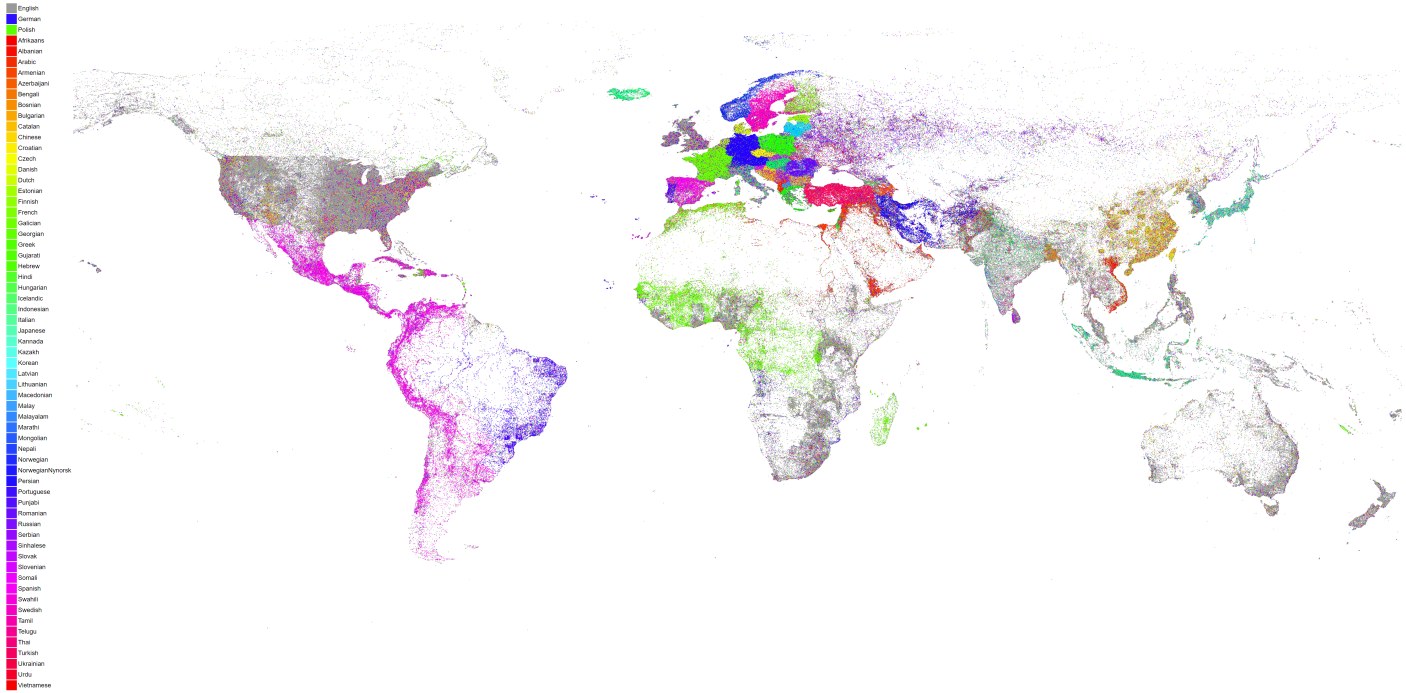
Projet Bigdata 2019: GDELT

Intro

> " *The Global Database of Events, Language, and Tone (**GDELT**) (https://www.gdeltproject.org/), est une initiative pour construire un catalogue de comportements et de croyances sociales à travers le monde, reliant chaque personne, organisation, lieu, dénombrement, thème, source d'information, et événement à travers la planète en un seul réseau massif qui capture ce qui se passe dans le monde, le contexte, les implications ainsi que la perception des gens sur chaque jour".*

Cette base de données a eu beaucoup d'utilisations, pour mieux comprendre l'évolution et l'impact de la crise financière du 2008 (Bayesian dynamic financial networks with time-varying predictors (https://arxiv.org/pdf/1403.2272v1.pdf)) ou analyser l'évolution des relations entre des pays impliquées dans des conflits (Massive Media Event Data Analysis to Assess World-Wide Political Conflict and Instability (http://www.gao.ece.ufl.edu/GXU/fun_reading/sbp_hurst.pdf)).

L'objectif du projet est de concevoir un système qui permet d'analyser le jeu de donnees GDELT et ses sources de donnees.



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))

Contexte

A. Jeu de données

GDELT est composé de trois types de fichiers CSV:

- les events (schema (https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.events?tab=schema), CAMEO Ontology (http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf), documentation (http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf))
- les mentions (schema (https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.eventmentions), documentation (http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf))
- le graph des relations ⇒ GKG, Global Knowledge Graph (schema (https://bigquery.cloud.google.com/table/gdelt-bq:gdeltv2.gkg), documentation (http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf))

La totalité des fichiers du jeu de données est indexé par deux fichiers:

- Master CSV Data File List – English (http://data.gdeltproject.org/gdeltv2/masterfilelist.txt)
- Master CSV Data File List – GDELT Translingual (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/)

Le jeu de données de GDELT v2.0 est disponible également sur Google BigQuery (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 repondre aux question suivantes:

a. afficher le nombre d'articles/évènements 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. dresser la cartographie des relations entre les pays d'après le ton des articles : pour chaque paire (pays1, pays2), calculer le nombre d'article, le ton moyen (aggrégations sur Année/Mois/Jour, filtrage par pays ou carré de coordonnées)

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.

Budget AWS à ne pas depasser: 300E par groupe.

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 (https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/)

GDELT Translingual (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/)

Une compilation des demos GDELT (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)

Last updated 2019-12-31 07:06:45 +0100