PROJET GDELT

Matyas Amrouche, Mathieu Bruniquel, Jacky Kaub,
Andre Macedo Farias, Benoit de Menthier, Theo Nazon

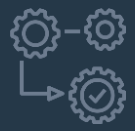
// BILAN

OBJECTIF

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

CAHIER DES CHARGES

- Système distribué et tolérant aux pannes
- 1 an de données GDELT
- Minimisation des couts
- 5 requêtes prédéfinies



RESULTAT

- ✓ SOLUTION => MONGODB (via EC2)
- ✓ 1 AN DE DONNÉES CHARGÉ SUR MONGO
- ✓ REQUÊTES 1/2/3/4/5 FONCTIONNELLES SUR 1 AN (+PARAMETRABLES POUR QUERY TIME <<<1min.)
- ✓ COLLECTIONS CHARGEES SUR MongoDB FLEXIBLES => pas de chargement de 1 table / 1 requête (plus lent en requêtage mais permettant de modifier 1 requête sur les mêmes champs)
- ✓ SYSTÈME RESILIENT (excl. Infra avec un arbitre qui devrait être ajouté)
- ✓ BUDGET FAIBLE -> $53 vs. $340 DE BUDGET (16% DU BUDGET ALLOUÉ UTILISÉ)

Sommaire

EXPLORATION DES DONNEES

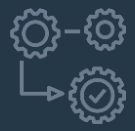
ARCHITECTURE AWS / MongoDB

PRE-PROCESSING AVANT IMPORT

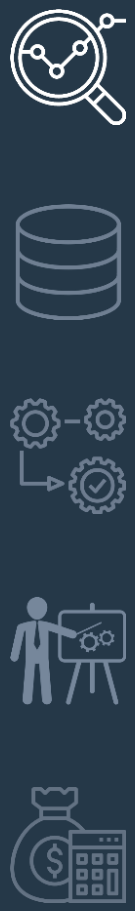
RESULTATS DES REQUETES

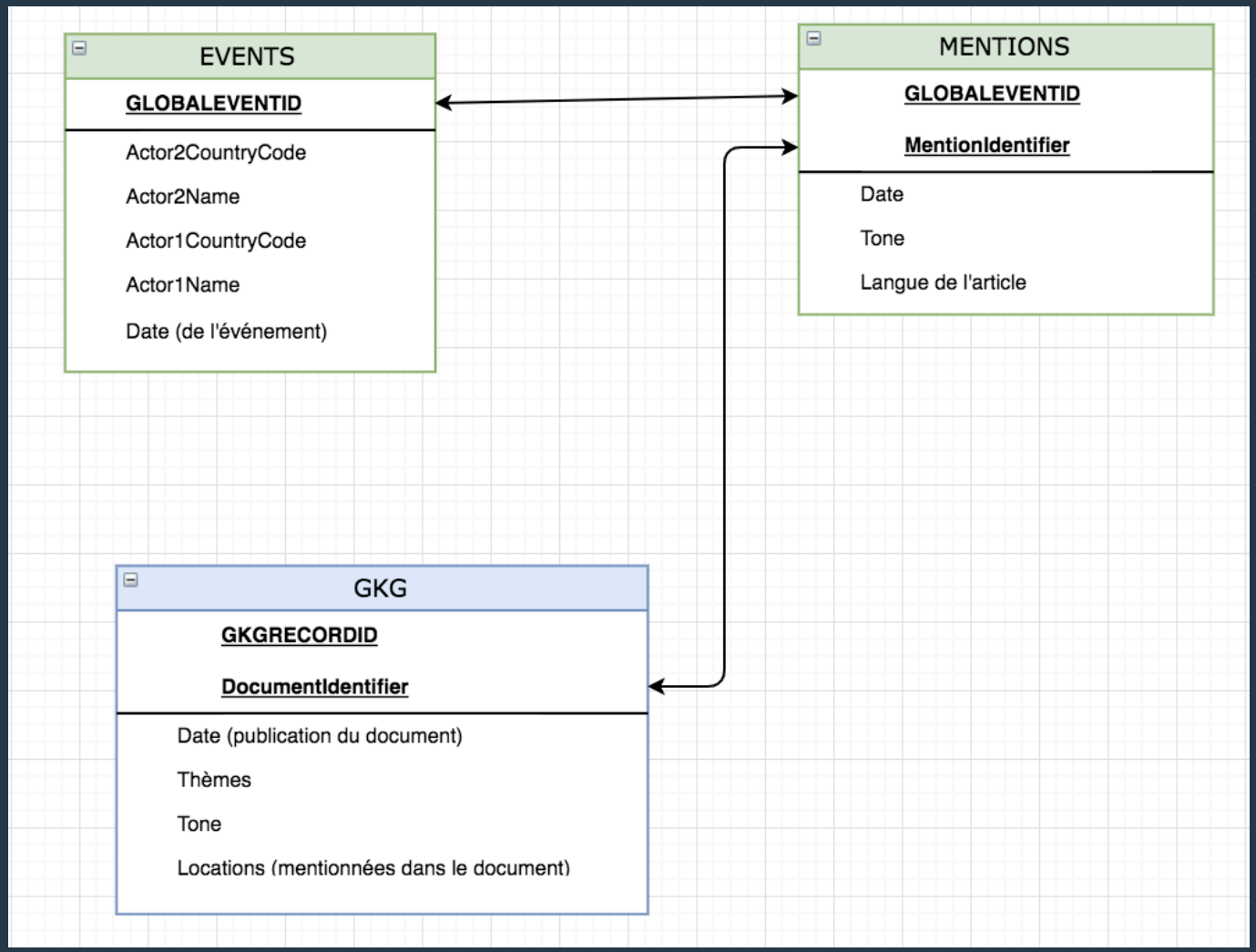
BUDGET FINAL

// Exploration des données | Général

	Description de la donnée	Requête(s)
 EVENTS	Table de tous les évènements et de leur description	[1, 2, 3, 4]
 MENTIONS	Table des mentions dans un article faisant référence à un événement (date publication, source, langue…)	[1, 2, 3, 4]
 GKG	Table des articles et de leur relation avec des Thèmes/Acteurs/Lieux… -> réseau de connaissance	[5]
		

// Exploration des données | Schéma DB





// Architecture | Solution ➔ MongoDB

Rationnel

- Structure de données stable
- Besoins clairs et définis
- Peu d`écritures & MAJ nécessaires

+

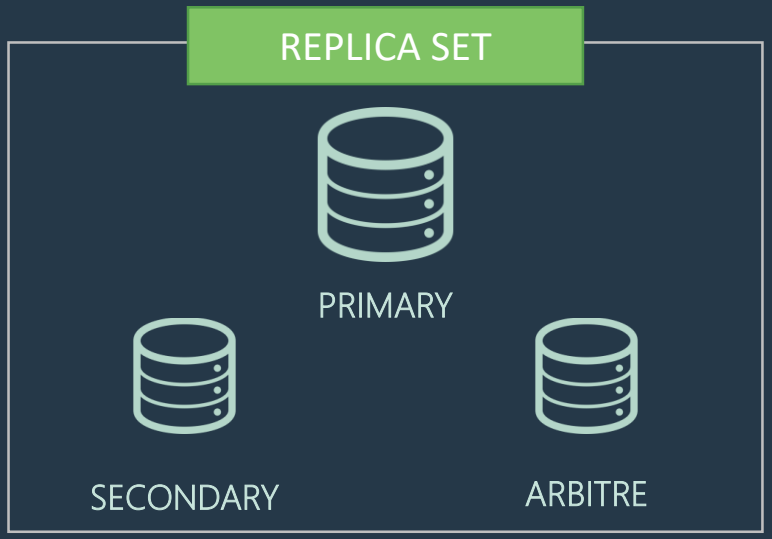
- Flexibilité (gestion des données sparses)
- Rapidité des requêtes [à confirmer]
- Résilience du système

-

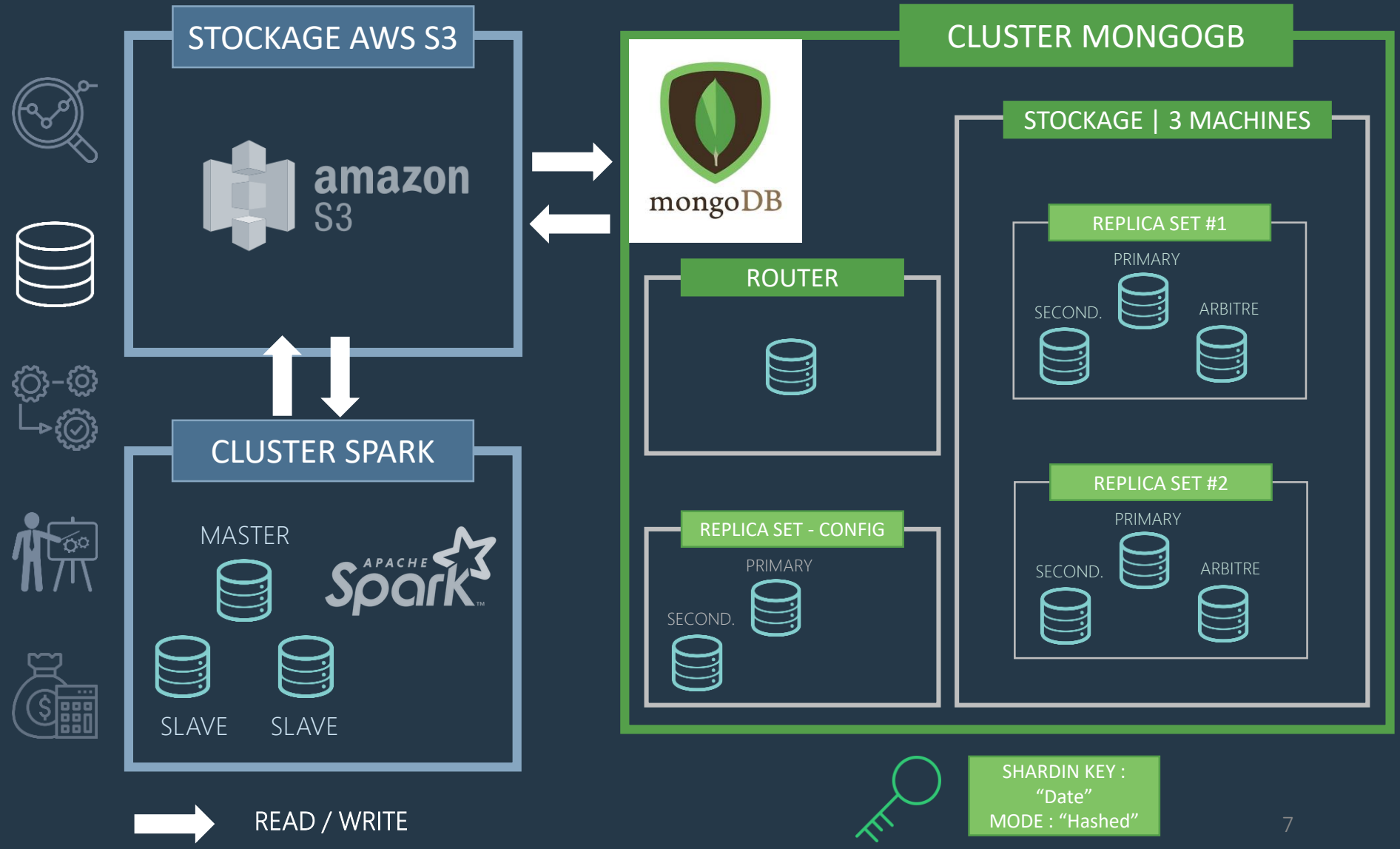
- Pas de jointure => nécessité de pre-procésser les données
- Utilisation de la mémoire (stockage des noms de clés de chaque document)

Approche / Paramètres

- Pre-processing des données pour minimiser la taille des fichiers
- Déploiement sur Amazon EC2 pour solution sur mesure
- Fault-tolérance : 1 (pour minimiser les coûts)
- Facteur de réplication : 1
- Nombre de machines : 3

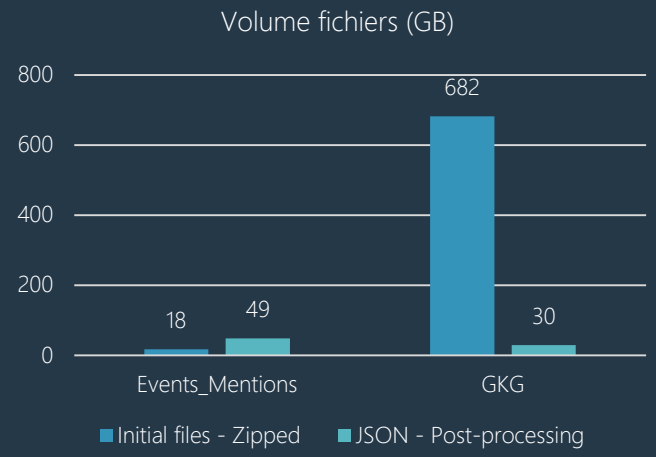


// Architecture | Schéma cible sur EC2



// Pre-Processing

- Concaténation des informations Events + Mentions dans un seul JSON / Event [Requête 1/2/3/4]
- Filtrage des informations GKG pour répondre à la requête additionnelle [Requête 5]




```
{ "EventID" : "719031141 ",
  "Date" : "20180101",
  "Year" : "2018",
  "Month" : "01",
  "Actor1Name" : "NIGERIA" ,
  "Actor1Country" : "NGA" ,
  "Actor2Name" : "" ,
  "Actor2Country" : "" ,
  "Contains_2actors" :0,
  "Mention" :[
        { "MentionDate" : "20180101003000",
          "MentionYear" : "2018",
          "MentionMonth" : "01 ",
          "Tone" : "4.8051948051948",
          "Language" : "eng" },
        { "MentionDate" : "20180101034500",
          "MentionYear" : "2018",
          "MentionMonth" : "01",
          "Tone" : "-4.8051948051948",
          "Language" : "eng" }]}
```

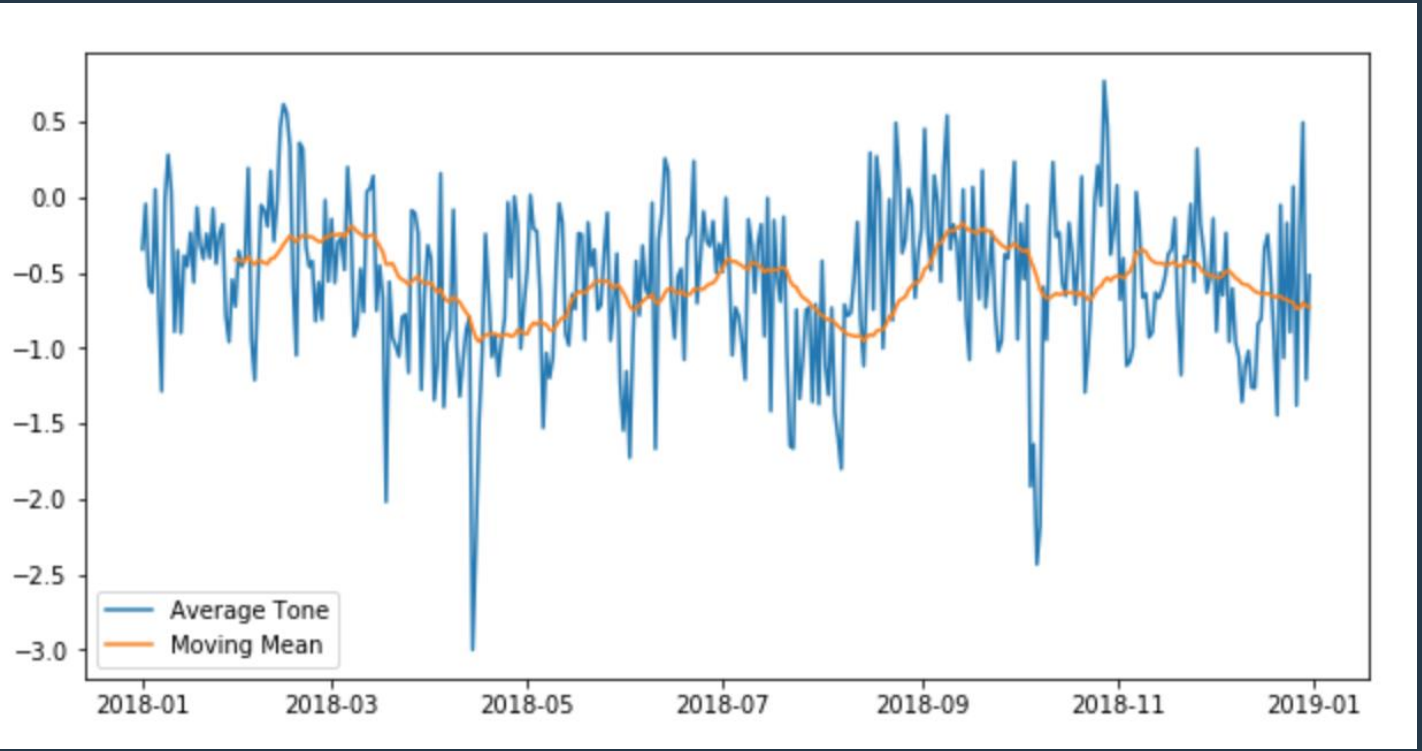
```
{ "ArticleID":"20180101000000-0",
  "Date":"20180101",
  "LocationsList":["Puerto Rico","Virgin Islands"],
  "Category":[ "security_violence", "society"],
  "GkgTone":"-3.64583333333333"}
```

// Résultats des requêtes | Requête Bonus (1/2)

EVOLUTION DU TON DES ARTICLES IMPLIQUANT 2 PAYS [Country_1, Country_2], AVEC POSSIBLITE DE FILTRER PAR 5 CATEGORIES (Diplomacy, Society, Security Violence, Economy, Other)

Evolution du ton des articles mentionnant [France, China]

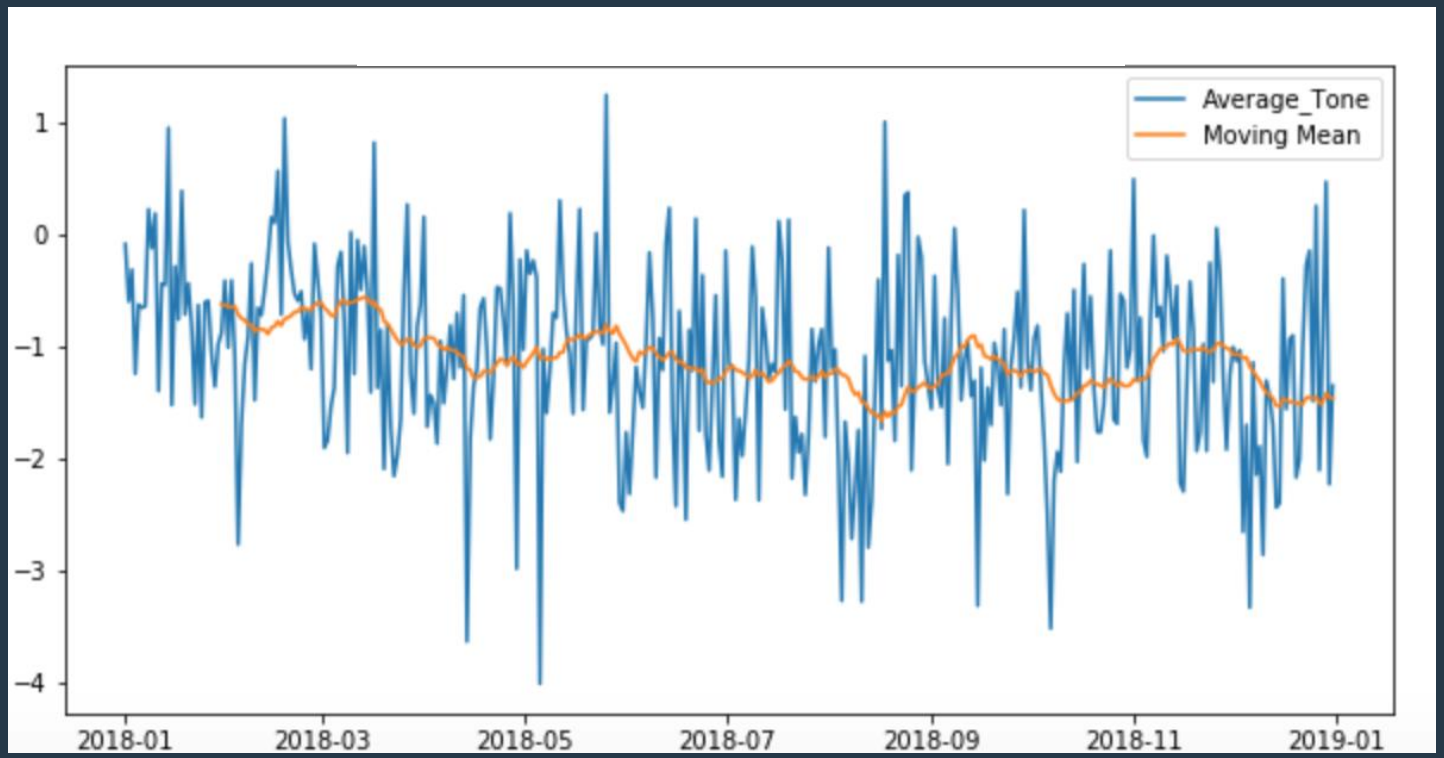
(1/50eme des données, période de 30 jours)



// Résultats des requêtes | Requête Bonus (2/2)

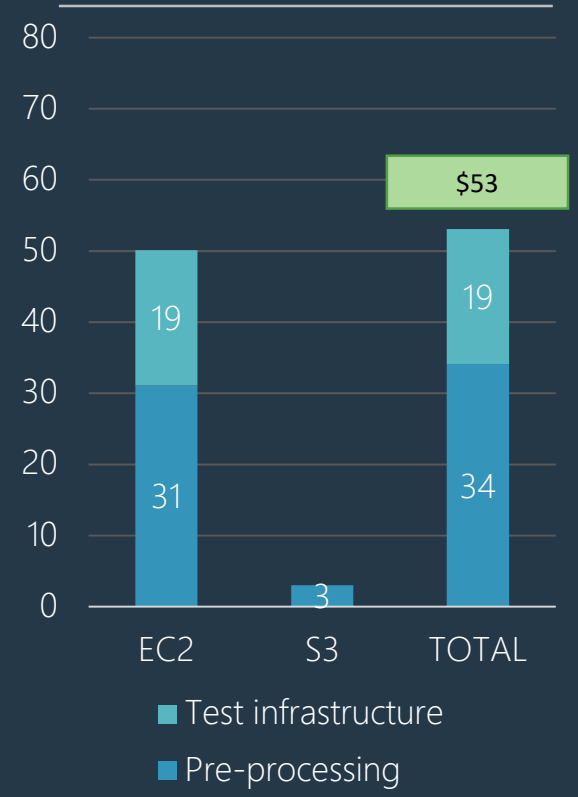
Evolution du ton des articles mentionnant [France, China]

(1/50eme des données, période de 30 jours) - THEME => DIPLOMACY

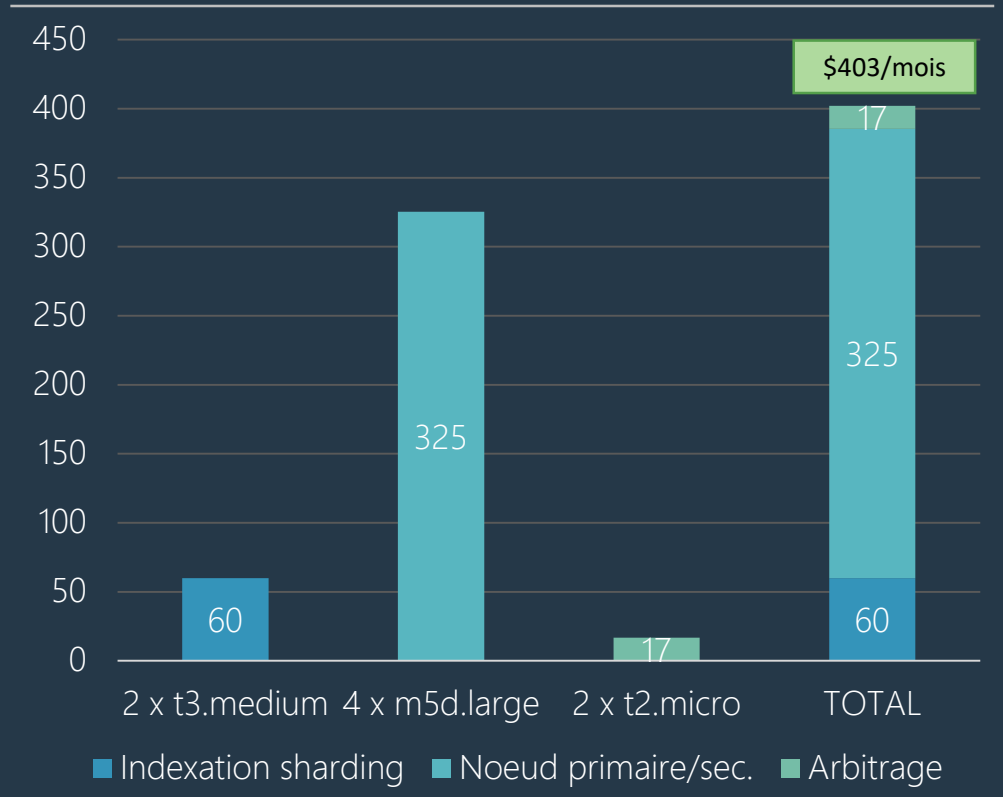


// Budget Final | $53 vs. $340 => 16% du total

Budget final de la partie pre-processing



Budget / mois Infrastructure
($/mois extrapolé à partir du cout par heure des machines)




DEMONSTRATION