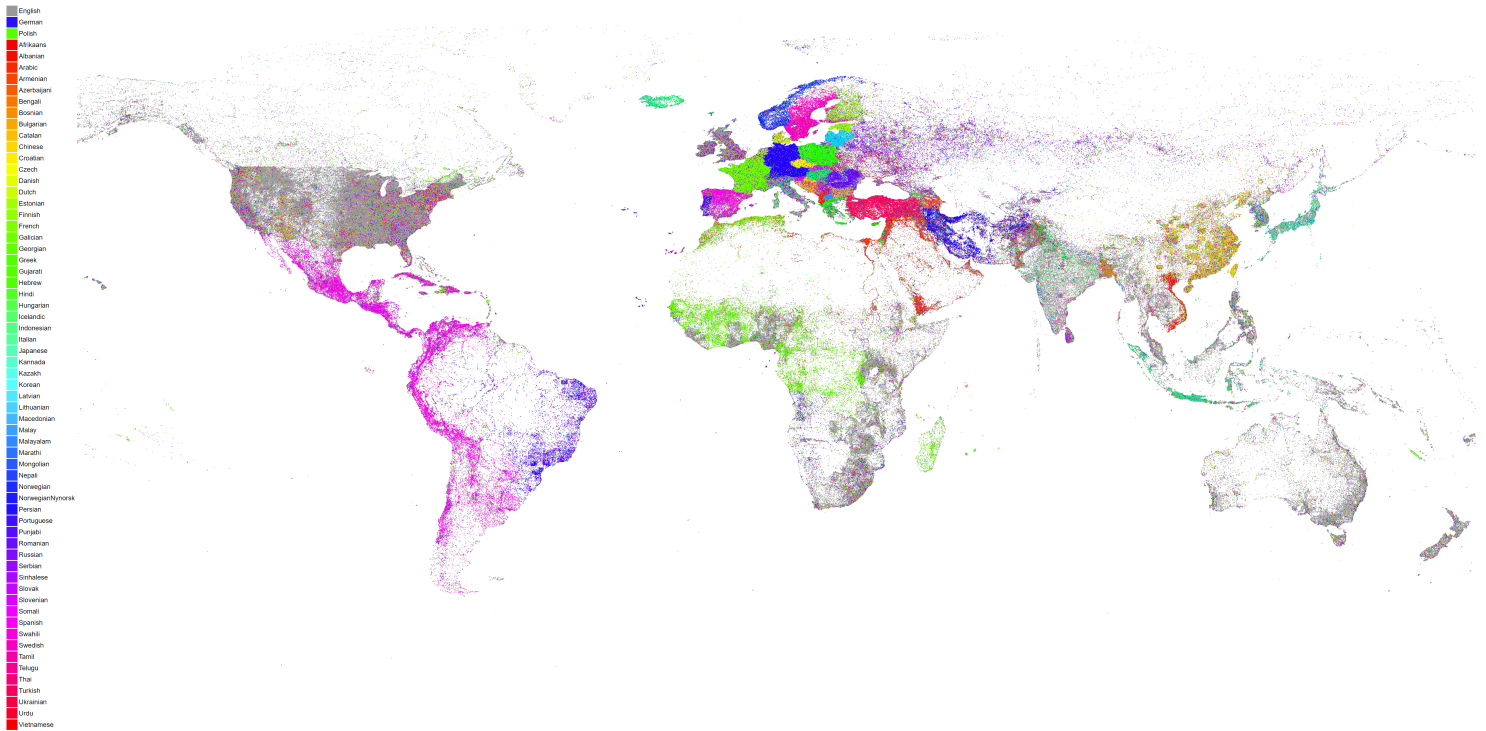


Projet Bigdata 2020-2021: Etudier l'évolution de la pandémie COVID19 via son impact media

Intro

“ *The Global Database of Events, Language, and Tone (GDELT)* (<https://www.gdeltproject.org/>) monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

L'objectif du projet est de concevoir un système qui permet d'analyser l'évolution de la pandémie COVID19 via sa couverture médiatique. Dans ce but nous allons utiliser le jeu de données GDELT.



Locations mentioned in global news coverage monitored by GDELT 2015-2018, colored by the primary language of coverage mentioning each location (*Seeing The World Through The Eyes Of Others: Mass Machine Translation*, [KALEV LEETARU](https://www.forbes.com/sites/kalevleetaru/2018/11/24/seeing-the-world-through-the-eyes-of-others-mass-machine-translation/#3ae501bd2c8a) (<https://www.forbes.com/sites/kalevleetaru/2018/11/24/seeing-the-world-through-the-eyes-of-others-mass-machine-translation/#3ae501bd2c8a>)

Contexte

A. Jeu de données

Nous allons utiliser:

- les events ([CAMEO Ontology](http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf) (<http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf>), [documentation](http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf) (<http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf>))
- les mentions ([documentation](http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf) (<http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf>))
- le graph des relations = GKG, Global Knowledge Graph ([documentation](http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf) (<http://data.gdeltproject.org/documentation/GDELTEventCodebookV2.0.pdf>))

Les fichiers du jeu de données sont indexé par deux fichiers:

- [Master CSV Data File List – English](http://data.gdeltproject.org/gdeltv2/masterfilelist.txt) (<http://data.gdeltproject.org/gdeltv2/masterfilelist.txt>)
- [Master CSV Data File List – GDELTE Translingual](http://data.gdeltproject.org/gdeltv2/masterfilelist-translation.txt) (<http://data.gdeltproject.org/gdeltv2/masterfilelist-translation.txt>)

Pour plus d'infos consulter la [documentation](https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/). (<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>)

Le jeu de données de GDELTE v2.0 est disponible également sur [Google BigQuery](https://www.gdeltproject.org/data.html#googlebigquery).

(<https://www.gdeltproject.org/data.html#googlebigquery>). Cependant vous ne devez pas l'utiliser directement pour votre projet. Vous pouvez cependant l'utiliser pour explorer la structure des données, la génération des types de données ou utiliser des données connexes (ex codes pays etc...) .

Objectif

L'objectif de ce projet est de proposer un système de stockage distribué, résilient et performant sur AWS pour répondre aux questions suivantes:

- a. afficher le nombre d'articles/événements qui parlent de COVID qu'il y a eu pour chaque triplet (jour, pays de l'évènement, langue de l'article)
- b. pour un pays donné en paramètre, affichez les événements qui y ont eu place triés par le nombre de mentions (tri décroissant); permettez une agrégation par jour/mois/année
- c. pour une source de données passée en paramètre (gkg.SourceCommonName) affichez les thèmes, personnes, lieux dont les articles de cette source parlent ainsi que le nombre d'articles et le ton moyen des articles (pour chaque thème/personne/lieu); permettez une agrégation par jour/mois/année.
- d. est-ce qu'on observe des patterns dans l'évolution qui pourraient nous permettre d'identifier la prochaine vague/pandémie

C. Contraintes

1. Vous devez utiliser **au moins 1 technologie vue en cours** en expliquant les raisons de votre choix (SQL/Cassandra/MongoDB/Spark/Neo4j).
2. Vous devez concevoir **un système distribué et tolérant aux pannes** (le système doit pouvoir continuer après la perte d'un nœud).
3. Vous devez pré-charger **une année de données** dans votre cluster
4. Vous devez utiliser **AWS** pour déployer le cluster.

D. Les livrables

Vous devrez fournir:

- une archive avec votre code source (ou un lien sur github...)
- une courte présentation de votre architecture, modélisation, les avantages et inconvénients, des choix de modélisation et d'architecture, volumétrie, limites et contraintes (max 10 slides de présentation)

F. Organisation

Vous travaillerez par groupe de 4-5 personnes. La soutenance se déroulera de la manière suivante:

1. Présentation: 10 minutes
2. Démo: 10 minutes
3. Questions & Réponses : 10 minutes



Lors de la soutenance, les données devront être préalablement chargées dans votre cluster. Vous devez démontrer la résilience de votre système de stockage en tuant un noeud de votre cluster.

Ressources

[GDELT v2.0 dataset description](https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/) (https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/)

[GDELT Translingual](https://blog.gdeltproject.org/gdelt-translingual-translating-the-planet/) (https://blog.gdeltproject.org/gdelt-translingual-translating-the-planet/)

[Mapping the Linguistic Geography Of GDELT: 2015-2018](https://blog.gdeltproject.org/mapping-the-linguistic-geography-of-gdelt-2015-2018/)
(https://blog.gdeltproject.org/mapping-the-linguistic-geography-of-gdelt-2015-2018/)

[Une compilation des demos GDELT](https://blog.gdeltproject.org/a-compilation-of-gdelt-bigquery-demos/) (https://blog.gdeltproject.org/a-compilation-of-gdelt-bigquery-demos/)

[Article original sur la creation du dataset GDELT](http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf) (http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf)

Last updated 2020-12-18 11:54:15 +0100