

INSTITUT
POLYTECHNIQUE
DE PARIS

INF728 - Bases de données non relationnelles

GDELT - Les événements marquants de l'année 2021

GALLAY Nicolas, MAXWELL Olivier, De MIRIBEL Paul, MECCHIA Pierre, FORTORA Toni

11/02/2022

Introduction

Objectif

Concevoir un système de stockage distribué résilient et performant pour réaliser 4 requêtes sur les données « The Global Database of Events, Language and Tone (GDELT) ».

Ressources

- 5 machines openstack
- 5 cerveaux
- Support du cours INF728

Contraintes

- Utiliser une des technologies suivantes : MongoDB, Spark, Cassandra, SQL ou Neo4j
- Concevoir un système distribué tolérant à une panne
- Charger une année de donnée GDELT
- Utiliser les ressources openstack

Sommaire

Présentation

- Choisir une technologie
- Implémenter la solution
- Intégrer la solution dans une architecture en streaming

Démonstration

Questions

Présentation

Choisir une technologie

Choisir une technologie

Etudier les données et spécifier le besoin

Données

- Types variés
- Volumes importants ~ 34k .csv
- Structurées en table
- Nombre important de « null »

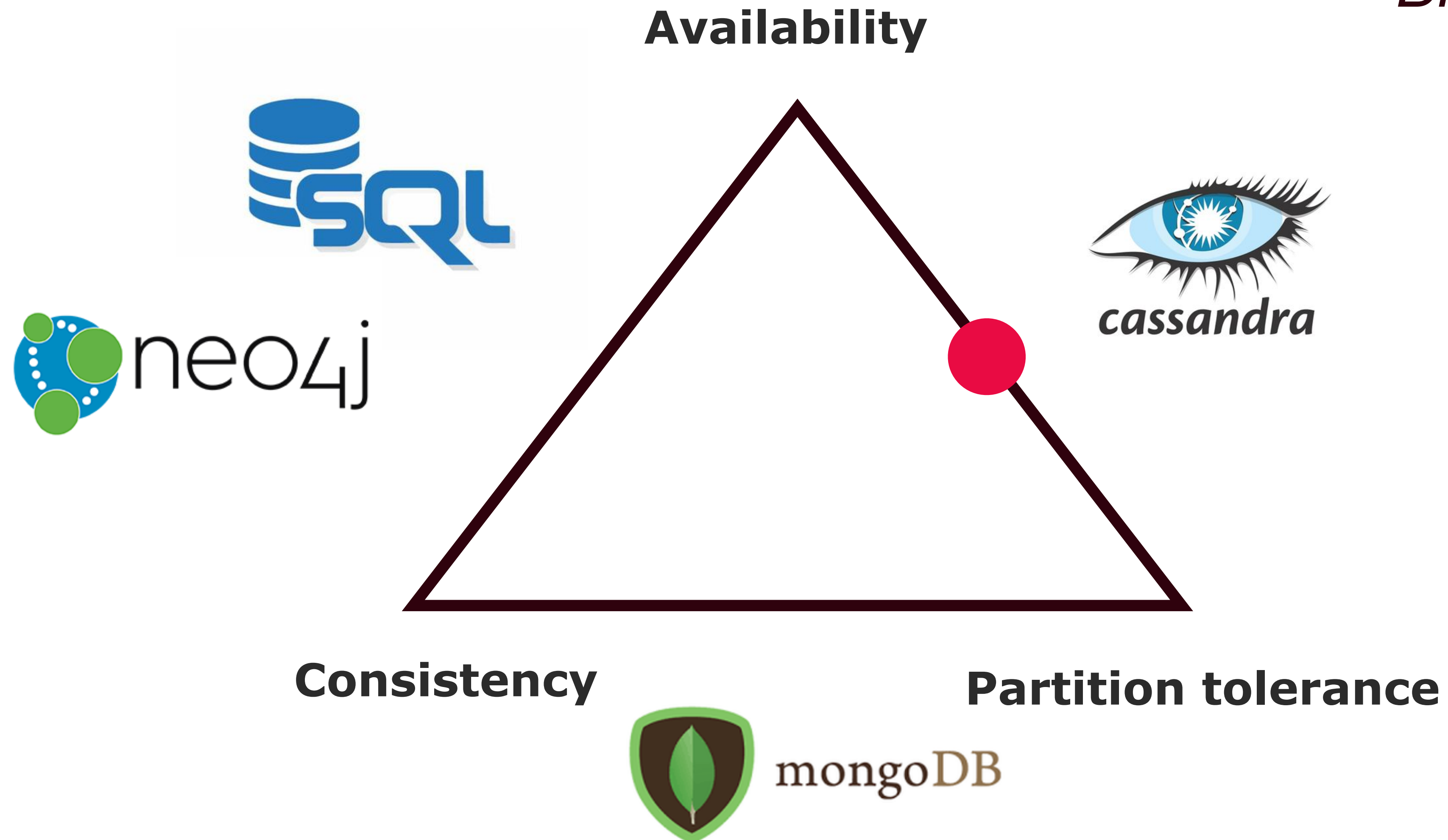
Spécifications

- Faible latence (quelques dizaines de secondes par requêtes)
- Disponibilité malgré une panne (no SPOF*)
- Pas d'écriture (hors insertion des données dans la base)

Choisir une technologie

Solution retenue

Spark
Big Data Tools

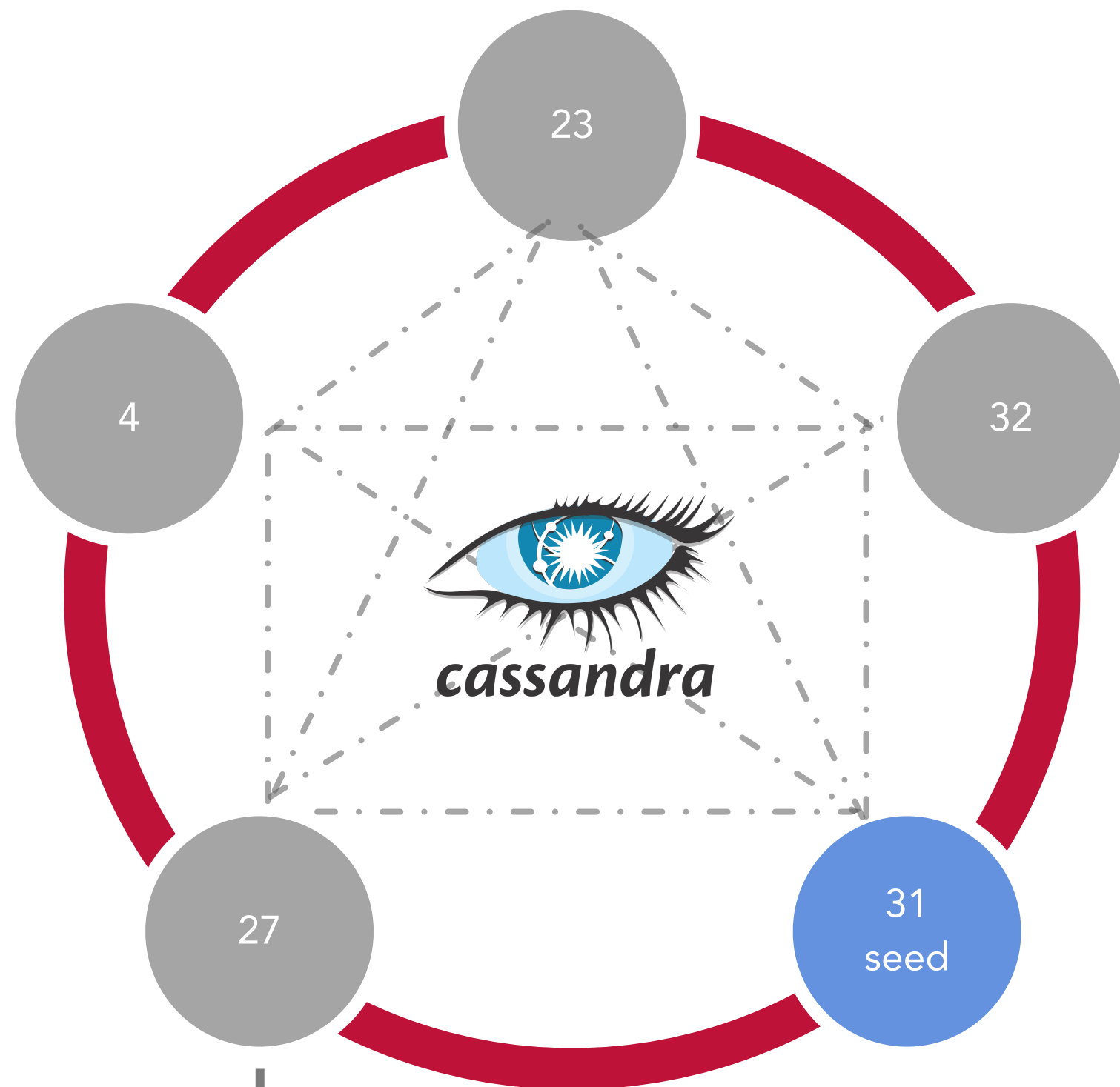


Présentation

Implémenter la solution

Implémenter la solution

Créer une architecture distribuée



- ✓ Intel Core Procesor
- ✓ 40 GB hard drive
- ✓ 6 GB RAM

```
1 cluster_name: 'MyCluster'
2 allocate_tokens_for_local_replication_factor: 3
3 partitioner: org.apache.cassandra.dht.Murmur3Partitioner
4 native_transport_port: 9042                2
5 seeds: "192.168.3.207"
6 listen_address: 192.168.3.139
7 rpc_address: 192.168.3.139
```

Extrait du fichier de configuration cassandra.yaml

```
ubuntu@tp-hadoop-32:~/NoSQLProject/apache-cassandra-4.0.1/bin$ ./nodetool status
Datacenter: datacenter1
=====
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address            Load        Tokens     Owns    Host ID                               Rack
UN  192.168.3.235       14.15 GiB   256        ?       1ed4a723-736f-4508-82cf-001a89690eee rack1
UN  192.168.3.74        13.1 GiB    256        ?       4bfa2de5-99dc-42bd-8cec-df76165bcbf6 rack1
UN  192.168.3.108       11.79 GiB   256        ?       125b60bf-eea5-4503-b55d-ce1019dda178 rack1
UN  192.168.3.207       12.55 GiB   256        ?       b3347412-b9ce-45bb-af40-4b09e2c6cccd rack1
UN  192.168.3.139       13.02 GiB   256        ?       c7b0cd78-cd0e-4d59-a72e-4445c260f0bd rack1
```

Statut du cluster

Implémenter la solution

Modélisation des données

« *Request first* » = *Dénormalisation*

- Définir des requêtes  **1 requête = 1 table**
- Définir une modélisation des données

Sélectionner les attributs pour chaque requête

Requête 1 : afficher le **nombre d'articles/événements** qu'il y a eu pour chaque triplet (**jour**, **pays de l'évènement**, **langue de l'article**).

events.csv

- GlobalEventID
- Day
- NumArticles
- ActionGeoCountryCode

mentions.csv

- GlobalEventID
- MentionDocTranslational-Info

Implémenter la solution

Modélisation des données

« *Request first* » = *Dénormalisation*

- Définir des requêtes  ***1 requête = 1 table***
- Définir une modélisation des données

Définir les primary keys, partition keys, clustering keys et types

nb_articles_events

globaleventid

day

mentionid

action_geocountrycode

mentiondoctranslationalinfo

Type

→ Minimiser le volume (arrondi)

Partition

→ Minimiser le nombre de partition

→ Lignes / partition < 100k

→ Volumes de données / partition < 100 Mo

Implémenter la solution

Modélisation des données

Définir une modélisation physique

Gdelt

nb_articles_events		
globaleventid	int	C
day	int	P
mentionid	text	C
action_geocountrycode	text	P
mentiondoctranslationalinfo	text	P

countries_events		
globaleventid	int	C
day	int	C
month	int	C
year	int	
nummentions	int	
action_geocountrycode	text	P

data_source		
day	int	C
month	int	C
sourcecommonname	text	P
documentidentifier	text	C
themes	set	
persons	set	
locations	set	
tone	float	

relationship		
sourceurl	text	C
day	int	C
month	int	C
averagetone	float	
actor1_geocountrycode	text	P
actor2_geocountrycode	text	P
themes	set	

Implémenter la solution

Evaluer et affiner le modèle

Evaluation	nb_articles_events	countries_events	data_source	relationship
75 %	0,15	6	0,8	0,4
95 %	0,5	31	7	1,7
98 %	4	475	132	26
99 %	10	1183	475	111
Nb partitions	47361	174	14268	12126

Taille des partitions par table en méga-octets







Implémenter la solution

Evaluer et affiner le modèle

Affinage		Requête 1 nb_articles_events	Requête 2 countries_events	Requête 3 data_source	Requête 4 relationship
Implémentation 1	Paramètre	NA	FR et US	lefigaro.com, bbc.com	US + CN
	Temps (s)	90	15 et 127	0.7, X	X

Améliorations

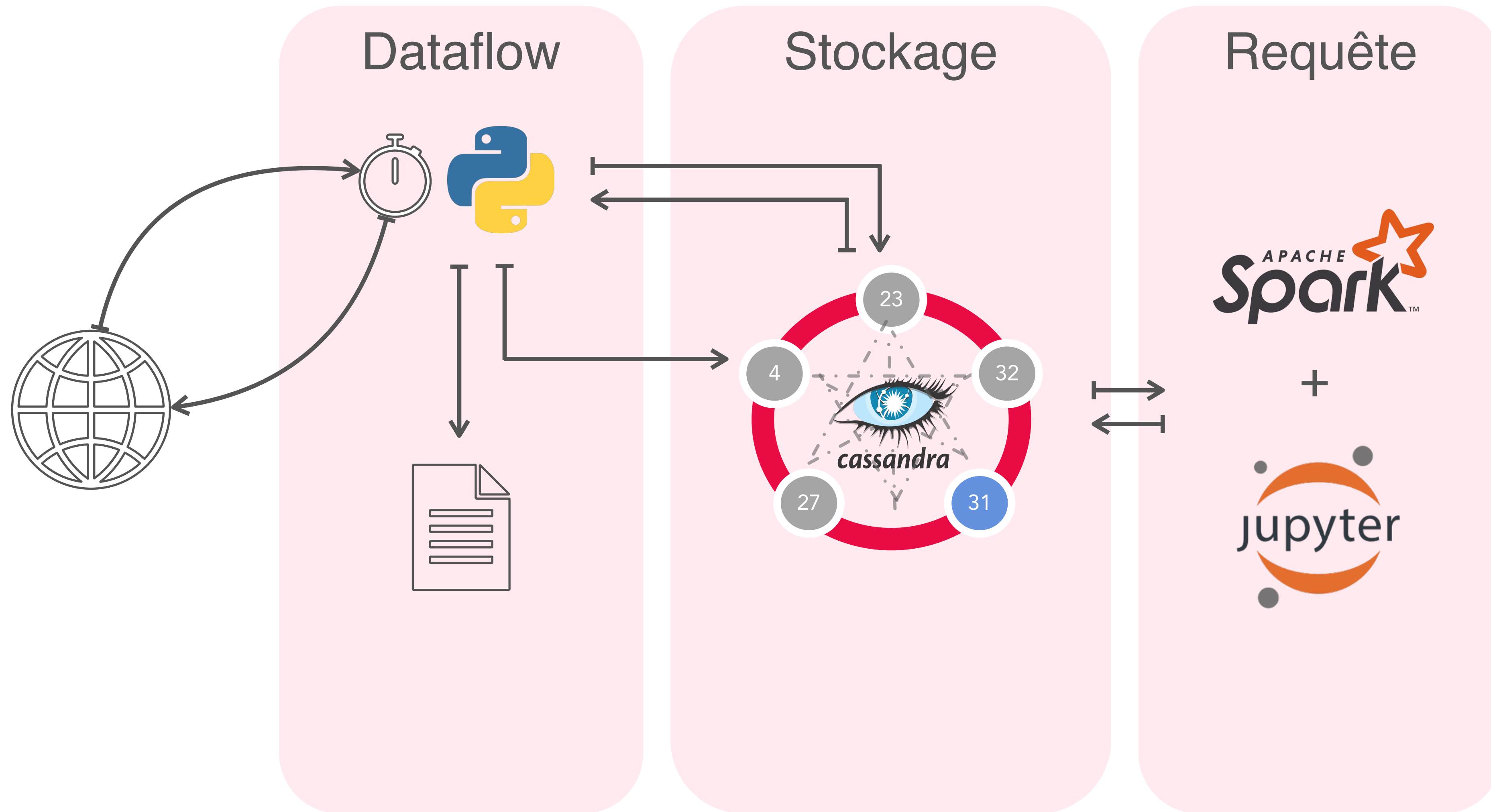
- Spark
 - Bucketing
 - Bloom filter
 - Read-repair

Présentation

Intégrer la solution dans une architecture streaming

Intégrer la solution dans une architecture en streaming



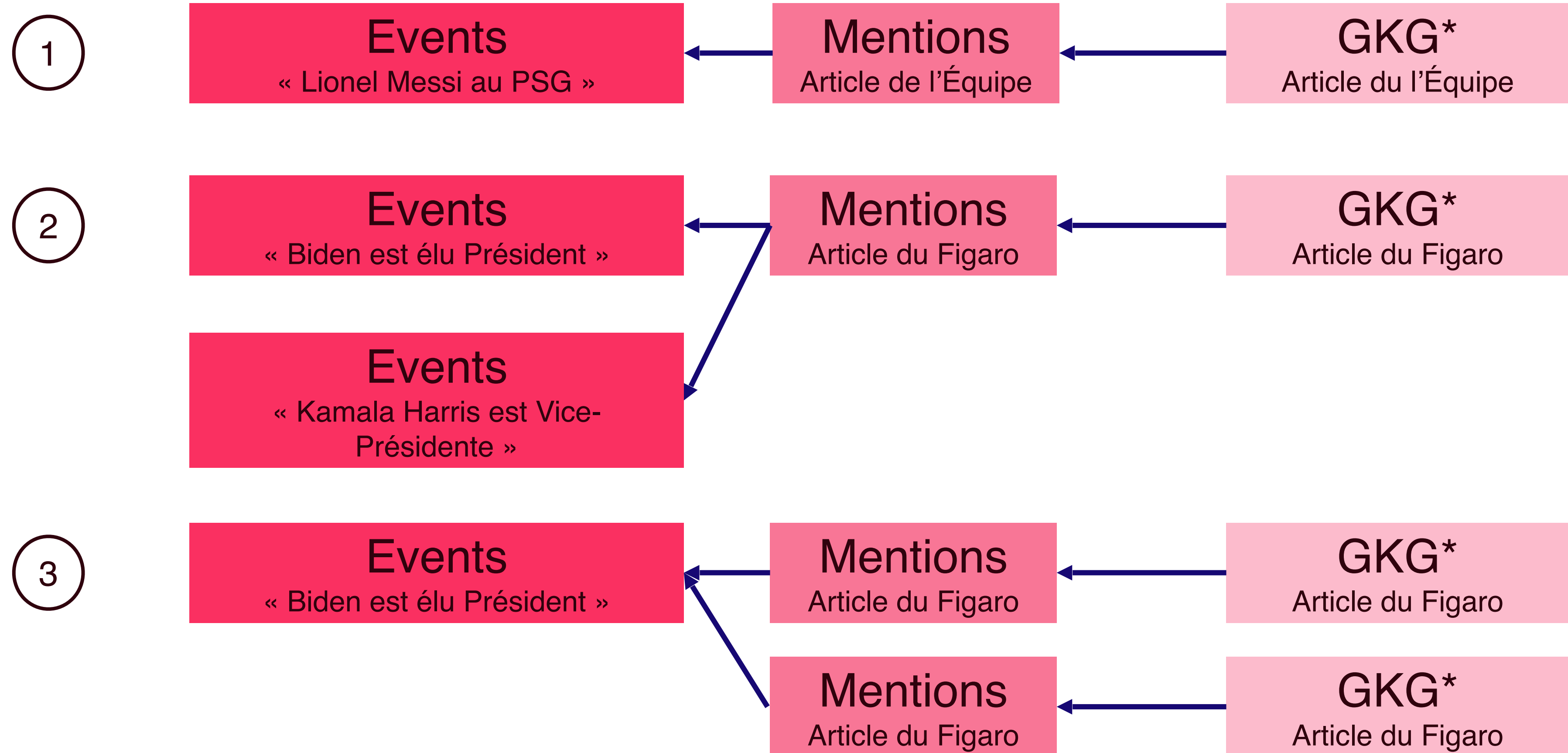
Démonstration

Questions

Annexes

Annexe

Entités et relations



Implémenter la solution

Partition Keys, Clustering Keys

1. afficher le nombre d'articles/événements qu'il y a eu pour chaque triplet (jour, pays de l'évènement, langue de l'article).

- day = 365
- action_geocountrycode = 300
- mentiondoctranslationalinfo = 50
- globaleventid = $1500 \times 4 \times 24h \times 365 = 35\,000\,000$
- mentionid = 35 000 000

PRIMARY KEY ((day, action_geocountrycode, mentiondoctranslationalinfo) globaleventid, mentionid)
WARNING on NULL values in PK

Implémenter la solution

Partition Keys, Clustering Keys

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

- globaleventid = 35 000 000
- day = 365
- month = 12
- year = 1
- nummentions = 0 – infini
- action_geocountrycode = 300
- mentiondoctranslationalinfo = 300

PRIMARY KEY ((action_geocountrycode) month, days, globaleventid)

WARNING on NULL values in PK

Implémenter la solution

Partition Keys, Clustering Keys

3. pour une source de données 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.

- day = 365
- month = 12
- sourcecommonname = infini
- themes = liste infinie
- persons = liste infinie
- locations = plusieurs milliers
- tone = -10 à 10 (presque infini à la vue de la précision du tone)
- documentidentfier = >35 000 000

PRIMARY KEY ((Sourcecommonname) month, day, documentidentfier)

WARNING on NULL values in PK

WARNNG SETS CANNOT BE PRIMARY KEYS

Implémenter la solution

Partition Keys, Clustering Keys

4. étudiez l'évolution des relations entre deux pays (specifies en paramètre) au cours de l'année. Vous pouvez vous baser sur la langue de l'article, le ton moyen des articles, les thèmes plus souvent citées, les personnalités ou tout élément qui vous semble pertinent.

- sourceurl = 1000 * 35 000
- day = 365
- month = 12
- averagetone = -10 à 10 (infini)
- actor1_geocountrycode = 300
- actor2_geocountrycode = 300
- themes = infini

PRIMARY KEY ((actor1_geocountrycode , actor2_geocountrycode), month, day, sourceurl)

WARNING on NULL values in PK

WARNNG SETS cannot be PRIMARY KEYS