Séminaire du pôle eLearning de

l'Institut de la Société Numérique (ISN)

24-25 mai 2016

Analyse et visualisation de traces (log) issues du MOOC efSUP

exemple d'utilisation de la suite Elasticsearch-Kibana

Tony DOAT

---

stef — ENS CACHAN — université PARIS-SACLAY — ENS ÉCOLE NORMALE SUPÉRIEURE DE LYON — ifé INSTITUT FRANÇAIS DE L'ÉDUCATION

Contexte

stef

- Des traces d'activité (*log*)
  - Les MOOCs sont accessibles par un site web
  - Historique des événements dans un fichier *log*
  - Le fichier *log* contient des traces de l'activité des utilisateurs de MOOC
  - L'analyse de ces fichiers peut fournir des informations sur le comportement de/des utilisateur(s)

  - Contraintes
    - Données générées en dehors de tout modèle d'apprentissage (non prescrit et non exhaustif)
    - Format et signification des données non connus *a priori*
    - Données potentiellement seulement semi-structurées
    - Données potentiellement massives

- Outil choisi pour les tests
  - Pile logicielle Elasticsearch (utilisé par : Blablacar, Facebook, Ebay, Orange, etc.)

Objectifs et problématiques de l'étude

stef

- Objectif
  - Tester la suite Logstash-Elasticsearch-Kibana (ELK) sur des données du MOOC Enseigner et Former dans le Supérieur (efSUP) accessible *via* France Université Numérique (FUN)

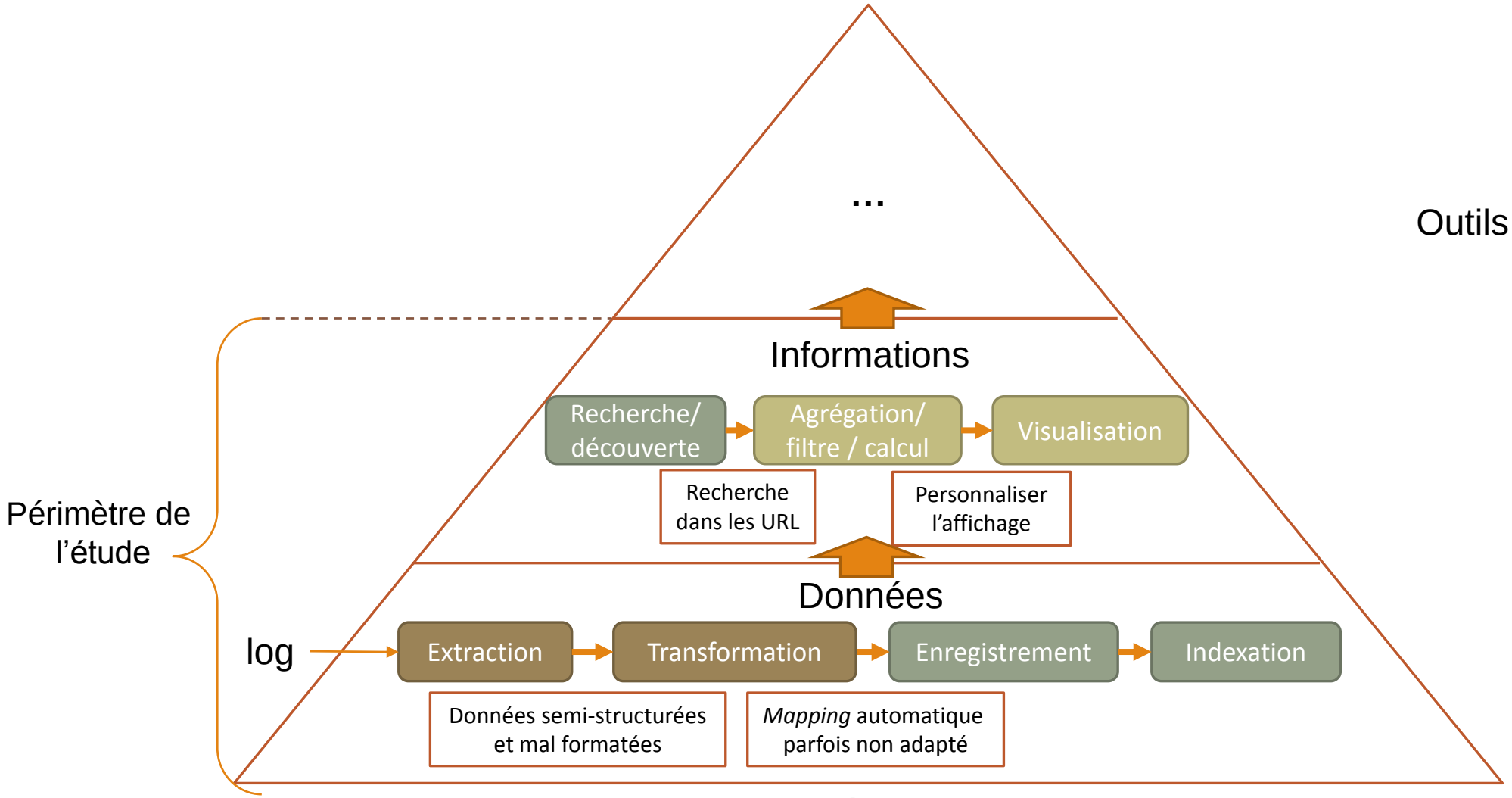
- Problématiques
  - Comment insérer les événements de log dans Elasticsearch (phase ETL) ?
  - L'outil est-il efficient pour la découverte de données ?
  - Quelles analyses et visualisations sont disponibles ?

Caractéristiques de cette étude

stef

- Les données
  - Semaine 1 du MOOC efSUP (plateforme EdX, du 7 mars au 13 mars 2016)
  - Format : *json*
  - Particularités des données : incomplètes
- La suite ELK
  - Elasticsearch fût créé en 2004 par Shay Banon
  - Les propriétés
    - Moteur de recherche distribué (gère la montée en charge)
    - Orienté document (schéma de données flexible)
    - Outil web interrogeable par requêtes HTTP (API REST)
    - Gère la collecte, le stockage, la découverte, l'analyse et la visualisation des données
  - Versions utilisées
    - Elasticsearch : version 2.3.1 – 4 avril 2016
      - serveur utilisant Lucene (version 5.5.0) pour l'indexation et la recherche des données
    - Kibana : version 4.5.0 – 30 mars 2016

Démarche

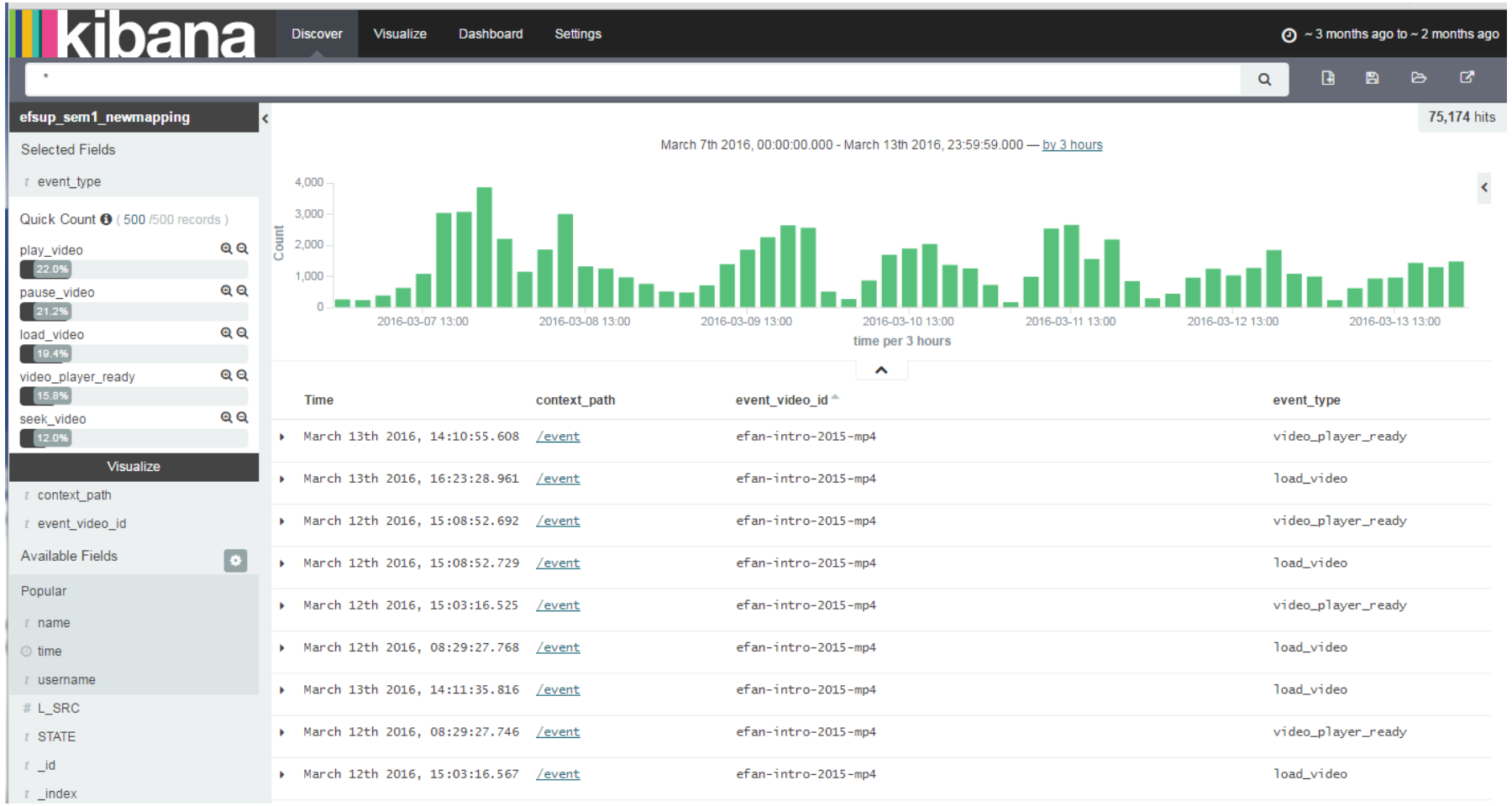


Outils utilisés

- Kibana
- Elasticsearch
- Logstash / algorithme (Python)
- Difficultés rencontrées

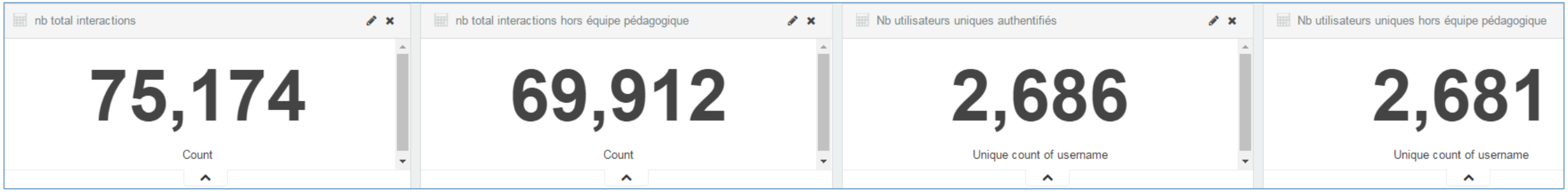
-> Logstash non adapté (version 1.5.1 – 16 juin 2015)

Résultats : exemple de découverte de données

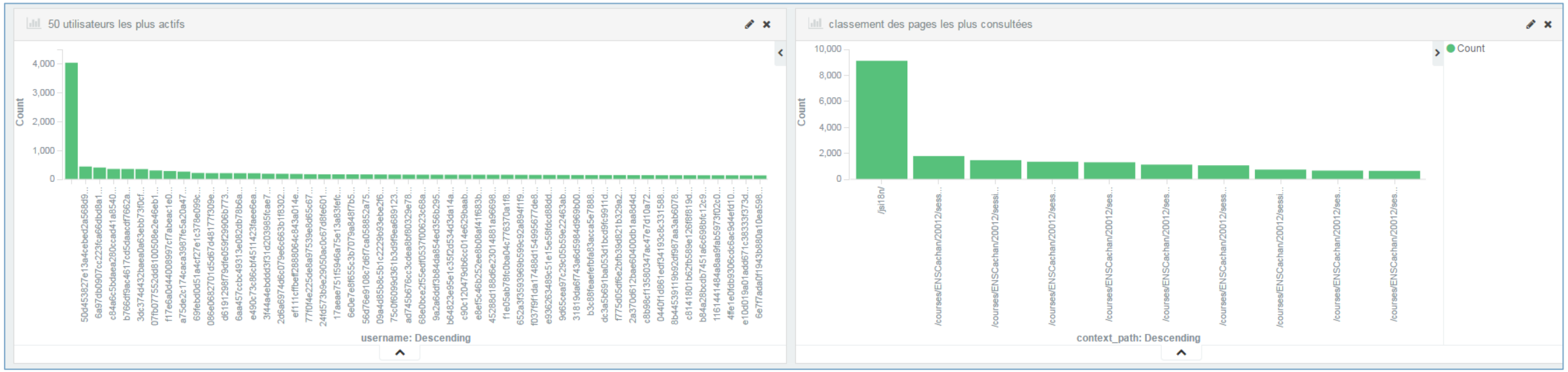


découverte du lien entre les identifiants de vidéo, les types d'événements et le contexte à partir des données

Résultats : exemples de rendus possibles



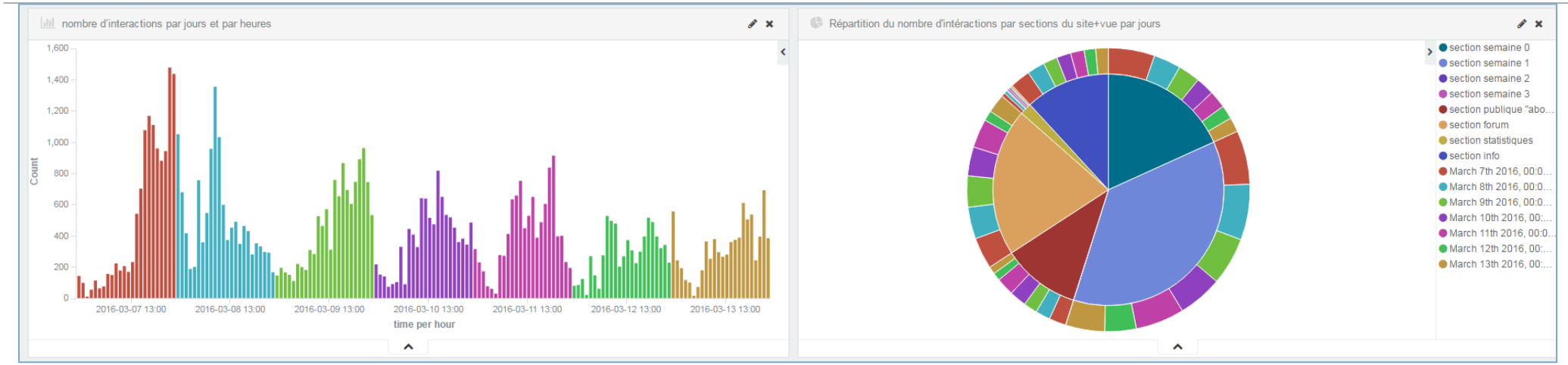
exemples de métriques



classement des utilisateurs les plus actifs

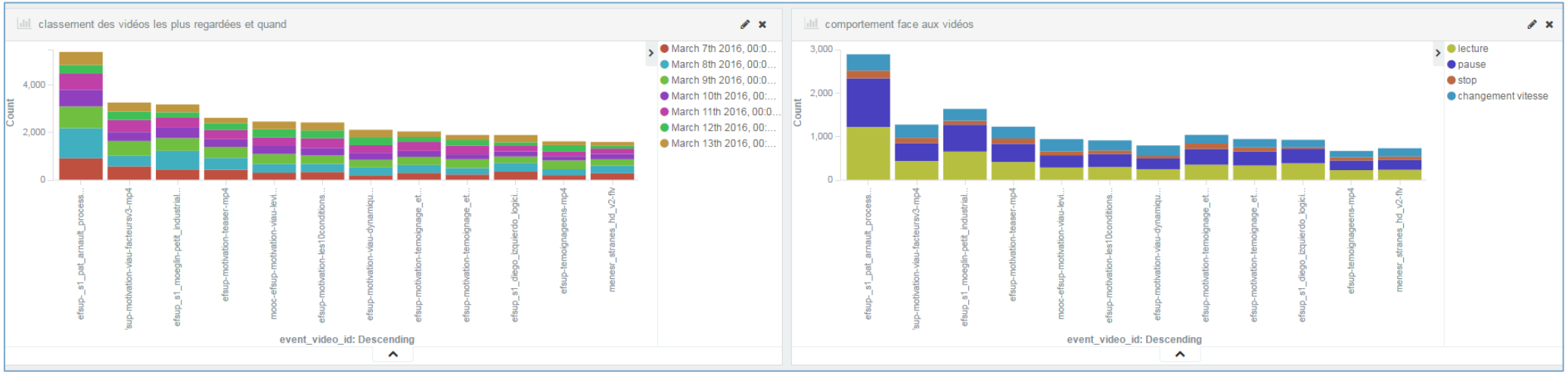
classement des pages les plus consultées

Résultats : exemples de rendus possibles



évolution du nombre d'interactions par heure et par jour

répartition du nombre d'interactions par section du site puis par jour



évolution du nombre d'interactions par vidéo et par jour

évolution du nombre d'interactions par vidéo et par type

En bref

stef

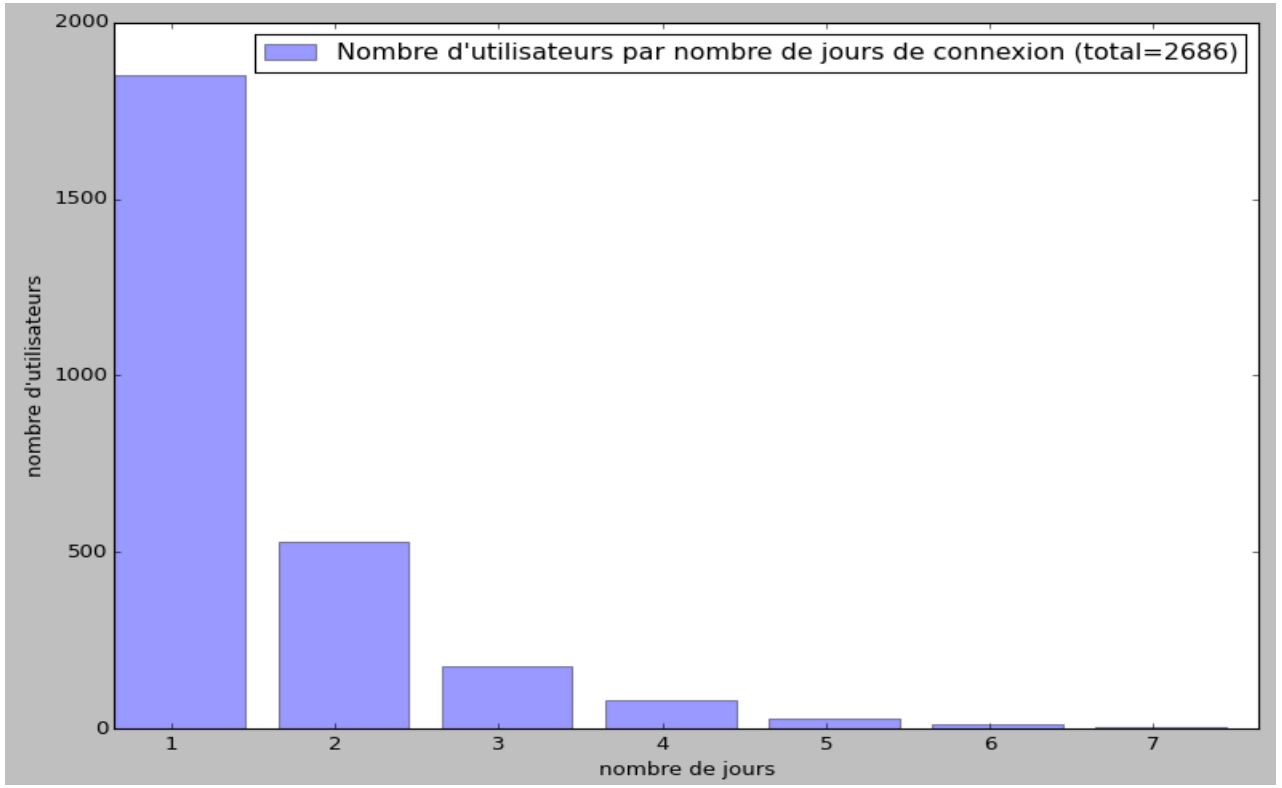
- Avantages d'ELK
  - Efficacité pour interroger les données (découverte)
  - Efficacité pour générer des graphiques à plusieurs niveaux de valeurs
  - Les possibilités de requêtes sur des valeurs de champs (« expressions régulières »)
  - Possibilité de mise à jour de graphiques en temps-réel
  - Possibilité de créer des tableaux de bord actionnables
  - Analyse simple des données (classement, filtre, min/max)

- Désavantages d'ELK
  - Impossibilité de réaliser des analyses complexes sur les données (pas d'écriture d'algorithmes ni de fonctions statistiques avancées)
  - Elasticsearch n'est pas une base de données (ni un entrepôt de données ; des données peuvent être perdues durant l'insertion)

Pour aller plus loin

stef

- Couplage hybride
  - Exemple avec ELK + Jupyter (script Python, rendu avec Matplotlib)



Nombre d'utilisateurs pour chaque nombre de jours d'interaction

Nota : toutes les interactions sont prises en compte, et non uniquement la section des vidéos de cours

Conclusion et perspectives

- Conclusion
  - Outil efficient pour la découverte de données (massives)
  - Permet de générer des graphiques contenant plusieurs niveaux d'information
  - Création de tableaux de bord temps-réel partageables et actionnables

- Perspectives
  - Cas de l'enrichissement des données
  - L'outil serait-il compatible avec l'utilisation d'un modèle d'apprentissage ?
  - Description des données au format RDF ?

FIN

stef

Merci de votre attention !

Références

- MOOC efSUP : https://www.fun-mooc.fr/courses/ENSCachan/20012/session01/about
- FUN : https://www.fun-mooc.fr/
- Suite logicielle ELK : https://www.elastic.co/
- Performances d'Elasticsearch : https://benchmarks.elastic.co/index.html