Plateforme Elastic Stack dans le Cloud pour l'analyse de données issues de l'Open Data

ASR 2018-2019

Auteurs

Houssem TOUANSI

Ghailene SOUISSI

Encadrants

Cédric JOLY

Amel BOUZEGROUB

Technologies

Contexte , Problématique et Objectifs

Contexte

- L'Open Data devient de plus en plus un sujet incontournable
- Les données sont devenu une source d'innovation dans de nombreux domaines .
- Certains organismes mettent à disposition de très grande masse de données pour qu'elles soient explorées, manipulées…

Problématique

- Préparation des données : extraction, nettoyage, fusion …
- Stockage des données volumineuses
- Analyse et exploitation des données

Objectifs

- Conception d'une architecture qui prend en charge tout le processus ETL (Extraction , transformation , load) basé sur le web scraping
- Visualisation des données et optimisation de la consommation de ressources
- Cas d'utilisation : Offres d'emplois et de stages issues de nombreuses sources (Jobteaser , jobetudiant …..)

Architecture

❑ **Conception de trois clusters dans le Cloud :**

✓ Le cluster 1 comporte **10 instances (2 cœurs et 4 GO de RAM)**, il est responsable de la partie préparation de données (Dynamique)

✓ Le cluster 2 comporte **3 instances (2 cœurs et 4 GO de RAM)**, 2 nœuds master et 1 nœud esclave, il responsable du stockage et de la visualisation de données (Statique)

✓ Le cluster 3 comporte **une seule instance (12 Go de RAM et 6 cœurs),** dans laquelle est déployé **logstash** , il est responsable de charger, traiter, filtrer les fichiers csv et les envoyer vers **Elasticsearch** (Statique)

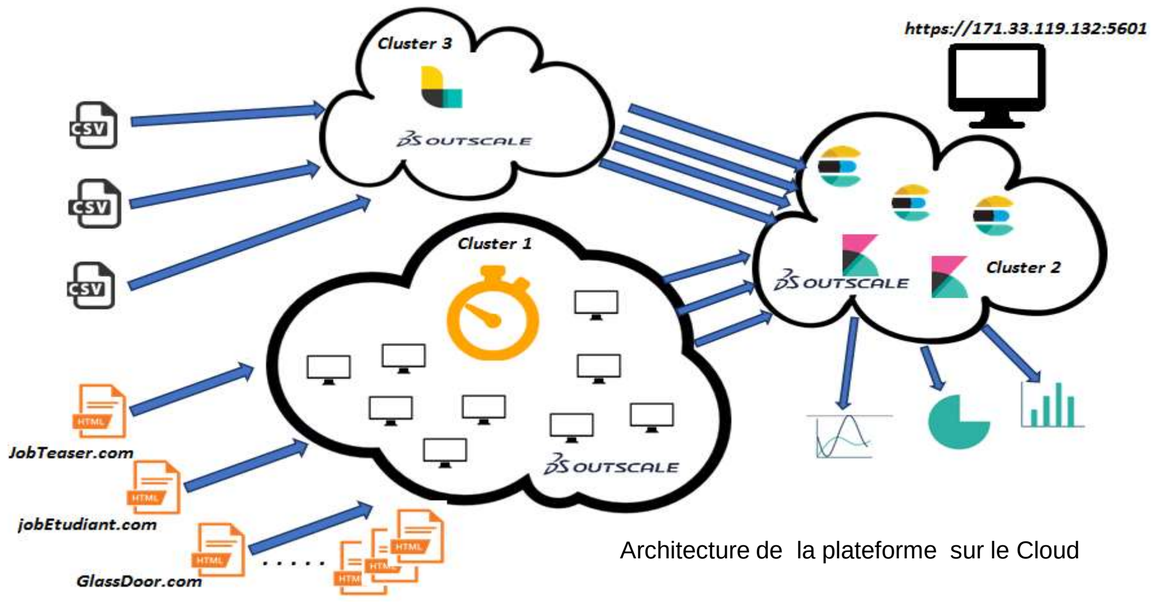
❑ **Automatisation** du processus de préparation de données pour optimiser la consommation de ressources cloud : création et terminaison des instances qui font le web scraping

3 phases :

✓ **Phase 1** : La création des instances

✓ **Phase 2** : L'exécution des scripts de web scraping par ces instances

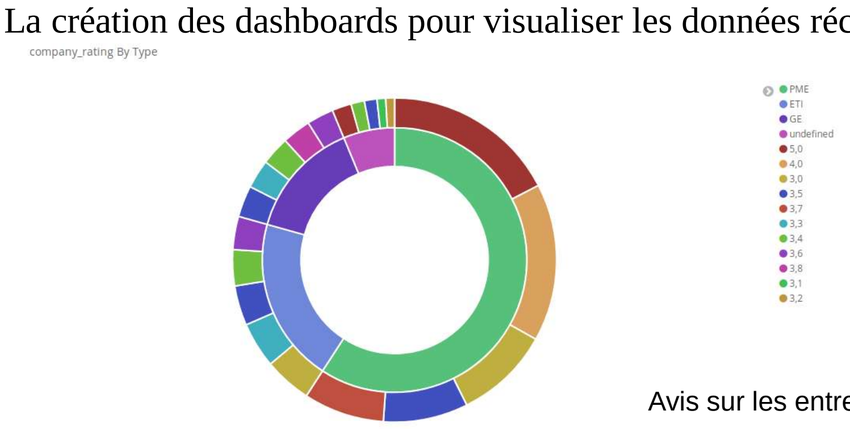
✓ **Phase 3** : La terminaison des instances



Architecture de la plateforme sur le Cloud

Visualisation des données

La création des dashboards pour visualiser les données récoltées et montrer la création de valeur.

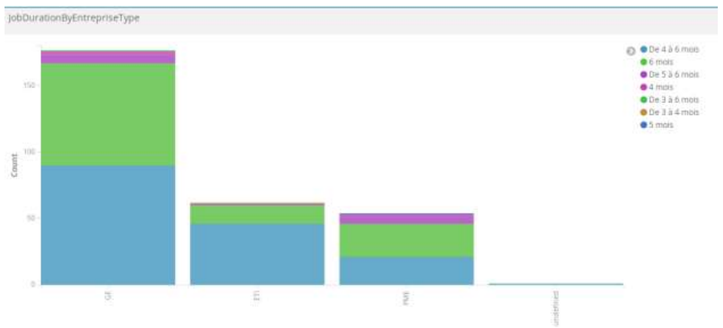


Avis sur les entreprises par type (GE,PME ..)

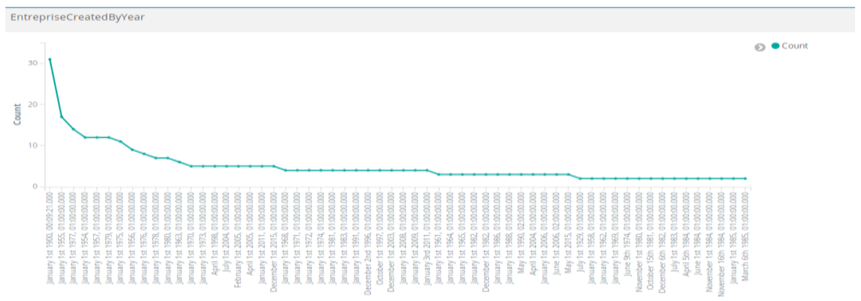
JobsByEntrepriseAndRating

company_name.keyword: Descending	category.keyword: Descending	rating.keyword: Descending	Count
dassault systemes	GE	3,8	54
atos	ETI	3,4	32
valeo	GE	3,1	26
cgi france	GE	3,4	19
sopra steria	GE	3,3	17
ibm france	GE	3,6	13
accenture	GE	3,7	10
avanade	GE	3,5	9
dxc technology	ETI	2,7	9
alten	GE	3,1	8

Nombre de stages par entreprise et score d'entreprise



Durée de stage par type d'entreprise



Nombre d'entreprises créées par an

Contact Houssem.touansi@telecom-sudparis.eu
Ghailene.souissi@telecom-sudparis.eu

Site web