



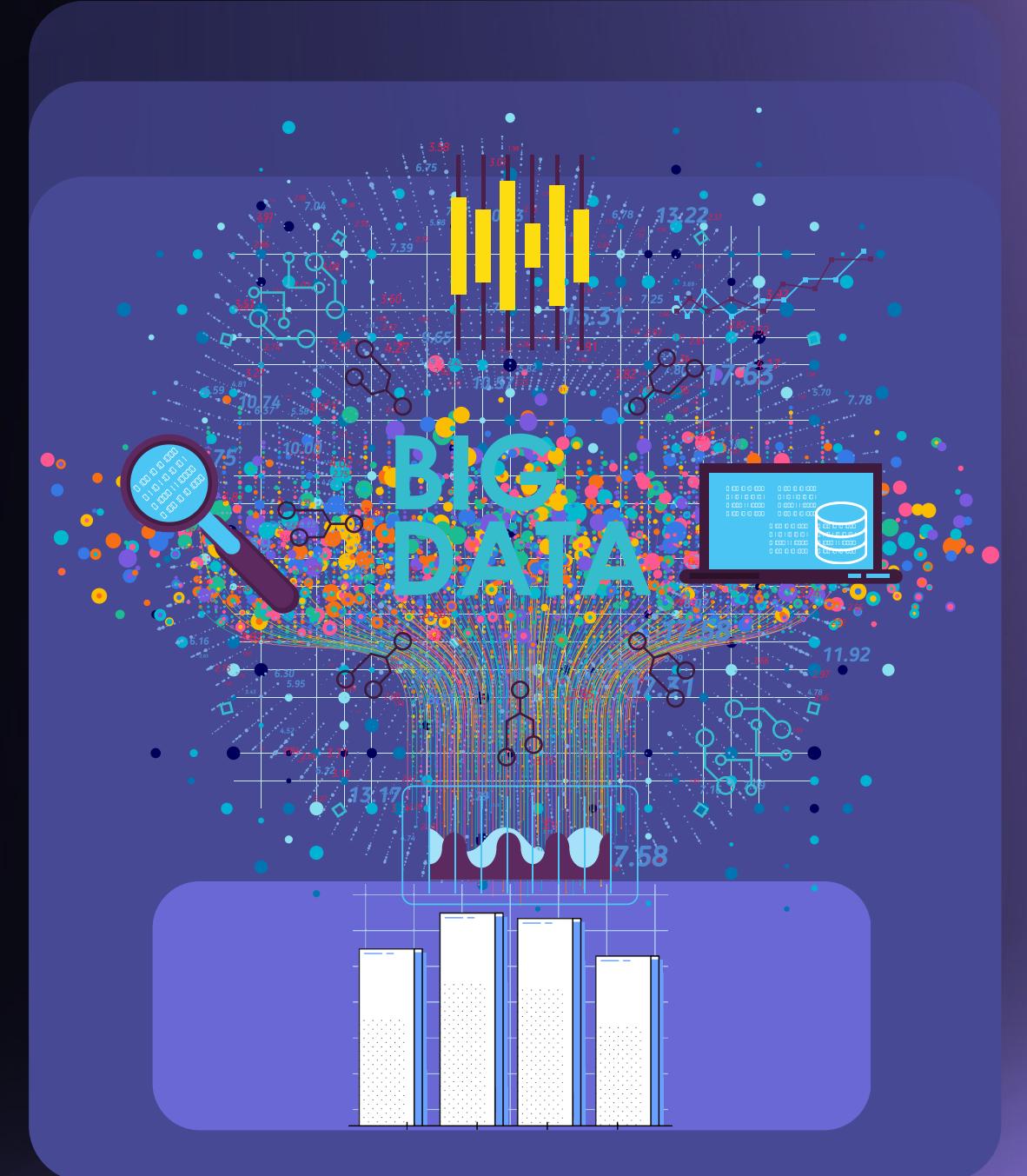
The GDELT Project :

The Global Database of Events, Language and Tone

Présentation du 29 janvier 2024

Laurent GAYRAUD - Célia GIOTTI - Alban PEREIRA - Guillaume PIOL

BGD705 - Bases de données non relationnelles



GDELT project

Sommaire

Objectif

- Construire un système de stockage distribué, résilient et performant pour traiter des données volumineuses

01

Présentation de la base de données choisie

02

Composition du cluster

03

Choix des technologies

04

Processus global du projet

05

Optimisation des entrées dans Cassandra

06

Les requêtes (1 à 4)

07

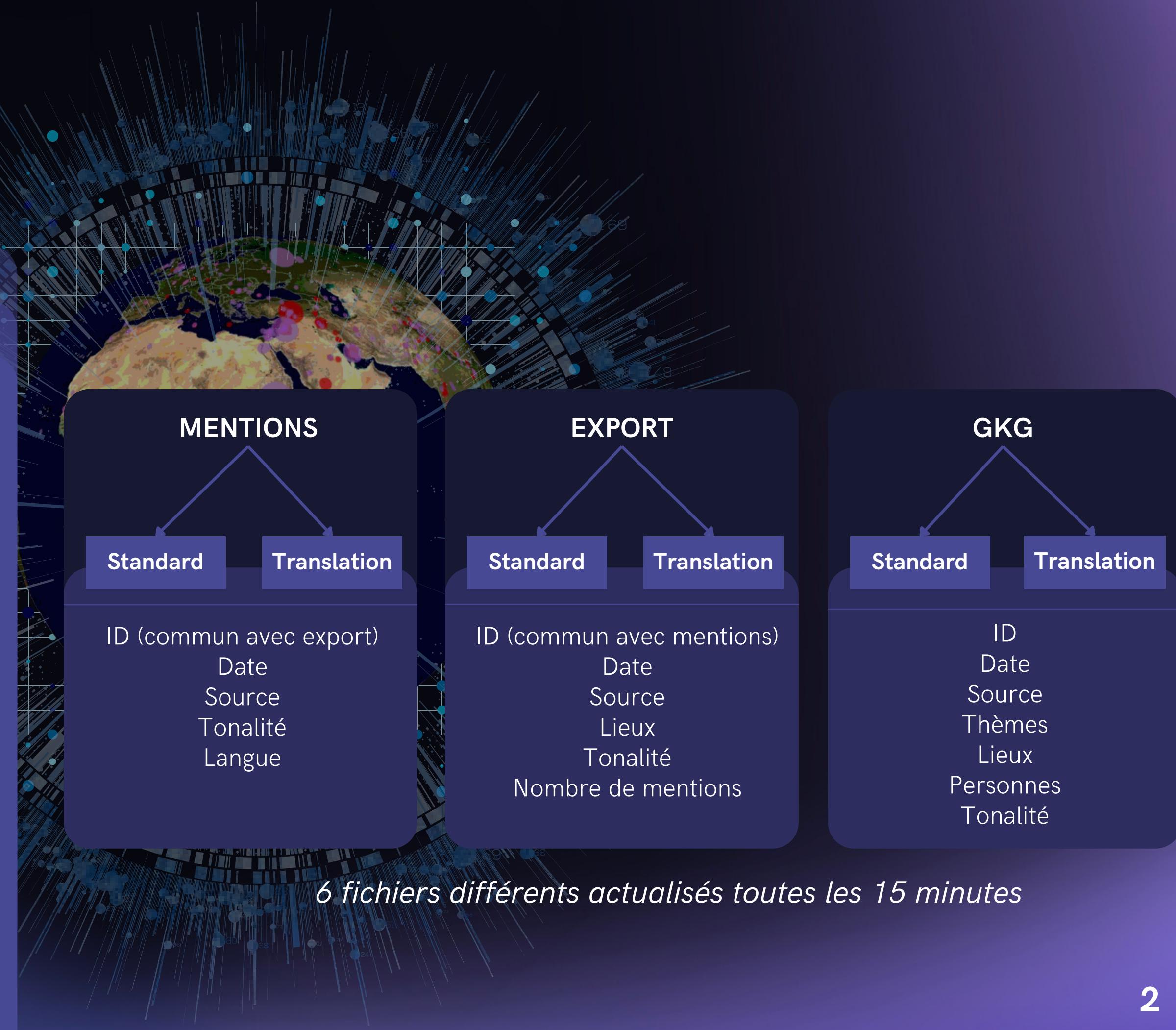
Conclusion

GDELT project

1 - Présentation de la base de données choisie

GDELT

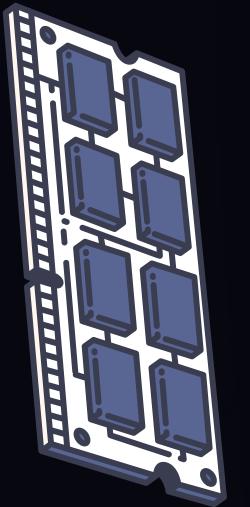
Grande ressource numérique qui surveille et analyse les événements mondiaux en temps réel, couvrant les actualités, les relations sociales et politiques à travers les médias dans plus de 100 langues différentes.



2 - Composition du cluster



8 machines



8 Go de RAM
par machine



70 Go de stockage
physique



6 fichiers générés chaque 15 minutes : soit plus de 200 000 fichiers sur une année



Taille des données
traitées : environ 18
millions de lignes

3 - Choix des technologies

GDELT project

● Choix 1

Téléchargement - Preprocessing

- Téléchargement réparti sur 8 machines (gain de temps)
- Résilience lors du téléchargement (ce qui n'est pas le cas avec Python)

Stockage des données

- Adapté pour un stockage sur de grands volumes de données
- Scalabilité
- Système distribué
- Résilience via sa configuration
- Très contraignant pour les requêtes demandant des agrégations

● Choix 2



● Choix 3

Exécution des requêtes

- Traitement des requêtes avec agrégation possible avec SparkSQL

Visualisation des résultats

- Visualisation claire des données
- Connecteur Spark-Zeppelin

● Choix 4



4 - Processus global

GDELT project

Téléchargement et
preprocessing des
données



Stockage des
données



Exécution des
requêtes



1



2



3



- Téléchargement des fichiers depuis les URLs
- Lecture et décompression des fichiers en mémoire
- Conversion en RDDs puis DataFrames
- Preprocessing, suppression des colonnes non utilisées et formatage des dates

- Import des données
- Stockage des DataFrames traités via Spark

- Requêtes réalisées via Spark SQL

5- Optimisation des entrées dans Cassandra

Nombre de requêtes : 4

Normalement, 4 DataFrames importés (1 par requête)

Mais pour un gain de temps, de traitement et d'espace : **création de 2 DataFrames** pour traiter les 4 requêtes

Fusion mentions et export

Requêtes 1 et 2

DataFrames 1 et 2 : 8 colonnes
Requête 1 : 7 colonnes utilisées
Requête 2 : 7 colonnes utilisées

Table GKG

Requêtes 3 et 4

DataFrames 3 et 4 : 10 colonnes
Requête 3 : 9 colonnes utilisées
Requête 4 : 7 colonnes utilisées

5- Optimisation des entrées dans Cassandra

Nombre de tables créées : 2

- Inutile de créer une table par requête (9 tables seraient nécessaires). L'accès aux données pour l'utilisateur est facilité
- Difficultés liées à l'optimisation des clés primaires évitées avec l'exécution des requêtes se fait avec PySpark et non pas avec CQL
- Eviter de stocker des données en double

6- 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)**

DATAFRAME CONCERNÉ : MENTIONS ET EXPORT

Import sur Cassandra :

```
CREATE TABLE request_1_2 (
    global_event_id INT,
    event_year INT,
    event_month INT,
    event_day INT,
    source_language TEXT,
    event_country TEXT,
    event_country_full_name TEXT,
    num_mentions INT,
    PRIMARY KEY (global_event_id)
);
```

Caractéristiques de la requête :

- Pour le triplet énoncé, on veut le nombre d'articles
- Argument optionnel : langue de l'article
Permet d'avoir pour une date fixée, pour un pays fixé, les langues les plus utilisées avec le comptage des articles

GDELT project

Requête 1 : Visualisation

```
+-----+-----+-----+-----+
|event_year|event_month|event_day|source_language|count|
+-----+-----+-----+-----+
|      2022|          1|       20|         eng|   656|
+-----+-----+-----+-----+
```

Temps d'exécution de la requête 1.1 : 35.743224143981934

```
+-----+-----+-----+-----+
|event_year|event_month|event_day|event_country|source_language|count|
+-----+-----+-----+-----+
|      2022|          1|       20|        FR|        ara|    52|
|      2022|          1|       20|        FR|       axe|     1|
|      2022|          1|       20|        FR|       bos|     2|
|      2022|          1|       20|        FR|      bul|     8|
|      2022|          1|       20|        FR|      cat|     3|
|      2022|          1|       20|        FR|      ces|     4|
|      2022|          1|       20|        FR|      deu|    24|
|      2022|          1|       20|        FR|      ell|    24|
|      2022|          1|       20|        FR|      eng|  656|
|      2022|          1|       20|        FR|      est|     2|
|      2022|          1|       20|        FR|      fas|     3|
|      2022|          1|       20|        FR|      fra|   460|
|      2022|          1|       20|        FR|      heb|     6|
|      2022|          1|       20|        FR|      hin|     2|
|      2022|          1|       20|        FR|      hrv|     3|
|      2022|          1|       20|        FR|      ind|     8|
|      2022|          1|       20|        FR|      ita|    48|
|      2022|          1|       20|        FR|      kor|     7|
|      2022|          1|       20|        FR|      lit|     3|
|      2022|          1|       20|        FR|      mar|     1|
+-----+-----+-----+-----+
```

only showing top 20 rows

Temps d'exécution de la requête 1.2 : 28.532270908355713

6- Requête 2

Import sur Cassandra :

```
CREATE TABLE request_1_2 (
    global_event_id INT,
    event_year INT,
    event_month INT,
    event_day INT,
    source_language TEXT,
    event_country TEXT,
    event_country_full_name TEXT,
    num_mentions INT,
    PRIMARY KEY (global_event_id)
);
```

→ *Pour un pays donné en paramètre, afficher les évènements qui ont eu lieu. Les trier par le nombre de mentions (tri décroissant). Permettre une agrégation par jour/mois/année*

DATAFRAME CONCERNÉ : FUSION DE MENTIONS ET EXPORT

Caractéristiques de la requête :

- Retourner les événements les plus mentionnées par pays en ordre décroissant
- Une agrégation par date doit permettre d'afficher les nombres de mentions les plus élevés par ordre décroissant
 - pour chaque année
 - pour chaque couple (année, mois)
 - pour chaque triplet (année, mois, jour)

Requête 2 : Visualisation

event_year	event_country	global_event_id	num_mentions
2023	FR	1078345217	184
2022	FR	1025728912	350
2021	FR	1060502877	110
2012	FR	1072771052	14

Temps d'exécution de la requête 2 (agrégation par année) : 36.644227743148804

GDEL project

Requête 2 : Visualisation

event_year	event_month	event_day	event_country	total_mentions
2023	1	1	FR	6721
2022	12	31	FR	8357
2022	12	30	FR	10262
2022	12	29	FR	12310
2022	12	28	FR	13190
2022	12	27	FR	10729
2022	12	26	FR	12886
2022	12	25	FR	10743
2022	12	24	FR	17406
2022	12	23	FR	20645
2022	12	22	FR	14306
2022	12	21	FR	6408
2022	12	20	FR	5879
2022	12	19	FR	12821
2022	12	18	FR	10483
2022	12	17	FR	9999
2022	12	16	FR	16764
2022	12	15	FR	7441
2022	12	14	FR	7867
2022	12	13	FR	8600

only showing top 20 rows

Temps d'exécution de la requête 2 (agrégation par jour) : 28.91463851928711

event_year	event_month	event_country	total_mentions
2023	1	FR	6721
2022	12	FR	398210
2022	11	FR	430768
2022	10	FR	463770
2022	9	FR	413372
2022	8	FR	382199
2022	7	FR	424554
2022	6	FR	380245
2022	5	FR	310224
2022	4	FR	345038
2022	3	FR	260850
2022	2	FR	306422
2022	1	FR	360189
2021	12	FR	3307
2021	11	FR	2117
2021	10	FR	1733
2021	9	FR	1546
2021	8	FR	2034
2021	7	FR	2079
2021	6	FR	1547

only showing top 20 rows

Temps d'exécution de la requête 2 (agrégation par mois) : 26.64337134361267

event_year	event_country	total_mentions
2023	FR	6721
2022	FR	4475841
2021	FR	20497
2012	FR	427

Temps d'exécution de la requête 2 (agrégation par année) : 33.69100069999695

6- Requête 3

Import sur Cassandra :

```
CREATE TABLE request_3_4 (
    global_event_id TEXT,
    internet_source TEXT,
    event_year INT,
    event_month INT,
    event_day INT,
    event_themes TEXT,
    person_list TEXT,
    event_tone DOUBLE,
    first_country TEXT,
    second_country TEXT,
    PRIMARY KEY (global_event_id) );
```

→ *Pour une source de données passée en paramètre (gkg.SourceCommonName) afficher 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).*

Permettre une agrégation par jour/mois/année.

DATAFRAME CONCERNÉ : GKG

6- Requête 3

Caractéristiques des sous requêtes :

- Requête 3.1 : Pour une source donnée, retourner thème, personnes, lieux, ton moyen
- Requête 3.2 : Pour une source donnée, retourner le ton moyen utilisé ainsi que le nombre d'articles associés
- Requête 3.3 : Pour une source donnée, retourner les thèmes les plus récurrents ainsi que le nombre d'articles associés et le ton moyen
- Requête 3.4 : Pour une source donnée, retourner les lieux les plus récurrents ainsi que le nombre d'articles associés et le ton moyen
- Requête 3.5 : Pour une source donnée, retourner les personnes les plus récurrentes ainsi que le nombre d'articles associés et le ton moyen

GDEL project

Requête 3 : Visualisation

internet_source	event_year	person	average_tone	person_count
msn.com	2023	kaylee goncalves	-5.126739239914556	10
msn.com	2023	eric adams	-4.466575762491335	9
msn.com	2023	ethan chapin	-4.822210841681	9
msn.com	2023	madison mogen	-4.822210841681	9
msn.com	2023	xana kernodle	-4.822210841681	9
msn.com	2023	anita pointer	1.663481941865202	8
msn.com	2023	los angeles	-1.2692896343506175	8
msn.com	2023	vladimir putin	-3.5819137746300984	7
msn.com	2023	bryan kohberger	-4.968904801277043	7
msn.com	2023	mike driscoll	-4.44653631541549	6
msn.com	2022	vladimir putin	-3.997599531012869	7552
msn.com	2022	joe biden	-3.5788010509873573	7522
msn.com	2022	donald trump	-3.321102604219602	3961
msn.com	2022	volodymyr zelensky	-4.229189863852053	2456
msn.com	2022	los angeles	-3.0587938053654464	2302
msn.com	2022	boris johnson	-3.249824866090651	2171
msn.com	2022	volodymyr zelenskyy	-4.059589710550768	1985
msn.com	2022	olaf scholz	-4.059589710550768	1985
msn.com	2022	volodymyr zelenskiy	-4.059589710550768	1985
msn.com	2022	antony blinken	-3.592263227911818	69
internet_source	event_year	first_country	average_tone	location_count
msn.com	2023	United States	-3.592263227911818	69
msn.com	2023	Germany	-4.620539624463687	13
msn.com	2023	Russia	-4.839172499017749	9
msn.com	2022	United States	-3.830439794118379	38324
msn.com	2022	Germany	-3.4719561375115457	5579
msn.com	2022	Russia	-4.421835527909091	4071

Temps d'exécution de la requête 3.5 (agrégation par année): 7.4756622314453125

6- Requête 4

Import sur Cassandra :

```
CREATE TABLE request_3_4 (
    global_event_id TEXT,
    internet_source TEXT,
    event_year INT,
    event_month INT,
    event_day INT,
    event_themes TEXT,
    person_list TEXT,
    event_tone DOUBLE,
    first_country TEXT,
    second_country TEXT,
    PRIMARY KEY (global_event_id) );
```

→ *Étudier l'évolution des relations entre deux pays (spécifiés 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 cités, les personnalités ou tout élément qui vous semble pertinent.*

DATAFRAME CONCERNÉ : GKG

Caractéristiques des sous requêtes :

- Requête 4.1 : Pour un couple de pays donné, retourner l'évolution du ton moyen
- Requête 4.2 : Pour un couple de pays donné, retourner les thèmes principaux avec un comptage par thème et le ton moyen associé

GDEL project

Requête 4 : Visualisation

first_country	second_country	event_year	average_tone	event_count
France	United States	2023	-0.9910945640116993	11
France	United States	2022	-2.567167745262044	11073

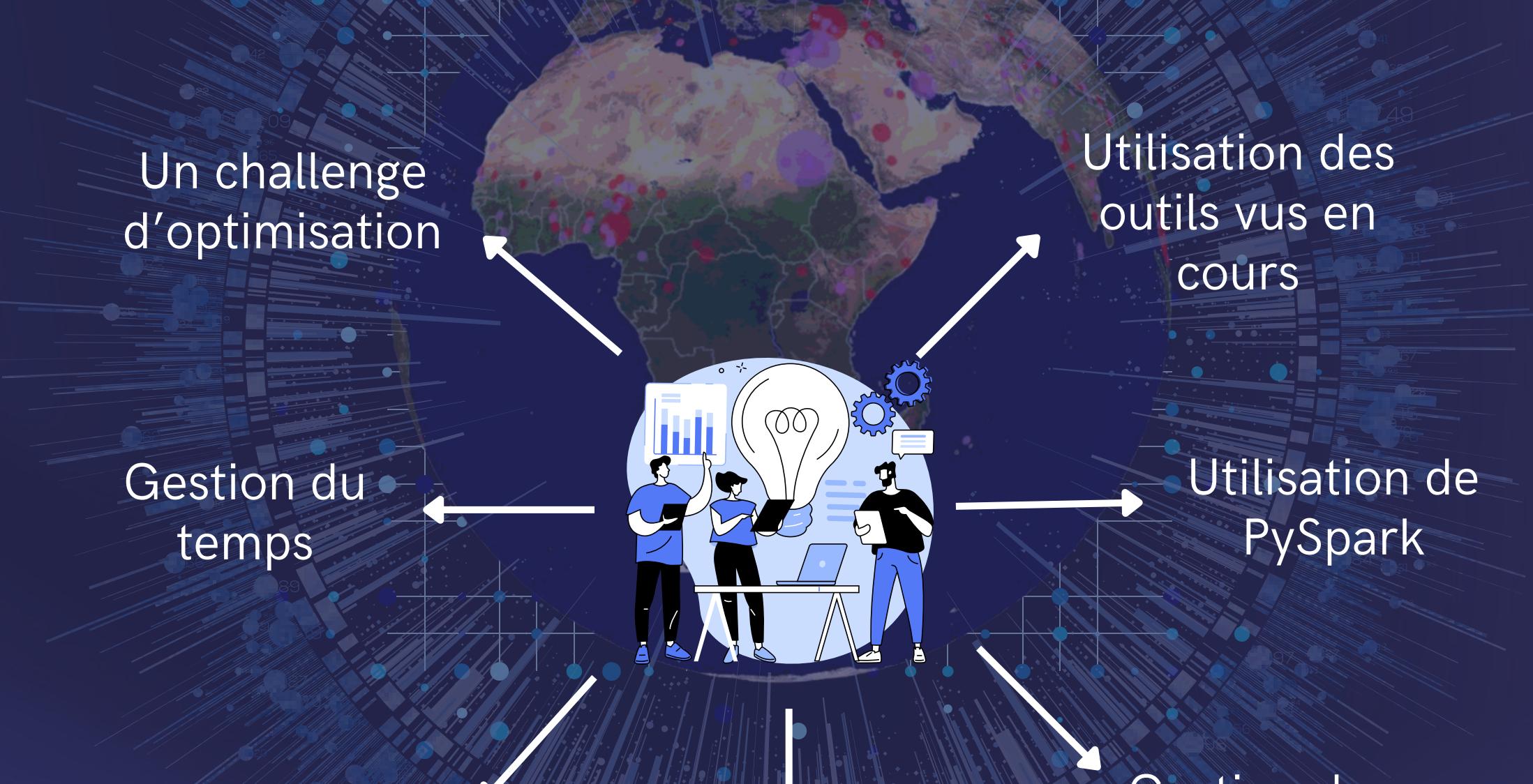
Temps d'exécution de la requête 4.1 (agrégation par année) : 14.083222150802612

first_country	second_country	event_year	event_month	event_day	event_themes	average_tone	theme_count
France	United States	2023	1	1	KILL	0.7793621150214001	4
France	United States	2023	1	1	AFFECT	-0.15928653322015357	4
France	United States	2023	1	1	SICKENED	-4.22535211267606	1
France	United States	2022	12	31	KILL	-0.2646934672650857	8
France	United States	2022	12	31	AFFECT	-2.7869638739043636	3
France	United States	2022	12	31	ARREST	-7.05521472392638	1
France	United States	2022	12	30	KILL	-1.6093009941878187	10
France	United States	2022	12	30	PROTEST	-1.3002144917141827	6
France	United States	2022	12	30	ARREST	-2.155754549794424	5
France	United States	2022	12	29	KILL	-4.148582878895634	24
France	United States	2022	12	29	AFFECT	0.42531605029568187	17
France	United States	2022	12	29	ARREST	-5.626298801786265	2
France	United States	2022	12	28	AFFECT	0.3099062867907563	28
France	United States	2022	12	28	KILL	-0.5235895960691346	13
France	United States	2022	12	28	ARREST	-3.8372656211579974	4
France	United States	2022	12	27	KILL	-1.329026504256349	13
France	United States	2022	12	27	AFFECT	-2.768865445129903	10
France	United States	2022	12	27	ARREST	-8.843971631205683	2
France	United States	2022	12	26	KILL	-6.335770871312488	37
France	United States	2022	12	26	AFFECT	-4.776173578983997	18

only showing top 20 rows

Temps d'exécution de la requête 4.2 (agrégation par jour) : 8.285163879394531

7 - Conclusion



The slide features a central illustration of a globe with a complex network of blue lines and dots representing data connections. In the foreground, three stylized human figures are shown working together around a laptop computer. A large glowing lightbulb is positioned above the laptop, symbolizing ideas or innovation.

Surrounding the central image are six text boxes, each connected to the globe by a white arrow:

- Un challenge d'optimisation
- Gestion du temps
- Des points à améliorer
- Utilisation des outils vus en cours
- Utilisation de PySpark
- Gestion des ressources

At the bottom center, there is additional text:

Compréhension des systèmes distribués

GDELT project

Merci pour votre
attention !

