TELECOM PARISTECH - 2019

NOSQL - GDELT

Anthony Houdaille
Alexandre Bec
Raphael Lederman
Anthony DeBradke
Thomas Binetruy
Maël Fabien

TELECOM ParisTech

GDELT

SUJET

GDELT

SUJET

PROBLÉMATIQUE

- Proposer un système de stockage distribué, résilient et performant sur AWS pour les données de GDELT.

- Afficher :
  - Le nombre d'articles / événements pour chaque (jour, pays de l'événement, langue de l'article).
  - Pour un acteur (pays/organisation…) ⇒ afficher les évènements qui y font reference.
  - Les sujets (acteurs) qui ont eu le plus d'articles positifs/négatifs (mois, pays, langue de l'article).
  - Acteurs/pays/organisations qui divisent le plus.

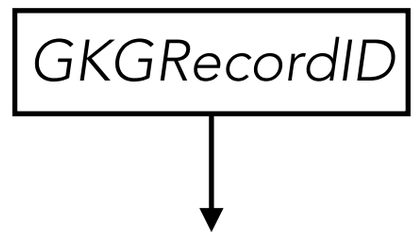
LES DONNÉES

Mentions

- *Date of mention*
- *Language of mention*
- *Type of mention*
- *Tone of mention*

Event

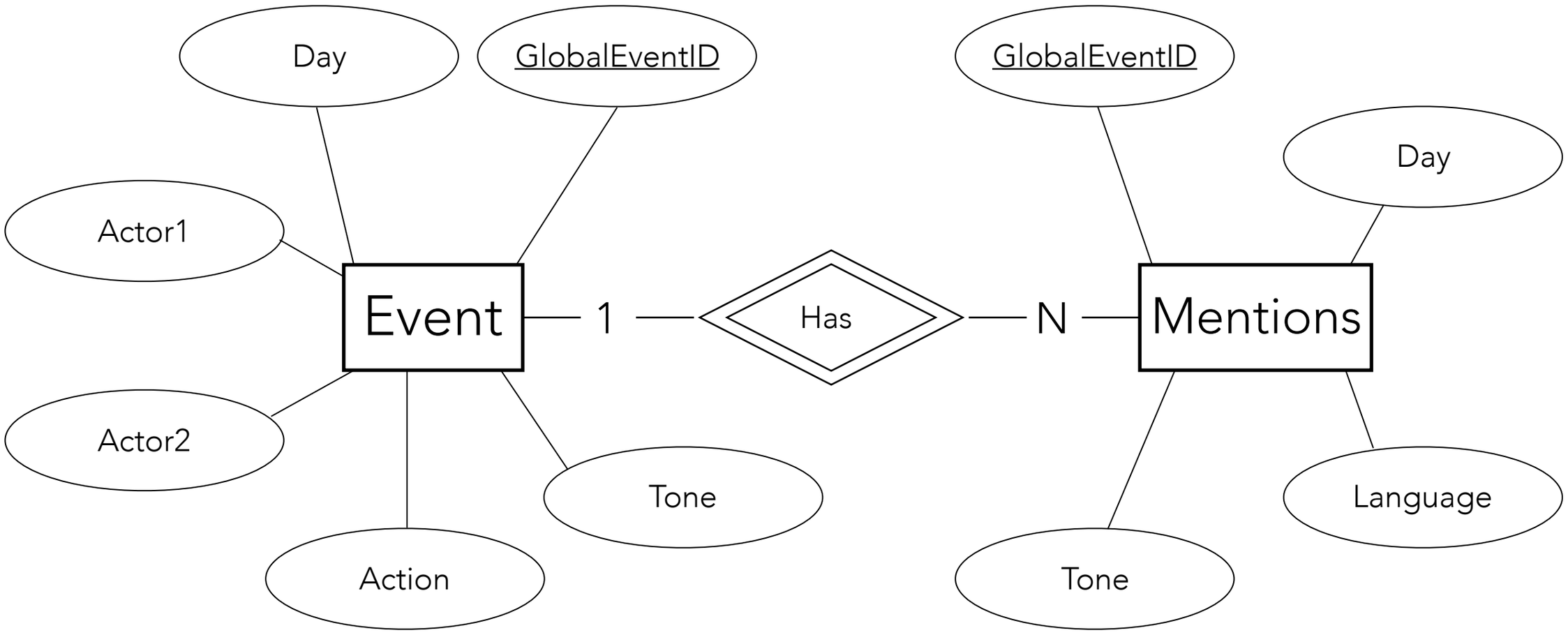
- *Date of event*
- *Actor 1 Code + Country*
- *Actor 1 Name*
- *Actor 2 Code + Country*
- *Actor 2 Name*
- *Number of Articles*
- *Average Tone*
- *Action Country*
- *Action Longitude*
- *Action Latitude*



GKG

- *Date*
- *Theme*
- *Collection*
- *Persons*

MODÈLE CONCEPTUEL



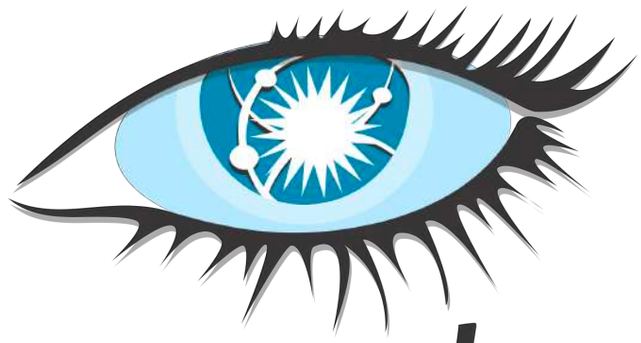
GDELT

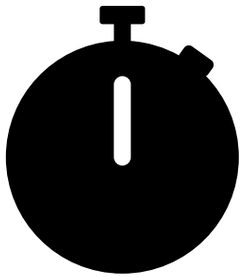
APPROCHE

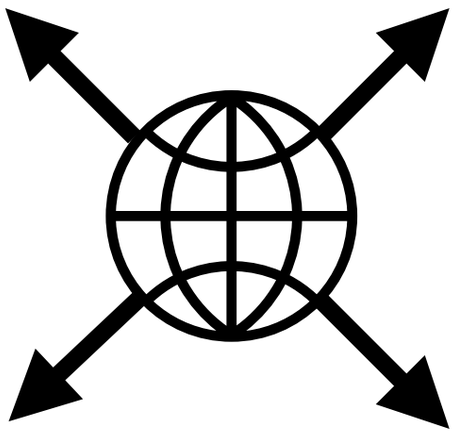
GDELT

APPROCHE

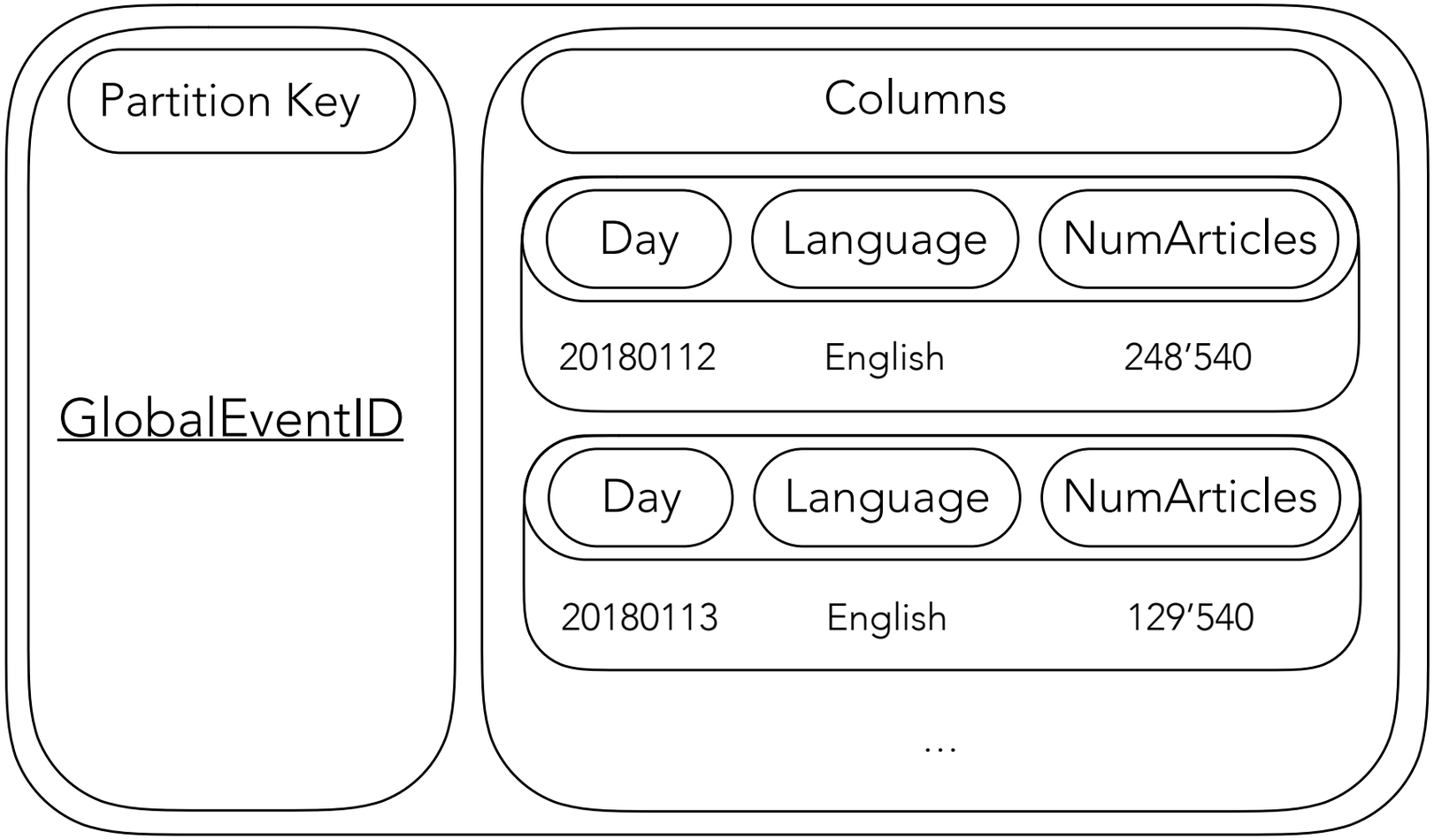
BASE DE DONNÉES





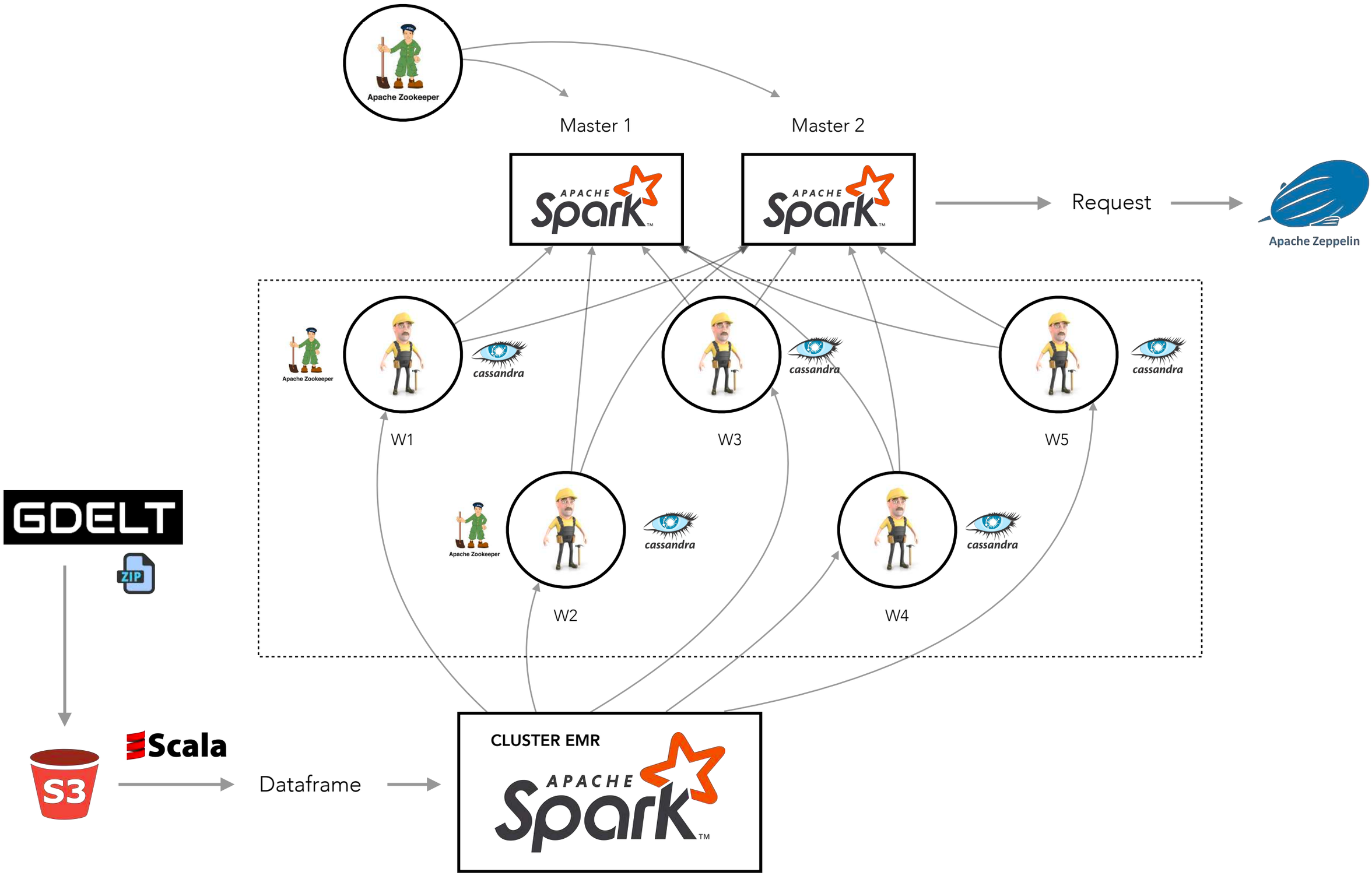


Exemple :

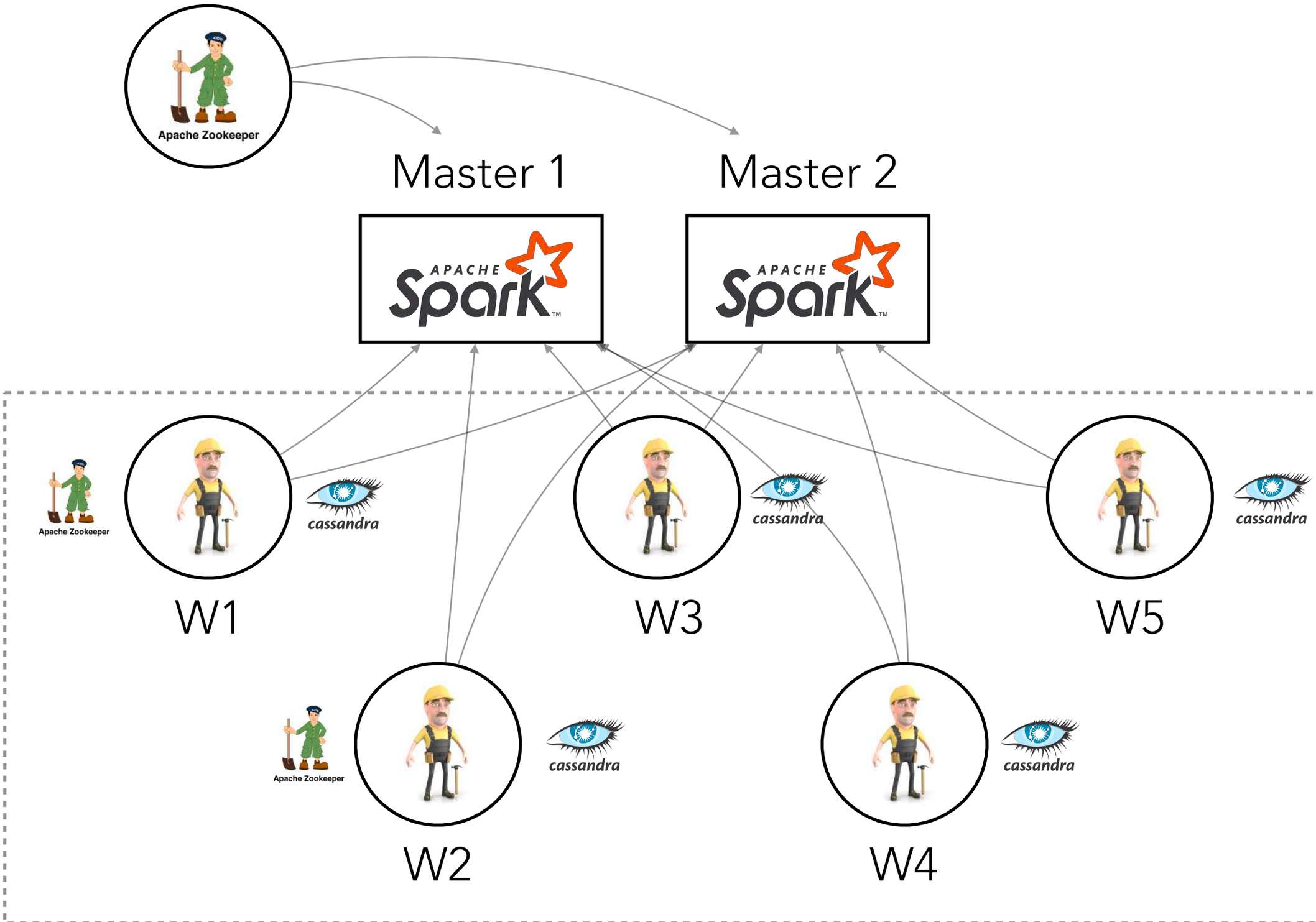


Replication Factor : **3**

ARCHITECTURE

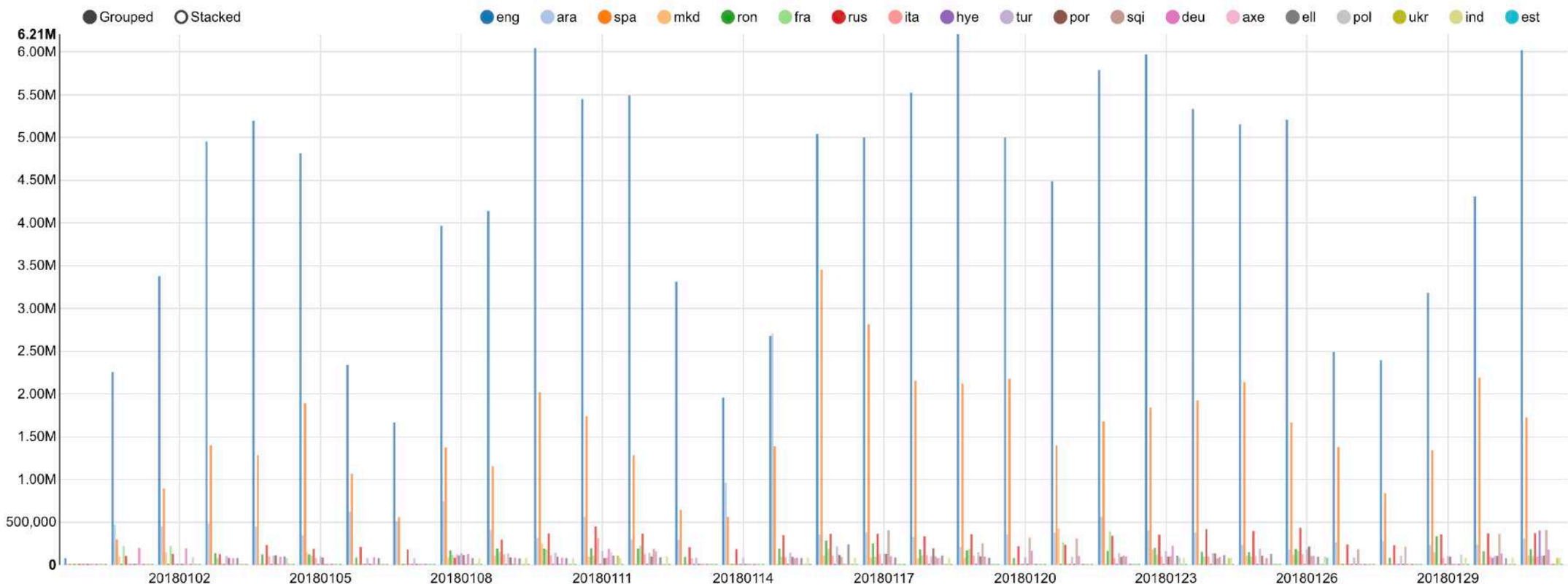


ARCHITECTURE



DATA VISUALIZATION

- Zeppelin



- Helium Zeppelin Leaflet

DATA VISUALIZATION

GDELT

PERFORMANCES

GDELT

PERFORMANCES

PERFORMANCES

- Une année zippée sur S3 : 698.9 Gb
- Besoin de séparer les tables en amont des requêtes

PERFORMANCES

- Requête d’un mois en Spark SQL :

```
z.show(spark.sql(""" SELECT * FROM q3 ORDER BY SumTone ASC LIMIT 100 """))
```

SPARK JOB FINISHED

settings

Language	ActionCountry	Month	Actor1Country	Actor1Code	SumTone
eng	US	201801	US	COP	-1365577.0567040334
eng	US	201801	US	JUD	-1037184.958702697
eng	US	201801	US	GOV	-1029943.3650900614
eng		201801		COP	-696337.9742912925
ara	IS	201801	IS	ISR	-694516.3164537457
eng	IS	201801	IS	ISR	-599651.2468628696
eng	IR	201801	IR	IRN	-476389.1384875841
eng	RS	201801	RS	RUS	-460193.67098307423

Took 10 min 48 sec. Last updated by anonymous at January 23 2019, 10:41:24 AM.

- Requête d’un mois en Cassandra :

Took 2 sec. Last updated by anonymous at January 23 2019, 9:10:10 PM.

GDELT

DIFFICULTÉS

GDELT

DIFFICULTÉS

DIFFICULTÉS

- Problème avec EC2 la veille de la présentation

DIFFICULTÉS

- Problème avec EC2 la veille de la présentation

    - Impact :

        - Passage sur EMR

        - Préparation des données avant passage dans Cassandra à refaire

DIFFICULTÉS

- Problème avec EC2 la veille de la présentation
  - Impact:
    - Passage sur EMR
    - Préparation des données avant passage dans Cassandra à refaire

```
java.lang.NoSuchMethodError: com.amazonaws.services.s3.transfer.TransferManager.<init>(Lcom/amazonaws/services/
  at org.apache.hadoop.fs.s3a.S3AFileSystem.initialize(S3AFileSystem.java:287)
  at org.apache.hadoop.fs.FileSystem.createFileSystem(FileSystem.java:2669)
  at org.apache.hadoop.fs.FileSystem.access$200(FileSystem.java:94)
  at org.apache.hadoop.fs.FileSystem$Cache.getInternal(FileSystem.java:2703)
  at org.apache.hadoop.fs.FileSystem$Cache.get(FileSystem.java:2685)
  at org.apache.hadoop.fs.FileSystem.get(FileSystem.java:373)
  at org.apache.hadoop.fs.Path.getFileSystem(Path.java:295)
  at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.setInputPaths(FileInputFormat.java:500)
  at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.setInputPaths(FileInputFormat.java:469)
  at org.apache.spark.SparkContext$$anonfun$binaryFiles$1.apply(SparkContext.scala:921)
  at org.apache.spark.SparkContext$$anonfun$binaryFiles$1.apply(SparkContext.scala:916)
  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
  at org.apache.spark.SparkContext.withScope(SparkContext.scala:693)
  at org.apache.spark.SparkContext.binaryFiles(SparkContext.scala:916)
  ... 62 elided
```

```
ClassNotFoundException: com.amazonaws.services.s3.AmazonS3Client
```

or similar errors related to another `com.amazonaws` class means that one or more of the `aws-*-sdk` JARs are missing.

Solution: Add the missing JARs to the classpath.

Missing Method in `com.amazonaws` Class

This can be triggered by incompatibilities between the AWS SDK on the classpath and the version with which Hadoop was compiled.

The AWS SDK JARs change their signature between releases often, so the only way to safely update the AWS SDK version is to recompile Hadoop against the later version.

There is nothing the Hadoop team can do here: if you get this problem, then you are on your own. The Hadoop developer team did look at using reflection to bind to the SDK, but there were too many changes between versions for this to work reliably. All it did was postpone version compatibility problems until the specific codepaths were executed at runtime. This was actually a backward step in terms of fast detection of compatibility problems.

Troubleshooting Amazon S3

Common problems that you may encounter while working with Amazon S3 include:

1. Classpath Related Errors
2. Authentication Failures
3. S3 Inconsistency Side-Effects

Classpath Related Errors

DIFFICULTÉS

- Problème avec EC2 la veille de la présentation

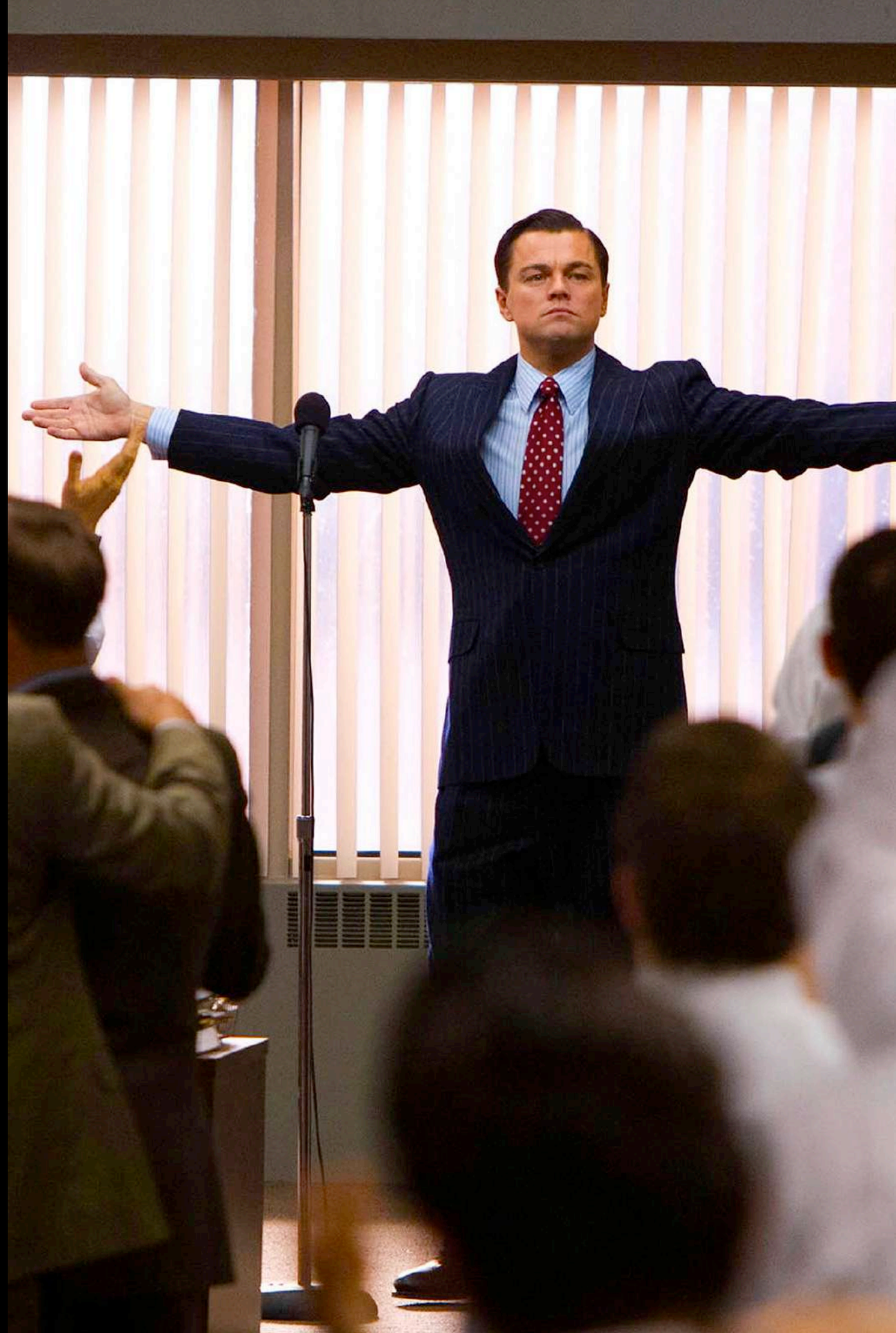
    - Impact :

        - Passage sur EMR

        - Préparation des données avant passage dans Cassandra à refaire

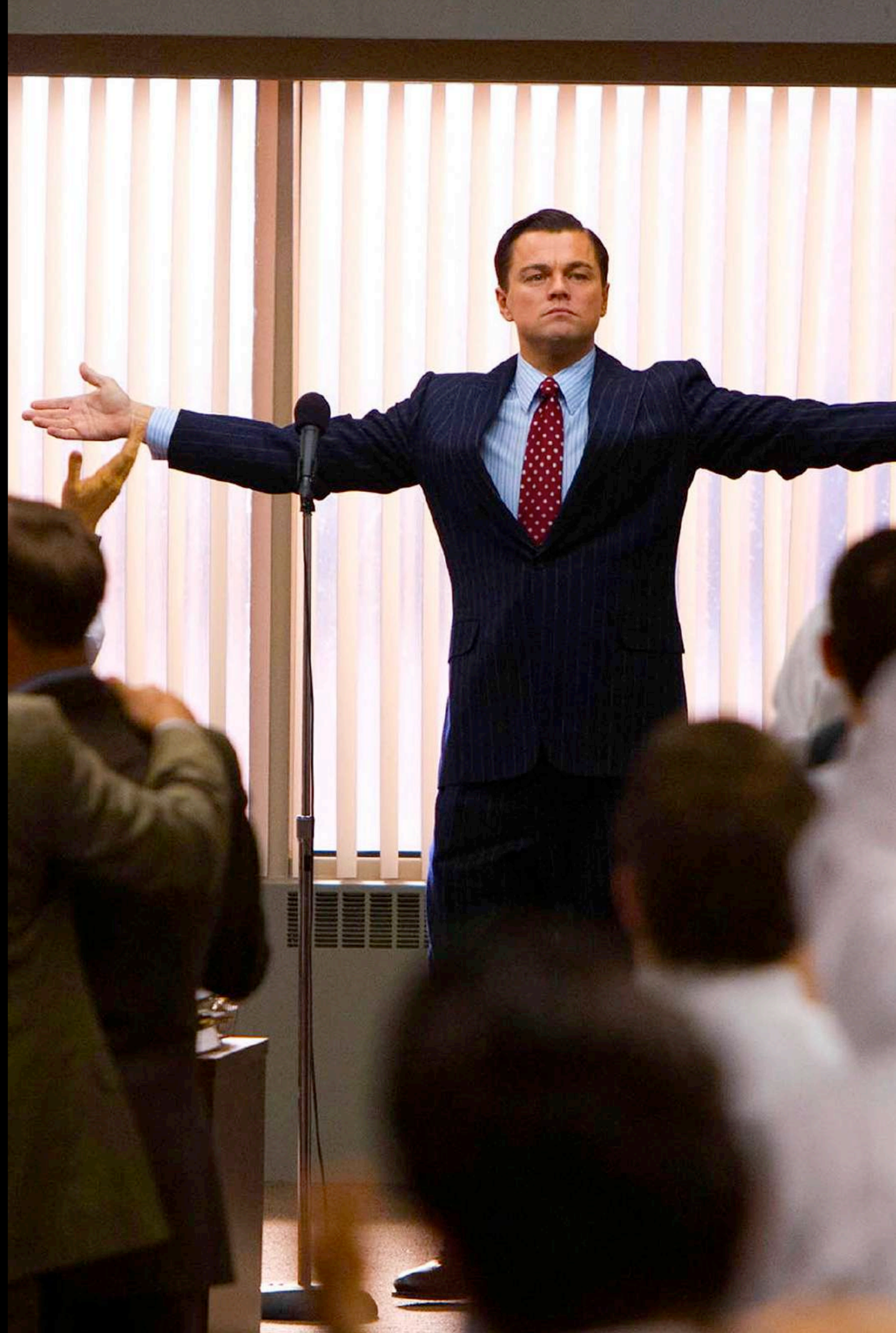
GDELT

BUDGET



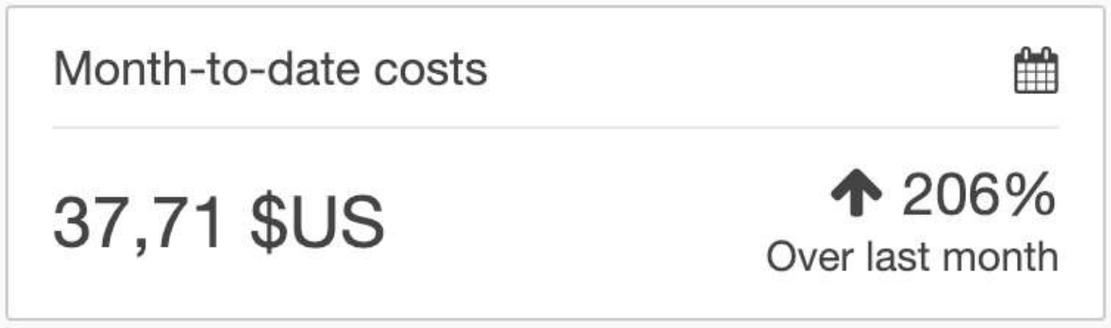
GDELT

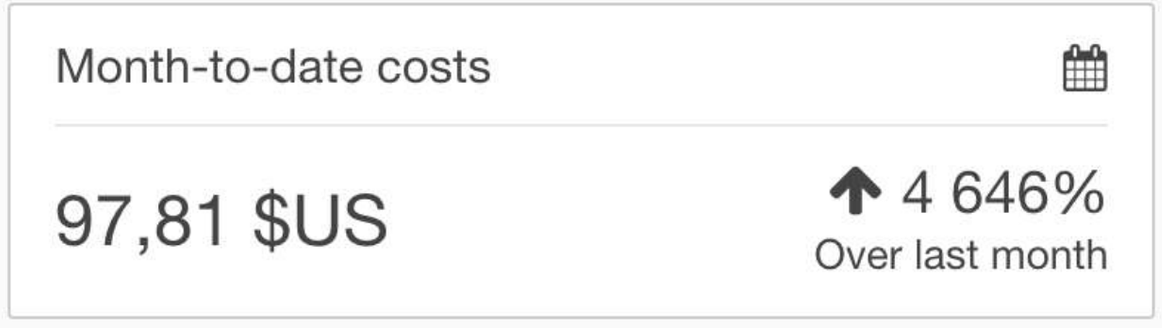
BUDGET



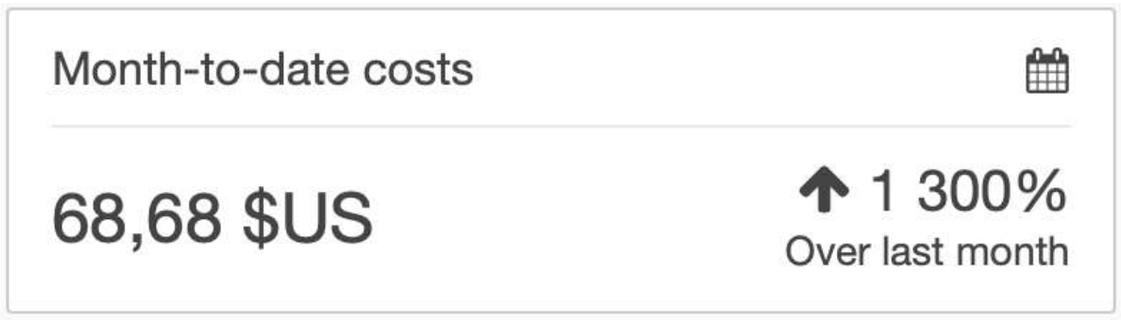
BUDGET

- Charger et stocker les données dans S3, pre-processing vers Cassandra :





- Architecture EC2 :



204.2$

GDELT

AMÉLIORATIONS

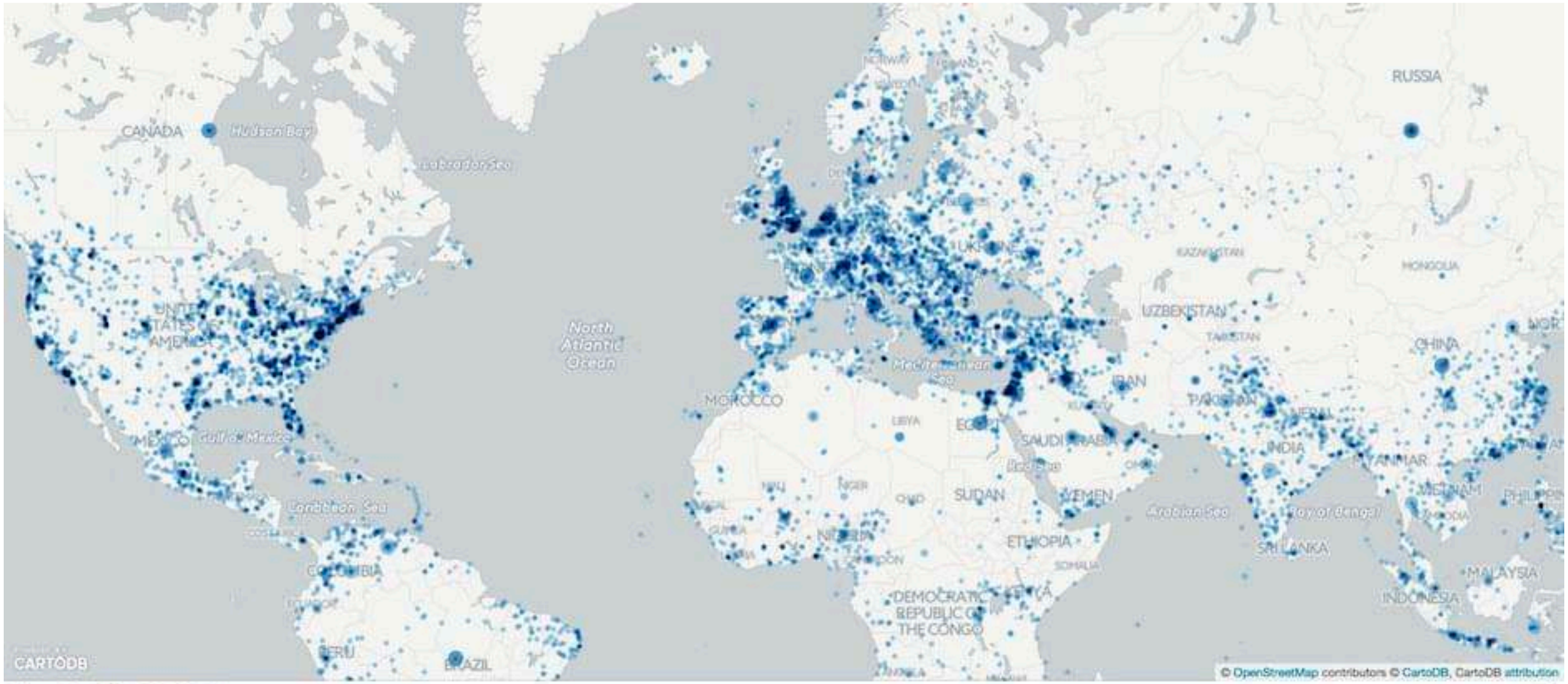
GDELT

AMÉLIORATIONS

AMÉLIORATIONS POSSIBLES

- Données sur une année complète
- Streaming : Chargement toutes les 15mn
- Exploration approfondie (ML, DL) des données
- Dashboard + CartoDB
- Automatiser le déploiement (Ansible, Docker…)

AMÉLIORATIONS POSSIBLES



GDELT

DEMONSTRATION

TELECOM PARISTECH - 2019

TELECOM PARISTECH - 2019

GDELT - NOSQL

MERCI DE VOTRE ATTENTION !