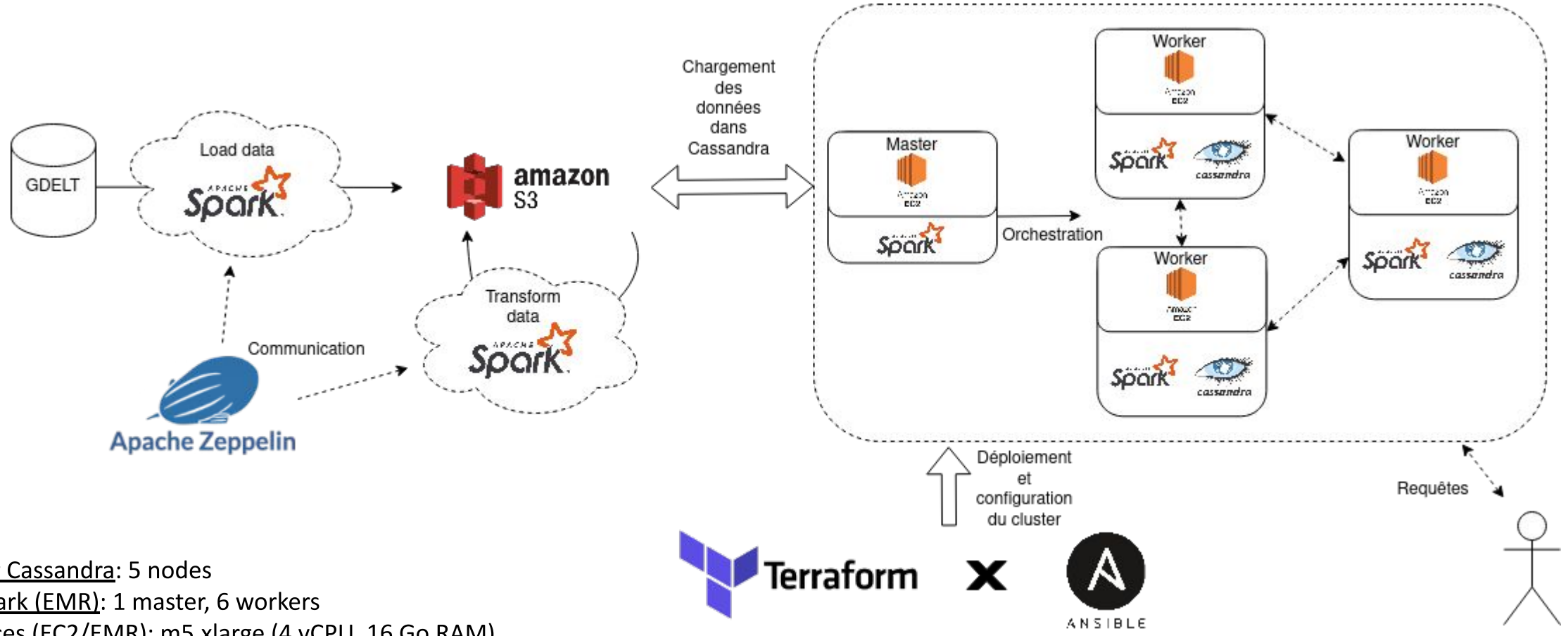
Projet GDELT

Morgan FASSIER, Theo LEFIEVRE, Ernest MAJDALANI, Fabrice PONT,
Adrien ROBILIARD

Plan

I. Description de l’architecture
II. Choix : BD, facteur de réplication
III. ETL & Requêtes
IV. Temps, stockage et coût
V. Question ouverte
VI. Problèmes rencontrés
VII. Démonstration

I. Description de l'architecture



Cluster Cassandra: 5 nodes
ETL Spark (EMR): 1 master, 6 workers
Instances (EC2/EMR): m5.xlarge (4 vCPU, 16 Go RAM)

II. Choix : BD, facteur de réplication

BD NoSQL : Cassandra (5 noeuds)

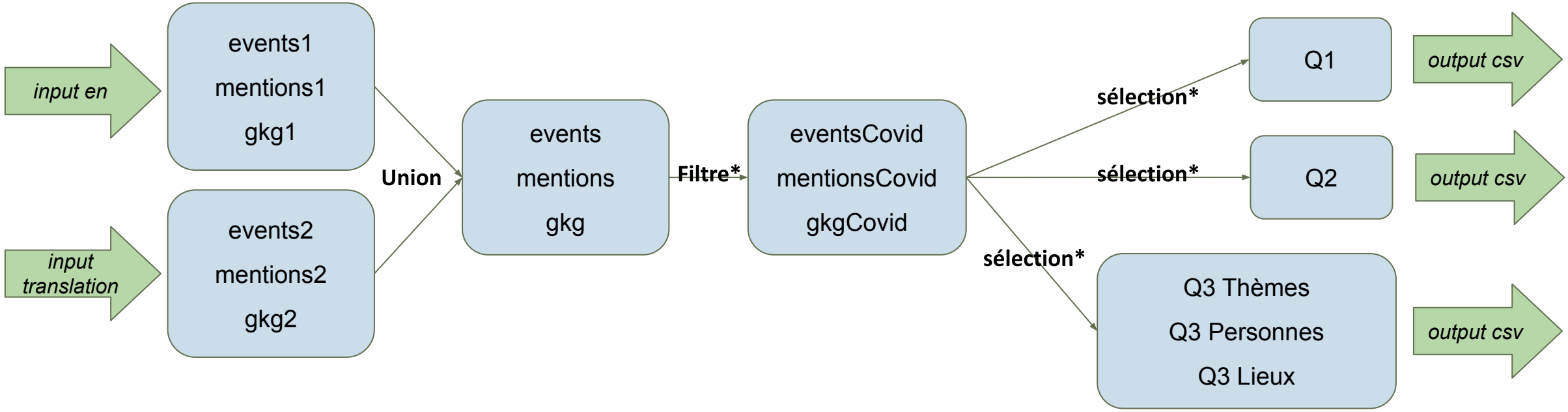
- Adaptée au volume de données
- Facilement requêtable : CQL
- Basée sur la JVM -> intégration facile à d'autres applications (ex : Spark)
- Ecriture rapide

Facteur de réplication : 3, Quorum en lecture+écriture

→ on peut perdre un nœud sans impacter l'application.

III. ETL & Requêtes

Preprocessing Spark : création des 5 dataframes pour les requêtes



** voir code spark pour plus de détails*

III. ETL & Requêtes

Requêtes CQL

Question 1 : *afficher le nombre d'articles/évènements qui parlent de COVID qu'il y a eu pour chaque triplet (jour, pays de l'évènement, langue de l'article)*

Table T1

```
CREATE TABLE T1 (
        date date,
        pays text,
        langue text,
        id_event text,
        id_mention text
        PRIMARY KEY ((date, pays, langue), id_event,
                        id_mention)
)
```

Requête 1

```
SELECT date, pays, langue, COUNT(DISTINCT id_event),
            COUNT(id_mention)
FROM T1
GROUP BY date, pays, langue
```

Champs utilisés

MentionTimeDate, ActionGeo_CountryCode, MentionDocTranslationInfo, GlobalEventID, MentionIdentifier

III. ETL & Requêtes

Requêtes CQL

Question 2 : *pour un pays donné en paramètre, affichez les évènements qui y ont eu place triées par le nombre de mentions (tri décroissant); permettez une agrégation par jour/mois/année*

Table T2

```
CREATE TABLE T2 (
        pays text,
        id_event text,
        date date,
        id_mention text,
        PRIMARY KEY((pays), id_event, date, id_mention)
)
```

Requête 2

x = ''FR''
y = ''%Y-%M-%d'' OU ''%Y-%M'' OU ''%Y''

```
SELECT id_event, DATE_FORMAT(date, y) AS d,
            COUNT(id_mention)
FROM T2
WHERE pays = x
GROUP BY id_event, d
ORDER BY COUNT(id_mention) DESC
```

Champs utilisés: ActionGeo_CountryCode, GlobalEventID, EventDate, MentionIdentifier

III. ETL & Requêtes

Requêtes CQL

Question 3 : *pour une source de donnés passée en paramètre (gkg.SourceCommonName) affichez les thèmes, personnes, lieux dont les articles de cette source parlent ainsi que le nombre d'articles et le ton moyen des articles (pour chaque thème/personne/lieu); permettez une agrégation par jour/mois/année.*

Table T3_p

p ∈ {thème, pays, personne}

```
CREATE T3_p (
        source text,
        p text,
        id_mention text,
        tone float,
        date date,
        PRIMARY KEY(source, date, p, id_mention)
)
```

Requête 3_p

s = ''lemonde.fr''

y = ''%Y-%M-%d'' OU ''%Y-%M'' OU ''%Y''

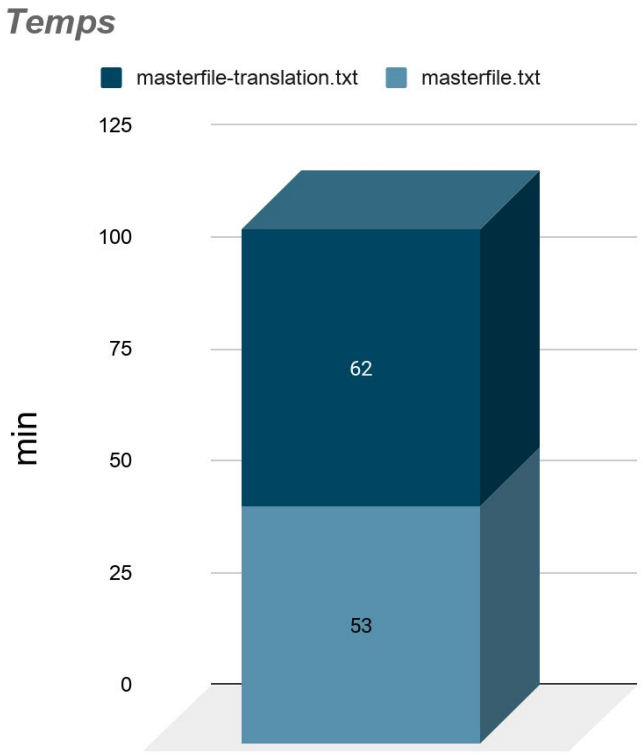
```
SELECT p, AVG(tone), COUNT(id_mention),
            DATE_FORMAT(date, y) AS d
FROM T3_p
WHERE source = s
GROUP BY p, d
```

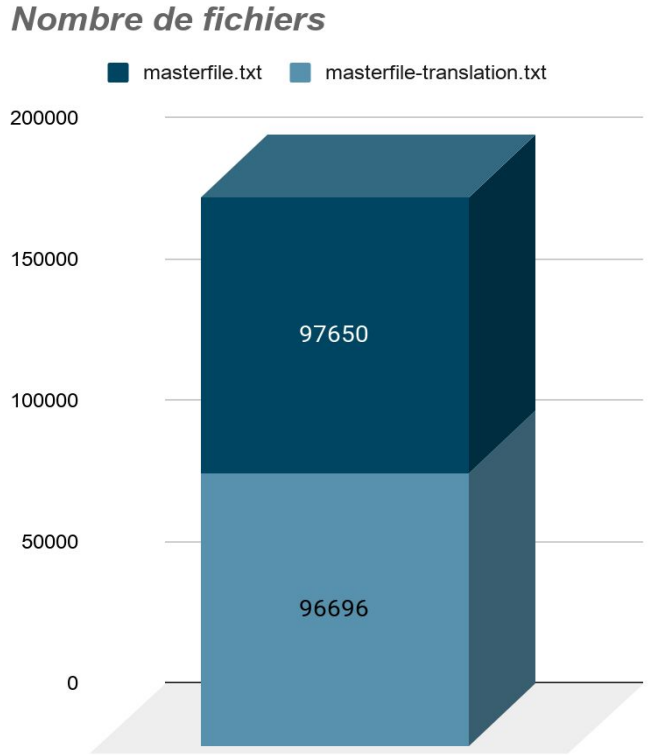
Champs utilisés: SourceCommonName, GkgRecordID, V1.5Tone, V1Themes/V1Persons/V1Locations, V2.1Date

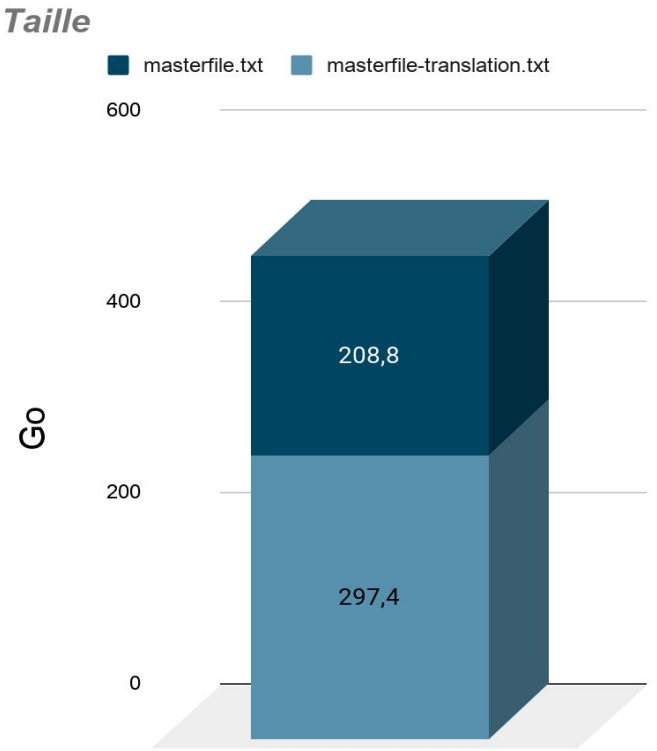
IV. Temps & prix

- Prix de stockage négligeable pour le peu de temps qu'on l'utilise (calculer / mois de stockage)
- Prix pour mettre les fichiers :
  ~ 1$ pour les 180 000 fichiers (0.005$/1000)

Téléchargement des données



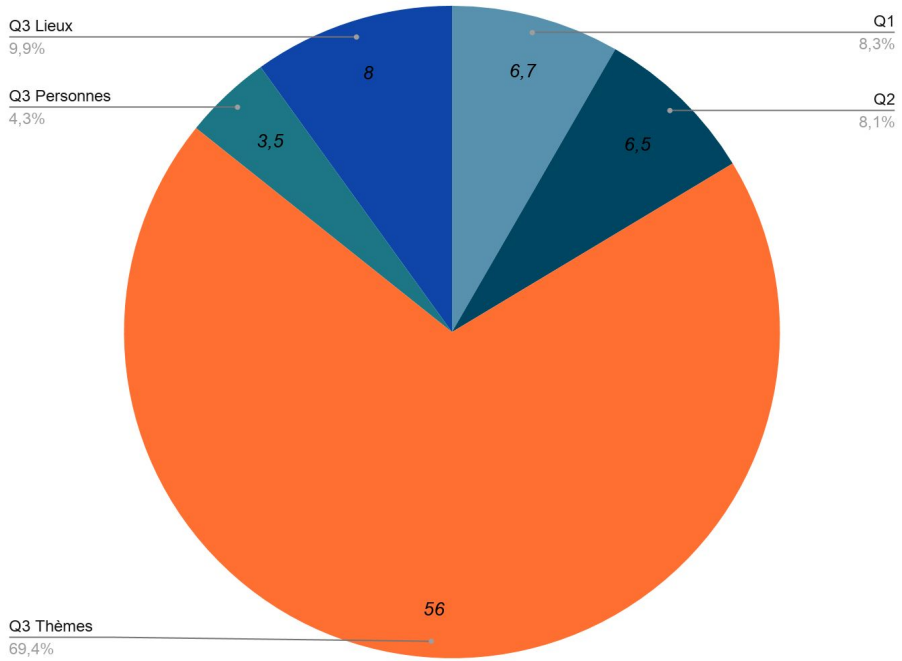




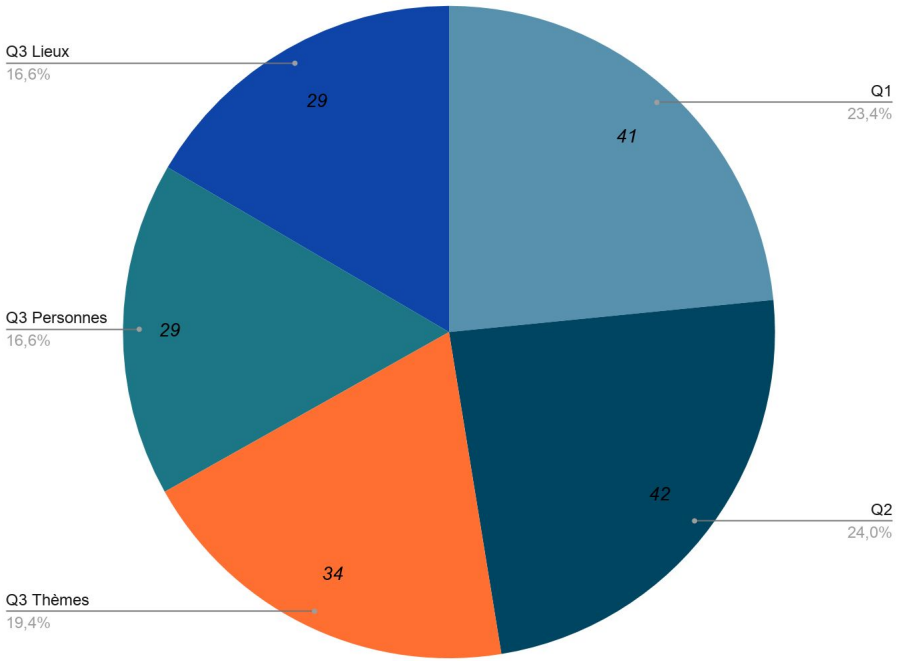
IV. Temps & prix

ETL (zip bruts S3 -> transformation -> csv voulus S3)

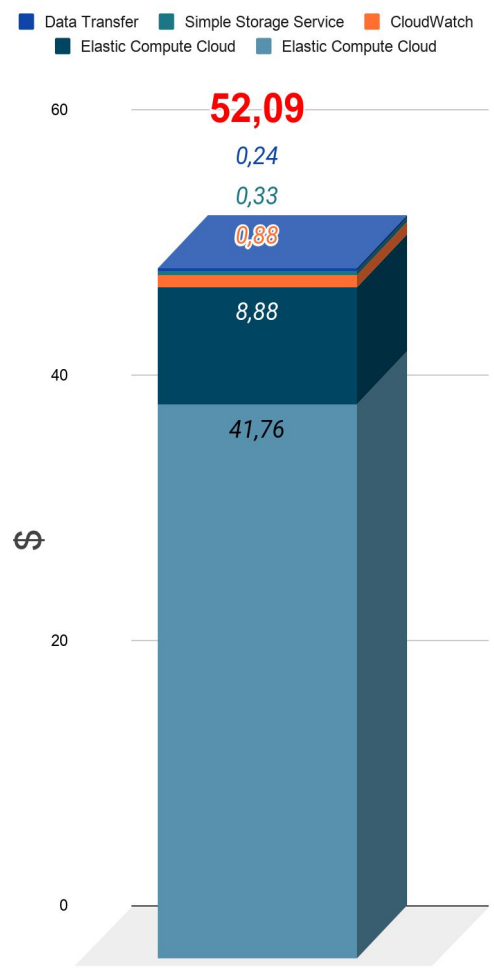
Taille (Go)



Temps (minutes)

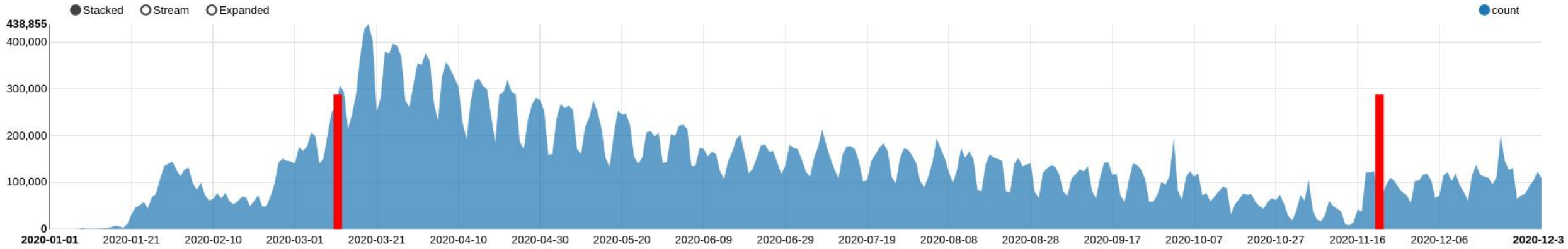


Prix

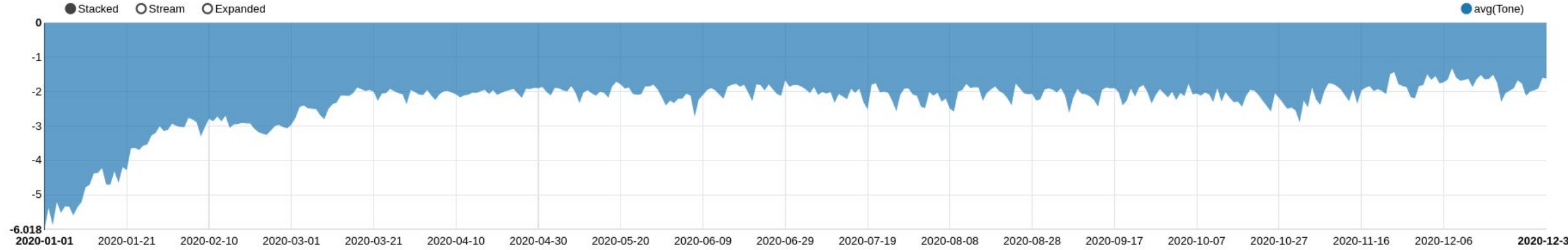


V. Question ouverte

Évolution temporelle du nombre total d'articles parlant de Coronavirus dans le monde



Évolution temporelle du ton moyen des articles parlant de Coronavirus dans le monde



VI. Problèmes rencontrés

- Impossibilité de créer des IAMs => impossibilité d'utiliser AWS Keyspace
- Configuration des instances EC2 complexe
- Erreur dans la lecture de deux fichiers zip : `java.io.EOFException: Unexpected end of ZLIB input stream`
- Limitations liées aux comptes educates (nombre de noeuds, mémoire des noeuds, etc...)

VII. Démonstration

- Requêtes
- Tolérance aux pannes

Merci !

Vous avez des questions ?