GDELT : exploration de données massives grâce au NoSQL

Pierre DAL BIANCO
Aurélien KOUADIO
Jia LIAO
Philéas SAMIR
Gwladys SANCHEZ
Pooran SHAHDI

Projet INF728

Plan

Plan

Introduction

- **GDELT**
  une base de connaissances en temps réel constituée à partir d'articles de presse
- **Problématique**
  Comment exploiter une base de données si volumineuse et riche en informations ?
- **Méthode**
  Allers-retour entre objectif et méthode, des ajustements permanents

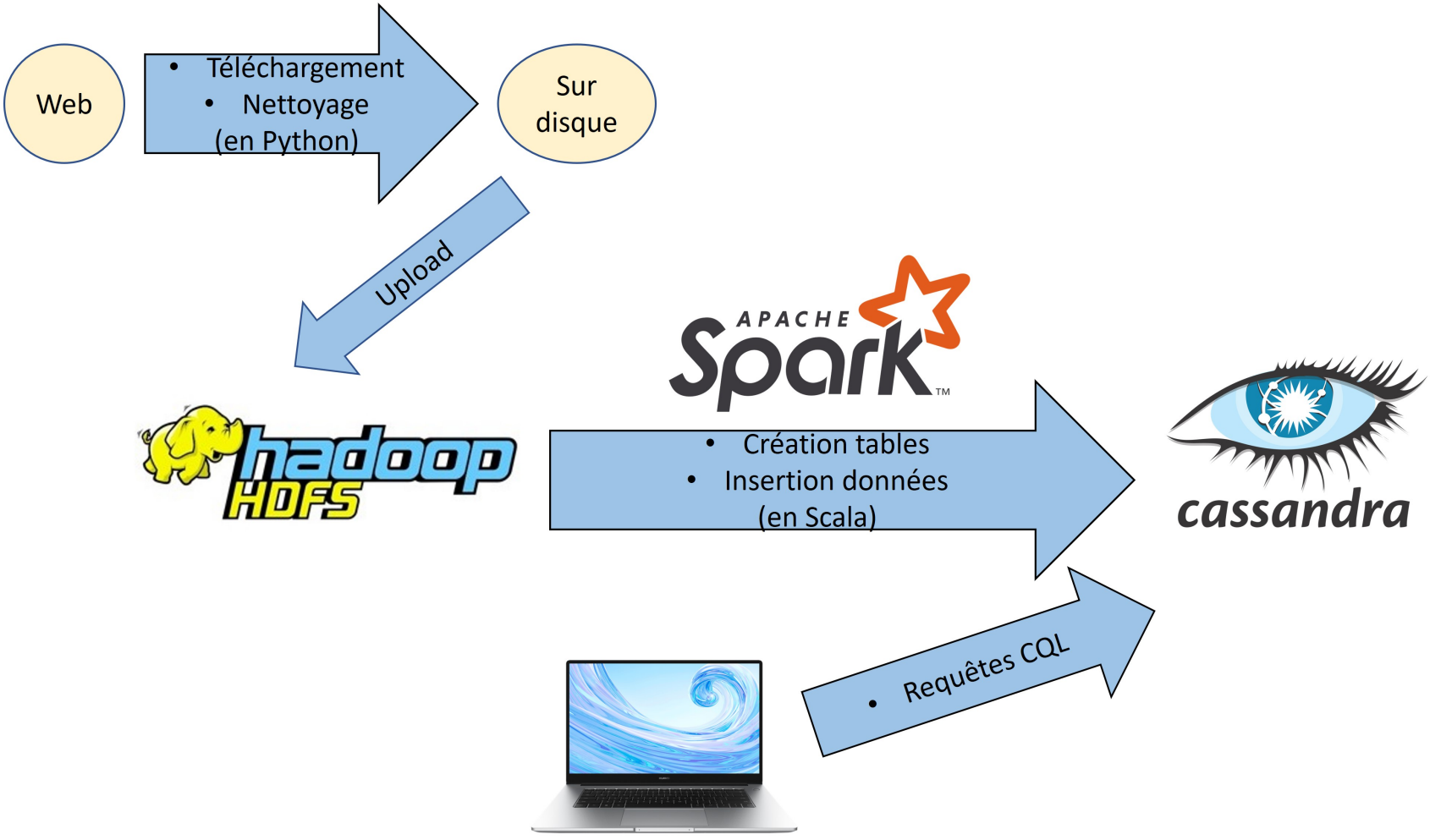
Plan

Méthode - Structure

Architecture et modélisation

- Choix de la technologie : **HDFS, Cassandra & Spark SQL**
  - Cassandra : Requêtes pré-définies et données bien structurées $\rightarrow$ base orientée "colonnes"
  - Spark SQL et HDFS : Volumétrie
- Distribution et réplication
  - 6 nœuds
  - RF = 2 sur HDFS
  - RF = 3 sur Cassandra
- Structure des tables

Méthode - ETL


Web
• Téléchargement
• Nettoyage
(en Python)
Sur disque
Upload
APACHE Spark
hadoop HDFS
• Création tables
• Insertion données
(en Scala)
cassandra
• Requêtes CQL

Méthode - ETL

Traitement des données

- Récupération des données : entre **traitement séquentiel** des fichiers et **parallélisation des tâches**
  - Scraping, téléchargement et Nettoyage : Python
  - Enregistrement sur disque (local) ou distribué (HDFS)
  - Pré-traitement et insertion dans Cassandra : Spark / Scala
  - Requêtes : CQL

Avantages et inconvénients

- Rigidité des requêtes : pré-traitement lourd
- Scalabilité, rapidité et résilience aux pannes

Plan

Tables constituées

Volumétrie

```
ubuntu@tp-hadoop-7:~$ nodetool status
Datacenter: dc1
===============
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address         Load       Tokens       Owns (effective)  Host ID                               Rack
UN  192.168.3.41    2.44 GiB   256          47.7%             f99bbe8b-9f0a-4600-b3c2-1dd15dd7153b  rack1
UN  192.168.3.250   2.15 GiB   256          45.9%             4928bb7c-0407-4613-8e9b-fc36f7e37972  rack1
UN  192.168.3.90    1.69 GiB   256          51.1%             ca36896f-65a1-4486-b71f-b228cb1b5d75  rack1
UN  192.168.3.179   2.32 GiB   256          51.4%             a211c6ab-aac0-44fa-8e11-5af269ec63f7  rack1
UN  192.168.3.99    2 GiB      256          52.1%             bb43daa1-8701-412c-91fe-716ddec383bb  rack1
UN  192.168.3.134   605.95 MiB  256          51.8%             7aa8810c-99e0-4ba0-b131-3530133bbf70  rack1
```

- Volume total des tables : 3,74 GiB (x3)
- Volume sur HDFS : 10 GiB (x2) (avant agrégation)

Requête 1 : nombre d'articles par événement, par jour, pays et langue

Agrégation par jour de mention

```
cqlsh:production> select event_id, jour_event, jour_mention, pays, langue,
sum(total) from table_ab group by pays, event_id, annee_event, mois_event,
jour_event, annee_mention, mois_mention, jour_mention;

 event_id  | jour_event | jour_mention | pays | langue | system.sum(total)
-----------+------------+--------------+------+--------+-------------------
 962221207 |   20210101 |     20210101 |   JE |    eng |                 1
 962221747 |   20210101 |     20210101 |   JE |    eng |                 1
 962223104 |   20210101 |     20210101 |   JE |    eng |                 6
 962224030 |   20210101 |     20210101 |   JE |    fra |                 1
 962253642 |   20210101 |     20210101 |   JE |    eng |                 1
 962253794 |   20210101 |     20210101 |   JE |    fra |                 4
 962255674 |   20210101 |     20210101 |   JE |    kor |                 1
 962255681 |   20210101 |     20210101 |   JE |    kor |                 1
 962271187 |   20210101 |     20210101 |   JE |    eng |                 1
 962271188 |   20210101 |     20210101 |   JE |    eng |                 3
 962271274 |   20210101 |     20210101 |   JE |    eng |                 1
 962278013 |   20210101 |     20210101 |   JE |    eng |                 1
 962278015 |   20210101 |     20210101 |   JE |    eng |                 1
 962281339 |   20210101 |     20210101 |   JE |    eng |                 1
 962281366 |   20210101 |     20210101 |   JE |    eng |                 1
 962281400 |   20210101 |     20210101 |   JE |    eng |                 1
 962282926 |   20210101 |     20210101 |   JE |    eng |                 1
 962283006 |   20210101 |     20210101 |   JE |    eng |                 1
 962283301 |   20210101 |     20210101 |   JE |    eng |                 1
 962283349 |   20210101 |     20210101 |   JE |    eng |                 1
 962287288 |   20210101 |     20210101 |   JE |    eng |                 3
 962287289 |   20210101 |     20210101 |   JE |    eng |                 3
```

Par jour de l'événement

```
cqlsh:production> select event_id, jour_event, pays, langue,
sum(total) from table_ab group by pays, event_id, annee_event
, mois_event, jour_event;

 event_id  | jour_event | pays | langue | system.sum(total)
-----------+------------+------+--------+-------------------
 962221207 |   20210101 |   JE |    eng |                 1
 962221747 |   20210101 |   JE |    eng |                 1
 962223104 |   20210101 |   JE |    eng |                 6
 962224030 |   20210101 |   JE |    fra |                 1
 962253642 |   20210101 |   JE |    eng |                 1
 962253794 |   20210101 |   JE |    fra |                 4
 962255674 |   20210101 |   JE |    kor |                 1
 962255681 |   20210101 |   JE |    kor |                 1
 962271187 |   20210101 |   JE |    eng |                 1
 962271188 |   20210101 |   JE |    eng |                 3
 962271274 |   20210101 |   JE |    eng |                 1
 962278013 |   20210101 |   JE |    eng |                 1
 962278015 |   20210101 |   JE |    eng |                 1
 962281339 |   20210101 |   JE |    eng |                 1
 962281366 |   20210101 |   JE |    eng |                 1
 962281400 |   20210101 |   JE |    eng |                 1
 962282926 |   20210101 |   JE |    eng |                 1
 962283006 |   20210101 |   JE |    eng |                 1
 962283301 |   20210101 |   JE |    eng |                 1
 962283349 |   20210101 |   JE |    eng |                 1
 962287288 |   20210101 |   JE |    eng |                 3
 962287289 |   20210101 |   JE |    eng |                 3
```

TELECOM Paris

Requête 2 : événements d'un pays triés par nombre de mentions

Version 1 : agrégation sur la date de l'*événement*

```
cqlsh:production> SELECT jour_event, event_id, SUM(total) as compte FROM table_ab WHERE pays = 'FR' G
ROUP BY event_id, annee_event, mois_event, jour_event;

 jour_event | event_id  | compte
------------+-----------+--------
   20210101 | 962219618 |     45
   20210101 | 962219620 |      2
   20210101 | 962219635 |     12
   20210101 | 962220031 |      5
   20210101 | 962220032 |      5
   20210101 | 962220242 |      1
   20210101 | 962220315 |     45
   20210101 | 962220321 |     35
   20210101 | 962220322 |      8
   20210101 | 962220323 |     63
   20210101 | 962220326 |     27
   20210101 | 962220328 |     71
   20210101 | 962220331 |     23
   20210101 | 962220332 |      2
   20210101 | 962220335 |      7
   20210101 | 962220336 |      1
   20210101 | 962220337 |     63
   20210101 | 962220338 |      7
```

Requête 2 : événements d'un pays triés par nombre de mentions

Version 2 : agrégation sur la date de la *mention*

```
cqlsh:production> SELECT jour_event, jour_mention, event_id, SUM(total) as compte FROM table_ab WHERE
pays = 'FR' GROUP BY event_id, annee_event, mois_event, jour_event, annee_mention, mois_mention, jour_
mention;

 jour_event | jour_mention | event_id  | compte
------------+--------------+-----------+--------
   20210101 |     20210101 | 962219618 |     45
   20210101 |     20210101 | 962219620 |      2
   20210101 |     20210101 | 962219635 |     12
   20210101 |     20210101 | 962220031 |      5
   20210101 |     20210101 | 962220032 |      5
   20210101 |     20210101 | 962220242 |      1
   20210101 |     20210101 | 962220315 |     45
   20210101 |     20210101 | 962220321 |     35
   20210101 |     20210101 | 962220322 |      8
   20210101 |     20210101 | 962220323 |     63
   20210101 |     20210101 | 962220326 |     27
   20210101 |     20210101 | 962220328 |     71
   20210101 |     20210101 | 962220331 |     23
   20210101 |     20210101 | 962220332 |      2
   20210101 |     20210101 | 962220335 |      7
   20210101 |     20210101 | 962220336 |      1
   20210101 |     20210101 | 962220337 |     63
```

Requête 3 : thèmes, personnes et lieux mentionnés par une source donnée

```
cqlsh:production> SELECT source, theme, personne, lieu, SUM(total) AS somme_total, SUM(somme_ton) AS
somme_ton, jour FROM table_c WHERE source = 'lemonde.fr' GROUP BY theme, personne, lieu, annee, mois,
 jour;

 source     | theme              | personne               | lieu | somme_total | somme_ton | jour
------------+--------------------+------------------------+------+-------------+-----------+----------
 lemonde.fr | ACT_MAKESTATEMENT |                    UNK |   UK |           1 | -0.411523 | 20210130
 lemonde.fr |            AFFECT |                    UNK |   AF |           1 |  -7.36842 | 20210503
 lemonde.fr |            AFFECT |                    UNK |   AF |           1 |   -1.9656 | 20210831
 lemonde.fr |            AFFECT |                    UNK |   BO |           1 |  -5.01089 | 20211117
 lemonde.fr |            AFFECT |                    UNK |   CD |           1 |  -3.07329 | 20210729
 lemonde.fr |            AFFECT |                    UNK |   CF |           1 |  -3.49345 | 20210818
 lemonde.fr |            AFFECT |                    UNK |   FR |           1 |  -1.75879 | 20210215
 lemonde.fr |            AFFECT |                    UNK |   FR |           1 |  -2.88625 | 20210223
 lemonde.fr |            AFFECT |                    UNK |   FR |           1 |  -4.46429 | 20210526
 lemonde.fr |            AFFECT |                    UNK |   FR |           1 |  -2.95699 | 20210719
 lemonde.fr |            AFFECT |                    UNK |   GM |           1 |   -1.9656 | 20210831
 lemonde.fr |            AFFECT |                    UNK |   LA |           1 |        -5 | 20211202
 lemonde.fr |            AFFECT |                    UNK |   TS |           1 |  -1.75879 | 20210123
 lemonde.fr |            AFFECT |                    UNK |  UNK |           1 | -0.842105 | 20210127
 lemonde.fr |            AFFECT |                    UNK |   US |           1 |        -5 | 20211202
 lemonde.fr |            AFFECT |            adama diop |  UNK |           1 |   4.34783 | 20210716
 lemonde.fr |            AFFECT |  alexander lukashenko |   LG |           1 |  -3.01724 | 20210829
 lemonde.fr |            AFFECT |        amico patrick  |   FR |           1 |   -4.5208 | 20210901
 lemonde.fr |            AFFECT |    andreano e piccini |   BR |           1 |  -2.11176 | 20210107
 lemonde.fr |            AFFECT |       angela merkel   |   GM |           1 |  -3.79147 | 20210903
 lemonde.fr |            AFFECT |      brigitte bardot  |   JM |           1 | -0.478469 | 20210801
```

Requête 4 : relations entre pays

```
cqlsh:production> SELECT langue, lieu, SUM(total) AS somme_total, SUM(somme_ton) AS somme_ton FROM ta
ble_d WHERE langue = 'tur' AND lieu = 'FR' GROUP BY annee, mois, jour;

 langue | lieu | somme_total | somme_ton
--------+------+-------------+-------------
    tur |   FR |         136 |  -330.14392
    tur |   FR |         199 |   -55.01159
    tur |   FR |         144 |  -273.24723
    tur |   FR |         308 |  -457.48843
    tur |   FR |         252 |  -172.70329
    tur |   FR |         180 |  -307.71869
    tur |   FR |         222 |  -507.62603
    tur |   FR |         197 |   -61.06843
    tur |   FR |         193 |  -382.02268
    tur |   FR |         110 |  -112.92646
    tur |   FR |         240 |  -536.85742
    tur |   FR |         254 |  -182.06695
    tur |   FR |         247 |  -303.03266
    tur |   FR |         246 |  -391.10367
    tur |   FR |         278 |   -435.9928
    tur |   FR |         267 |  -601.71279
    tur |   FR |         245 |  -407.67164
    tur |   FR |         394 |  -574.94082
```

Plan

Conclusion

- Complétude des données : un pré-traitement sans perte
- Réduction de la volumétrie :
  - plusieurs centaines de Go avant nettoyage
  - de 60Go *après nettoyage* à seulement 10Go sur disque
  - volume total des tables (avant réplication) : moins de 4Go
- Efficacité des requêtes : temps de réponse très rapides