INF 728 - NoSQL
Projet GDELT

Thomas Koch
Xavier Bracquart
Fayyaz Ali
Philippe Bénézeth


GDELT Project

Modélisation des données

Les données

EVENT

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

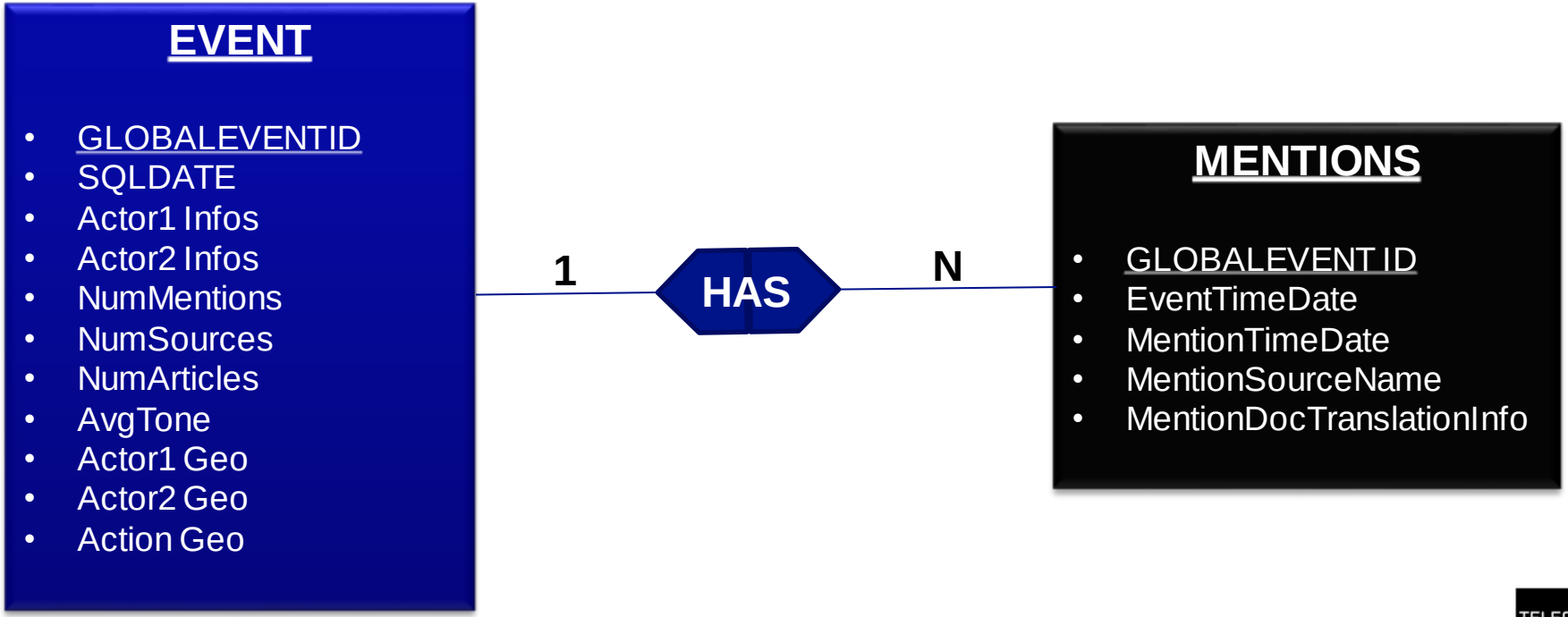
MENTIONS

- GLOBALEVENT ID
- EventTimeDate
- MentionTimeDate
- MentionSourceName
- MentionDocTranslationInfo

GKG

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

Les relations / modèle conceptuel



Modélisation des données

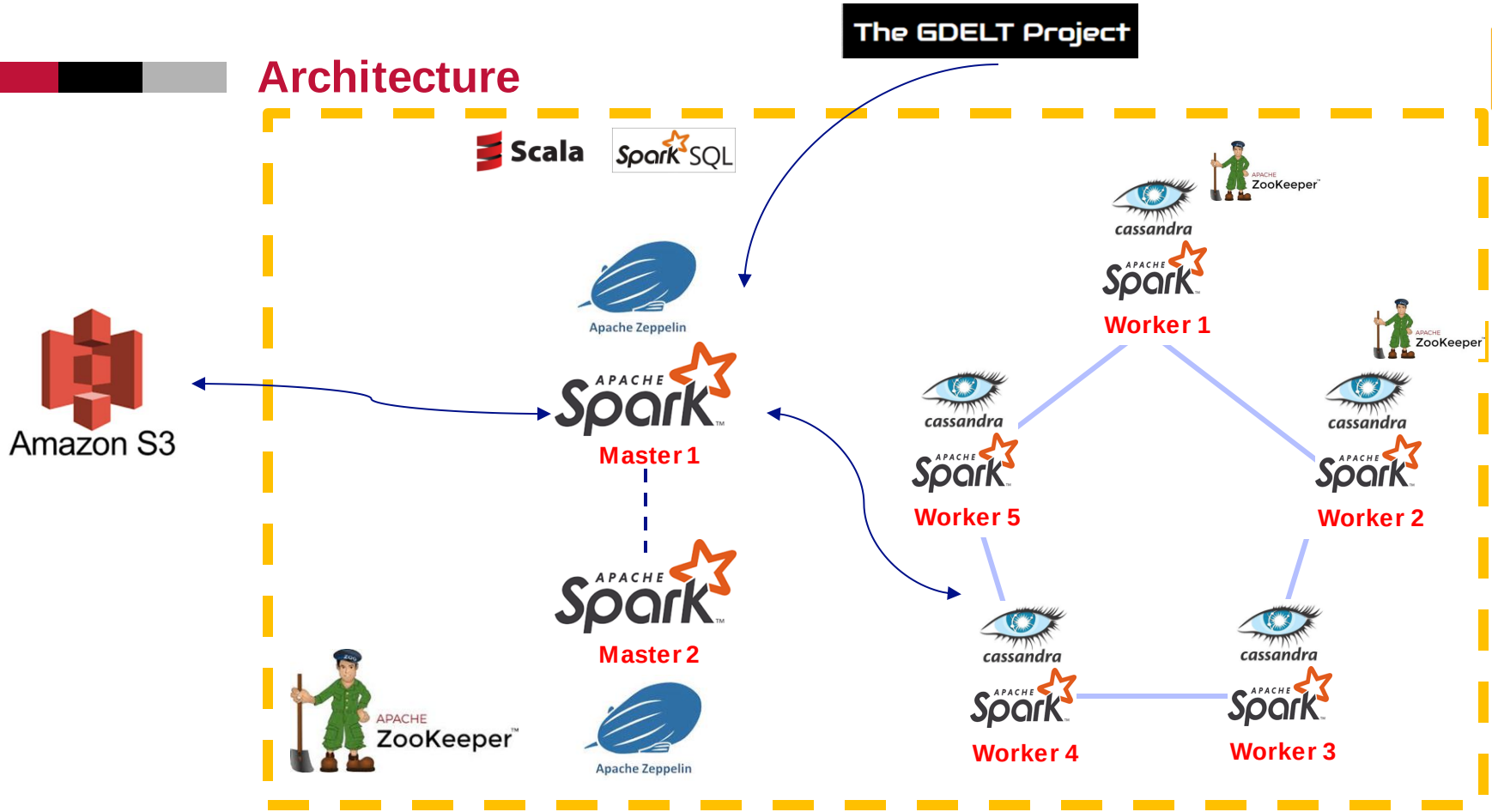
- **Traitement via UDF de certaines colonnes**
  - **Dates**
  - **Langues**

- **Beaucoup de valeurs NULL dans tables**

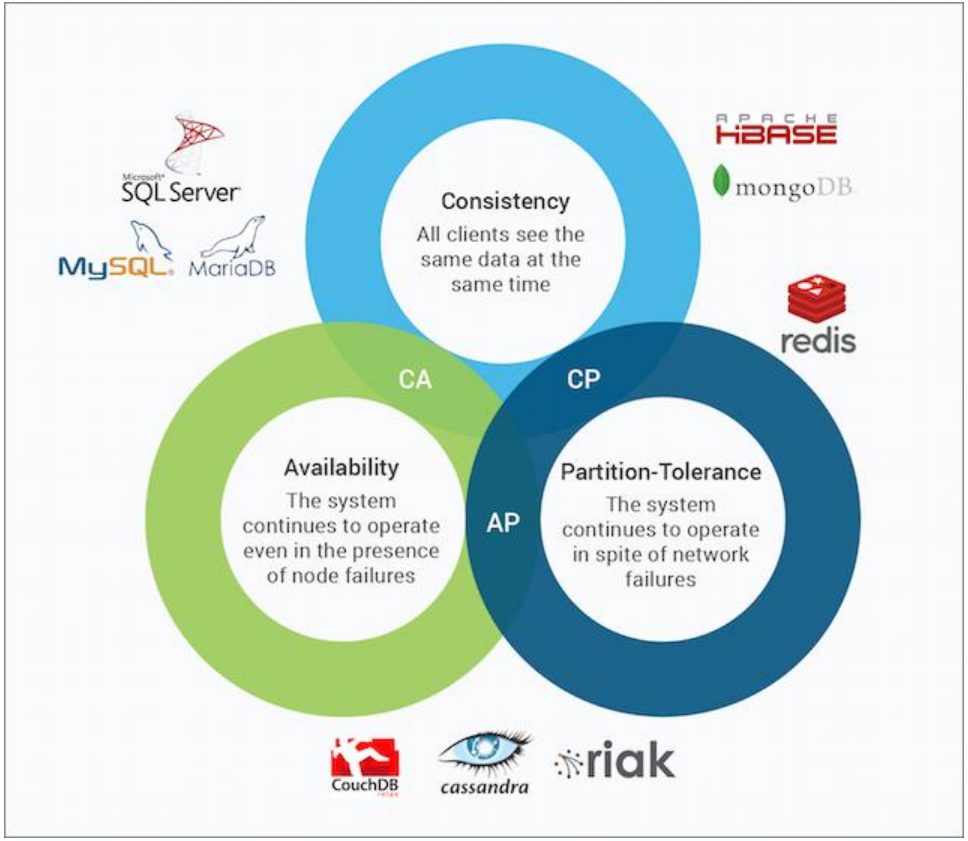
- **N fois GLOBALEVENTID => difficulté pour les jointures (duplication d'informations)**

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)

- **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