



Une école de l'IMT

Projet NoSQL GDELT

16 Janvier 2018

Rémi Ferreira
Maxime Poulain
Jean - Marc Sevin
Olivier Schultz



La loi du mort kilométrique



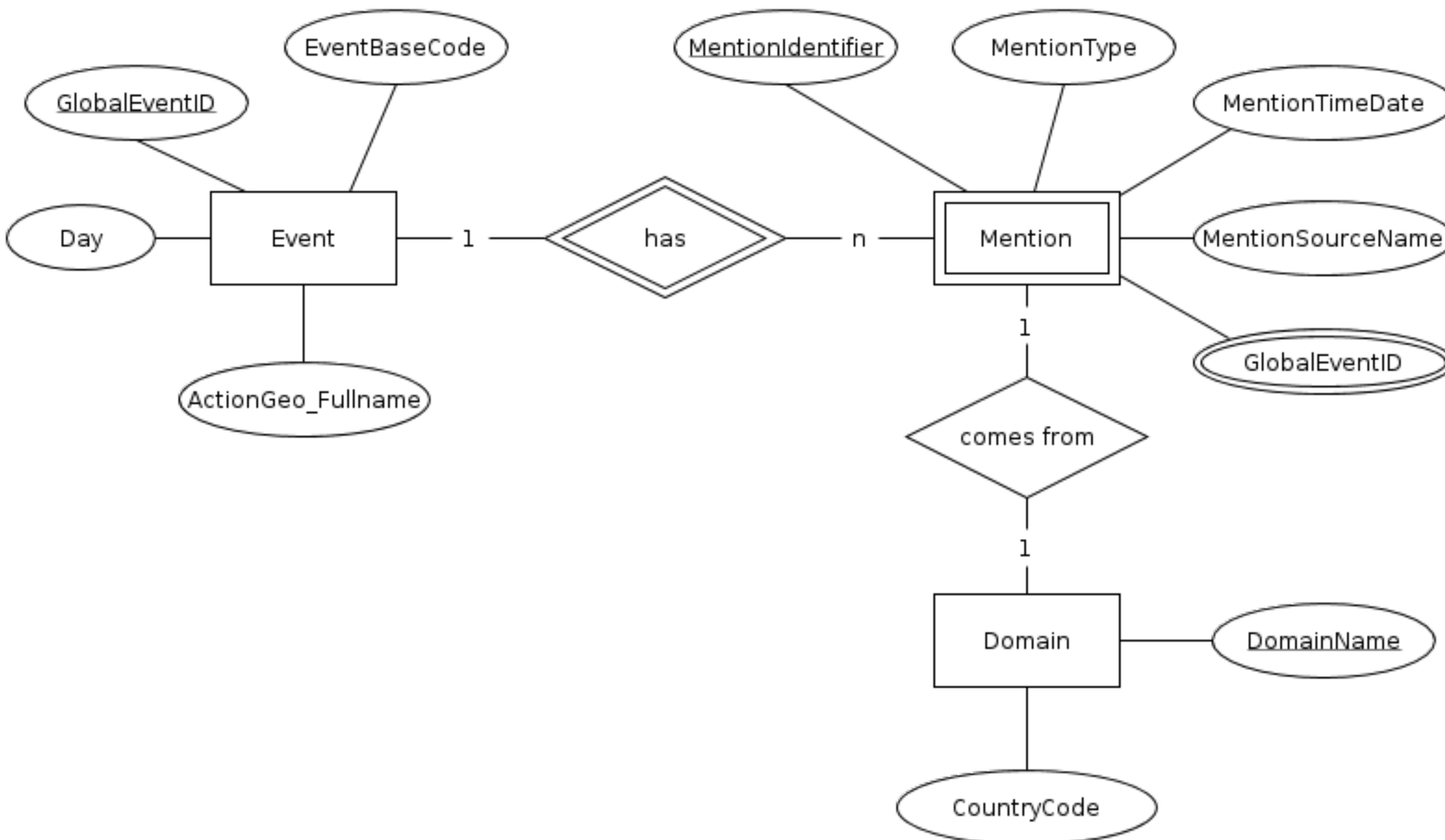
Notre problématique

....

- Identifier un attentat à partir de sa localisation, son code CAMEO et sa date (ex. Las Vegas, 180, 01/10/2010)
- Déterminer le nombre de fois où cet événement a été mentionné dans un article du web, par jour et par pays
- Afficher ces fréquences sur une carte du monde
- Comparer le phénomène de diffusion géographique et temporel pour différents attentats
- Vérifier la conformité avec la loi du mort kilométrique

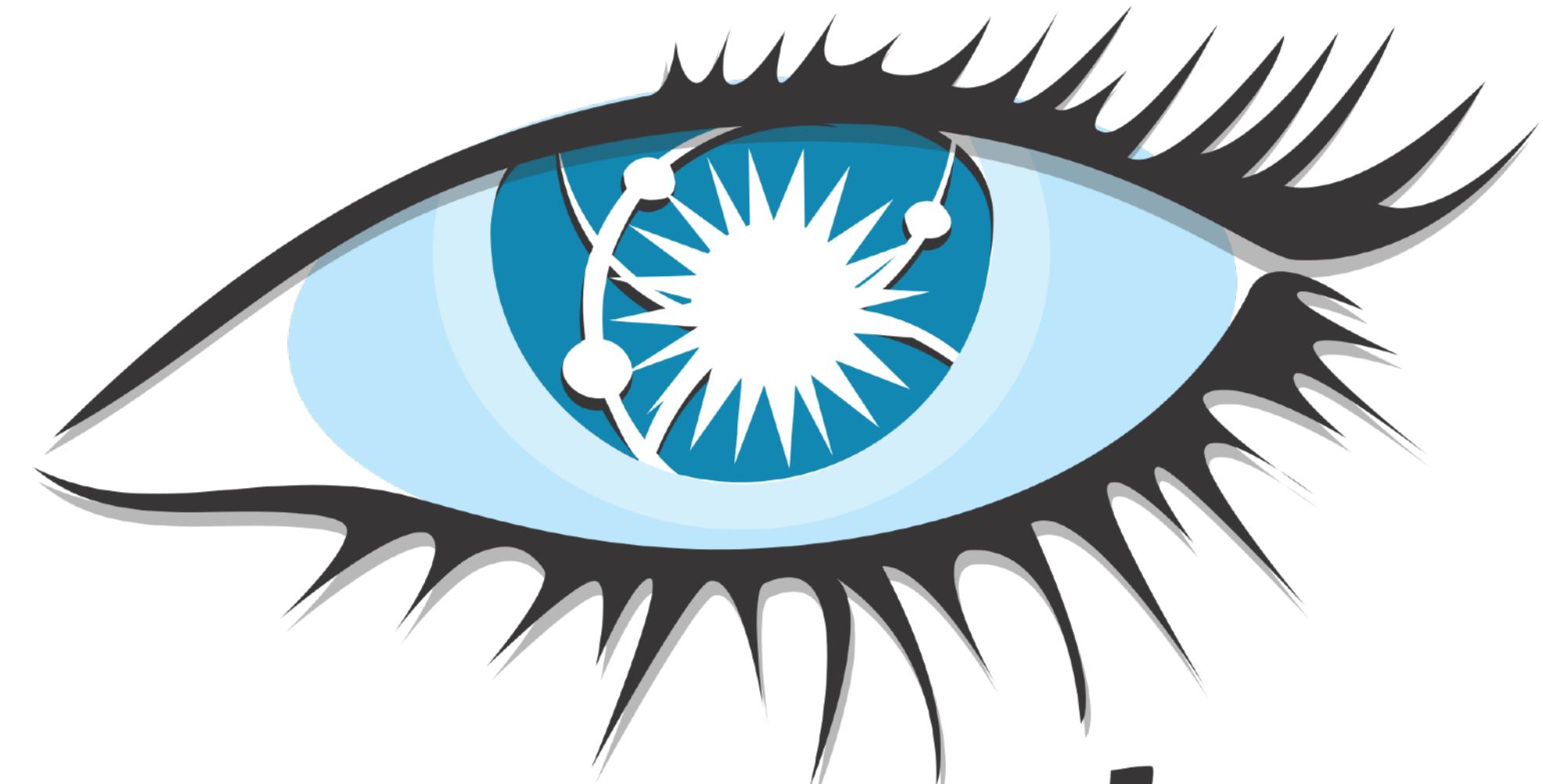
Modèle Conceptuel de Données

....

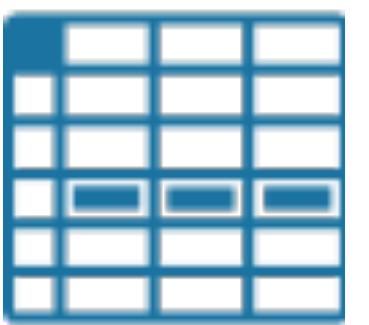


Pourquoi Cassandra ?

....



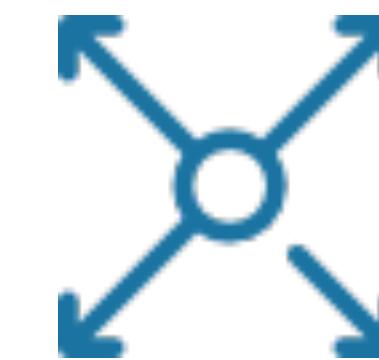
cassandra



Données
structurées



Rapidité



Passage à
l'échelle



Robustesse

Modèle Logique de Données

....



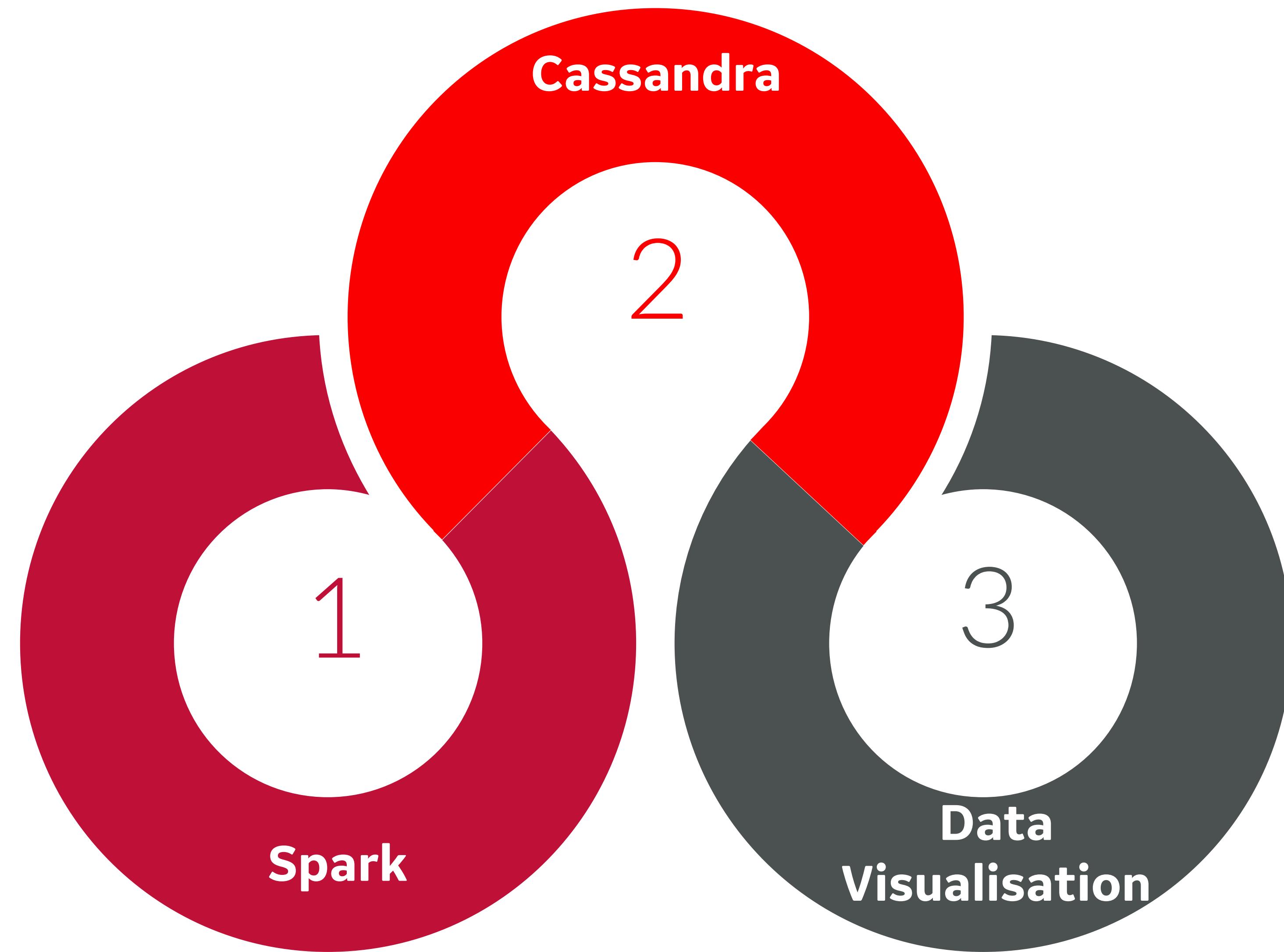
gdel

mentions_by_location_eventcode

Partition Key	Clustering columns / Columns (Cells)		
Las Vegas, 180	day	country	frequency
	20171001	FR	5
	day	country	frequency
	20171001	US	18
	day	country	frequency
	20171002	FR	7

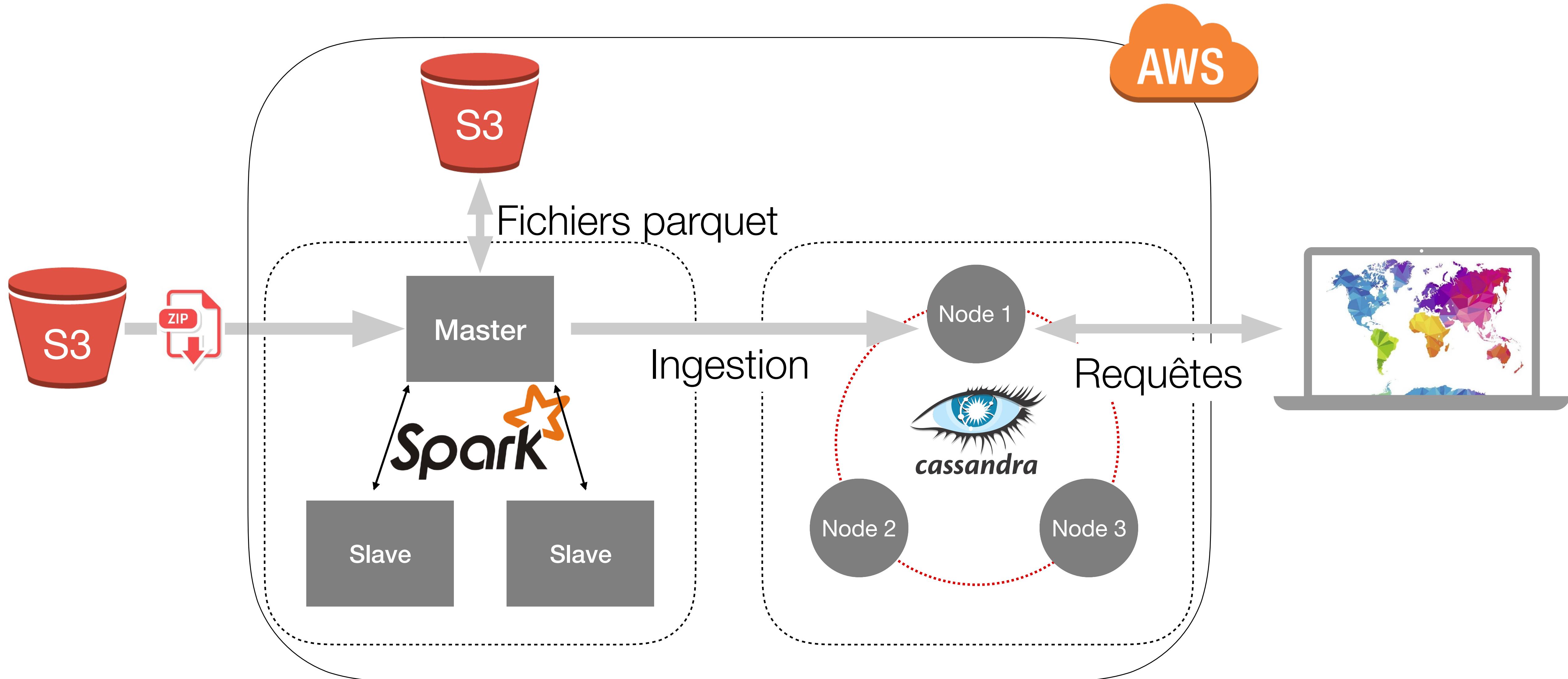
Les grandes Étapes

....



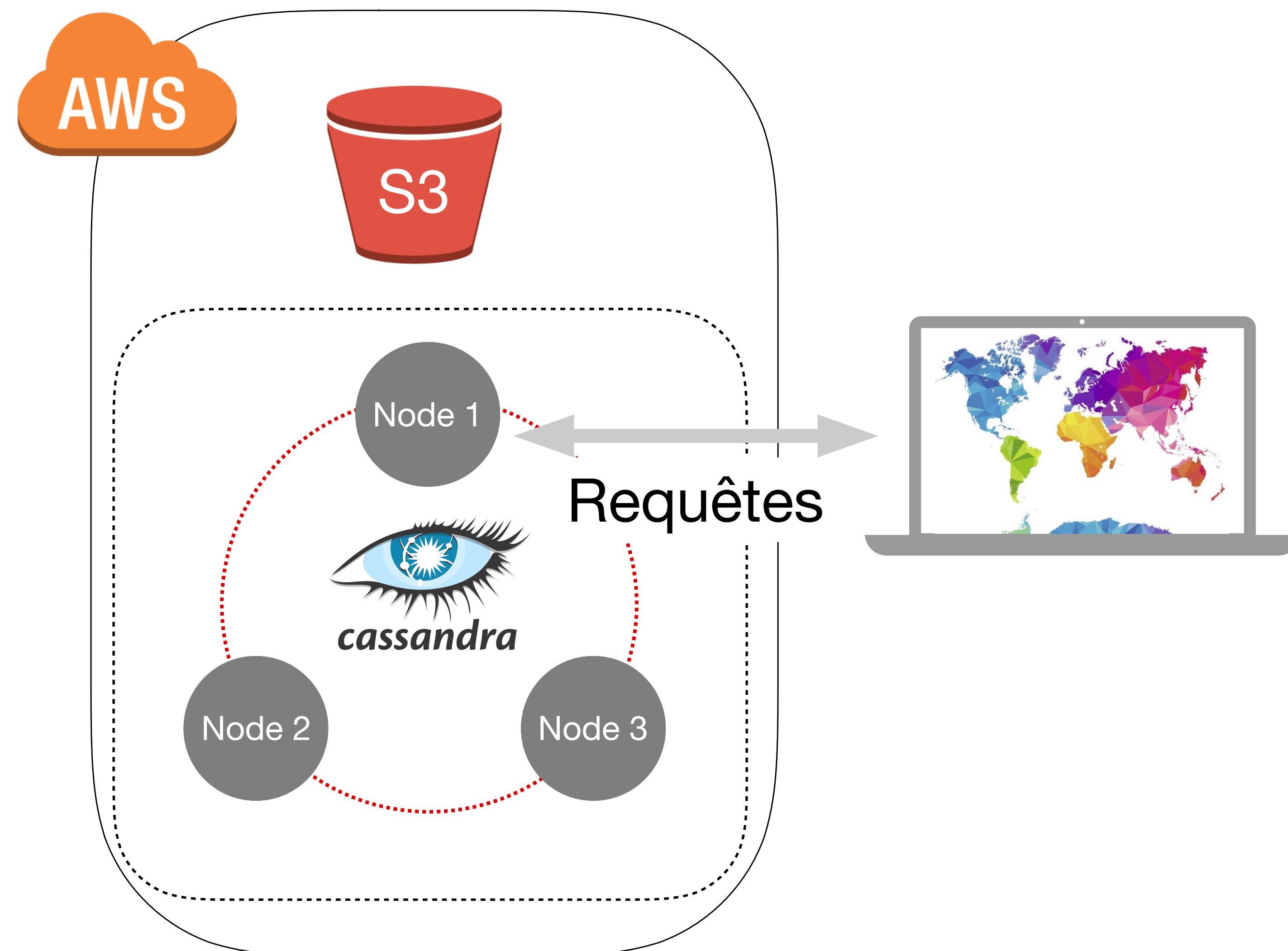
Architecture Utilisée

....



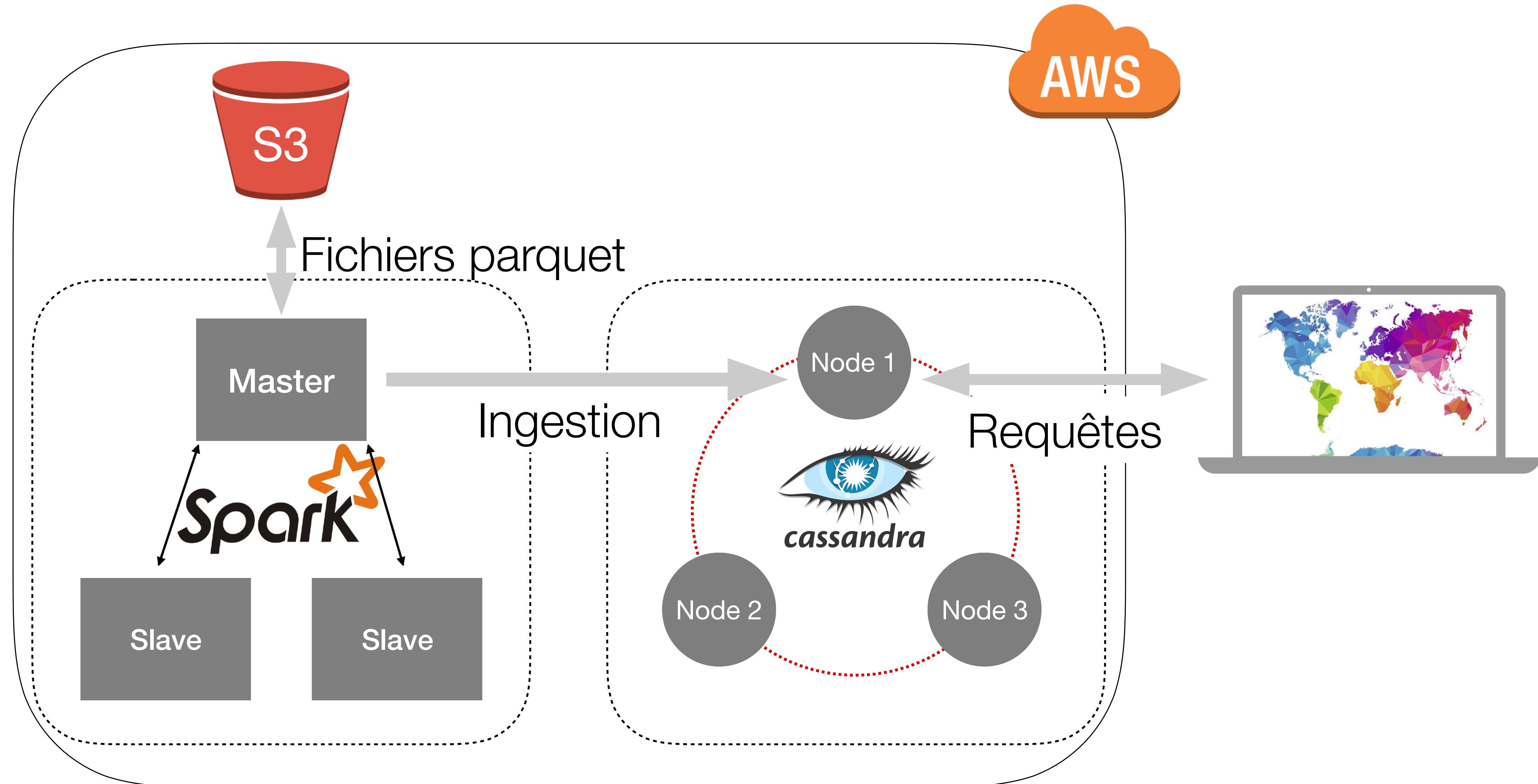
Architecture Actuelle

....



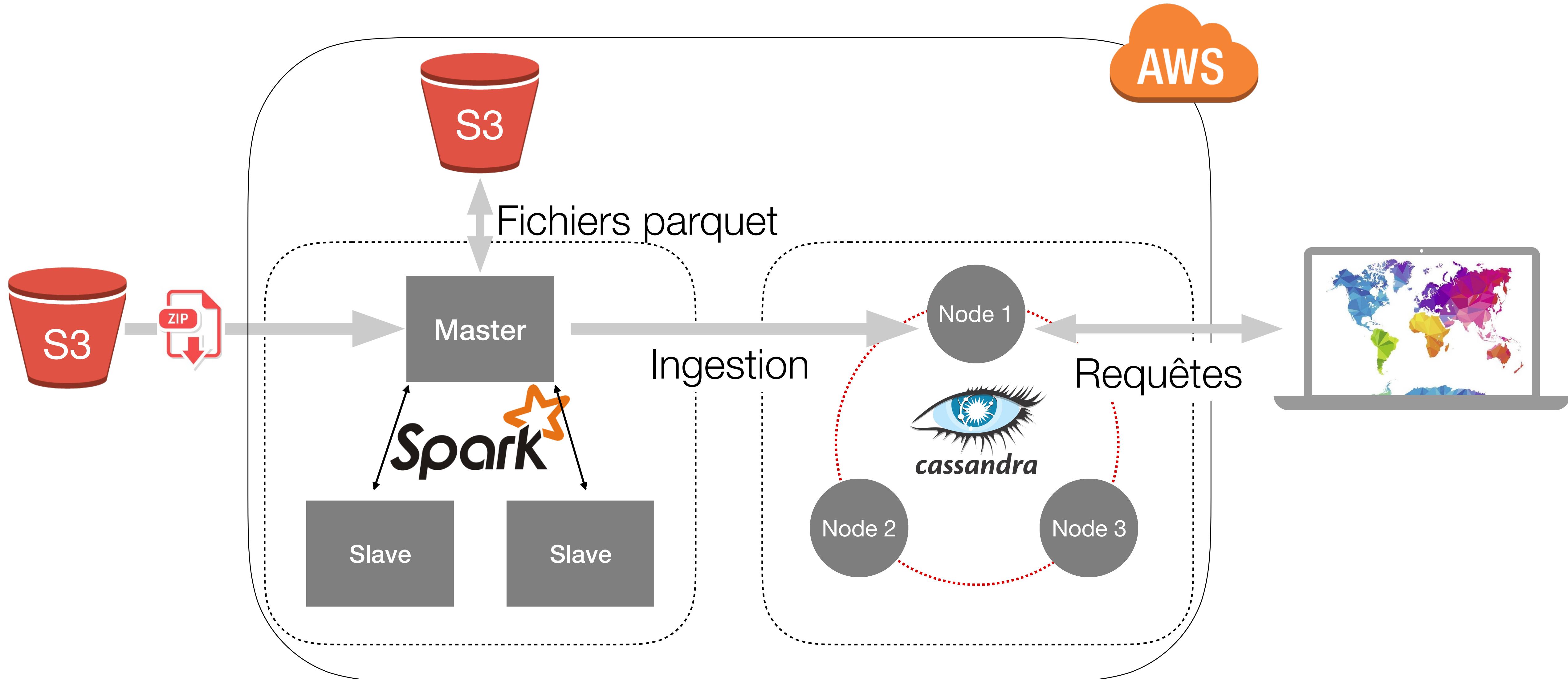
Nouveau Use Case

....



Nouveau Use Case

....



Spark
....

**Import des
données**
(Oct/Nov/Dec)

**Preprocess
Partie 1**

**Sauvegarde
dans S3**

**Preprocess
Partie 2**

**Injection
dans
Cassandra**

Cassandra

....

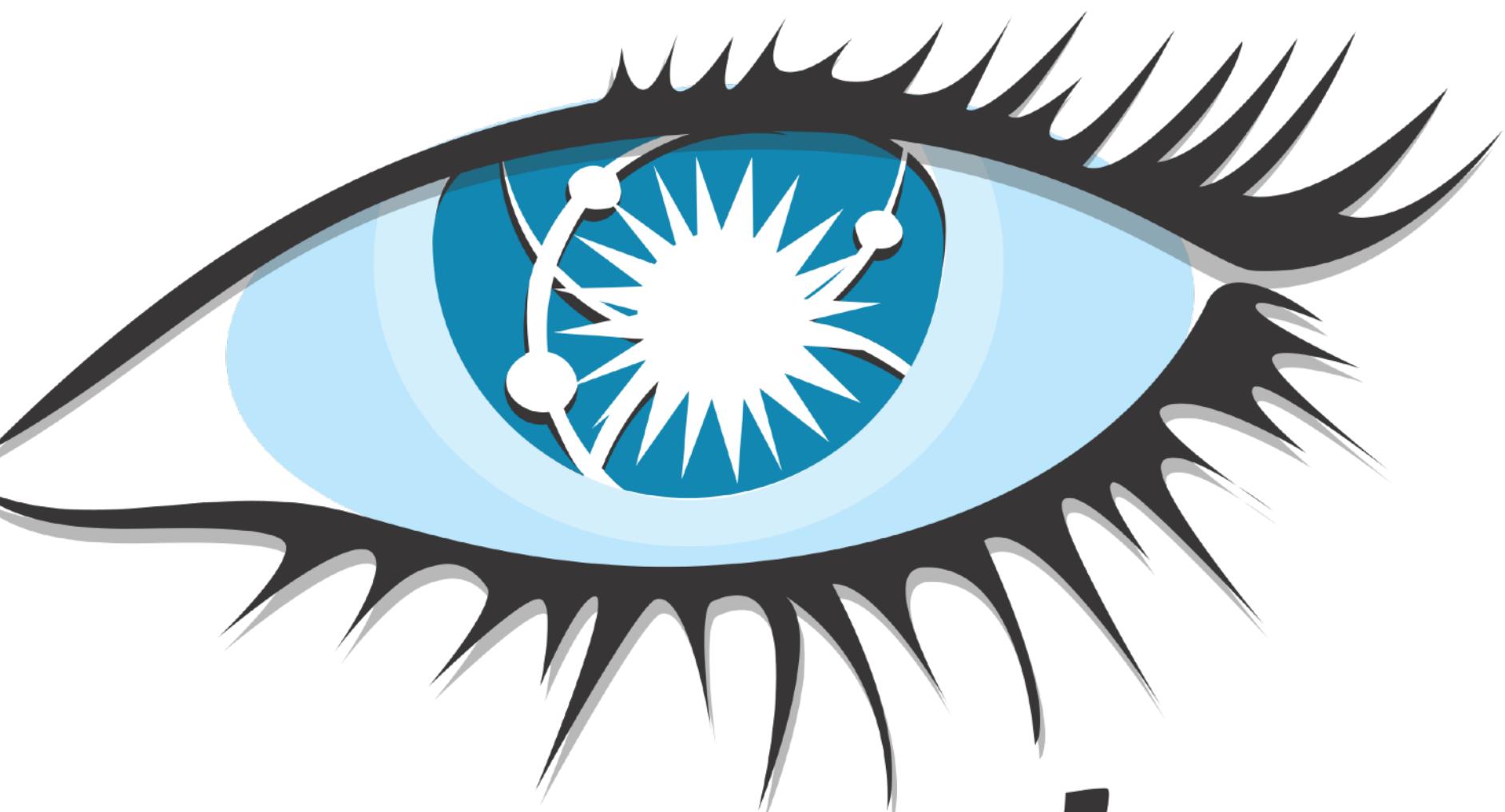
| 2,5 mois de données chargés en base

| 3 Noeuds

| RéPLICATION Factor de 3

| 100% des données sur les 3 noeuds

| Tolérant à la perte d'un serveur



Visualisation

....

Code en python, connecteur
cassandra

Libraries: Folium, Tkinter

```
SELECT * FROM gdelt WHERE  
location = input AND CAMEOCode =  
input
```



Démo



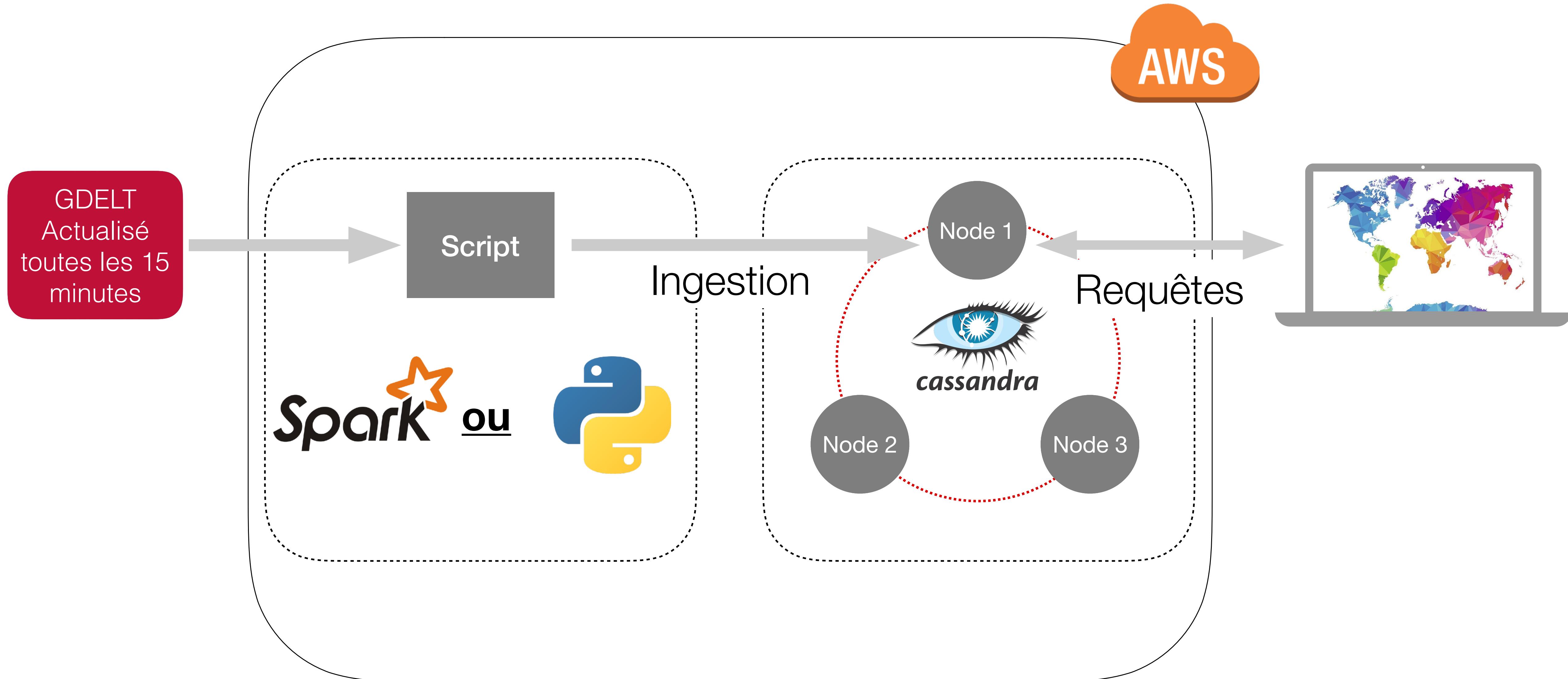
Améliorations possibles

....

- Etendre notre visualisation sur d'autres types d'événements que les attentats.
- Mieux filtrer les villes lors du pré-processing pour pouvoir requêter un évènement efficacement
- Affiner le moyen de détecter un événement en particulier.
- Faire pointer la visualisation sur une web app (Flask, Django, ...)
- Adapter l'infrastructure pour accueillir des données par batch toutes les 15 minutes

Architecture possible en « Streaming »

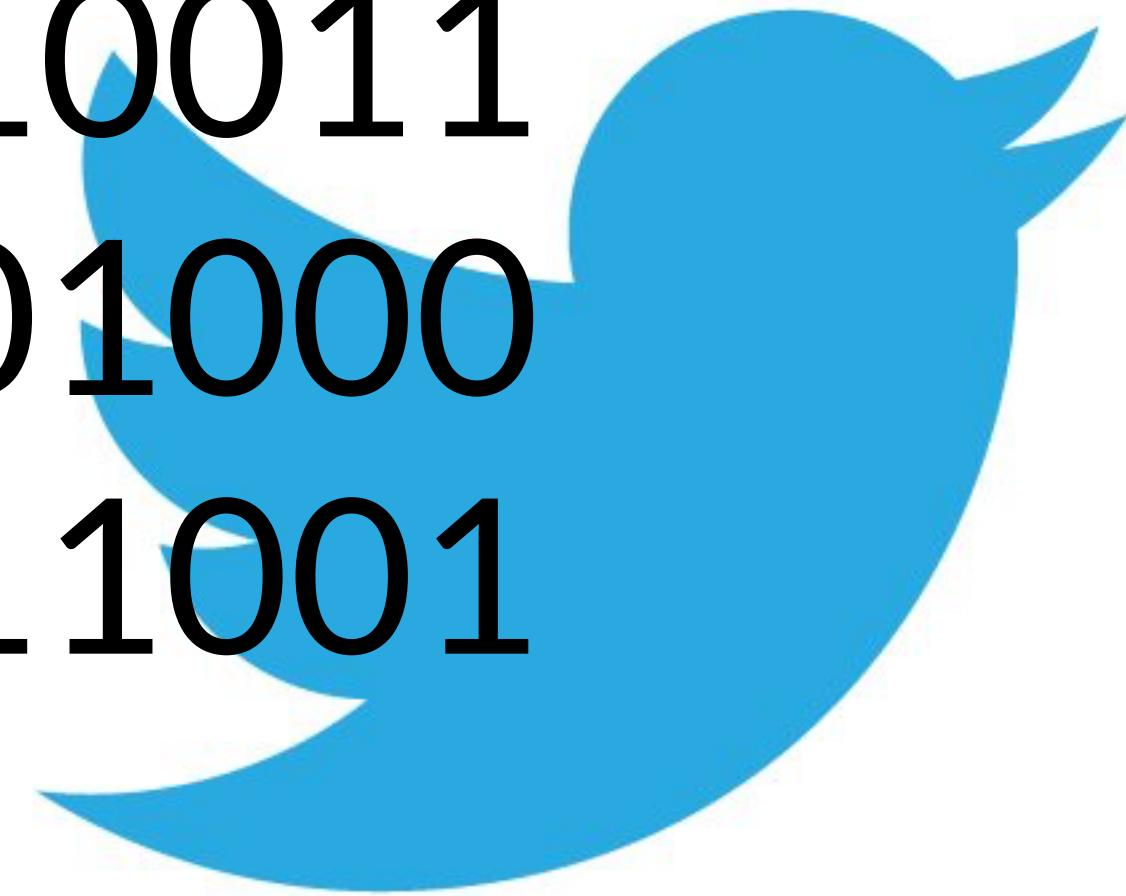
....



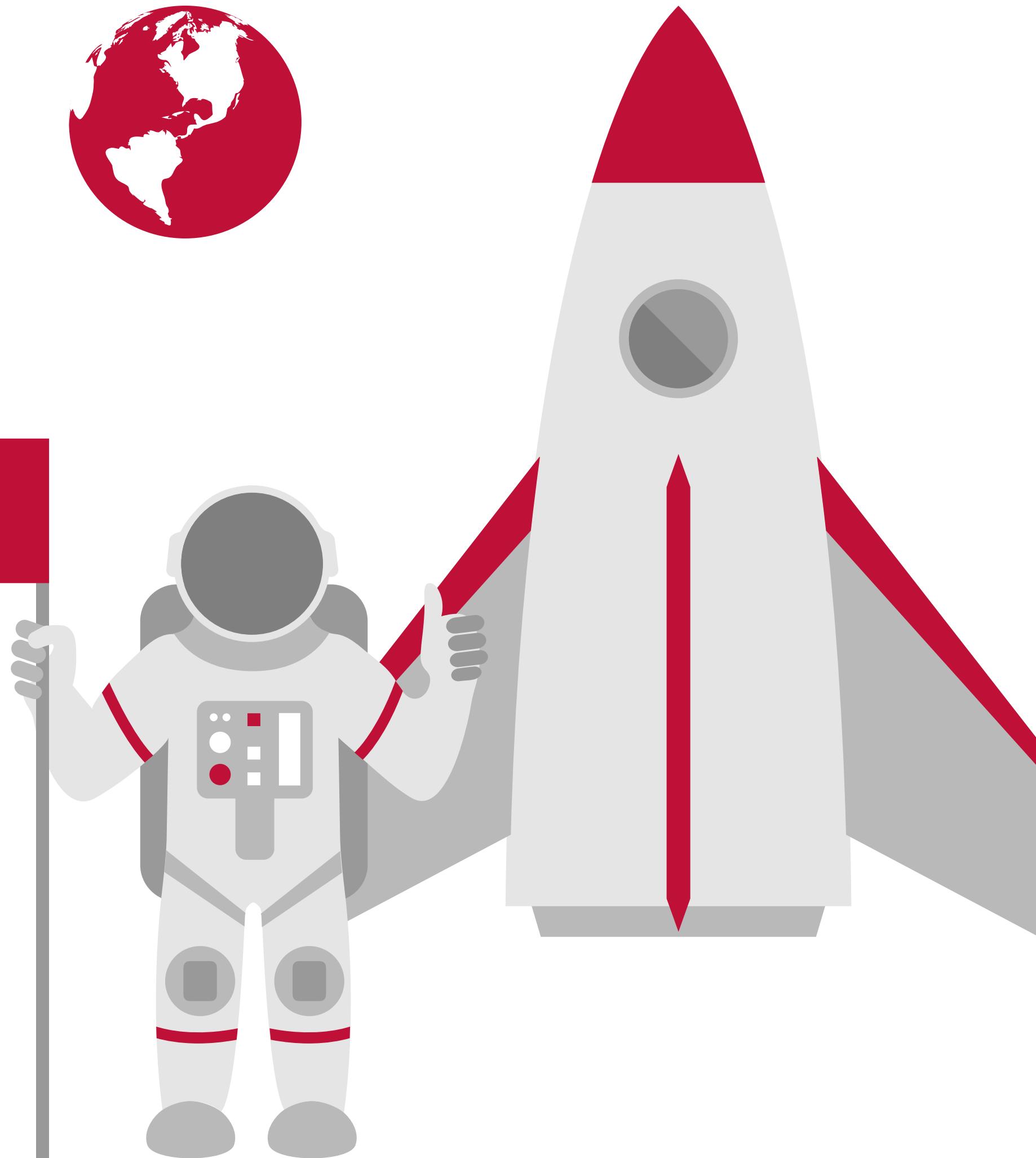
Conclusion

....

1001010
1110011
0101000
1111001



FINISH



A sepia-toned photograph of a desk setup. In the foreground, a spiral-bound notebook with horizontal ruling lies diagonally. Behind it, a portion of a laptop keyboard is visible. In the bottom right corner, the dark, rectangular screen of a smartphone is partially visible. The lighting is warm and focused on the center of the frame.

Questions