Projet No-SQL

lesfurets

X

TELECOM Paris

IP PARIS

Étudier l'évolution de la pandémie COVID-19 via son impact média

BARDONNET – BREBOIN – DELARUE – NOURRY – PANNIER

Sommaire

- Équipe
- Contexte
- Architecture
- Volumétrie
- Modélisation
- Avantages et Inconvénients
- Limites du modèle et contraintes
- Annexes

Équipe

Vincent BARDONNET

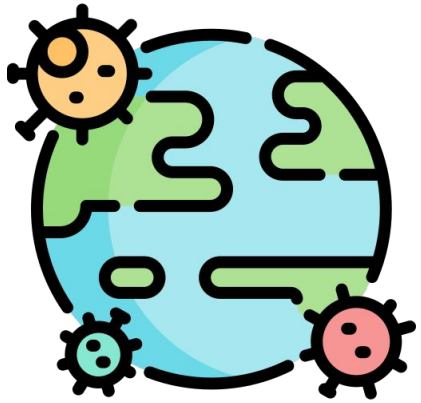
Alexandre BREBOIN

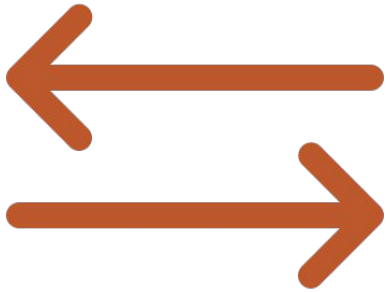
Simon DELARUE

Mathias NOURRY

Valentin PANNIER

Contexte

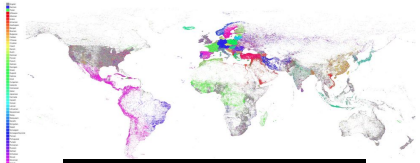




Pandémie COVID-19

Couverture médiatique au sujet la pandémie

Architecture



GDELT 2.0

EMR

APACHE Spark

Scala

ETL 1
Importing
Raw data

ETL 2
Raw data =>
preprocessed
data

Amazon S3

Amazon S3

Amazon S3

EC2

cassandra

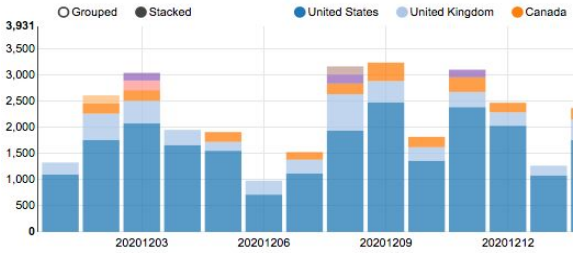
cassandra

cassandra

cassandra

cassandra

Apache Zeppelin



Modélisation

Type des machines AWS utilisées :

- EMR (ETL1 + ETL2) :
  - m4.xlarge x 7 = 28 CPUs
- EC2 (Cassandra) :
  - m4.large x 5 = 10 CPUs
  - Configuration :
    - Replication factor = 3
    - Read consistency = One | Write consistency = Quorum

Méthode de distribution des coûts AWS :

Répartition de la partie EMR (ETL1 + ETL2) en 3, afin de limiter les coûts AWS :

- Chacunes des 3 personnes à chargé, via son propre EMR, 4 mois de données Gdelt (~~ 130Go/personne) sur son propre bucket S3
- Une 4eme personne a récupéré les données nettoyées sur les 3 buckets S3 et les a concaténé sur Cassandra

Volumétrie

- Volume total cumulé des .zip sur AWS S3 : 410 Go
  - 3 Go pour les fichiers .zip **export**
  - 5 Go pour les fichiers .zip **mentions**
  - 400 Go pour les fichiers .zip **gkg**

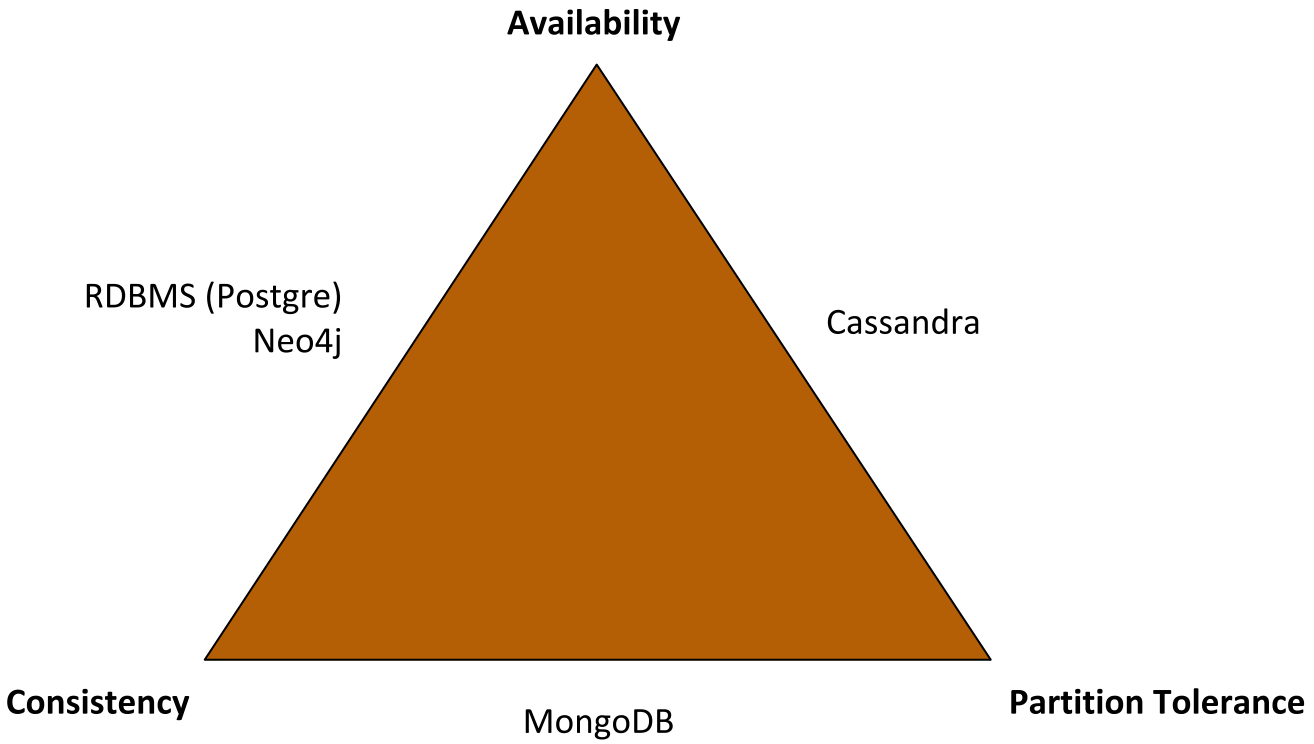
- Volume total cumulé des .parquet sur AWS S3 : 21 Go
  - 160 Mo pour les fichiers .parquet **export**
  - 350 Mo pour les fichiers .parquet **mentions**
  - 20.5 Go pour les fichiers .parquet **gkg** (1.7 Go par mois)

=> Impossibilité de charger le gkg sur le ring Cassandra par manque d'espace disque (50Go)

Environ 500 Mo de données chargées sur le ring cassandra (0,16+0,35)

(avec un facteur de réplication de 3 : 3 x 0.5 = 1.5 Go)

Avantages et inconvénients



Limites du modèle et contraintes

Certains articles sont référencés
avec un numéro d'article
=> Potentiellement un article sur le Covid non détecté

https://thejewishnews.com/2020/09/16/hope-emerges-ethiopias-hidden-jews-begin-to-receive-aid-from-israel-and-abroad/
https://www.leaderlive.co.uk/news/18725656.hundreds-young-adults-flintshire-wrexham-swell-ranks-universal-credit-claimants/
https://www.hellomagazine.com/royalty/2020091697355/the-queen-prince-philip-leave-balmoral-for-sandringham/
https://www.hindustantimes.com/world-news/indian-who-killed-estranged-partner-in-uk-jailed-for-life/story-SIkmrTD1GlqNQdLuhQ5ppJ.html
https://www.premierleague.com/news/1832276

Certains articles n'ont pas de Pays associés
=> Lors d'une requête par Pays, Sous-évaluation du nombre d'article

957212052	20201124	United State	https://www.pjstar.com/story/news/2020/11/30/illinois-lawmakers-demand-hearing-into-lasalle-veterans-hom
957212053	20201130	Sri Lanka	http://www.lankaweb.com/news/items/2020/11/30/phis-warn-covid-19-spread-in-colombo-beyond-control/
957212079	20201201		https://www.cbc.ca/news/canada/hamilton/krown-kafe-business-covid-19-violations-1.5822639
957212099	20201201	United State	https://www.5newsonline.com/article/news/health/dr-jos-romero-and-state-leaders-address-covid-19-vaccine-
957212113	20201201	United State	https://www.siasat.com/moderna-asking-us-european-regulators-to-greenlight-covid-19-vax-2035361/

Cassandra ne supporte ni les jointures ni les groupby car les jointures sur des machines différentes dans un système distribué ne sont pas performantes. Pour résoudre cela, nous avons dû passer par Spark SQL ce qui rajoute des étapes supplémentaires.

Merci

DES QUESTIONS?

Annexes

Avantages et inconvénients

Critères / Technologie	Type de Stockage	Language	Fault Tolerance	Complexité
Neo4J	Orienté graphe −	Cypher graph query language −	−	Déconseillé −
Cassandra	Large +	CQL (≈ SQL) +	Multiple Master Nodes +	Configuration difficulty −
PostgreSQL	-Need structure -Small −	SQL +	Manual failover −	−
MongoDB	Large +	Queries over JSON structure +	-Single Master -Slaves read-only −	Mostly easy language +