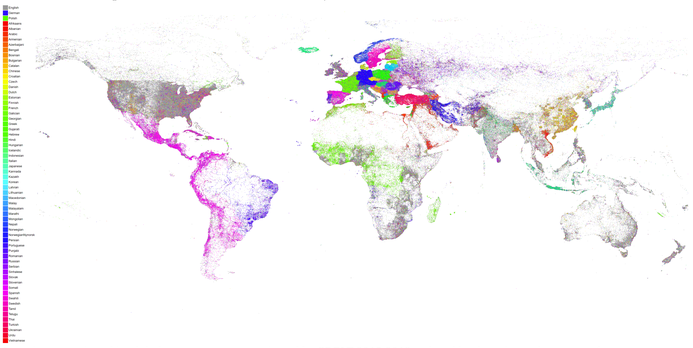
Projet Bigdata 2020-2021: *Etudier l'evolution de la pandemie COVID19 via son impact media*

Intro

> " *The Global Database of Events, Language, and Tone (**GDELT**) 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'evolution de la pandemie COVID19 via sa couverture mediatique. Dans ce but nous allons utiliser le jeu de donnees 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)

Contexte

A. Jeu de données

Nous allons utiliser:

- les events (CAMEO Ontology, documentation)
- les mentions (documentation)
- le graph des relations ⇒ GKG, Global Knowledge Graph (documentation)

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

- Master CSV Data File List – English
- Master CSV Data File List – GDELT Translingual

Pour plus d'infos consulter la documentation.
Le jeu de données de GDELT v2.0 est disponible également sur Google BigQuery. 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 repondre aux question suivantes:

a. afficher le nombre d'articles/évènements 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 évènements qui y ont eu place triées par le nombre de mentions (tri décroissant); permettez une agrégation par jour/mois/année
c. pour une source de donnés passée en paramètre (gkg.SourceCommonName) affichez les thèmes, personnes, lieux dont les articles de cette sources parlent ainsi que le 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'evolution qui pourraient nous permettre d'identifier la prochaine vague/pandemie

---

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 noeud).
3. Vous devez pre-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 stocquage en tuant un noeud de votre clusteur.

Ressources

GDELT v2.0 dataset description
GDELT Translingual
Mapping the Linguistic Geography Of GDELT: 2015-2018
Une compilation des demos GDELT
Article original sur la creation du dataset GDELT