COLLECTE DE DONNÉES SUR

Projet BIG DATA

Etudiants:
Ibrahima DIATTARA
Rednic AMOUZOUN
Abdoul SY
Yasmine MARICAR
Elhadj Boubacar SOW

M2 Innovation, Market and Data Science

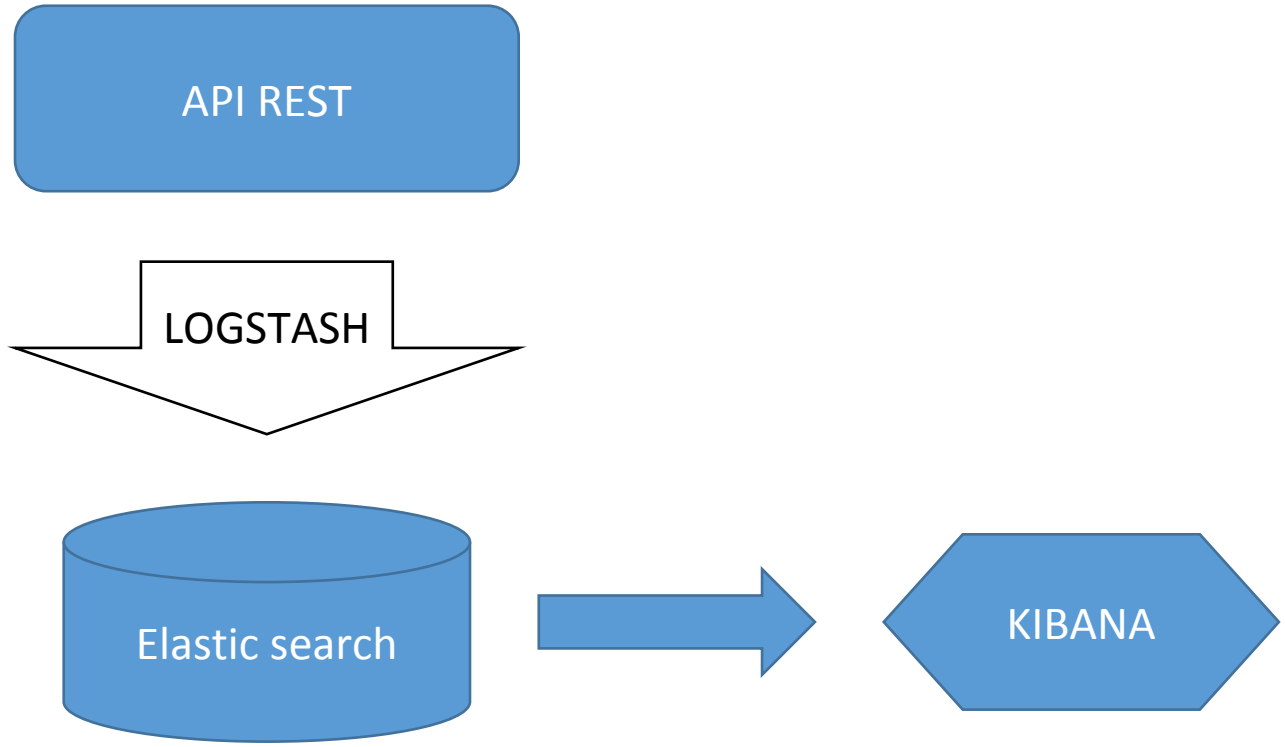
Professeur:
Aymen GHADGHADI

Plan

Introduction

- Récupérer les données des opérateurs mobiles FR sur Twitter
- Stocker en temps réel
- Visualiser les données
- Archiver pour les analyses de sentiment

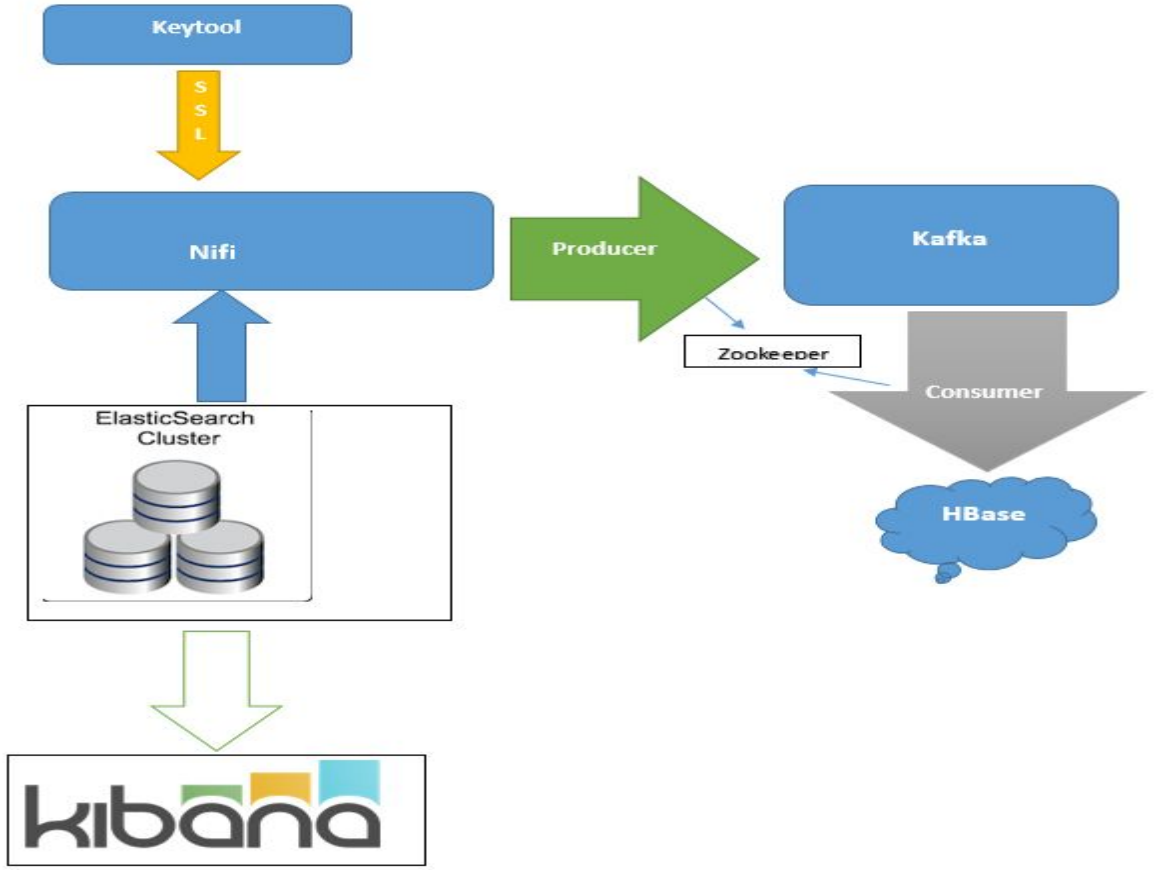
Architecture streaming



Estimation données par jour 1Go de log par jour => 30 Go par mois et 1 shard

Suppression des données par mois et 1 réplica => 30*2 = 60GO/mois stockage dans Elasticsearch

Architecture Data in REST



Elastic search

- Recherche en quasi temps réel
- Scalabilité, Haute disponibilité
- Automatiquement sauvegardé et repliqué
- Index
- shard

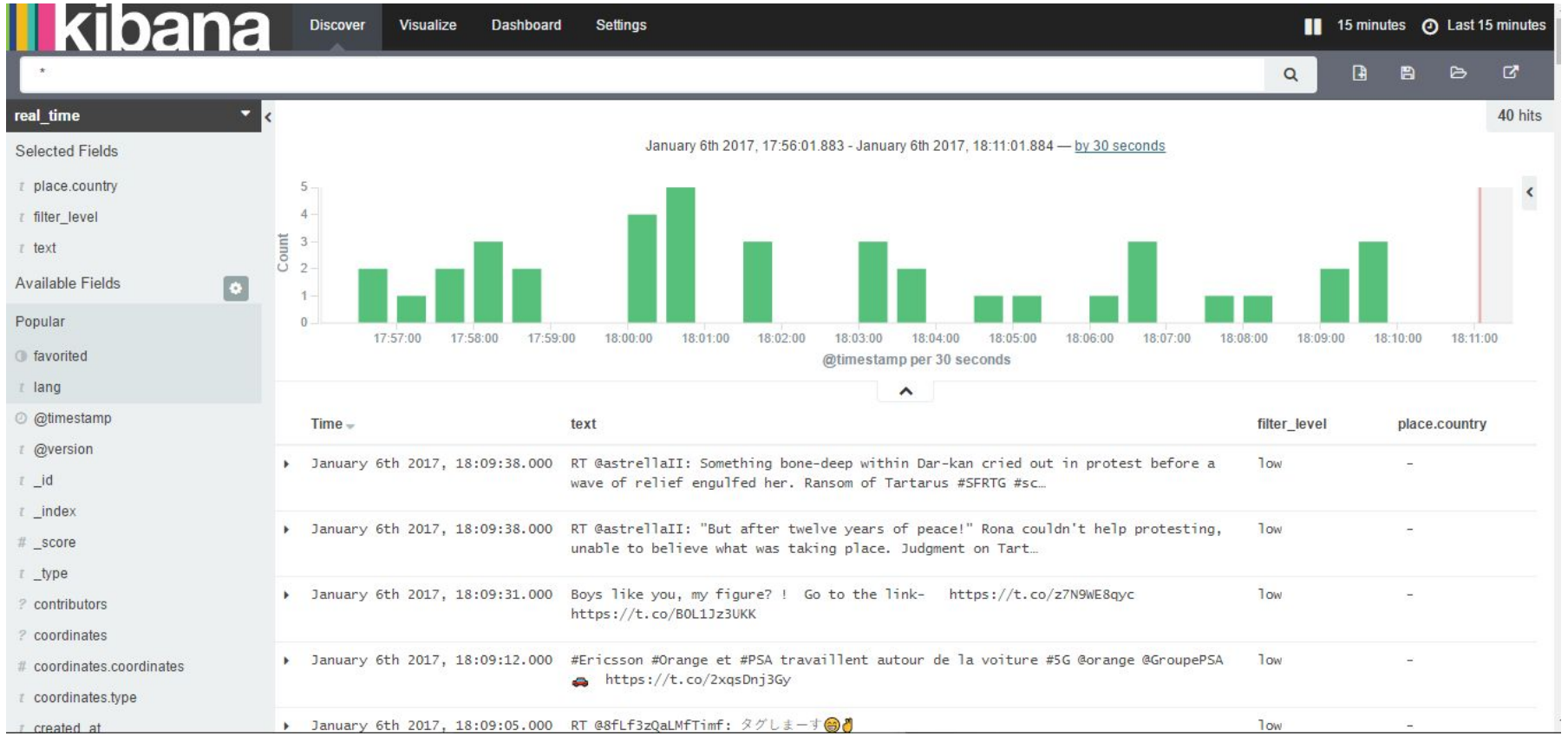
Logstash

- Collecte de données multi-sources
- Structurées
- Semi-structurées
- Non-structurées

Métadonnées

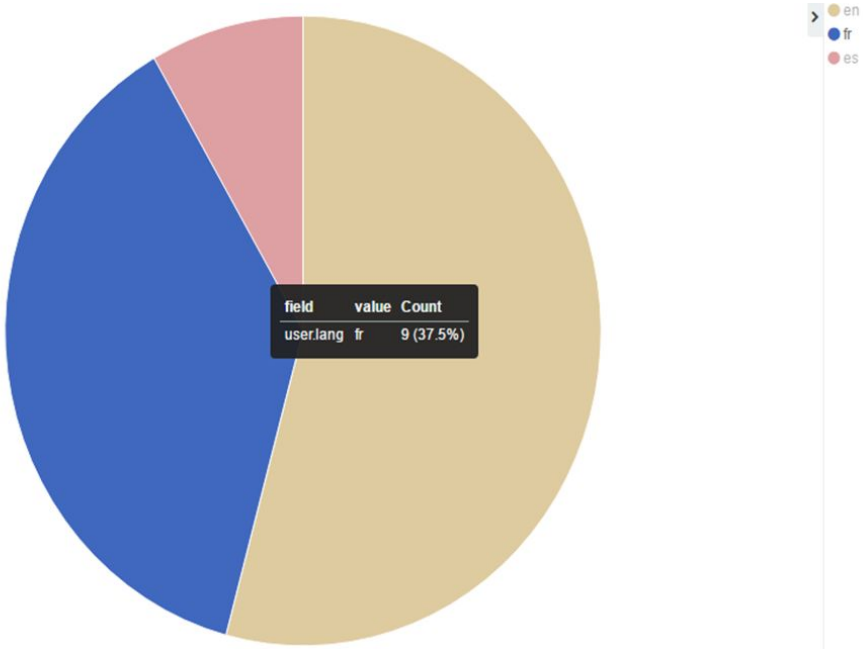
- .text: contenu du tweet
- .created_at: date et heure UTC de création du tweet
- .source: support utilisé
- .user.followers_count: nombre d'abonnés
- .user.friends_count: nombre de comptes suivis
- Etc…

Visualisation sur kibana

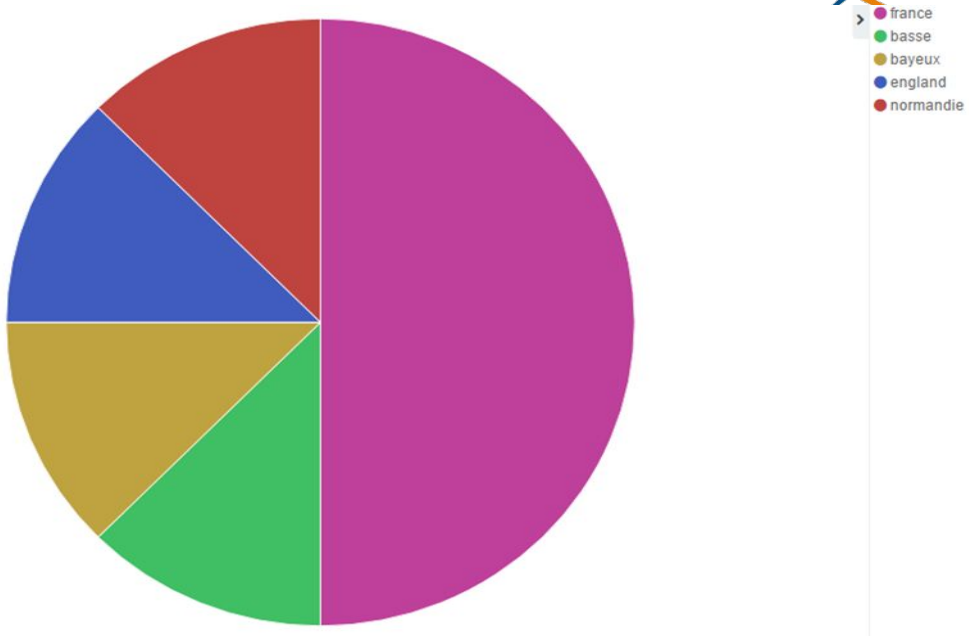


- 40 tweets récupérés sur les dernières 15 min le 6 janvier à 17h56

Visualisation sur kibana



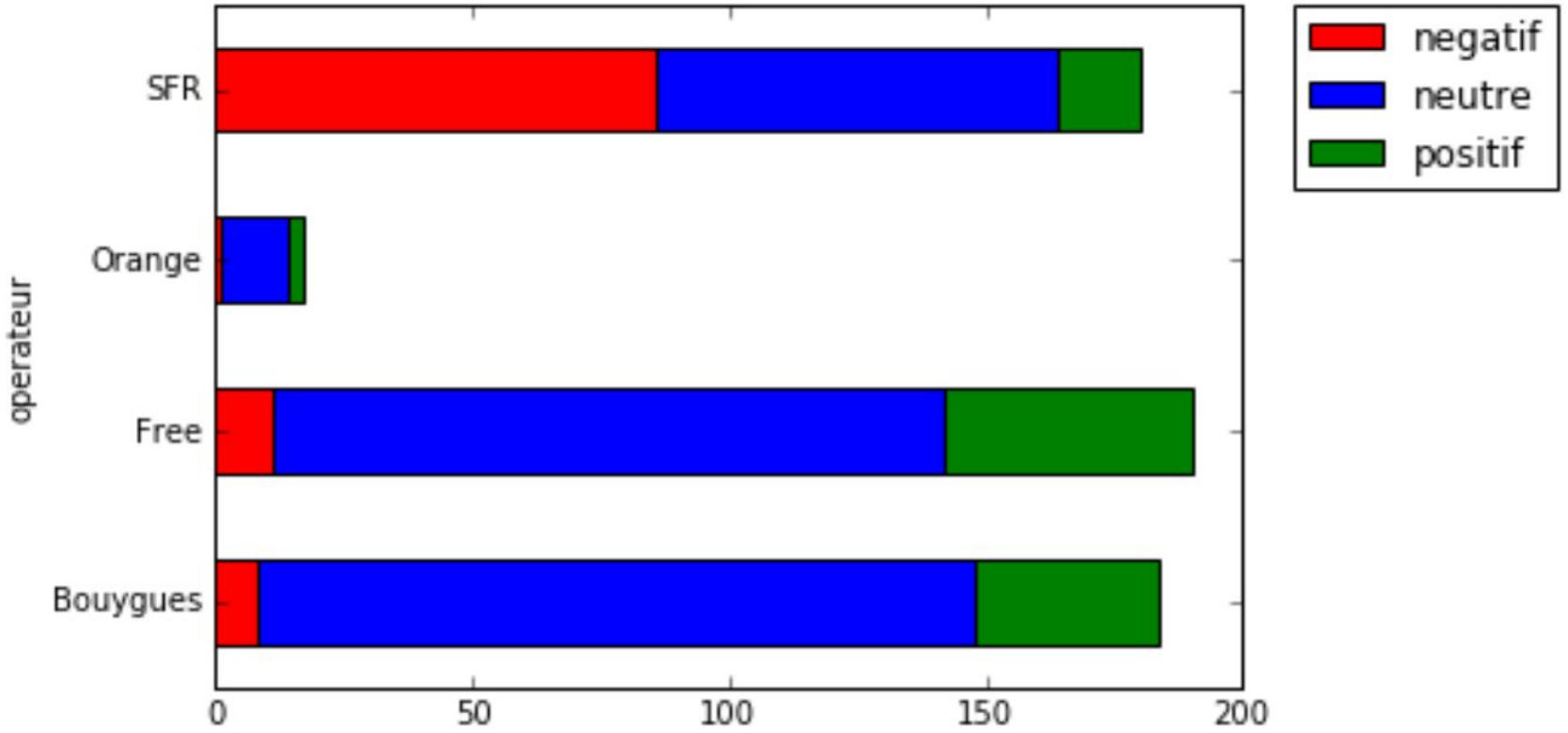
Ici nous avons la répartition des tweets par langue:
38% sont en français et un peu plus de 50% en Anglais



Ici nous avons la localisation des tweets:
Environ 60 % ont été tweetés en France

Analyse de sentiments sur les tweets

- Cette analyse, nous l'avons faite de manière statique sous Python: les données d'ES ont été exportées en Json puis importées sous python



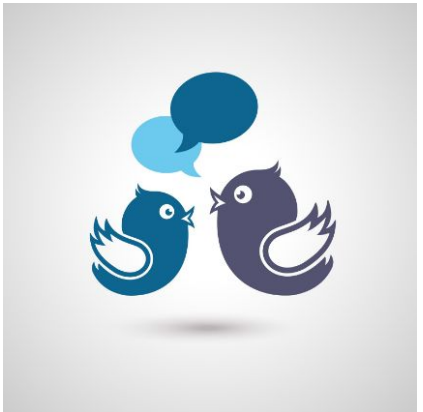
Tarification Elastic cloud

- **Summary**
- Reserved memory : **2 GB**
- Reserved storage : **48 GB**
- High availability : Yes
- **$173/month** ($147.05/month with 15% annual discount)
- Stack ELK : **Gratuit** ou <10000 €/an avec support

- **Cluster configuration**
- Total no. of data nodes: **2 (4 GB)**
- No. of primary nodes: **1 (2 GB)**
- No. of replica nodes: **1 (2 GB)**
- Tiebreaker node: 1
- Data centers: 2
- *https://www.elastic.co/fr/cloud/as-a-service/pricing*

Enjeux commerciaux

- Notre produit répond à des enjeux réels au sein des entreprises qui doivent contrôler de plus en plus leur e-réputation
- L’exploitation des données en temps réel est au cœur de la stratégie marketing de celles-ci.



Difficultés rencontrées & Extension Proposée

Étape	Description	Problèmes rencontrées	Résolution	Temps estimé	Temps réel	Statut
1	Architecture	-Estimer le nombre de logs par jour	-Faire un process background	10 jours	12 jours	OK
2	Développement	Environnement non stable	Réactualisation régulière	5 jours	7 jours	OK
3	Finance	Estimer le coût de la solution en production finale.	Chercher sur les sites revendeurs et contacts support des outils.	5 jours	5 jours	OK
4	Analyse	-Le #orange qui n'était pas uniquement relatif au réseau téléphonique (à la couleur ou au fruit par exemple) -l'incapacité de faire une analyse de sentiments en temps réel	Nous avons juste considéré le #orange en prenant le risque d'avoir des tweets n'étant pas liés à l'entreprise...	1 jour	1 jour	Ok

Conclusion