Projet INF-728

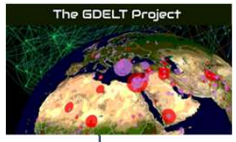
MS BGD 2020-2021 – 22 jan 2021

Nicolas Calligaro
Frederic Haykal
Julien Maksoud
Axel Michalewicz
Lingli Zhang

Architecture & Justification

Internet

The GDELT Project



= 0(Tb)

Stockage des raw data

Amazon **S3**

Accessibilité

Coût

Granularité de l'offre

Traitement des données

Amazon **EMR**
Auto Scaling

Cluster

M5.2XLarge
8 coeurs
32 Gb de RAM
1 Master + 2 Workers

jupyter Scala Apache Spark

~~**Chargement dans une base de données type NoSQL**~~

~~Amazon DocumentDB (MongoDB)~~

~~Amazon Keyspaces (Cassandra)~~

Chargement dans une base de données type NoSQL

Amazon **EC2 Auto Scaling !**

VPC

Primary

ubuntu

mongoDB Compass

Secondary

Secondary

Client

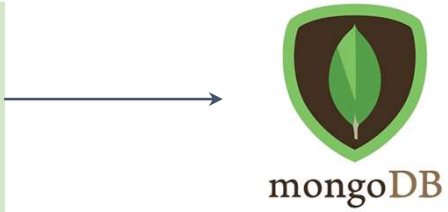
Requêtage

Cassandra VS MongoDB

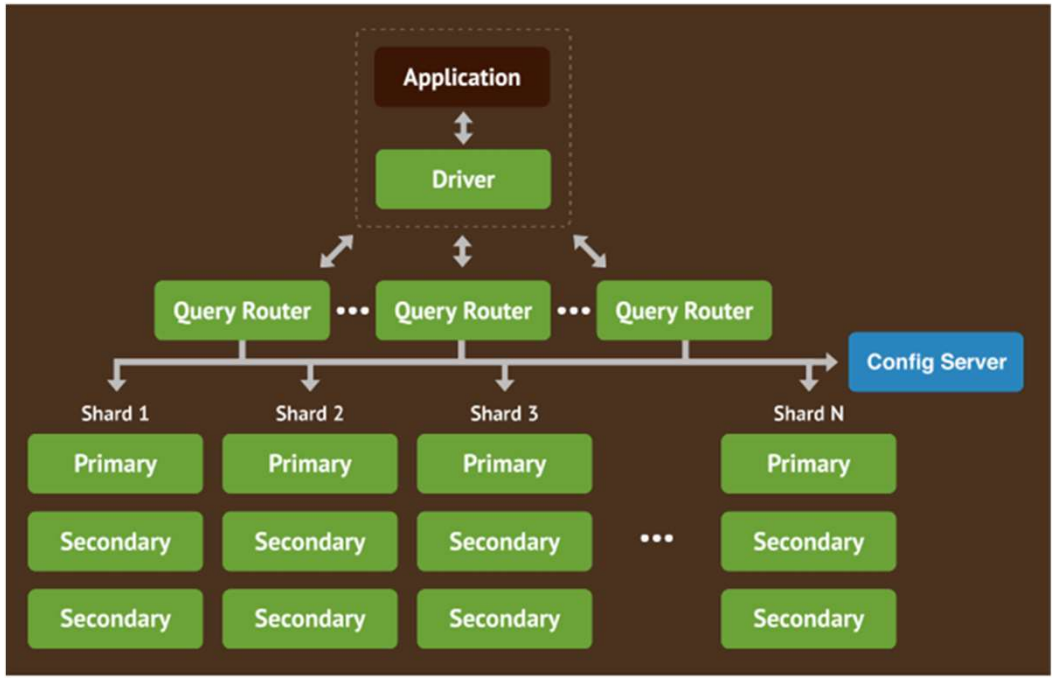
Caractéristique	Important pour ce projet	Cassandra	MongoDB
Modèle De Donnée	**Oui**	Colonnes	**+ Documents**
Langage De Requête	**Oui**	Léger et simple	**+ Mongo Query Langage (Puissant)**
Agrégation	**Oui**	Néant	**+ Aggregation Framework (Compass)**
Nœud Principal	Non	**+ Multiple**	Unique ou Sharding
Passage À L'échelle	Non	**+ Native**	Null ou Sharding
Schéma	Non	Fixe	**+ Schéma Modifiable à Chaud**
Chargement Des Données	Non	**+ Efficace**	Relative
Consistance Des Donnée	**Oui**	**+ Meilleur**	Relative (Réplicat Absorbe Crazy Monkey)

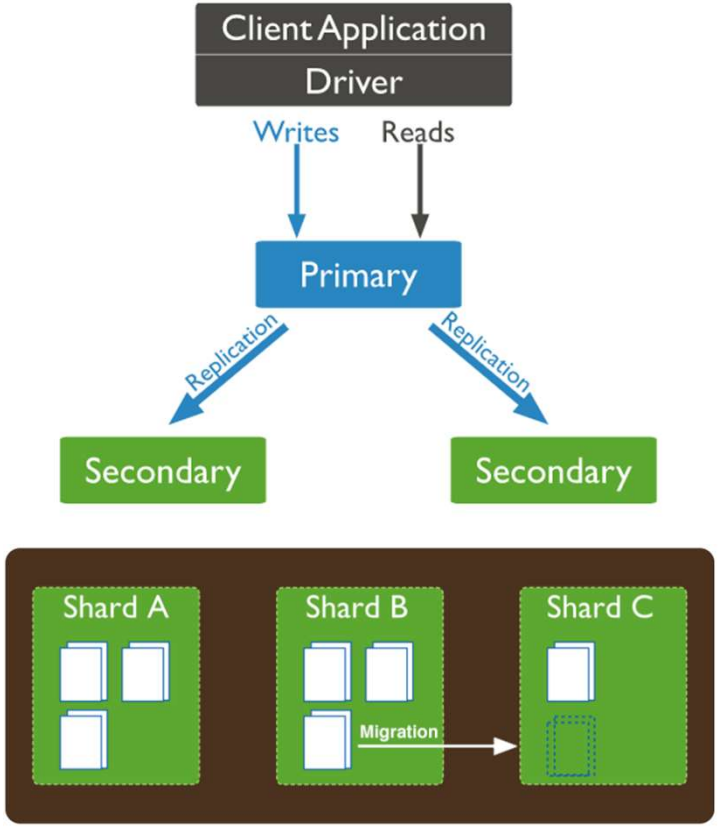
Besoins du projet GDELT

- *Pas de flux continu → Passage à l'échelle pas primordial*
- *Requêtes complexes de documents nécessaires*
- *Réduire la quantité de données / Minimiser le nombre de tables*



Architecture MongoDB (cf cours de Geoffrey)





Load Balancing

Stockage des raw data (S3)
Traitement (EMR)
Mise en DB NoSQL (EC2)
Requêtage (EC2)

Import des données dans AWS S3

1 fichier de chaque toutes les 15 minutes: masterfilelist.txt

masterfilelist.txt

```
117660  6504649 3e4882e7b5aa79b94b9fd800acabdb97 http://data.gdeltproject.org/gdeltv2/20210108120000.gkg.csv.zip
117661  68696 8392b1de9eaf3ce7b7ec2fc70839d767 http://data.gdeltproject.org/gdeltv2/20210108121500.export.CSV.zip
117662  125923 b876c5f98da95bc624e88f98fcb1af62 http://data.gdeltproject.org/gdeltv2/20210108121500.mentions.CSV.zip
117663  4748976 9cbfa559aab6f13f27511d02da1afd00 http://data.gdeltproject.org/gdeltv2/20210108121500.gkg.csv.zip
117664  92211 643075007162c647026540d136317b4b http://data.gdeltproject.org/gdeltv2/20210108123000.export.CSV.zip
117665  172181 a846ebffb98523423a8f6d658778e721 http://data.gdeltproject.org/gdeltv2/20210108123000.mentions.CSV.zip
117666  6265775 b3ad64294216ab836e8a0a3dd6647bb3 http://data.gdeltproject.org/gdeltv2/20210108123000.gkg.csv.zip
117667  86010 e04ec2ba8096a5dd91fe83e60fede02
117668  158964 fd818796b9198aa281ac01c72bdf27
117669  6310064 5ac32dd4e374b948579407697cee6
117670  87435 2ff9a44b8efa604a7400125f6daa683
117671  164973 84045cac0c96ae26d75e8c34add649
117672  6133097 4cc4d52d4070081536665168cf24d
117673  73182 9fdbe629b84f04bdde7908a488147d8
117674  161068 3aafd3d90e0691f94d4f22204a16bb
117675  6873515 c6c0313c4af0cd12eb8e11fcd5773
117676  85725 1c8bcccd8c92c81645ed55b986d360e
117677  162038 034e04dc8d9a0f4185fe6b8ccf9c3a
117678  6705445 e26bd8aa05d79c5377c47cf3ee468
117679  80973 179f442631181bd2160e4d0be6413ac
117680  172948 b485ef1c3d132174173b65149edbed
```

15 min de data

15 min de data

15 min de data

Lancement du téléchargement des fichiers de GDELT vers S3

Entrée [13]:

```
def StokeFileS3(URL: String) = {
    val fileName = URL.split("/").last
    val dir = "/mnt/tmp/"
    val localFileName = dir + fileName
    try {
            fileDownloader(URL,  localFileName)
            @transient val s3Client: AmazonS3 = AmazonS3ClientBuilder.defaultClient()
            val localFile = new File(localFileName)
            s3Client.putObject("testfuret/Master_Translate_file ", fileName, localFile)
            localFile.delete()

    } catch {
        case e: java.io.FileNotFoundException =>
        case e: com.amazonaws.SdkClientException =>

    }
```

```
+--------+--------------------------------+--------------------------------------------------------------------------------+
|size    |hash                            |url                                                                             |
+--------+--------------------------------+--------------------------------------------------------------------------------+
|6178636|72d0e5ca10362beca19a0510854f4721|http://data.gdeltproject.org/gdeltv2/20200201000000.translation.gkg.csv.zip|
|6133442|765b9d4a4dd32770732defd66de74be37|http://data.gdeltproject.org/gdeltv2/20200201001500.translation.gkg.csv.zip|
|6454334|209299355a96b7e5a9ecf3ec77d2f6de|http://data.gdeltproject.org/gdeltv2/20200201003000.translation.gkg.csv.zip|
|5654513|1ddd9159fce3a3ebf01be064bdbc4bfa|http://data.gdeltproject.org/gdeltv2/20200201004500.translation.gkg.csv.zip|
|5517145|5570a7f6cf093ec26464d07d0d849846|http://data.gdeltproject.org/gdeltv2/20200201010000.translation.gkg.csv.zip|
|5602121|2c41db3dc0ddb82d8f0fd034baca89bd|http://data.gdeltproject.org/gdeltv2/20200201011500.translation.gkg.csv.zip|
|6049830|22430c0bf433ec91e4fccb777d1f431e|http://data.gdeltproject.org/gdeltv2/20200201013000.translation.gkg.csv.zip|
|6369154|39a46116c40134a5db96beeab30f434c|http://data.gdeltproject.org/gdeltv2/20200201014500.translation.gkg.csv.zip|
|5541150|cbb2ef523e91f05a31f1f51f9fb8b516|http://data.gdeltproject.org/gdeltv2/20200201020000.translation.gkg.csv.zip|
|4966093|8d8112546a147bc3a9d1322176ab3f0e|http://data.gdeltproject.org/gdeltv2/20200201021500.translation.gkg.csv.zip|
|5464143|35e3d8448dd0079f049c686a6aafd06f|http://data.gdeltproject.org/gdeltv2/20200201023000.translation.gkg.csv.zip|
|5997714|7587d659be5c8844f9819f3e86ee2706|http://data.gdeltproject.org/gdeltv2/20200201024500.translation.gkg.csv.zip|
|5524668|e6d69c9bb0989902869cc2a975a1cadd|http://data.gdeltproject.org/gdeltv2/20200201030000.translation.gkg.csv.zip|
|5202802|44523fb855f67ad40d5dc19b3f9963ee|http://data.gdeltproject.org/gdeltv2/20200201031500.translation.gkg.csv.zip|
|4663946|3e06fe6d1fae36a72eeeafa3ad03a9a7|http://data.gdeltproject.org/gdeltv2/20200201033000.translation.gkg.csv.zip|
|6223891|0680afd7745c352c8caaeb8fd3270fdd|http://data.gdeltproject.org/gdeltv2/20200201034500.translation.gkg.csv.zip|
```

```
yle='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
l
ect("url").repartition(100).foreach( r=> {
g](0)
yle='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

Quelques Tb en CSV

Cluster EMR

jupyter Scala Apache Spark

Stockage des raw data

Amazon **S3**

S3 Standard Bucket

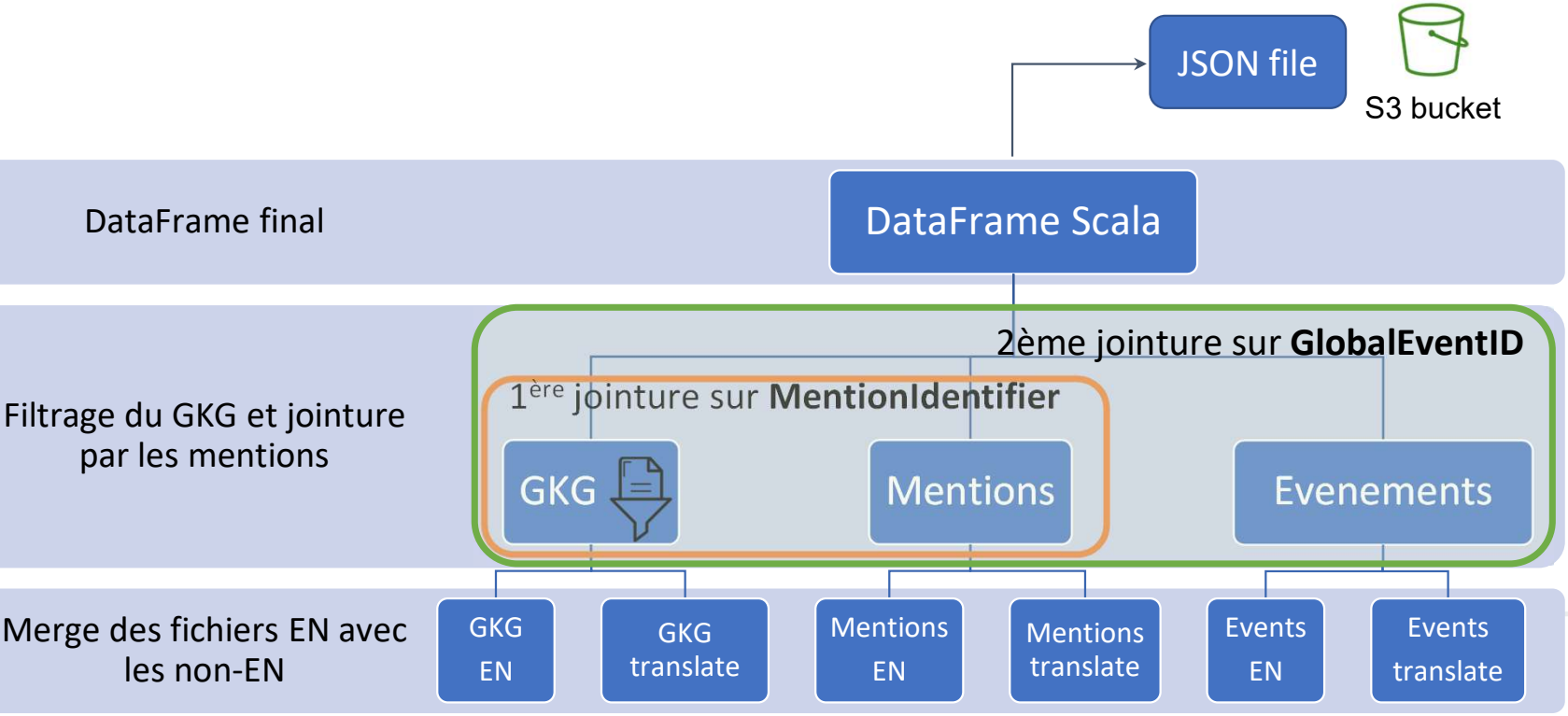
Events
Mentions
GKG

Traitement des données dans EMR

Mention
Quel évènement a été cité dans quel article ?

Evenement
Décrit un évènement

Articles de tout thèmes (GKG)



Traitement des données

Amazon **EMR**

Cluster

jupyter Scala Spark

Traitement des données