



# INF 728 - NoSQL

## *Projet GDELT*

Thomas Koch  
Xavier Bracquart  
Fayyaz Ali  
Philippe Bénézech

The background of the image is a composite of a satellite view of Earth and a global network diagram. The Earth is shown from space, with green landmasses and blue oceans. Overlaid on this is a dense web of thin, glowing yellow lines that represent a global network, possibly representing data connections or satellite orbits. These lines radiate from various points across the globe, creating a complex, interconnected pattern. The overall color palette is dominated by the blues and greens of the Earth, with the bright yellow of the network lines providing a high-contrast visual element.

# **GDELT Project**



# Modélisation des données

# Les données

## EVENT

- GBOLEVENTID
- SQLDATE
- Actor1 Infos
- Actor2 Infos
- NumMentions
- NumSources
- NumArticles
- AvgTone
- Actor1 Geo
- Actor2 Geo
- Action Geo

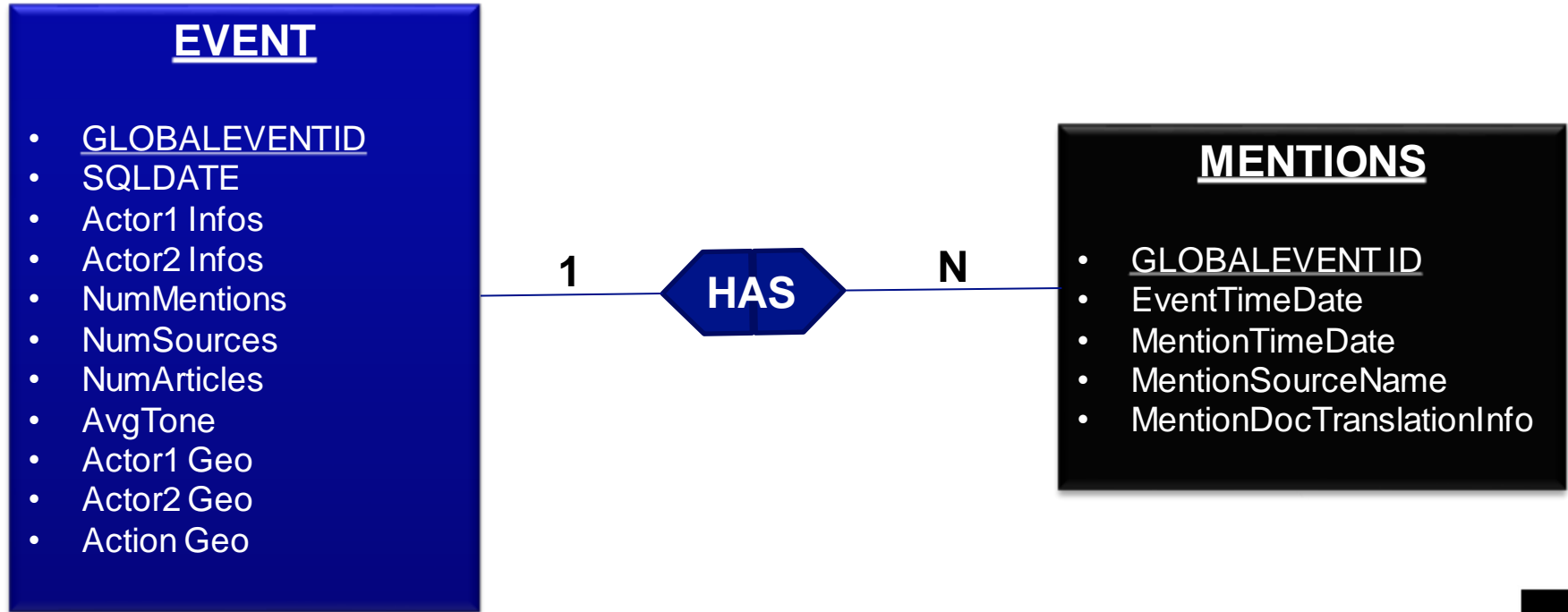
## MENTIONS

- GBOLEVENT ID
- EventTimeDate
- MentionTimeDate
- MentionSourceName
- MentionDocTranslationInfo

## GKG

- GKGRECORD ID
- DATE
- SourceCommonName
- Themes
- Locations
- Persons

## Les relations / modèle conceptuel



# Modélisation des données

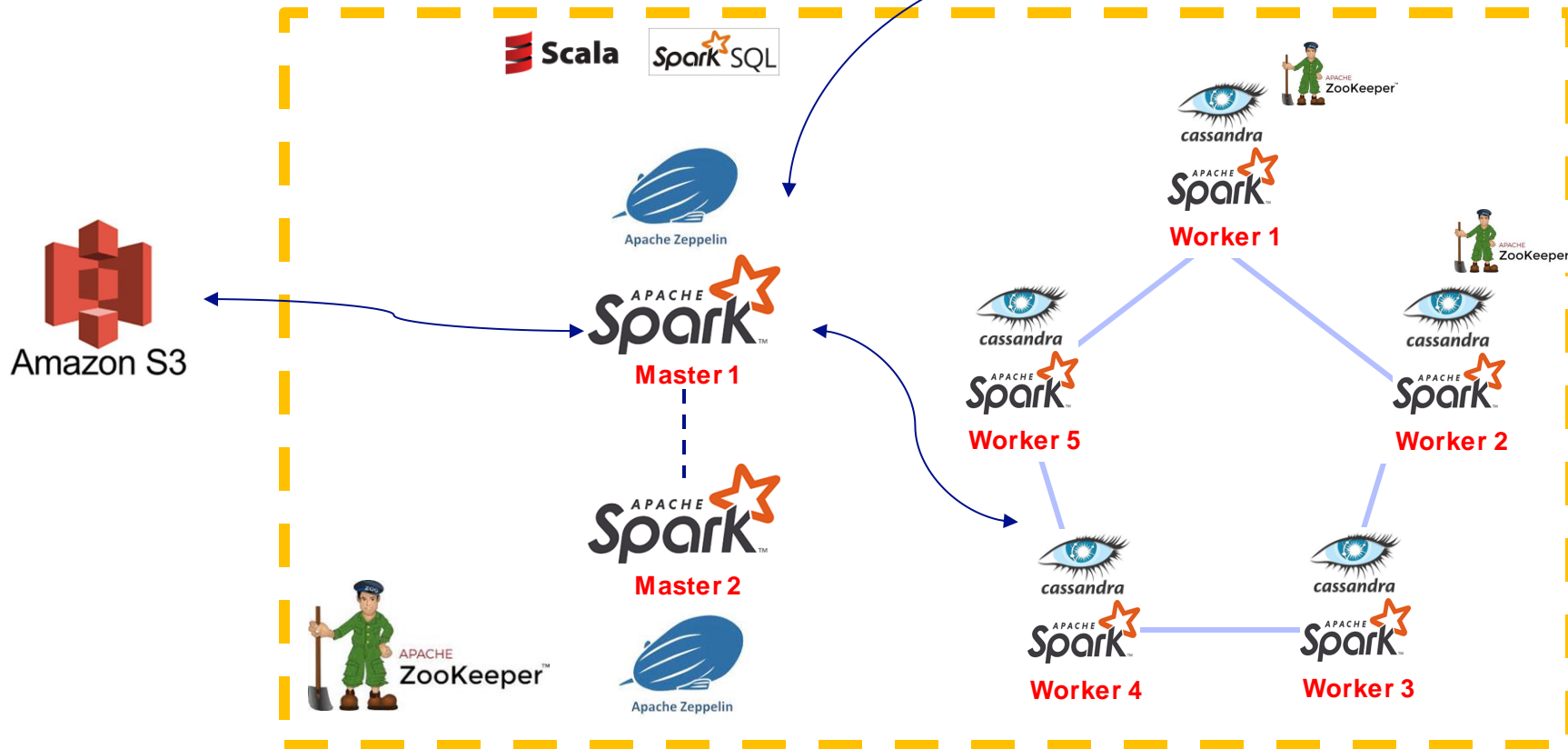
- **Traitement via UDF de certaines colonnes**
  - Dates
  - Langues
- **Beaucoup de valeurs NULL dans tables**
- **N fois GLOBALEVENTID => difficulté pour les jointures (duplication d'informations)**



# Architecture

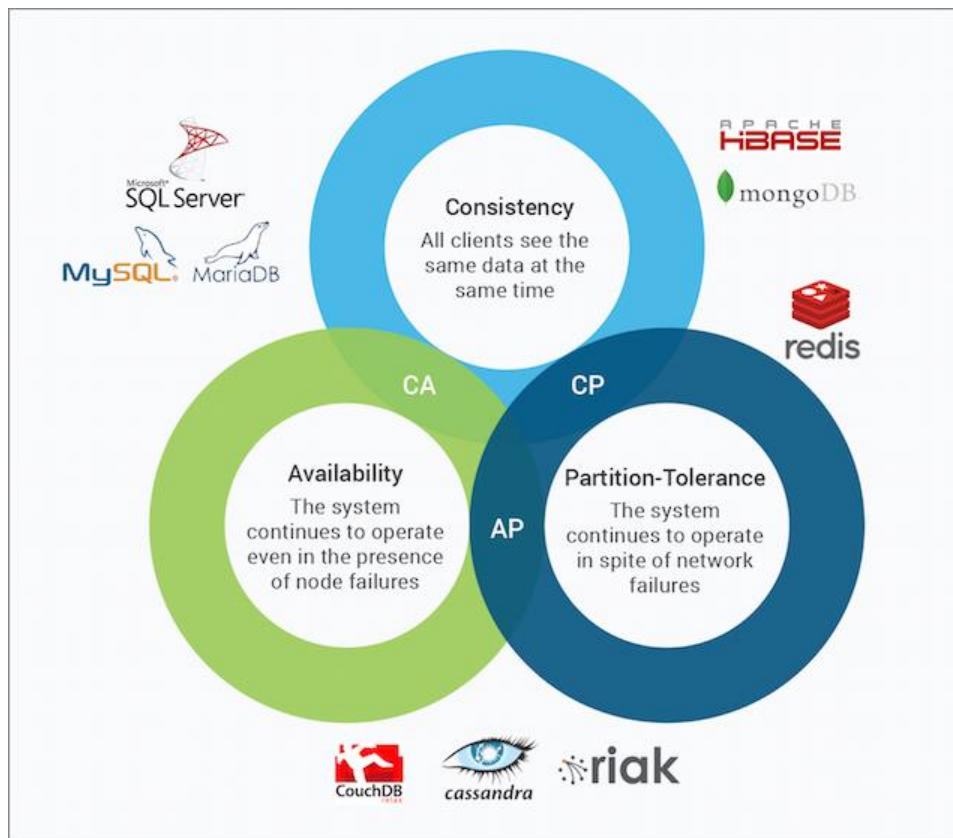
## *Avantages et inconvénients*

# Architecture





# Architecture



# Architecture

## *Avantages et inconvénients*

### ■ Avantages :

- **Configuration à la main**
  - Intérêt pédagogique
  - Maîtrise et compréhension des flux
  - Simplifie le débogage
- **Limitation des coûts**
- **Scalabilité**
  - coûts au juste besoin (t2.micro, t2.large, m5.xlarge)



35 €

### ■ Inconvénients :

- **Configuration à la main**
  - prend du temps / solution clé en main
  - Complexité, moins "user friendly"



# Volumétrie

## *Limites et contraintes*

# Volumétrie

## *Limites et contraintes*

- **Contrainte temporelle pour paramétrage de l'EC2**
  - **Manque de temps pour passer à l'échelle**
  - **Mais architecture capable de tenir la charge (si volumes disques appropriés sur les Workers)**
- **Travail sur un mois de données**
  - (sauf requête 3 pour cause de limitation redimensionnement disque)
- **Environ 58 Go de données sur S3**

## Démonstration

