



Une école de l'IMT




Erwan FLOCH  
Nicolas LOUIS  
Vincent MARTINEZ

Thomas RIVIERE  
Chloé YOUNES

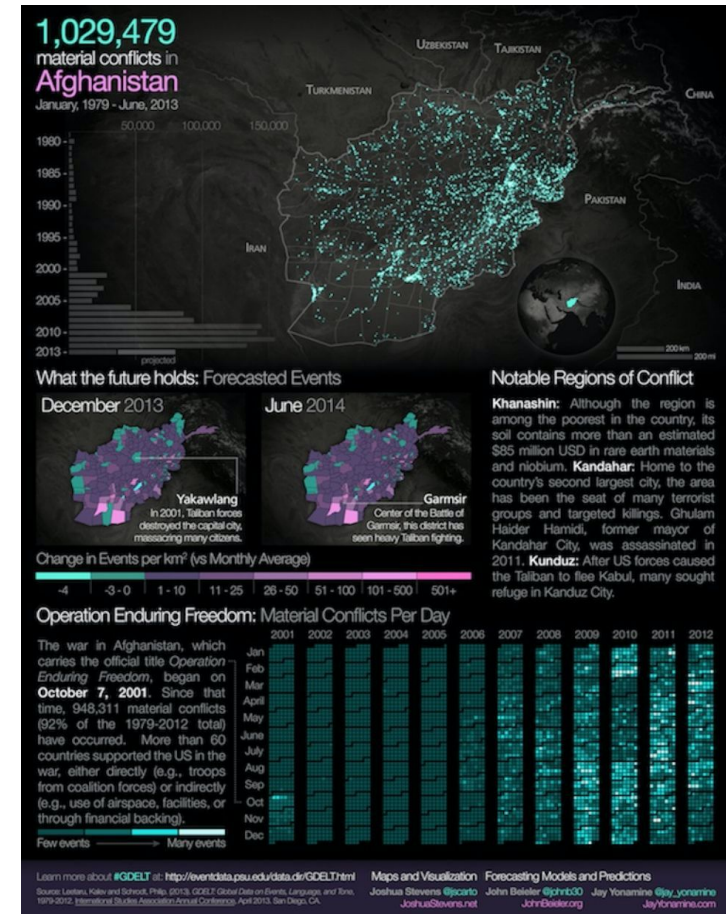


# Sommaire

- **Contexte**
- **Architecture proposée**
- **Modélisation et remplissage de la base de données**
- **Performances de la modélisation et budget**
- **Conclusion**
- **Démo**

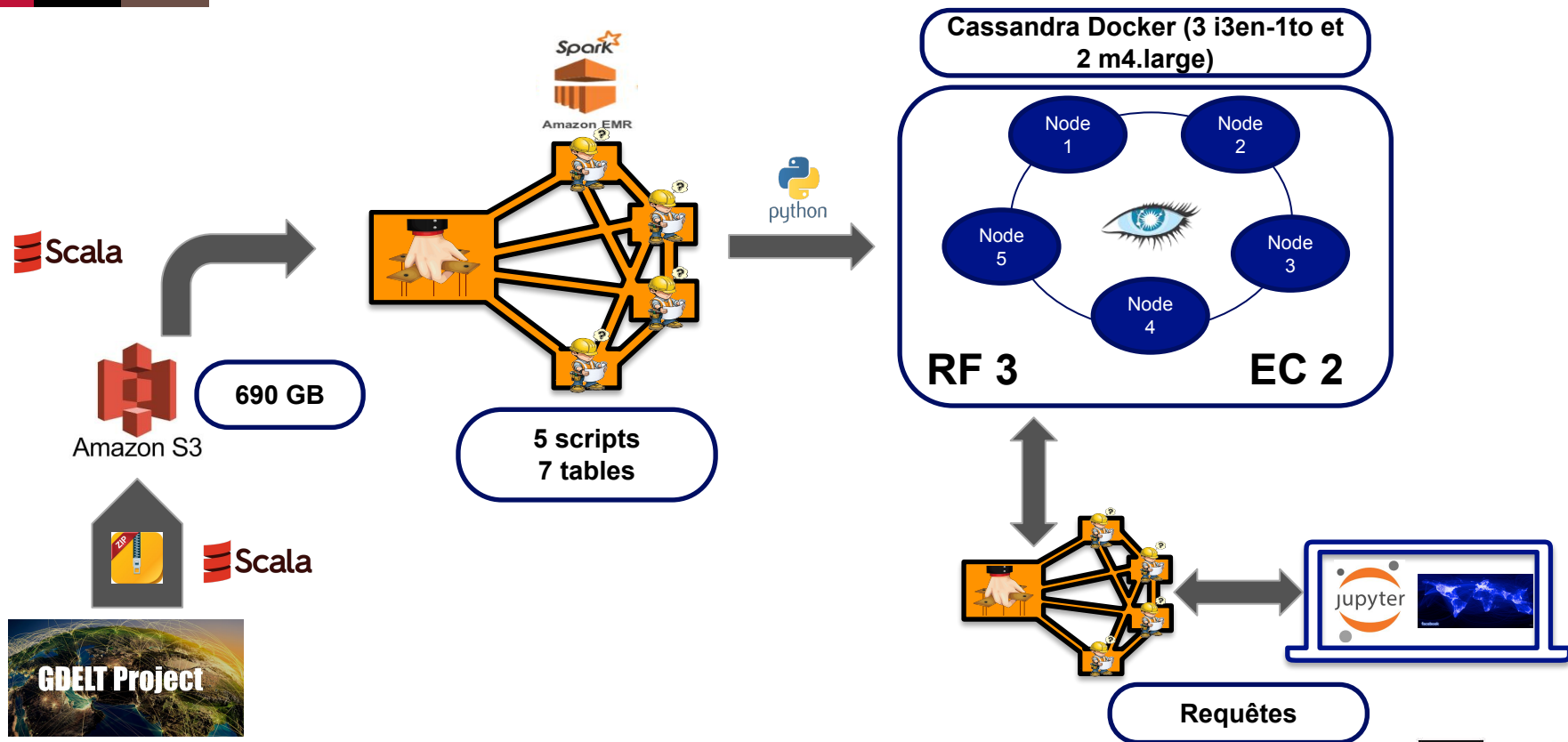
- 
- **Contexte**
  - Architecture proposée
  - Modélisation et remplissage de la base de données
  - Performances de la modélisation et budget
  - Conclusion
  - Démo

L'objectif du projet est de concevoir un système qui permet d'analyser le jeu de données GDELT et ses sources de données sur l'année 2019.

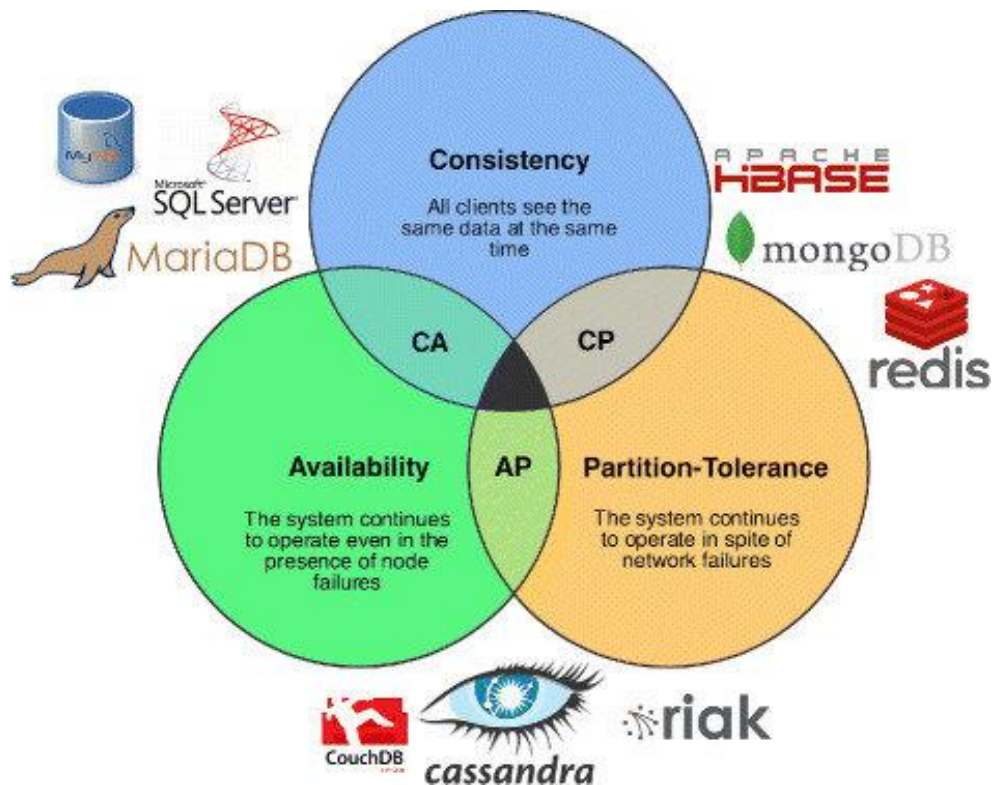


- Contexte
- **Architecture proposée**
- Modélisation et remplissage de la base de données
- Performances de la modélisation et budget
- Conclusion
- Démo

# Détail de l'architecture



# Pourquoi Cassandra ?



Source: Laurenço et al. Journal of Big Data

- 
- Contexte
  - Architecture proposée
  - **Modélisation et remplissage de la base de données**
  - Performances de la modélisation et budget
  - Conclusion
  - Démo



# Requête 1

- **TABLE Requête 1** obtenue avec un JOIN entre EVENT et MENTIONS

TABLE Requête 1
jour DATE pays TEXT langue TEXT count INT
PRIMARY KEY ((jour), pays, langue))

## Requête 2 - Backward propagation

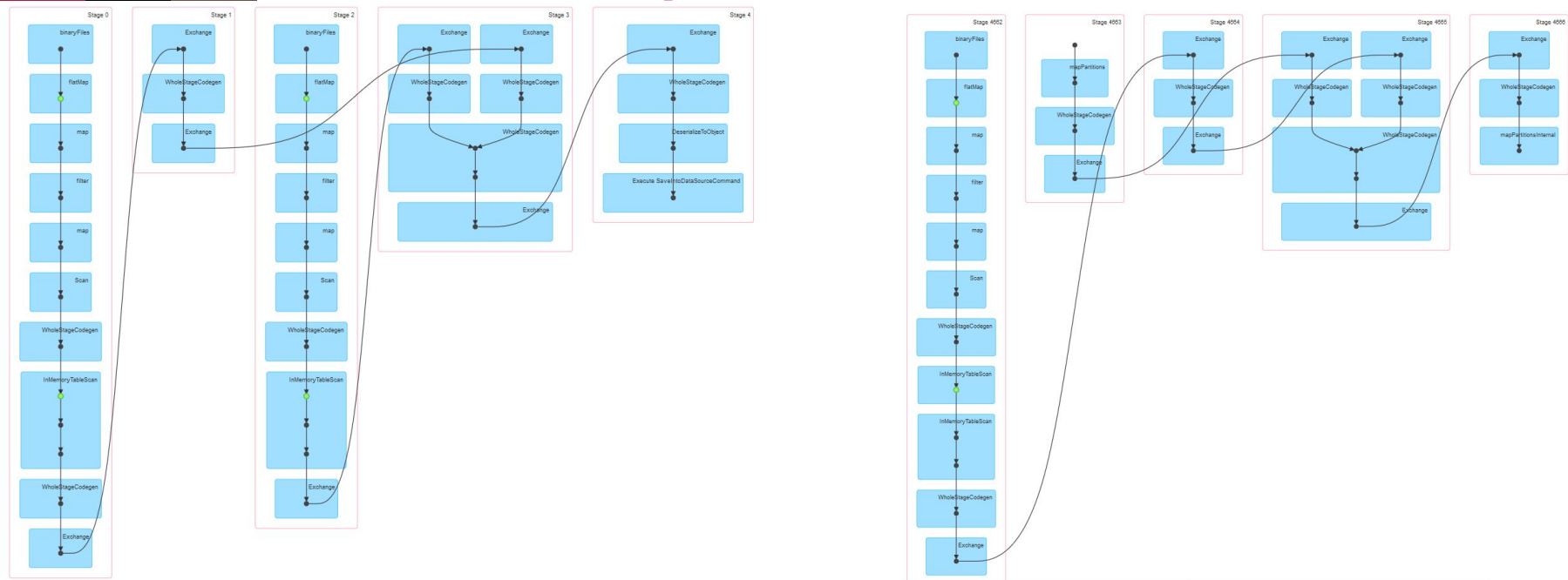
- 2 TABLES : 1 table des données et 1 table de mapping
- TABLE Requête 2 obtenue à partir d'un COUNT sur MENTIONS suivi d'un JOIN avec EVENT

TABLE Requête 2	
year INT monthyear INT day INT country TEXT count INT	eventid TEXT
PRIMARY KEY ((country), year, monthyear, day, eventid)	

TABLE Requête 2 mapping	
eventid TEXT day INT country TEXT count INT sumtone INT	actor1countrycode TEXT actor2countrycode TEXT actor1lat TEXT actor2lat TEXT actor1long TEXT actor2long TEXT
PRIMARY KEY (eventid)	

Ordonnée selon les années DESC, mois ASC, jour ASC et eventid DESC

# DAG de la requête 2



## Requête 3

- 3 TABLES : 1 pour les thèmes, 1 pour les personnes et 1 pour les lieux
- TABLE Requête 3 obtenue à partir de GKG
- Exemple TABLE 3 THEME :

TABLE 3 THEME	
year INT month INT day INT source TEXT count INT	theme TEXT tone DOUBLE
PRIMARY KEY ((source), year, month, day, count)	

Ordonnée selon les années DESC, mois ASC, jour ASC et count DESC


## Requête 4

- Pour cette table, les tables de la requête 2 sont réutilisées. Elles ont été optimisées en ce sens dès la requête 2.
- table de mapping permet d'incrémenter les champs nécessaires pour répondre spécifiquement à la requête 4.

TABLE 4	
year INT monthyear INT day INT pays1 TEXT pays2 TEXT	averagetone FLOAT numberofarticle INT
PRIMARY KEY ((pays1), year, monthyear, day, pays2)	

Ordonnée selon les années DESC, mois DESC et jour DESC

- **Remplissage de la base de données rapide :**
  - Tables fonctionnalité 1 + 2 (1 jour) : 3 min 15 sec
  - Table fonctionnalité 3 (1 jour) : 3 min
  - Table fonctionnalité 4 (1 jour) : 2,5 sec

- 
- Contexte
  - Architecture proposée
  - Modélisation et remplissage de la base de données
  - **Performances de la modélisation et budget**
  - Conclusion
  - Démo

# Performances

## ■ Volumétrie des tables Cassandra:

- Table 1 (1 AN): 5 Mo
- Table 2 (1 AN): 530 Mo + 3,2 Go
- Table 3 (1 AN): 1,34 Go
- Table 4 (1 AN): 9 Mo
- **TOTAL: ~5go**

## ■ Temps de réponse de requêtage:

- Table 1: ~ instantané
- Table 2: ~ entre 0'' et 6'' (si requête sur l'année)
- Table 3: ~ instantané
- Table 4: ~ instantané



# Problèmes rencontrés / solutions retenues

- **Déséquilibre entre les clusters Spark EMR et Cassandra:**
  - Problème : sous dimensionnement du cluster Cassandra / EMR = Cassandra saturé.
  - Solution : Augmentation du cluster Cassandra de 3 noeuds à 5 noeuds, sans perte des données déjà stockées.
- **Temps de chargement sur Cassandra trop long même après redimensionnement :**
  - Problème : hors délai pour la présentation
  - Solution : optimisation de la requête (facteur d'accélération x3)
- **Schéma des tables :**
  - Problème : multiplication des tables et temps de chargement trop important
  - Solution : post-processing

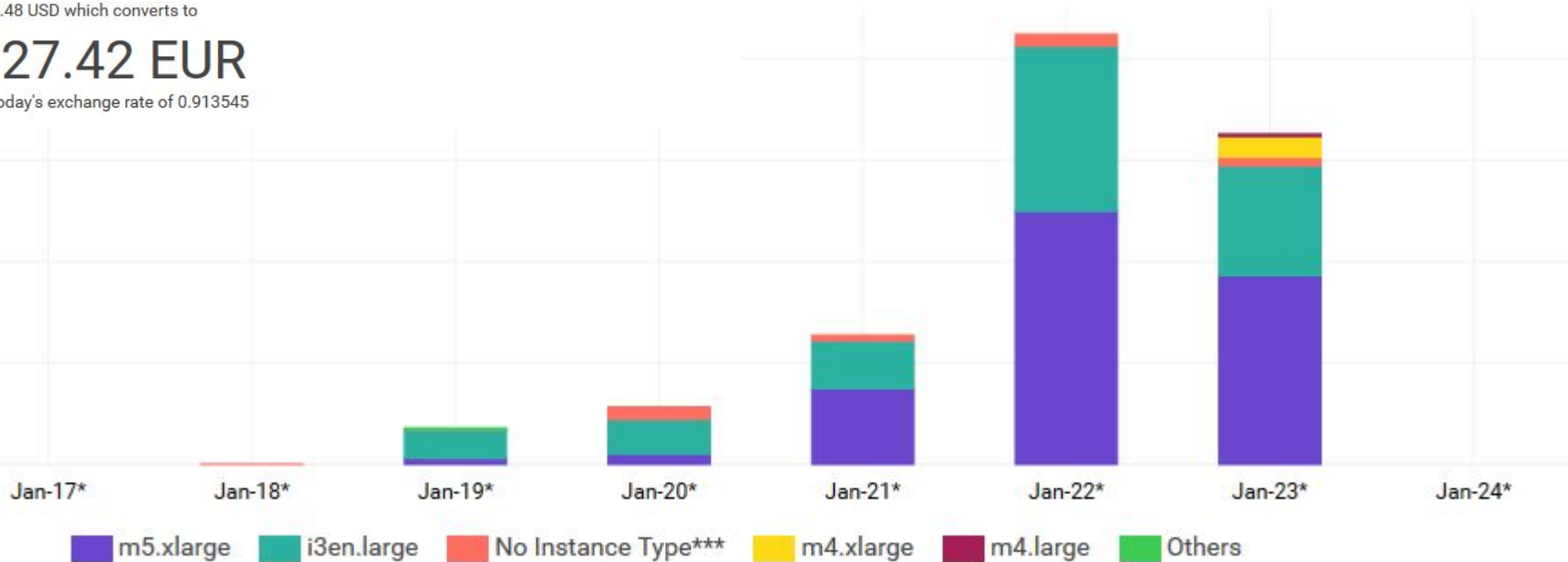
# Budget


Current month-to-date balance for January 2020, the exchange rate for the Payment Currency is estim:

139.48 USD which converts to

## 127.42 EUR

at today's exchange rate of 0.913545



- 
- Contexte
  - Architecture proposée
  - Remplissage de la base de données
  - Performances de la modélisation et budget
  - **Conclusion**
  - Démo

# Conclusion

- **Techno vue en cours utilisée:** Cassandra + Spark
- **Réponses aux requêtes :** quasi instantané
- **Chargement d'une année de données pour les 7 tables**
- **Résilience à la perte d'un noeud :** OK
- **Clusters (production & pré-production + backup) déployés sur AWS**

- 
- Contexte
  - Architecture proposée
  - Remplissage de la base de données
  - Performances de la modélisation et budget
  - Conclusion
  - **Démo**

