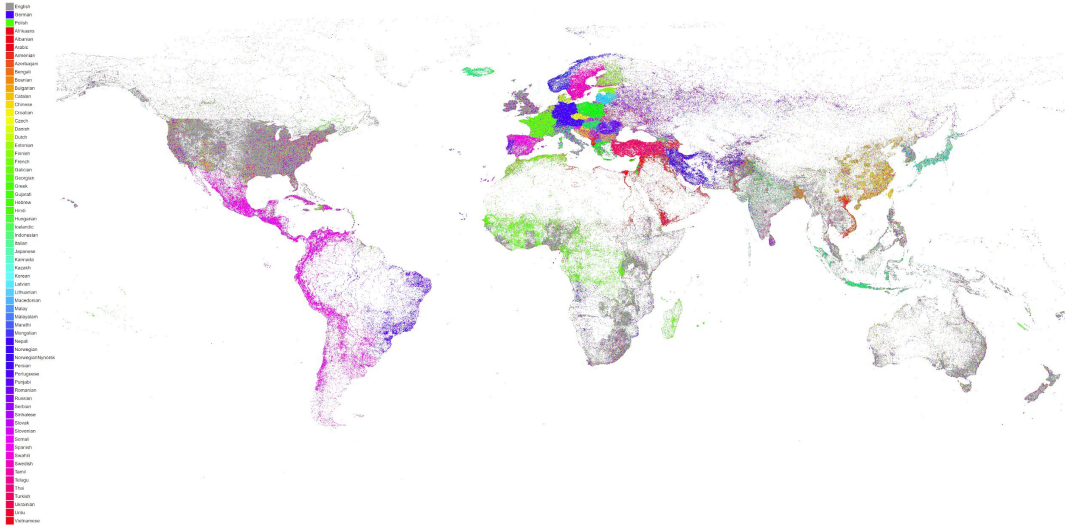
Projet Bigdata 2020: GDELT

INF 728



Jérémie PERES, Li XU, Benyang SUN et Kevin FERIN

Sommaire

I. Présentation du sujet

> *The Global Database of Events, Language, and Tone (**GDELT**), 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*

3 tables :

MENTIONS

- GLOBALEVENTID
- MentionDocTranslationInfo

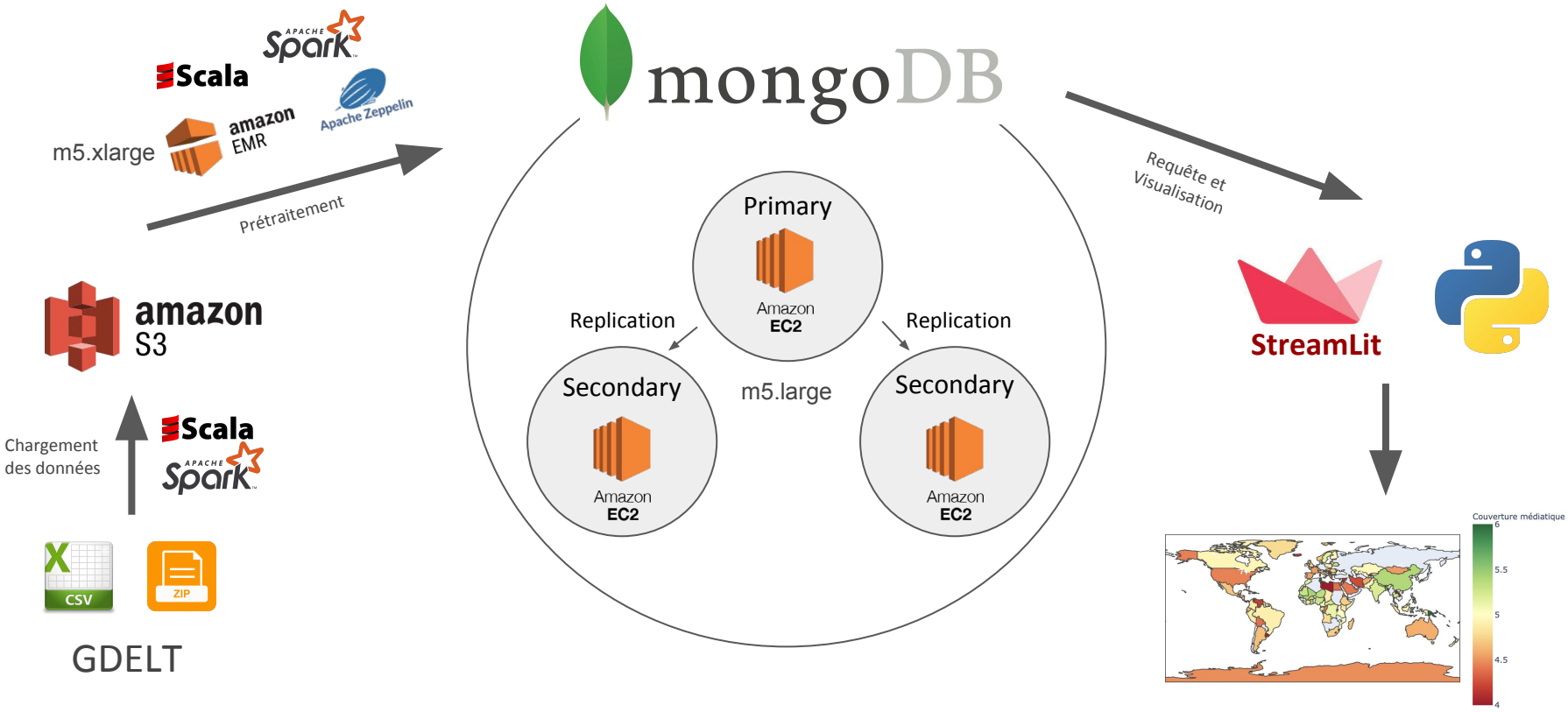
GKG

- GKGRECORDID
- DATE
- SourceCommonName
- Themes
- Locations
- Persons
- Tone

EVENTS

- GLOBALEVENTID
- SQLDATE
- NumArticles
- AvgTone
- ActorGeo
- ActionGeo

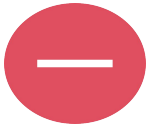
II. Choix de l'architecture



II. Choix de l'architecture

Pourquoi avoir choisi MongoDB ?

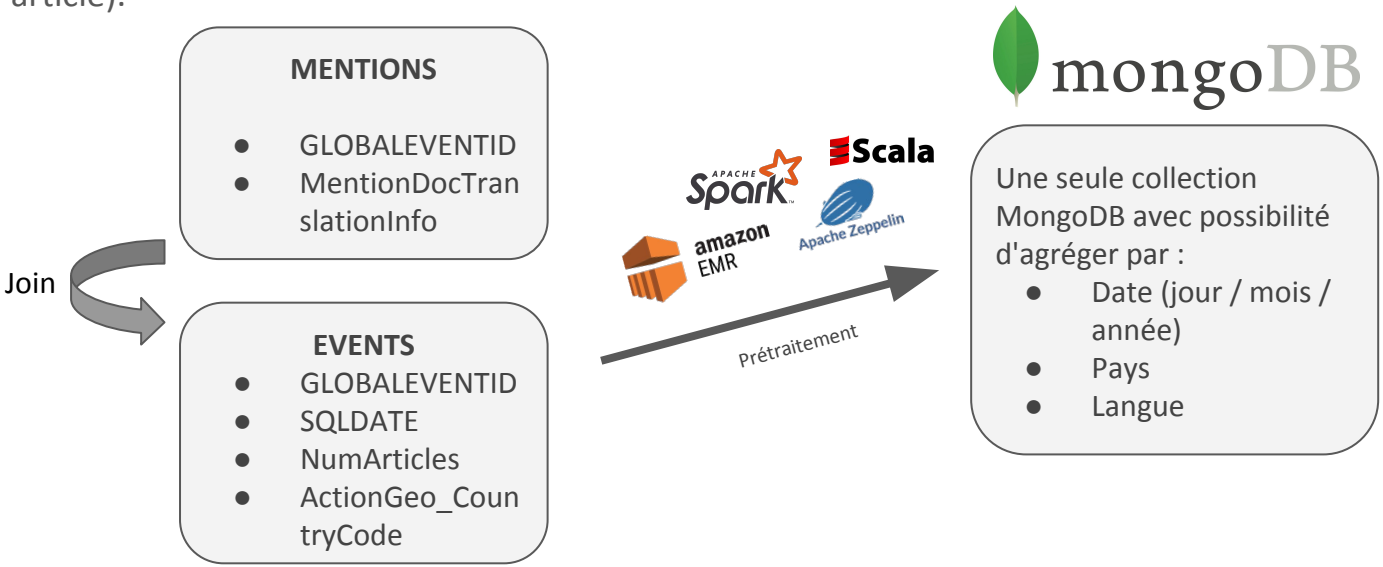
- Technologie montante, base NoSQL la plus populaire aujourd'hui
- Pas de pré-structuration des données : éléments Json dans les collections
- Facilité de prise en main
- Rapidité de requêtage
- Ecriture seulement sur primary et réplication automatique sur chacun des noeuds secondaires, puis lecture possible sur les deux secondary



- Pas de jointure
- Utilisation importante de la mémoire : MongoDB a tendance naturellement à utiliser plus de mémoire car il doit stocker les noms de clés dans chaque document

III. Requêtes

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

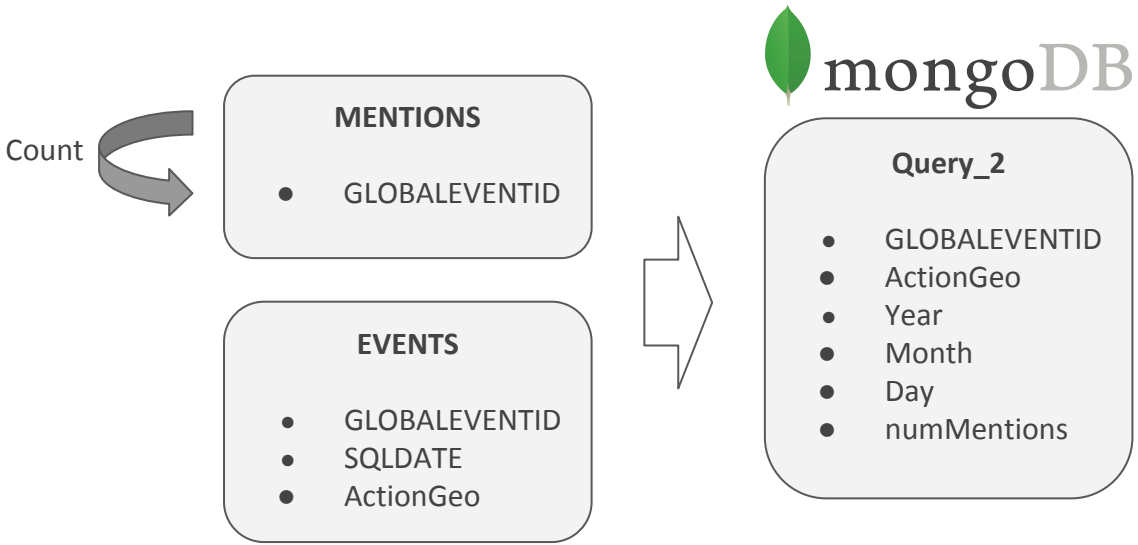


Performances :

- **3min** pour charger un mois de données sur MongoDB
- **<1s** en moyenne pour requêter MongoDB

III. Requêtes

Query 2 : 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

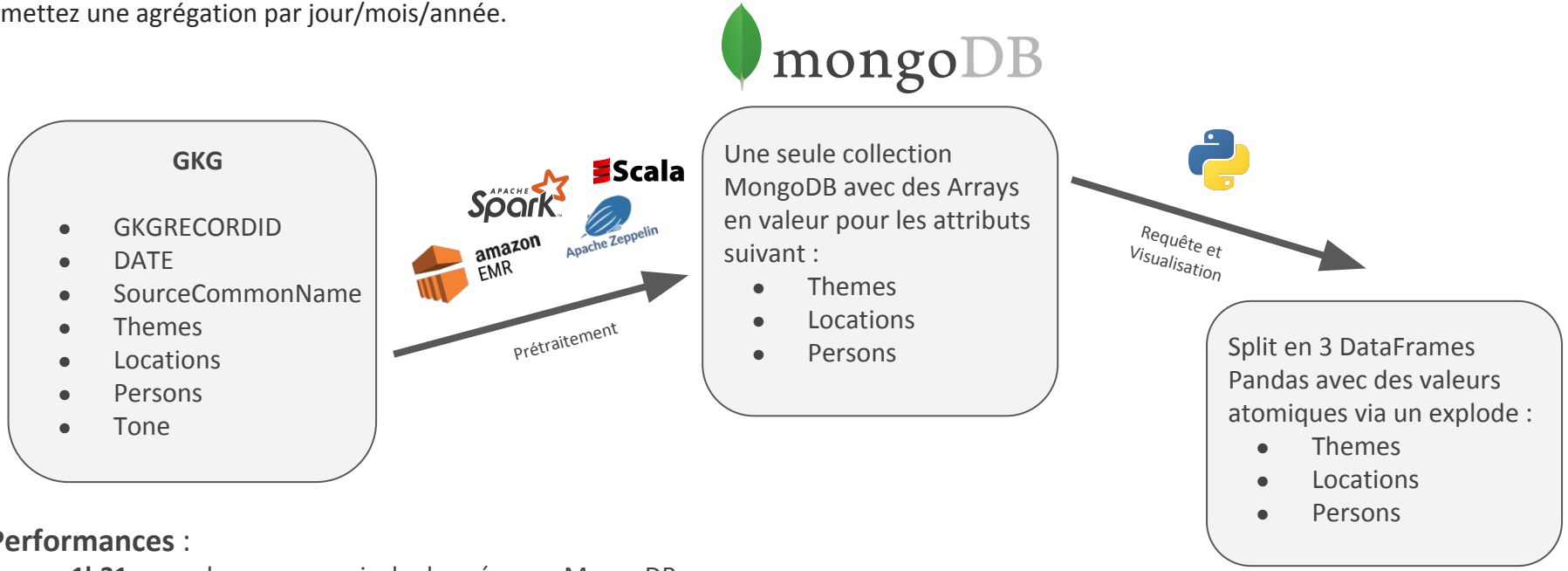


Performances :

- **8min20** pour charger un mois de données sur MongoDB
- **5s** en moyenne pour requêter MongoDB pour France passée en paramètre

III. Requêtes

Query 3 : Pour une source de donnés passée en paramètre (gkg.SourceCommonName) affichez les thèmes, personnes, lieux dont les articles de cette source 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.



Performances :

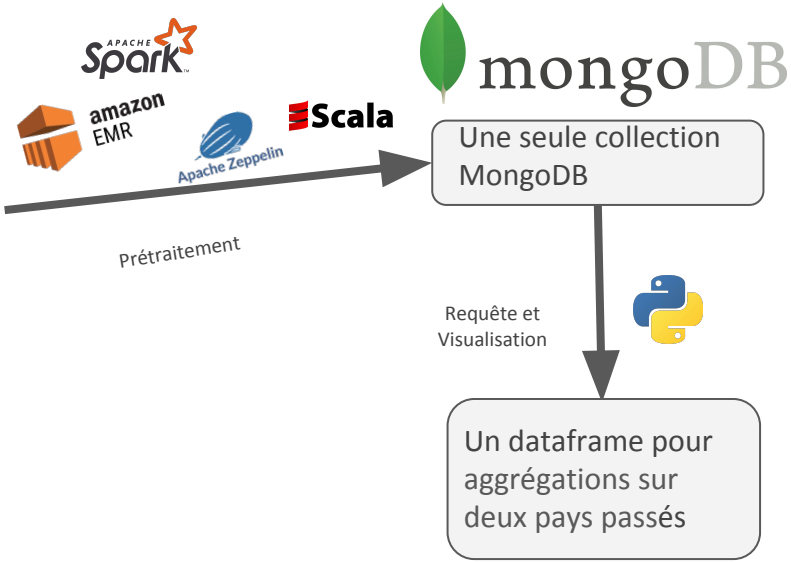
- **1h31** pour charger un mois de données sur MongoDB
- **1min** en moyenne pour requêter MongoDB pour une source passée en paramètre

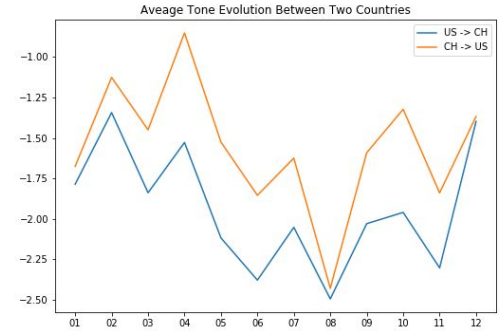
III. Requêtes

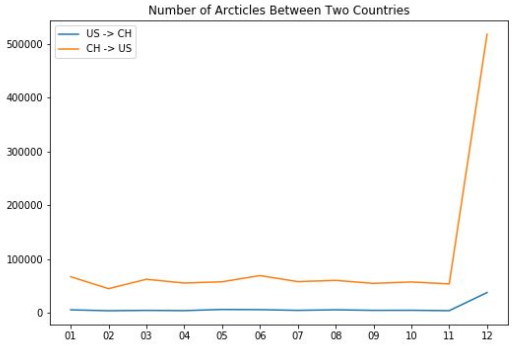
Query 4 : Dresser la cartographie des relations entre les pays d'après le ton des articles : pour chaque paire (pays1, pays2), calculer le nombre d'article, le ton moyen (aggrégations sur Année/Mois/Jour, filtrage par pays ou carré de coordonnées)

Events

- "SQLDATE"
- "Actor1Geo_CountryCode
- "Actor1Geo_Lat"
- "Actor1Geo_Long"
- "Actor2Geo_CountryCode
- "Actor2Geo_Lat"
- "Actor2Geo_Long"
- "AvgTone"
- "NumArticles"







Performances :

- **5min30** pour charger un mois de données sur MongoDB
- **<1s** en moyenne pour requêter MongoDB deux pays passés en paramètre

IV. Budget

EC2	$36.28
ElasticMapReduce	$4.46
DataTransfer	$1.21
S3	$0.75
Autres services	$0.00
Taxes	$8.54
Total	**$51.24**

V. Points d'amélioration

- Utilisation du sharding de mongoDB afin de pouvoir utiliser un an de données efficacement
- Exploration approfondie des données (ML, Data viz)
- Déploiement du cluster MongoDB de manière automatisée avec un script
- Meilleur dimensionnement des machines EC2

VI. Démonstration

1. WebApp (+ 3 notebooks Zeppelin)

+

2. Résilience