  file { … }
}
```

```
filter {
  if ( … ) {
    drop { … }
  }
  csv { … }
  mutate { … }
  mutate { … }
}
```

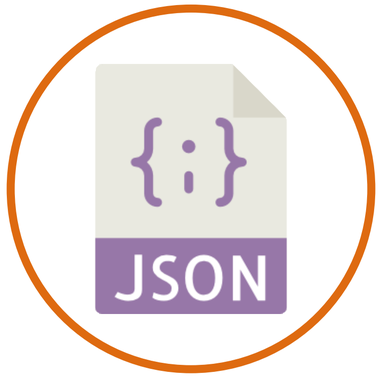
```
output {
  elasticsearch {
… }
}
```

elasticsearch

Performance

- ✓ « Near real-time »
- ✓ Haute disponibilité
- ✓ Scalabilité horizontale

Fonctionnement



- ✓ Stockage JSON
- ✓ Moteur d’indexation
- ✓ Moteur de recherche
- ✓ Moteur d’agrégation

Accessibilité

- ✓ API REST
- ✓ Clients TCP
- ✓ Multi-langages

Elasticsearch ça sert à quoi ?

elasticsearch

Massif

Stocker des données massives, qu'elles soient temporelles ou contextuelles

Analyse

Agréger des données volumineuses pour en faire une analyse statistique

Full Text

Réaliser des suggestions et de la recherche sur des éléments textuels

Détection

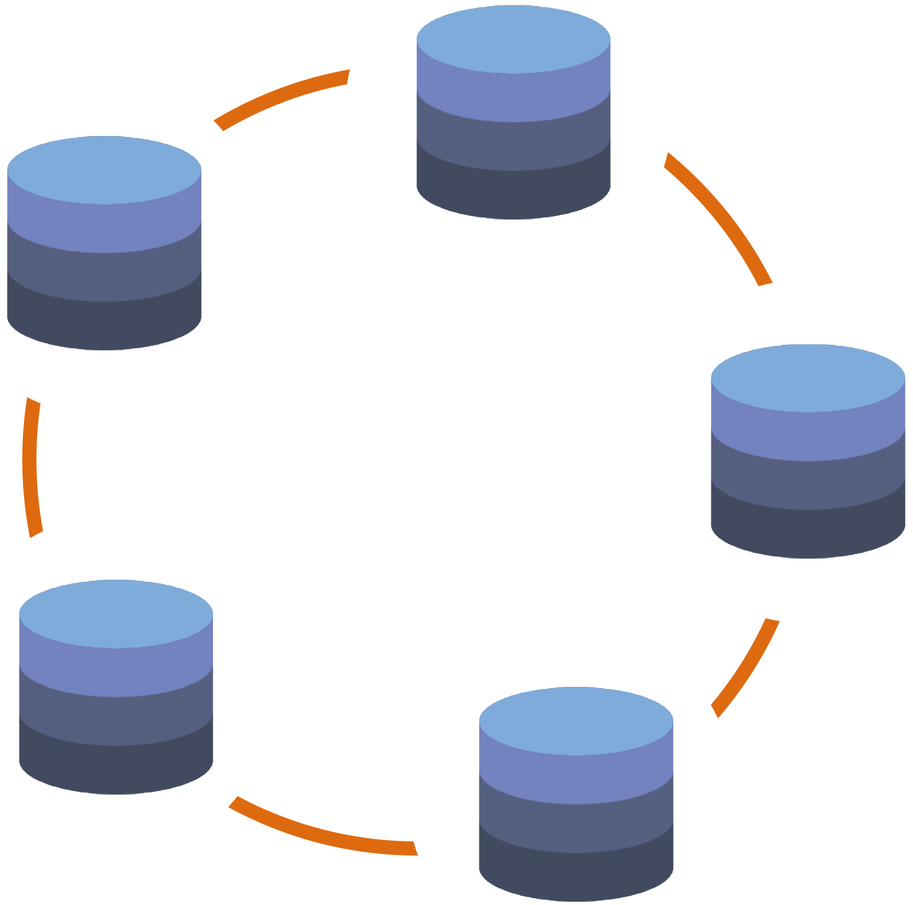
Générer des informations additionnelles pour en tirer de l'information ou des alertes

Un seul format :

- Format JSON ultra flexible
- Structuré en documents
- Avec des données typées
- Similaire à une ligne en base de données

```
{
    "prenom" : "Marine",
    "nom" : "MESNAGE",
    "age" : 23,
    "tags" : ["big data", "elastic",
"hadoop"],
    "email" : "mmesnage@aneo.fr",
    "geek" : true
}
```

L’infrastructure d’elasticsearch



Cluster

- ✓ Elasticsearch fonctionne en cluster
- ✓ Un cluster est formé de plusieurs nœuds. Généralement 1 nœud = 1 serveur
- ✓ Un cluster est identifié par un nom unique

Nœud

- ✓ Un nœud maitre : consistance et distribution des données
- ✓ Un nœud data : stocke une partie des données & exécute les requêtes clients

Les concepts d’elasticsearch

elasticsearch

Indice

- ✓ Un ensemble de données consistantes
- ✓ Lié à un groupe de données spécifiques ou à une date

Shard

- ✓ Partie de données d’un indice
- ✓ Il peut être primaire ou réplica

Cerebro :

- ✓ Outil communautaire
- ✓ Suivi temps réel du cluster
- ✓ Visualisation graphique des configurations
- ✓ Réalisation des tâches d’administration

https://github.com/lmenezes/cerebro

1er exemple (suite) : les bureaux de votes

Les données ont été chargées dans Elasticsearch sous un format propre.

Les données sont réparties sur 4 shards d'un indice du cluster.

La structure permet d'indexer et requêter les données sur l'ensemble des critères.

http://localhost:9000 pour vérifier avec Cerebro.

kibana

Discover

- ✓ Données brutes

Visualize

- ✓ Tableaux
- ✓ Graphiques
- ✓ Informations

Dashboard

- ✓ Vue complexe

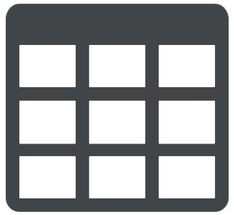
Beaucoup de visualisations

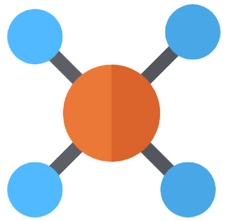
kibana



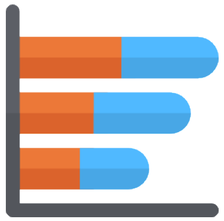





1,234

Small Medium Large Words






2ème exemple : l'éclairage public

La source de données est disponible ici :
https://opendata.paris.fr/explore/dataset/eclairage-public

Ce jeu de données contient la localisation de toutes les lampes de Paris.

Pour le charger avec Logstash, utiliser le script _StartLogstash_eclairage_publique.ps1.

Kibana permet d'explorer ces données pour mieux comprendre l'éclairage parisien.
http://localhost:5601

3ème exemple : les résultats électoraux

La source de données est disponible ici :
https://opendata.paris.fr/explore/dataset/resultats_electoraux

Ce jeu de données contient tous les derniers résultats électoraux.

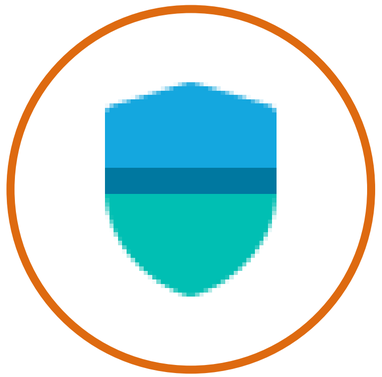
Pour le charger avec Logstash, utiliser le script _StartLogstash_resultats_electoraux.ps1.

Les données ne contiennent pas la géolocalisation. Elle est ajoutée à partir des bureaux du 1er exemple.

159 121 documents

x-pack

Security

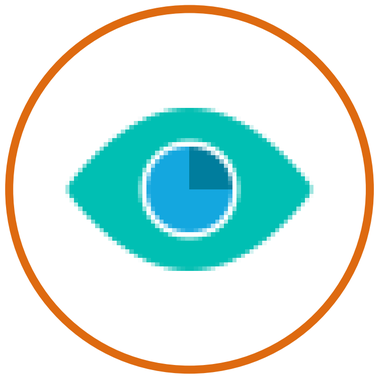


- ✓ Accès
- ✓ Cryptage

Alerting

- ✓ Surveillance

Monitoring

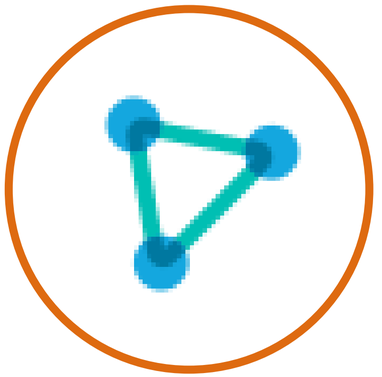


- ✓ Suivi de performance

Reporting

- ✓ Export PDF

Graph

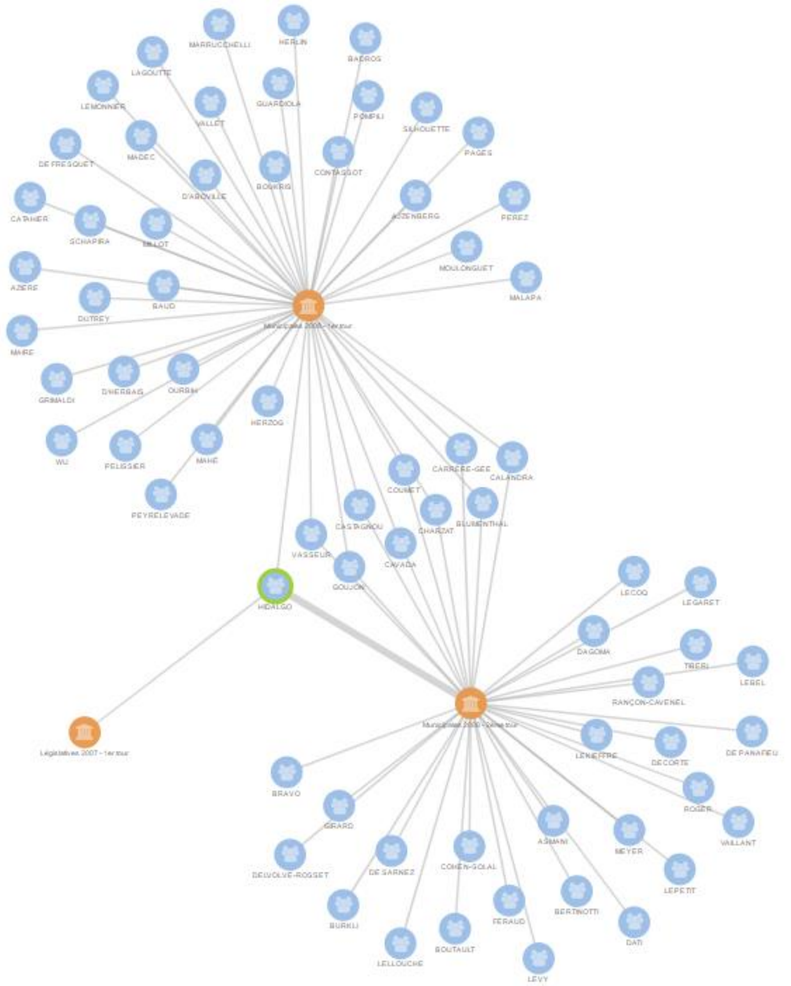


- ✓ Graphes

Machine Learning

- ✓ Détection
- ✓ Analyse

3ème exemple (le retour) : les résultats électoraux



A partir des mêmes données, on peut réaliser des graphes.

L’objectif, ici, va être de montrer les élections auxquelles s’est présentée Anne Hidalgo depuis 2007.

4ème exemple : le bilan électrique

La source de données est disponible ici :
https://data.enedis.fr/explore/dataset/bilan-electrique-demi-heure

Ce jeu de données contient toutes les informations sur le réseau électrique français par demi heure.

Pour le charger avec Logstash, utiliser le script _StartLogstash_bilan-electrique-demi-heure.ps1.

A partir de ces données, on peut utiliser le module de machine learning permettant de détecter les anomalies.

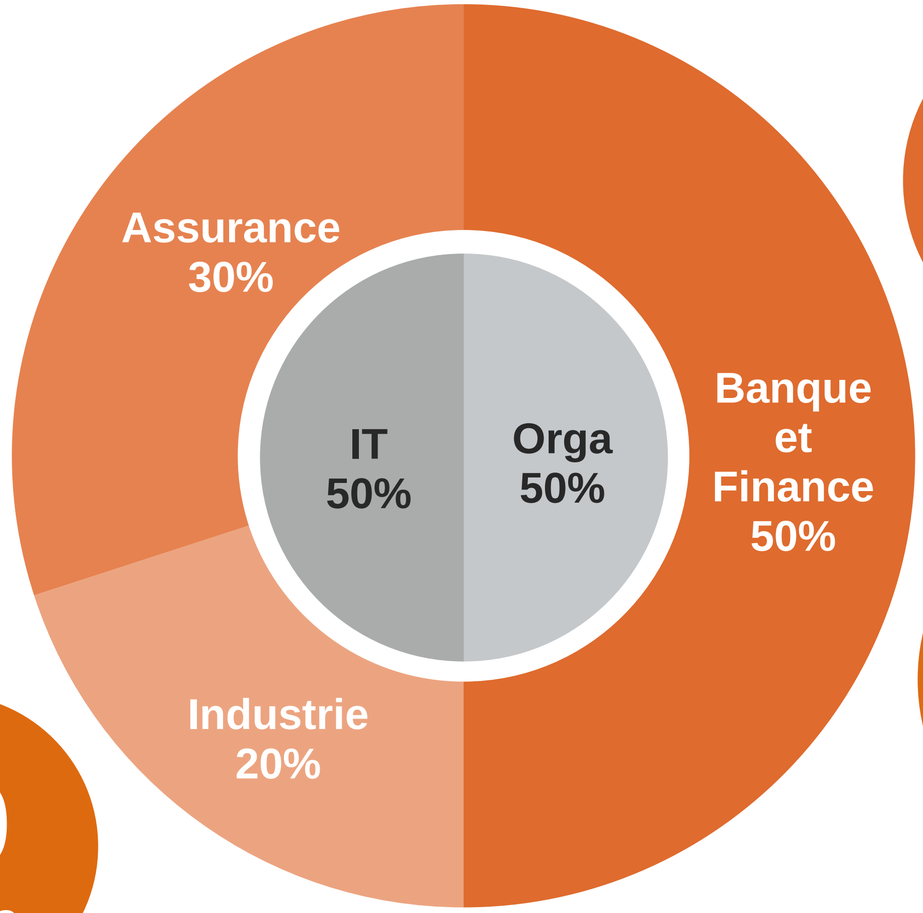
87 648
documents

En quelques chiffres

2002
Date de création

200
Personnalités

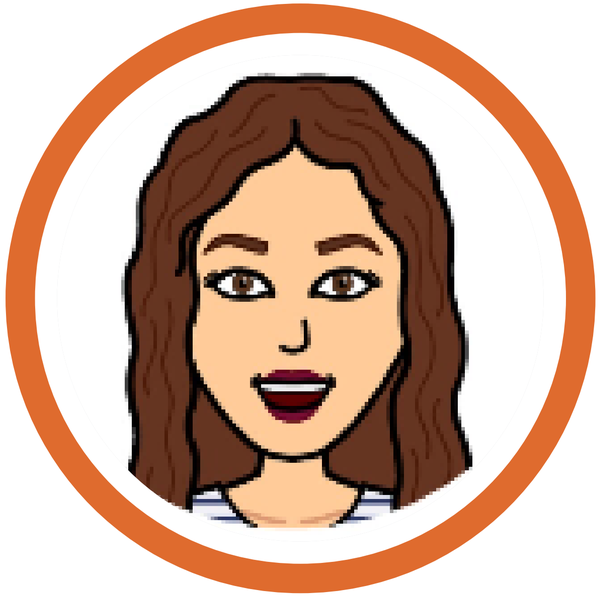
40
Clients

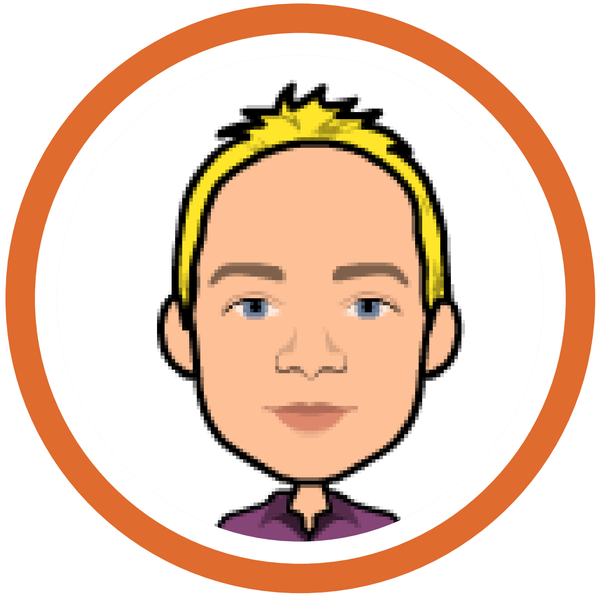


23
M€ de CA

35
Cercles

Si vous avez des questions





mmesnage@aneo.fr

cadam@aneo.fr