

ESUP-Days  
#19

05.02.2015  
Paris



# Présentation Elasticsearch

Nicolas CAN -- *Université Lille1*



# Présentation Elasticsearch

- ❖ Introduction / Présentation
- ❖ Installation / fonctionnement
- ❖ Indexation / mapping
- ❖ La recherche
- ❖ Le résultat
- ❖ Et plus encore...



# Introduction / Présentation

Moteur de recherche basé sur Lucene :

- ❖ Distribué
- ❖ RestFull
- ❖ Associé à Kibana / logstash
- ❖ OpenSource
- ❖ De plus en plus utilisé sur le web (wikipédia, stackoverflow, github...)



# Installation / Usage

- ❖ Ecrit en java (il faut une version 1.7 ou 1.8 pour l'utiliser)
- ❖ Disponible directement en téléchargement, il existe également des .deb ou .rpm
- ❖ Se clusterise par défaut pour partager les documents et requetes (nodes/shards port 9300)
- ❖ Tout un ensemble de plugin (kopf et marvel)
- ❖ API RESTful avec JSON via HTTP sur port 9200
- ❖ Nombreux clients officiels en java, php, js, python etc.



# Indexation / Mapping

- ❖ Refresh : tâche de fond qui ajoute les nouveaux documents à l'index de recherche
- ❖ Chaque document est découpé en différents champs
- ❖ Mapping : squelette du document
  - \* Type de champs (bool, date, number, string etc.)
  - \* Analyzer
  - \* Valeur exacte ou full-text
- ❖ Analyzer : traitement du champs (indexation/recherche)
  - \* Filtre
  - \* Tokenizer



# La recherche

- ❖ Query DSL : langage de recherche complet
  - ❖ Basé sur un objet json pour définir les requetes
  - ❖ Les requetes peuvent etre imbriquées et peuvent contenir des queries ou des filtres
  - ❖ Les requetes :
    - \* Recherche en texte intégral
    - \* Résultat dépend d'un score de pertinence
    - \* Type : match , multi match
  - ❖ Les filtres
    - \* Recherche oui/non
    - \* Recherche valeur exacte (term filter -> contient, range etc.)
- Sont mis en cache (sauf script filter et geo\_filter)



# Le résultat

- ❖ Les agrégations (anciennement appelés facets)
  - \* Regroupement de documents communs
  - \* Séparé en deux familles :
    - Les buckets
    - Les metrics
  - \* Tout un ensemble d'opération : min, max, sum, avg, date etc.
- ❖ Le scoring
  - \* Pertinence d'une recherche (query)
  - \* Basé sur :
    - Term Frequency
    - Inverse Document Frequency



# Mais aussi...

- ❖ Multilingue
- ❖ Auto-compléteur
- ❖ More Like This
- ❖ Recherche géographique
- ❖ Protection des données avec Shield
- ❖ Etc.

