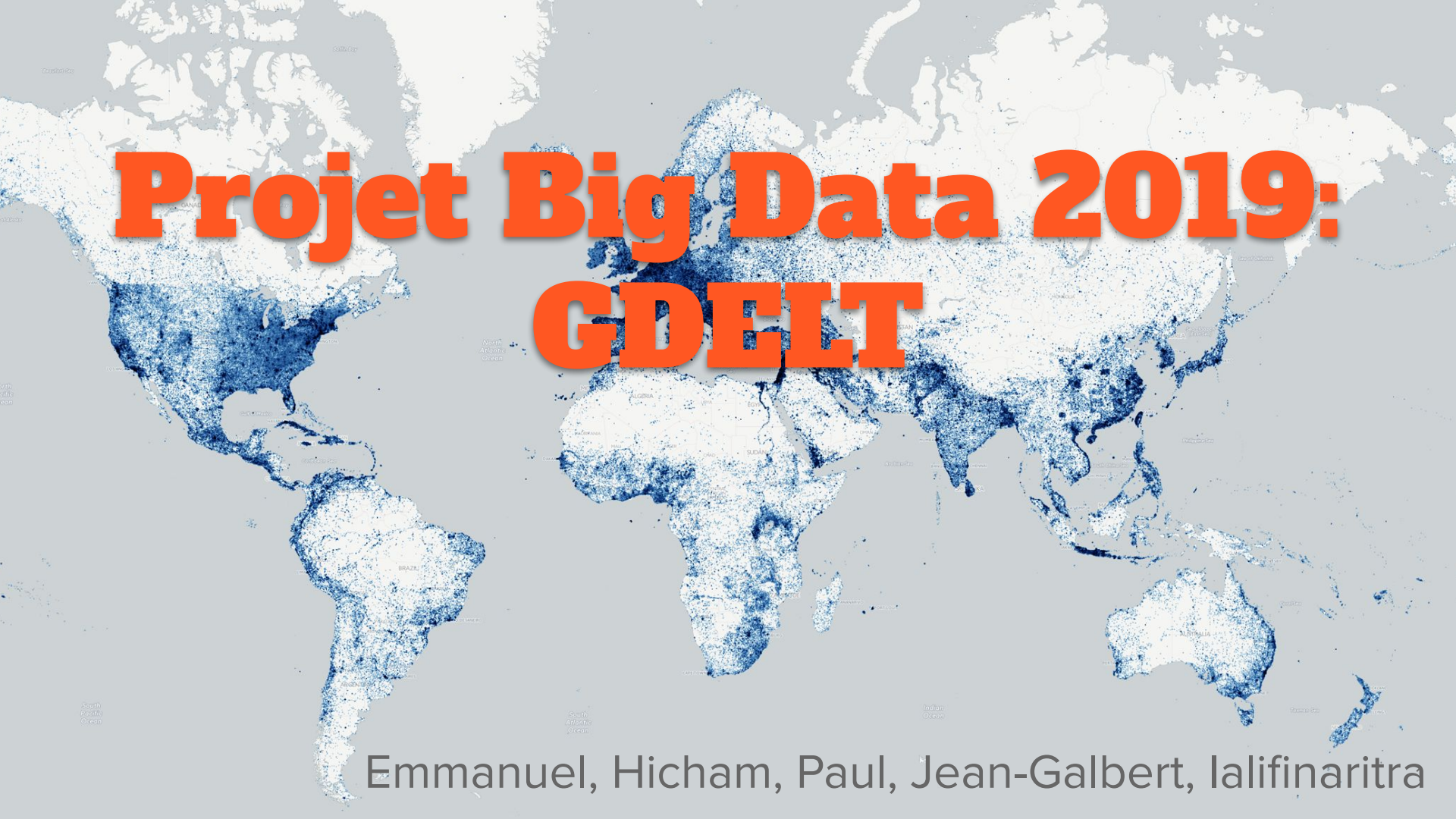


Projet Big Data 2019: GDELT

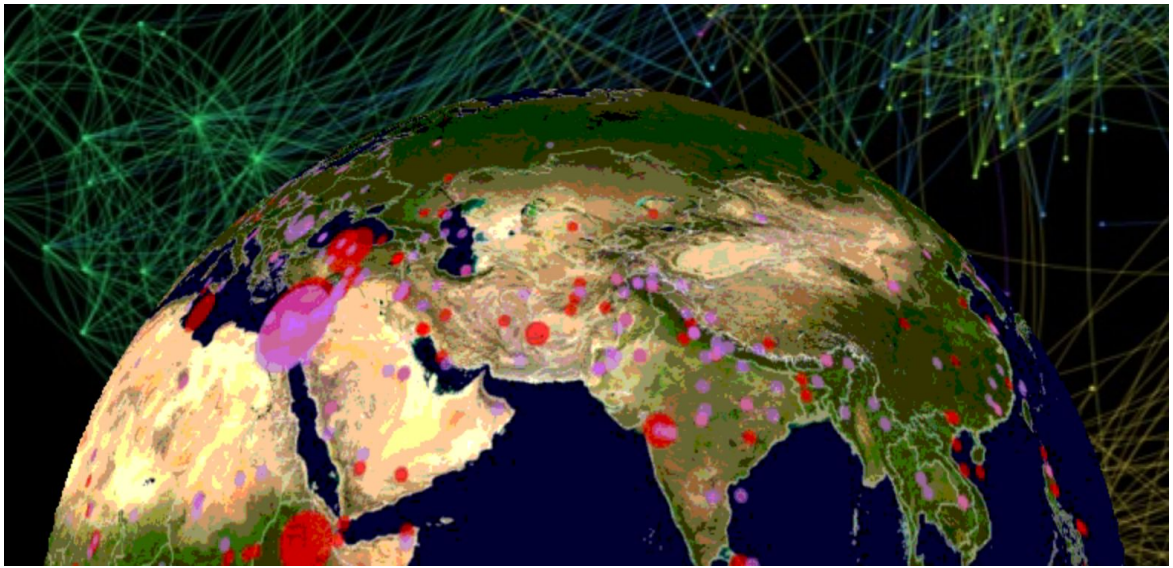
Emmanuel, Hicham, Paul, Jean-Galbert, Ialifinaritra

Sommaire

- Contexte
- Description des données
- Technologies choisies
- Architecture
- Requêtage
- Demo
- Améliorations

Objectif du projet

- **But :** Proposer un système de stockage distribué, résilient et performant sur AWS.
- **Base de données :** http://data.gdeltproject.org/gdeltv2/
    - Données d’articles de presse enregistrées toutes les 15 minutes et stockées en 3 CSV.



Description des données

Evénements

Evènemnt_Global_ID

Date

Pays

Nombre d'articles

Mentions

Ton moyen

Mentions

Evènemnt_Global_ID

Date

Source

Langue

GKG

Nom de la source

Date

Thème

Personne

Locations

Ton

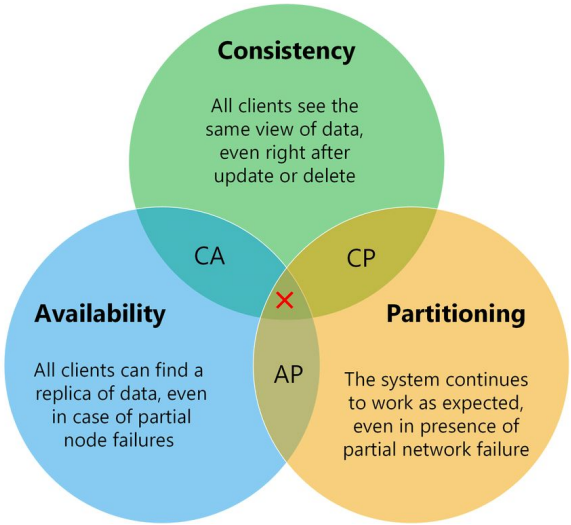
Nombre d'articles

Technologies

- AWS :

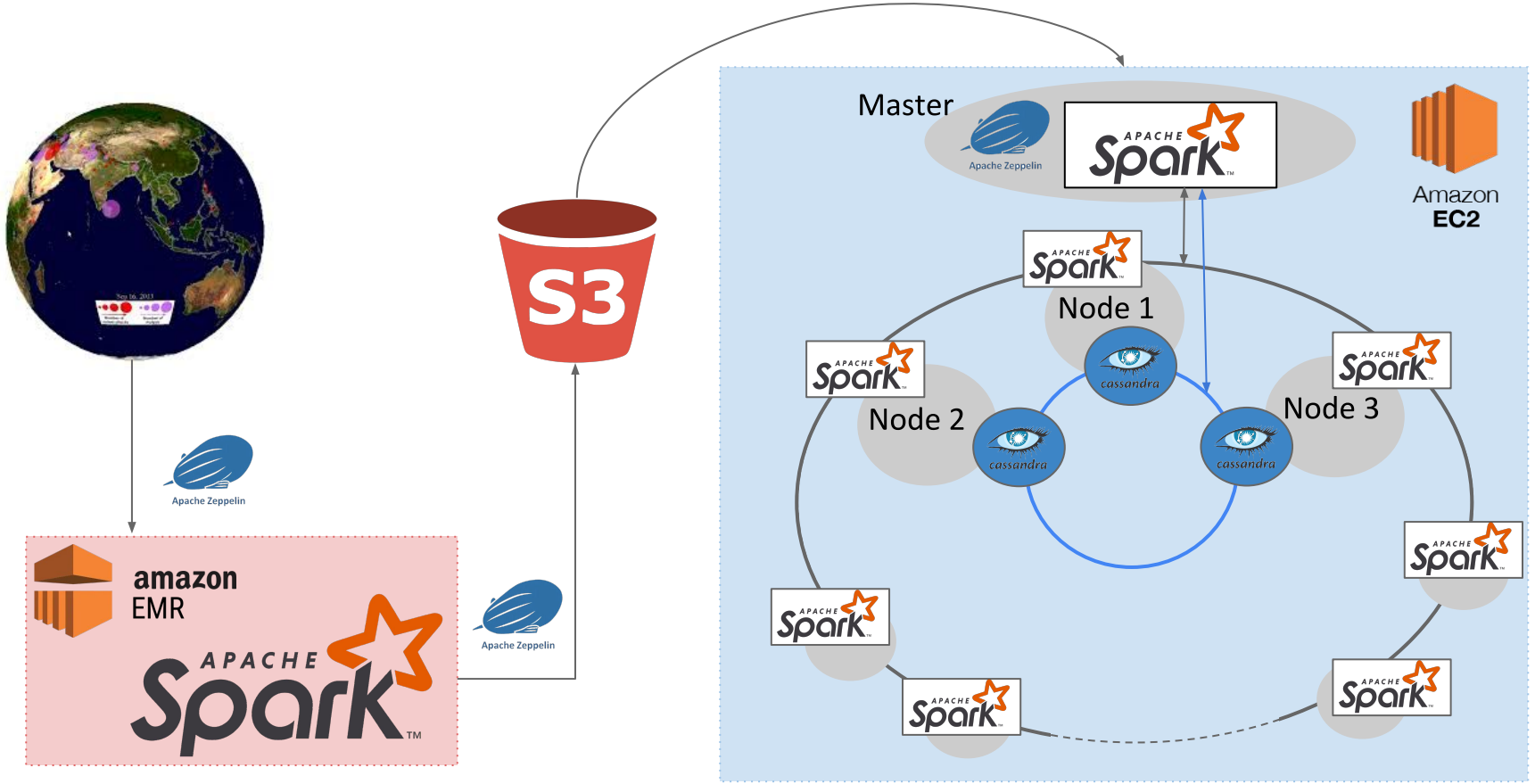
- Calcul distribué :

- Résilience du stockage des données:



- Notebook:

Visualisation de l'architecture



Choix des technologies :

Avantages :

- Flexibilité des installations sur EC2
- Configuration minimale de Cassandra
- Documentation fournie sur Cassandra
- Bon outil de visualisation sur Zeppelin

Contraintes :

- Compte Educate !
- Comportement global du système difficile à débugger

Modélisation

Requête 1 : Afficher le nombre d'articles/évènements qu'il y a eu pour chaque triplet (jour, pays de l'évènement, langue de l'article).

Evénements

Mentions

TABLE 1
Jour Pays Langue Mentions Evénements

Modélisation

Requête 2 : Pour un pays donné en paramètre, afficher les événements qui y ont eu place triés par le nombre de mentions (tri décroissant)

Evénements

TABLE 2
Evènement_Global_ID Pays Jour Mois Année Nombre Mentions

Modélisation

Requête 3 : Pour une source de donnés passée en paramètre (“gkg.SourceCommonName”), afficher 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)

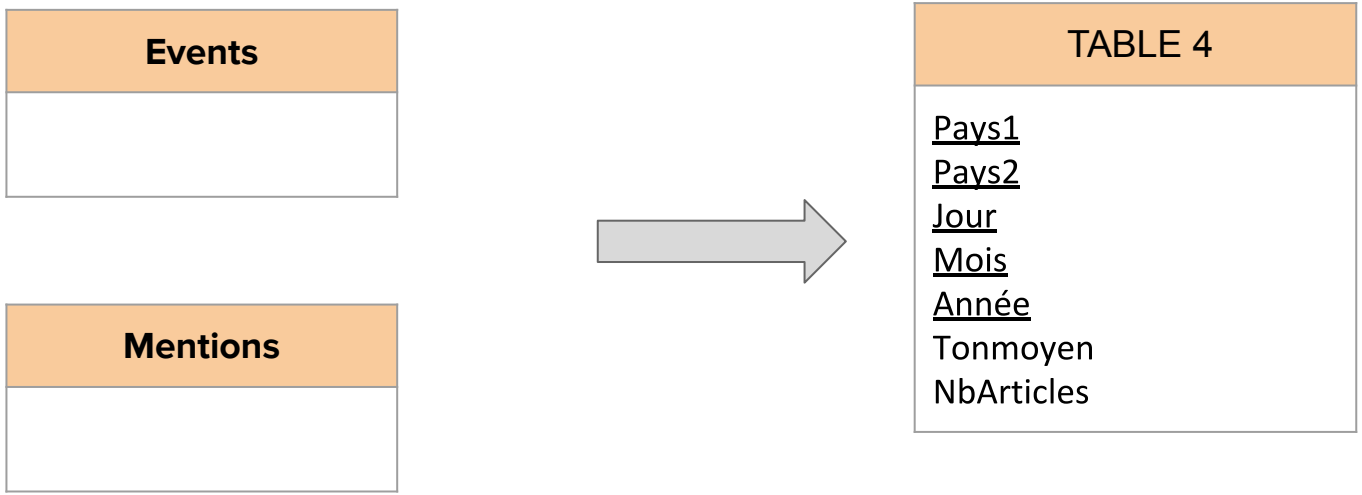
TABLE 3_1 : Thèmes
Source Commun Nom
Jour
Mois
Année
Thème
Ton moyen
Nombre Articles

TABLE 3_2 : Personnes
Source Commun Nom
Jour
Mois
Année
Personne
Ton moyen
Nombre Articles

TABLE 3_3 : Locations
Source Commun Nom
Jour
Mois
Année
Location
Ton moyen
Nombre Articles

Modélisation

Requête 4 : Dresser la cartographie des relations entre les pays d'après le ton des articles : pour chaque paire (Pays 1, Pays 2), calculer le nombre d'article, le ton moyen (agrégations sur Année/Mois/Jour, filtrage par pays ou carré de coordonnées)



Demo

Points d'amélioration

- Utilisation de EMR pour le chargement des données.
- Plus des noeuds Cassandra pour accélérer l’écriture.
- Requêtes optimisées, compilées en JAR pour accélérer le nettoyage et chargement les données