Rapport projet Framework Big Data

Membre du groupe:
Komlan DANTODJI
Oumar NIANG

Master 1 Big Data et Fouilles de données:
Université Paris 8

Plan:

1) Introduction

Le virus identifié en janvier 2020 en Chine est un nouveau coronavirus, nommé SARS-CoV-2. La maladie provoquée par ce coronavirus a été nommée COVID-19 par l'Organisation mondiale de la Santé - OMS. Depuis le 11 mars 2020, l'OMS qualifie la situation mondiale du COVID-19 de pandémie ; c'est-à-dire que l'épidémie est désormais mondiale.
Les coronavirus sont une famille de virus qui provoquent des maladies allant d'un simple rhume (certains virus saisonniers sont des coronavirus) à des pathologies plus sévères (comme les détresses respiratoires du MERS, du SRAS ou de la COVID-19).

En France, les premiers cas sont survenus en février 2020.Ainsi nous allons nous intéresser à la progression de la covid-19 depuis son arrivée jusqu'à aujourd'hui.

2) Problématique

Le but est d'analyser les données tweets qui ont été effectuées pendant la pandémie. D'autres données sont récupérées depuis Datagouv, OpenData pour les mettre en relation avec les données de twitter sur les thèmes covid 19 confinement,masques.

3) Objectifs

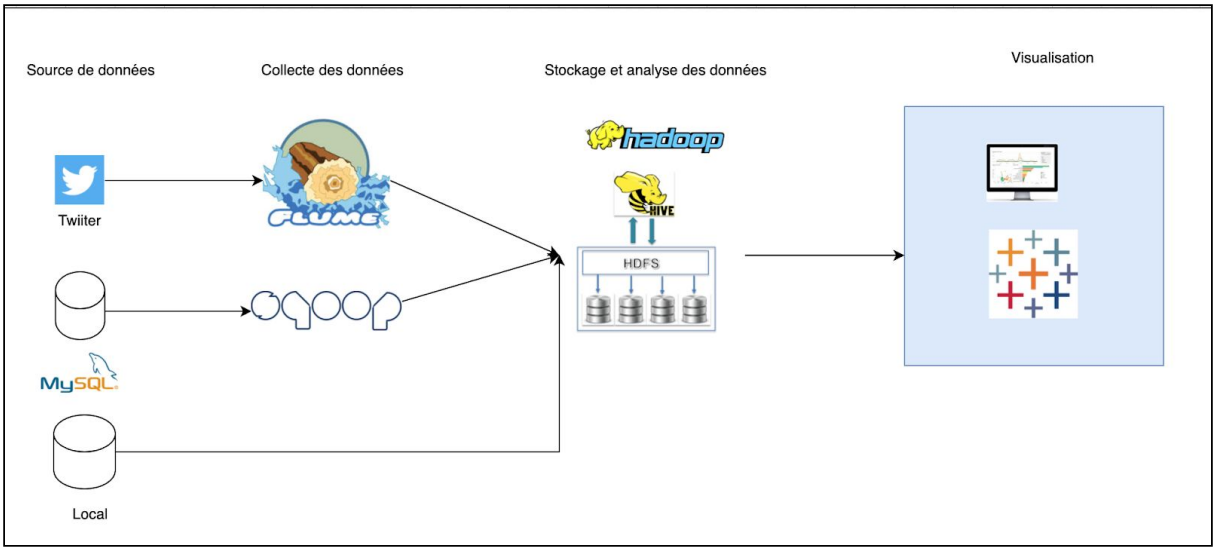
Sur ce projet ,il consiste à faire des analyses statistiques avec les données qui contiennent toutes informations relatives à la covid-19 en france et dans le monde.

Nous allons récupérer des données provenant de diverses sources notamment des données en temps réel de tweet, provenant d'une base de données sur l'évolution de la covid et des données en local .

Ainsi on pourra mettre en place des graphes de suivi selon le sexe, la classe d'age ,la periode, le nombre de personne hospitalisées , le nombre de personne en réanimation, le nombre de personnes décedées en fonction de l'age etc…

Pour les données de tweets on pourra calculer les tops 10 meilleurs utilisateurs qui ont retweeté des tweets.Ensuite classer les pays en fonction du nombre de tweets.

4)Technologies utilisées(Architecture)



Nous avons 4 étapes:

- Sources de données:
  1) **Twitter :**Les données sont récupérées en temps réel.
  2) **Mysql :** Les données proviennent de la base de données.
  3) **Local :** Les données sont stockées en local dans un fichier csv.
- La collecte des données:
  1) **Flume :** Pour récupérer les données de tweet en temps réel vers HDFS:
  2) **Sqoop :** Permet de récupérer les données depuis mysql vers HDFS.
  3) Pour les données en local on récupère les données vers HDFS avec la commande **put**.
- Analyse des données:
  On analyse les données avec hive puis charge les résultats dans un fichier csv pour la visualisation avec tableau desktop.

5) Mise en place de l'architecture:

Premièrement on a essayé d'utiliser la plateforme cloudera VM dans la réalisation de notre projet ,mais nous avons rencontré des problèmes avec la récupération des données avec flume.On a donc recours à mettre en place notre architecture personnelle en installant toutes les briques en local notamment hadoop, hive, sqoop, flume, mysql.

6) Données:

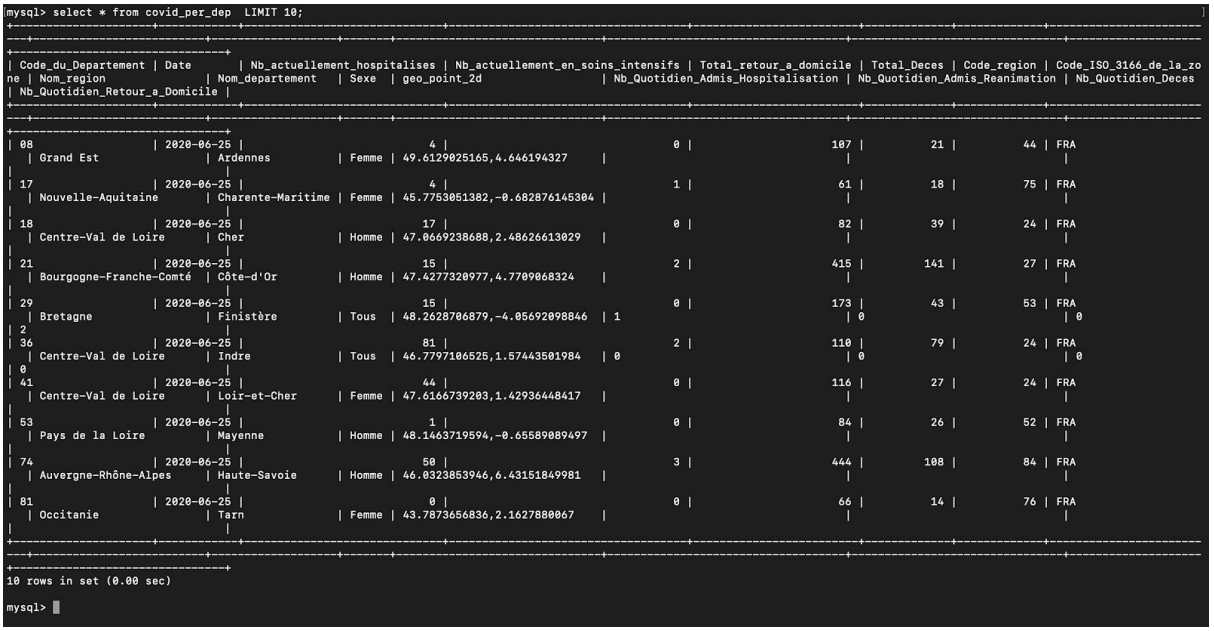
- **Les données de tweets**
  On a récupéré les données de tweeter à partir de son API après une demande auprès de tweeter.
  La commande ci-dessous charge les données tweets dans hdfs.

```
flume-ng agent -n TwitterAgent -Xmx512m -c conf -f
/Users/omzo/hadoop/apache-flume-1.9.0-bin/conf/flume.conf
-Dflume.root.logger=DEBUG.console
```

- Les données de Mysql

Ci-dessous le lien du site d'opendata contenant les données hospitalières relatives à l'épidémie de COVID-19 en France.
https://public.opendatasoft.com/explore/dataset/donnees-hospitalieres-covid-19-dep-france/table/?disjunctive.countrycode_iso_3166_1_alpha3&disjunctive.nom_dep_min.

On a créé une table correspondant à ce fichier csv dans mysql.



```
mysql> select * from covid_per_dep  LIMIT 10;
+----------------------+------------+-----------------------------+-----------------------------------+-----------------------+-------------+-------------+-----------------------+
| Code_du_Departement | Date       | Nb_actuellement_hospitalises | Nb_actuellement_en_soins_intensifs | Total_retour_a_domicile | Total_Deces | Code_region | Code_ISO_3166_de_la_zone | Nom_region                | Nom_departement    | Sexe  | geo_point_2d                | Nb_Quotidien_Admis_Hospitalisation | Nb_Quotidien_Admis_Reanimation | Nb_Quotidien_Deces | Nb_Quotidien_Retour_a_Domicile |
+----------------------+------------+-----------------------------+-----------------------------------+-----------------------+-------------+-------------+-----------------------+
| 08                   | 2020-06-25 |                           4 |                                 0 |                   107 |          21 |          44 | FRA                   | Grand Est                 | Ardennes           | Femme | 49.6129025165,4.646194327   |                                  0 |                            107 |                 21 |                             44 |
| 17                   | 2020-06-25 |                           4 |                                 1 |                    61 |          18 |          75 | FRA                   | Nouvelle-Aquitaine        | Charente-Maritime  | Femme | 45.7753051382,-0.682876145304 |                                  0 |                             61 |                 18 |                             75 |
| 18                   | 2020-06-25 |                          17 |                                 0 |                    82 |          39 |          24 | FRA                   | Centre-Val de Loire       | Cher               | Homme | 47.0669238688,2.48626613029 |                                  0 |                             82 |                 39 |                             24 |
| 21                   | 2020-06-25 |                          15 |                                 2 |                   415 |         141 |          27 | FRA                   | Bourgogne-Franche-Comté   | Côte-d'Or          | Homme | 47.4277320977,4.7709068324  |                                  0 |                            415 |                141 |                             27 |
| 29                   | 2020-06-25 |                          15 |                                 0 |                   173 |          43 |          53 | FRA                   | Bretagne                  | Finistère          | Tous  | 48.2628706879,-4.05692098846 | 1                                |                              0 |                  0 |                              2 |
| 36                   | 2020-06-25 |                          81 |                                 2 |                   110 |          79 |          24 | FRA                   | Centre-Val de Loire       | Indre              | Tous  | 46.7797106525,1.57443501984 | 0                                |                              0 |                  0 |                              0 |
| 41                   | 2020-06-25 |                          44 |                                 0 |                   116 |          27 |          24 | FRA                   | Centre-Val de Loire       | Loir-et-Cher       | Femme | 47.6166739203,1.42936448417 |                                  0 |                            116 |                 27 |                             24 |
| 53                   | 2020-06-25 |                           1 |                                 0 |                    84 |          26 |          52 | FRA                   | Pays de la Loire          | Mayenne            | Homme | 48.1463719594,-0.65589089497 |                                  0 |                             84 |                 26 |                             52 |
| 74                   | 2020-06-25 |                          50 |                                 3 |                   444 |         108 |          84 | FRA                   | Auvergne-Rhône-Alpes      | Haute-Savoie       | Homme | 46.0323853946,6.43151849981 |                                  0 |                            444 |                108 |                             84 |
| 81                   | 2020-06-25 |                           0 |                                 0 |                    66 |          14 |          76 | FRA                   | Occitanie                 | Tarn               | Femme | 43.7873656836,2.1627880067  |                                  0 |                             66 |                 14 |                             76 |
+----------------------+------------+-----------------------------+-----------------------------------+-----------------------+-------------+-------------+-----------------------+
10 rows in set (0.00 sec)

mysql>
```

On récupère les données depuis mysql vers hdfs grâce à cette commande ci dessous :

```
sqoop import-all-tables --connect jdbc:mysql://localhost:3306/covid_19_hospital
--username=root --password=omaryande --hive-import --create-hive-table
--hive-database covid19 -m 1
```

- Les données récupérées depuis le local

https://www.data.gouv.fr/en/datasets/coronavirus-covid19-evolution-par-pays-et-dans-le-monde-maj-quotidienne/#_

Ci-dessous le lien du site de DataGouv contenant les données relatives à l'évolution des cas de covid par pays dans le monde.

```
hdfs dfs -put /home/komlan/Downloads/coronavirus.politologue.com-pays-2021-01-08.csv /tweeterdata
```

```
hadoop@Komlan:~$ hdfs dfs -cat /user/hive/warehouse/covidtable/coronavirus.politologue.com-pays-2021-01-08.csv
2021-01-08;Andorre;8489;84;7724;0.99;90.99;8.02
2021-01-08;Émirats Arabes Unis;224704;697;201396;0.31;89.63;10.06
2021-01-08;Afghanistan;53332;2257;43440;4.23;81.45;14.32
2021-01-08;Antigua-et-Barbuda;167;5;150;2.99;89.82;7.19
2021-01-08;Albanie;62378;1230;36971;1.97;59.27;38.76
2021-01-08;Arménie;161415;2908;147961;1.80;91.66;6.53
2021-01-08;Angola;17974;413;11955;2.30;66.51;31.19
2021-01-08;Argentine;1690006;44122;1484794;2.61;87.86;9.53
2021-01-08;Autriche;376793;6641;349215;1.76;92.68;5.56
2021-01-08;Australie;28571;909;25817;3.18;90.36;6.46
2021-01-08;Azerbaïdjan;223417;2869;205406;1.28;91.94;6.78
2021-01-08;Bosnie-Herzégovine;114920;4285;80868;3.73;70.37;25.90
2021-01-08;Barbade;780;7;335;0.90;42.95;56.15
2021-01-08;Bangladesh;520690;7734;465279;1.49;89.36;9.16
2021-01-08;Belgique;658655;19936;0;3.03;0.00;96.97
2021-01-08;Burkina Faso;7713;89;5533;1.15;71.74;27.11
2021-01-08;Bulgarie;207259;8017;130191;3.87;62.82;33.32
2021-01-08;Bahreïn;94633;353;91630;0.37;96.83;2.80
2021-01-08;Burundi;885;2;773;0.23;87.34;12.43
2021-01-08;Bénin;3304;44;3185;1.33;96.40;2.27
2021-01-08;Saint Barthélemy;206;1;172;0.49;83.50;16.02
2021-01-08;Brunéi Darussalam;173;3;149;1.73;86.13;12.14
2021-01-08;Bolivie;168891;9304;136266;5.51;80.68;13.81
2021-01-08;Brésil;7961673;200498;7111558;2.52;89.32;8.16
2021-01-08;Bahamas;7969;175;6313;2.20;79.22;18.58
2021-01-08;Bhoutan;767;0;459;0.00;59.84;40.16
2021-01-08;Botswana;16050;48;12927;0.30;80.54;19.16
2021-01-08;Bélarus;208601;1498;190966;0.72;91.55;7.74
2021-01-08;Belize;11152;267;10344;2.39;92.75;4.85
2021-01-08;Canada;646782;16686;547143;2.58;84.59;12.83
2021-01-08;Rép. Dém. du Congo;18969;611;14743;3.22;77.72;19.06
2021-01-08;République Centrafricaine;4969;63;4885;1.27;98.31;0.42
2021-01-08;Congo (Brazzaville);7127;108;5846;1.52;82.03;16.46
2021-01-08;Suisse;477983;8219;317600;1.72;66.45;31.83
2021-01-08;Côte d'Ivoire;23254;139;22325;0.60;96.00;3.40
2021-01-08;Chili;633381;16974;595799;2.68;94.07;3.25
2021-01-08;Cameroun;26848;448;24892;1.67;92.71;5.62
2021-01-08;Chine;87336;4634;82176;5.31;94.09;0.60
2021-01-08;Colombie;1737347;45067;1580285;2.59;90.96;6.45
2021-01-08;Costa Rica;176407;2286;135475;1.30;76.80;21.91
2021-01-08;Cuba;13823;148;11531;1.07;83.42;15.51
2021-01-08;Cap-Vert;12146;113;11665;0.93;96.04;3.03
2021-01-08;Chypre;26208;140;2057;0.53;7.85;91.62
```

7) Analyse avec hive

- **Analyse des données tweets:**

voici le schémas de la table **tweets:**

```
CREATE EXTERNAL TABLE tweets(
       id BIGINT,
       created_at STRING,
       source STRING,
       favorited BOOLEAN,
       retweeted_status STRUCT<
           text: STRING,
           `user`: STRUCT<screen_name: STRING, name: STRING>,
           retweet_count: INT>,
       entities STRUCT<
           urls: ARRAY<STRUCT<expanded_url: STRING>>,
           user_mentions: ARRAY<STRUCT<screen_name: STRING, name: STRING>>,
           hashtags: ARRAY<STRUCT<text: STRING>>>,
       text STRING,
       `user` STRUCT<
           screen_name: STRING,
           name: STRING,
```

```
            friends_count: INT,
            followers_count: INT,
            statues_count: INT,
            location: STRING,
            verfied: BOOLEAN,
            utc_offset: INT,
            time_zone: STRING>,
        in_reply_to_screen_name STRING)
        ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
    LOCATION '/user/hive/warehouse/tweetertable';
```

Voici le schéma de la table au niveau de hive.

```
hive> describe tweets;
OK
id                      bigint                  from deserializer
created_at              string                  from deserializer
source                  string                  from deserializer
favorited               boolean                 from deserializer
retweeted_status        struct<text:string,user:struct<screen_name:string,name:string>,retweet_count:int>       from deserializer
entities                struct<urls:array<struct<expanded_url:string>>,user_mentions:array<struct<screen_name:string,name:string>>,hashtags:array<struct<text:string>>> from deserial
izer
text                    string                  from deserializer
user                    struct<screen_name:string,name:string,friends_count:int,followers_count:int,statues_count:int,location:string,verfied:boolean,utc_offset:int,time_zone:string
>       from deserializer
in_reply_to_screen_name string                  from deserializer
Time taken: 0.198 seconds, Fetched: 9 row(s)
hive>
```

On charge les données du fichier dans la table **tweets**.

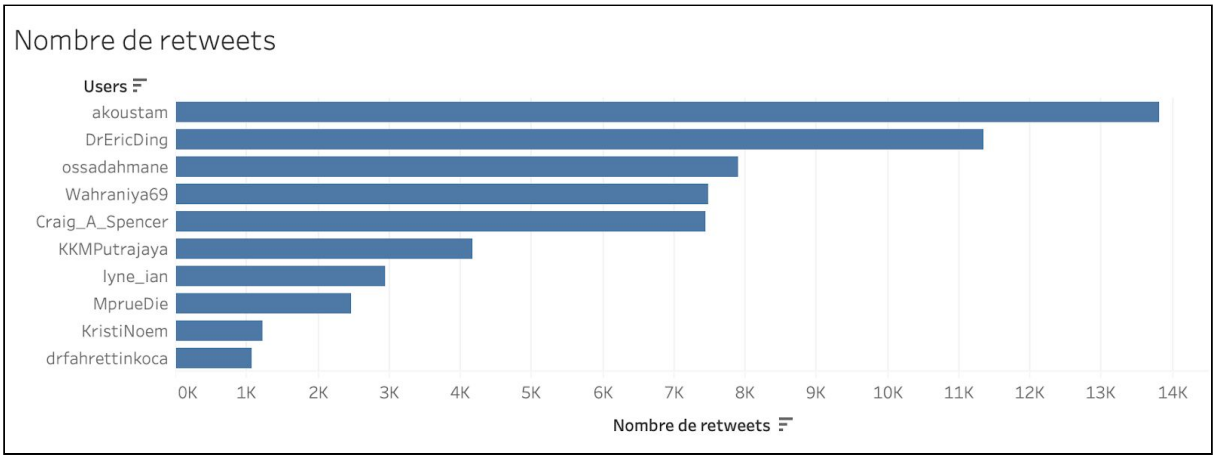
```
LOAD DATA INPATH '/user/flume/tweeter_data/FlumeData.1610225154985' INTO TABLE
tweets;
```

On calcule le top 10 des meilleurs utilisateurs qui ont retweeté sur la covid 19.

```
INSERT OVERWRITE LOCAL DIRECTORY '/Users/omzo/Desktop'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
SELECT t.retweeted_screen_name,
    sum(retweets) AS total_retweets,
    count(*) AS tweet_count FROM
          (SELECT   retweeted_status.`user`.screen_name   AS   retweeted_screen_name,
retweeted_status.text, max(retweeted_status.retweet_count) AS retweets FROM tweets
    GROUP BY retweeted_status.`user`.screen_name, retweeted_status.text) t
     GROUP BY t.retweeted_screen_name ORDER BY total_retweets DESC LIMIT 10;
```

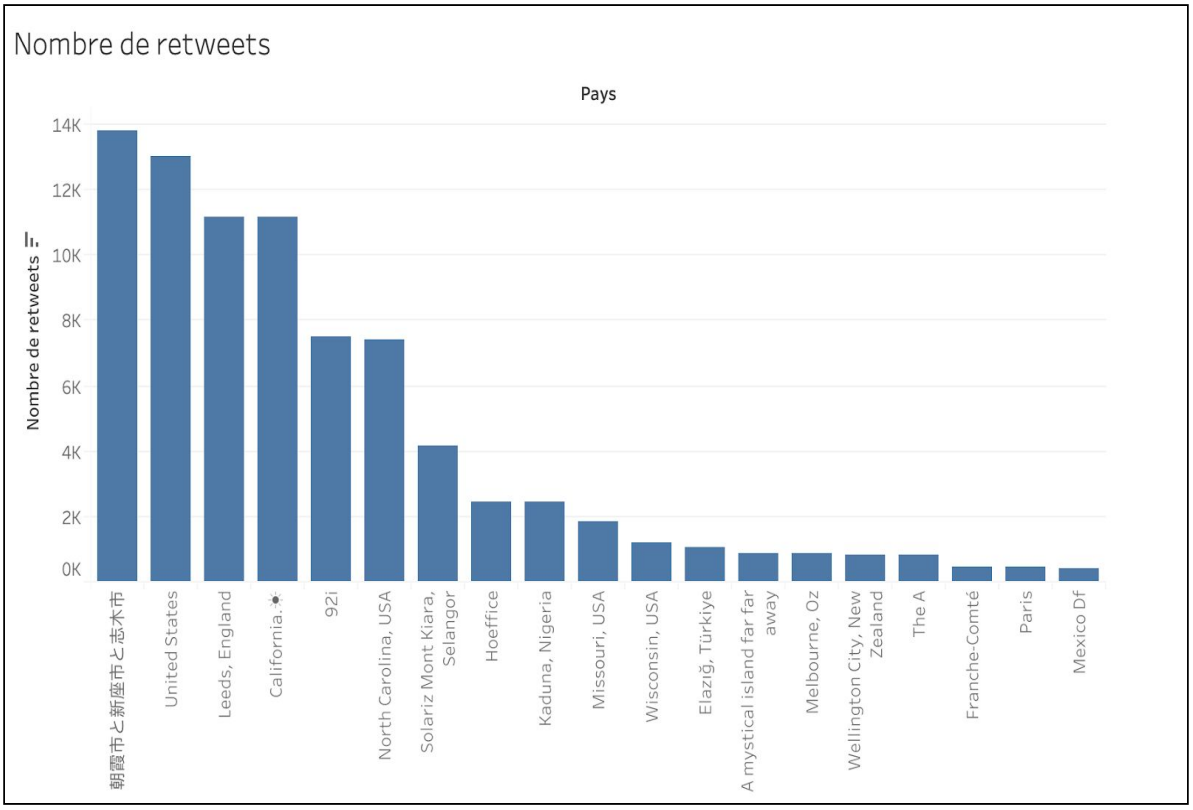
```
(base) omzo@MacBook-Pro-de-Oumar Desktop % cat top_10_retweet.csv
akoustam,13811,1
DrEricDing,11339,2
ossadahmane,7908,1
Wahraniya69,7487,1
Craig_A_Spencer,7435,1
KKMPutrajaya,4166,1
lyne_ian,2941,2
MprueDie,2473,1
KristiNoem,1221,1
drfahrettinkoca,1070,1
(base) omzo@MacBook-Pro-de-Oumar Desktop %
(base) omzo@MacBook-Pro-de-Oumar Desktop %
```

Le résultat de la requête est stocké dans un fichier csv qui sera envoyé dans tableau desktop pour la visualisation sous forme de diagrammes.



On calcule le top zone de localisation en fonction du nombre de retweets.

```
SELECT t.location,
   sum(retweets) AS total_retweets,
   count(*) AS tweet_count  FROM
   (SELECT `user`.location AS location, retweeted_status.text,
     max(retweeted_status.retweet_count) AS retweets FROM tweets
    GROUP BY `user`.location,retweeted_status.text) t
    GROUP BY t.location ORDER BY total_retweets DESC LIMIT 10;
```



- **Analyse des données provenant de mysql**

On créer une table **covid_stat_per_region.**

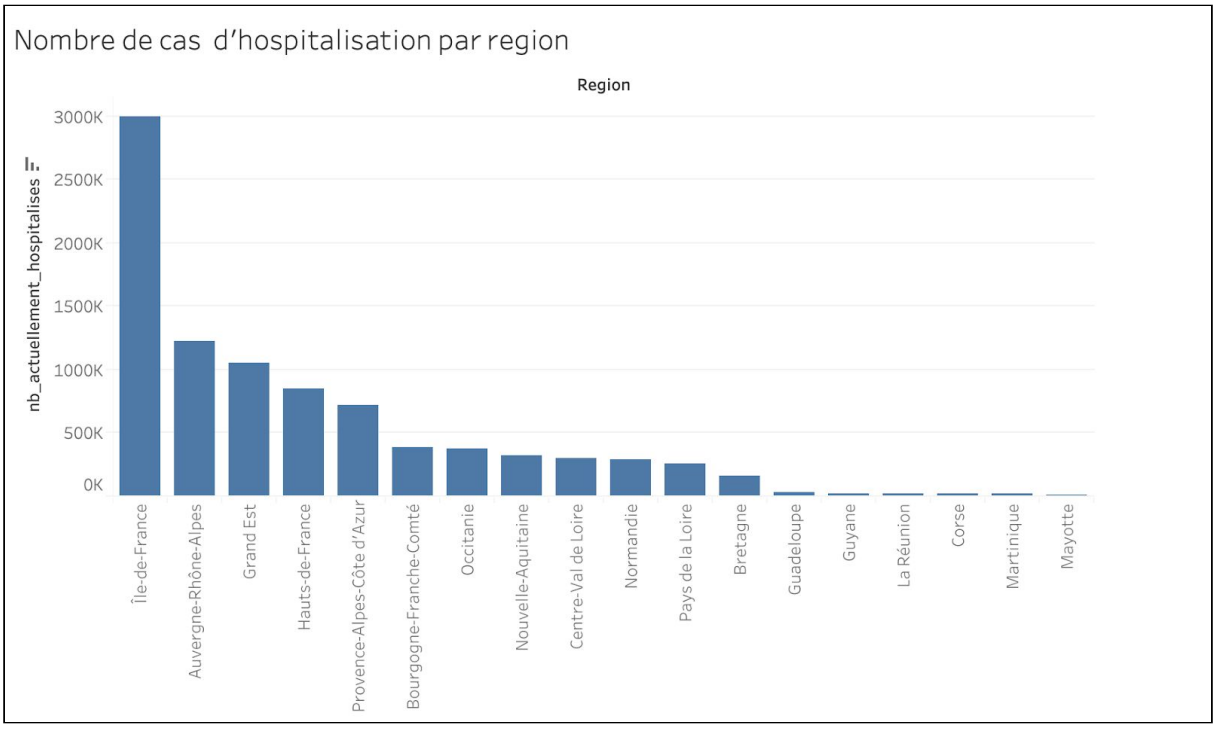
```
CREATE TABLE covid_stat_per_region (nom_departement
string,nb_actuellement_hospitalises int,nb_actuellement_en_soins_intensifs
int,total_retour_a_domicile int,total_deces int);
```

On calcule le nombre de cas d'hospitalisation,le nombre de personnes en soins intensifs,le nombre de personnes guéries et le nombre de personnes décédées dans chaque région ,puis ces résultats sont stockés dans la table **covid_stat_per_region**.

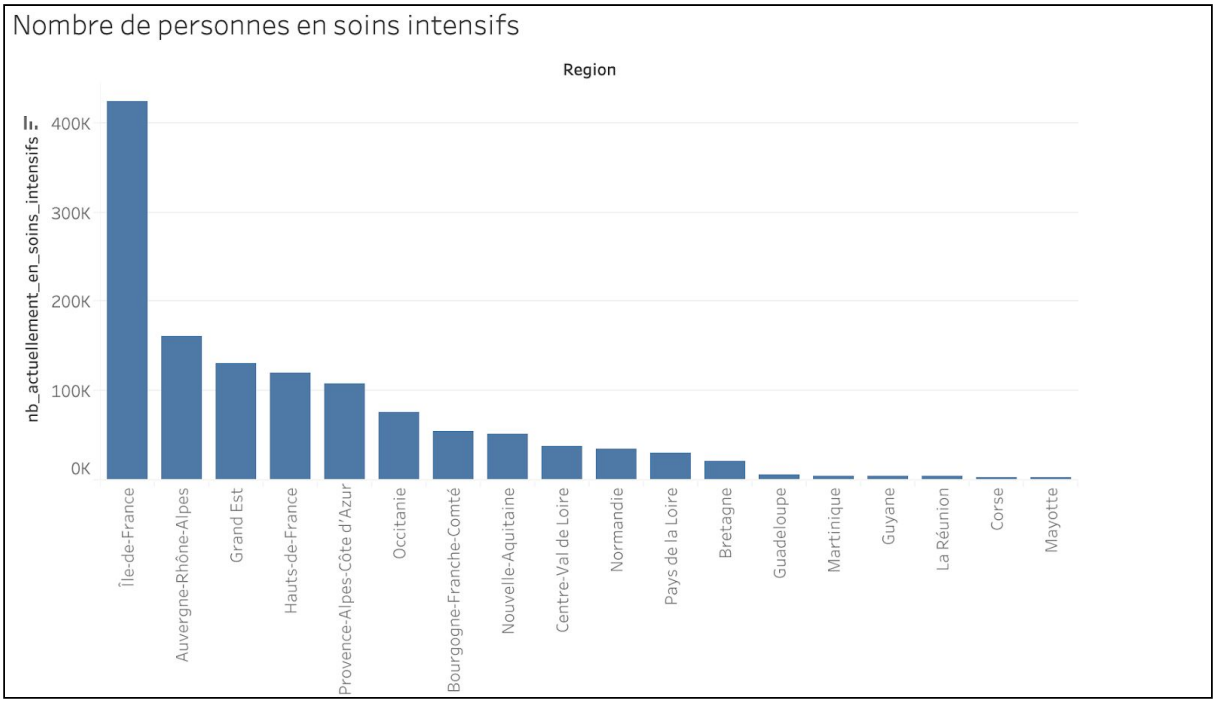
```
insert overwrite table covid_stat_per_region SELECT nom_region AS nom_region,
sum(nb_actuellement_hospitalises ) AS nb_actuellement_hospitalises ,
sum(nb_actuellement_en_soins_intensifs)AS
nb_actuellement_en_soins_intensifs,sum(total_retour_a_domicile)AS
total_retour_a_domicile,sum(total_deces) AS total_deces FROM covid_per_dep GROUP
BY nom_region ;
```

```
hive> select * from covid_stat_per_region;
OK
Auvergne-Rhône-Alpes    1224436 161719  5821065 1319631
Bourgogne-Franche-Comté 388598  54297   2433243 652970
Bretagne        158682  21070   854259  173434
Centre-Val de Loire     297534  37845   1355345 340499
Corse   21225   3720    158934  37610
Grand Est       1047489 130153  7015107 2037291
Guadeloupe      31625   5515    147978  33436
Guyane  26601   4193    573613  20818
Hauts-de-France 849232  119593  4379693 1174459
La Réunion      22019   4119    160518  8516
Martinique      16570   4763    79577   11674
Mayotte 10205   2327    198816  13898
Normandie       293148  35601   1243037 314055
Nouvelle-Aquitaine      318278  51716   1627217 324582
Occitanie       371070  76467   2231597 413882
Pays de la Loire        252767  31081   1448683 308292
Provence-Alpes-Côte d'Azur      723258  108592  4292723 719758
Île-de-France   2996340 424025  17280826        4327335
Time taken: 0.275 seconds, Fetched: 18 row(s)
```

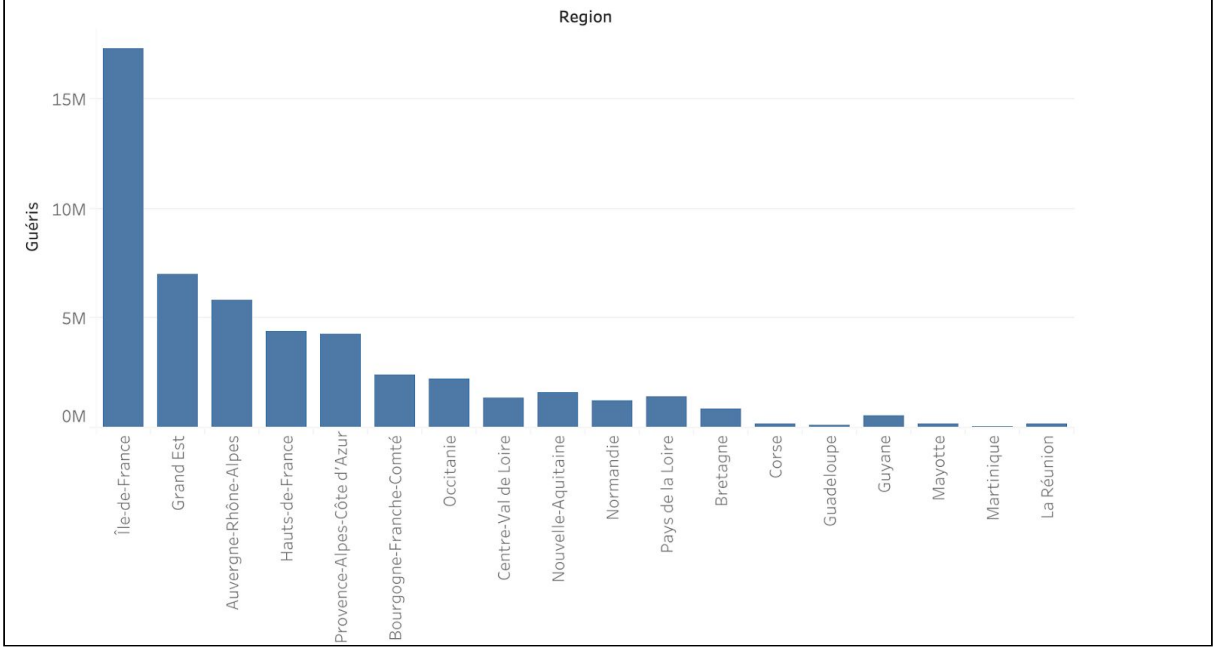
Le résultat est visualisé au niveau de Tableau Desktop:

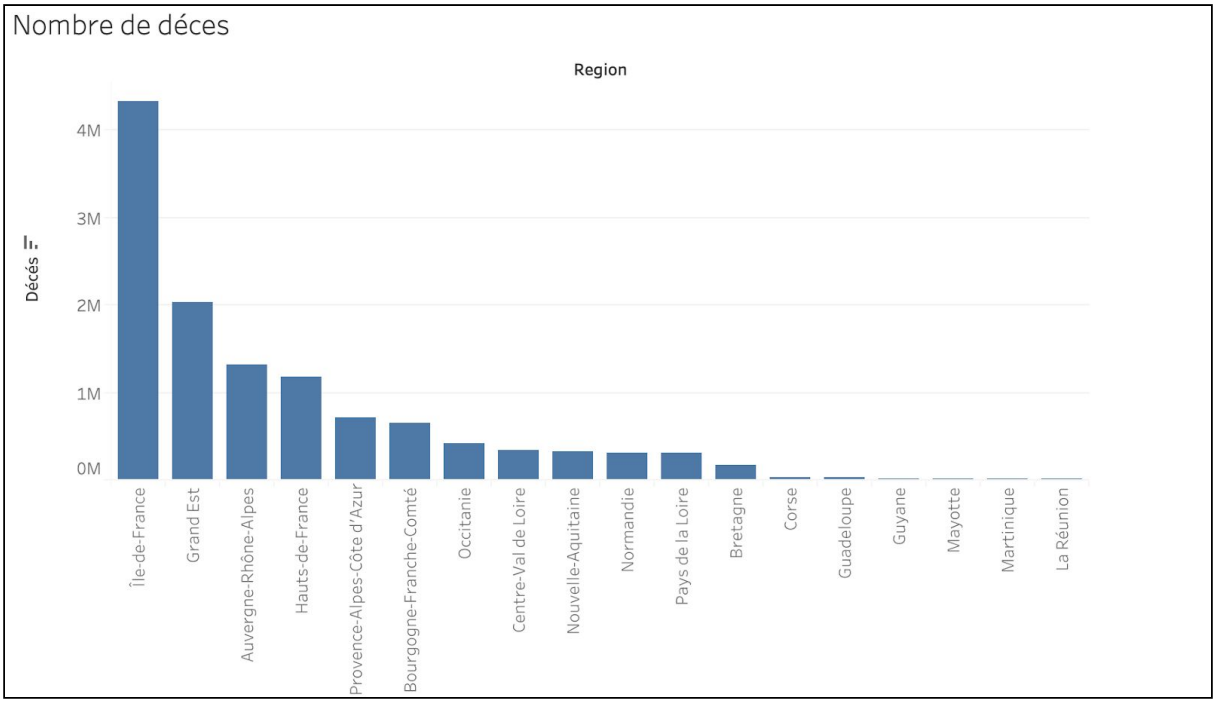


Nombre de personnes en soins intensifs



Nombre de personnes guéries





On créer une table **covid_stat_per_sexe.**

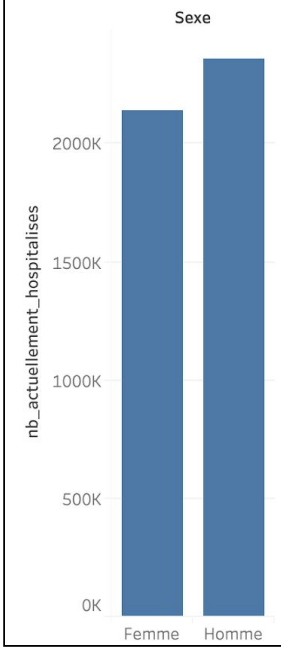
```
CREATE TABLE covid_stat_per_sexe (sexe string,nb_actuellement_hospitalises int,nb_actuellement_en_soins_intensifs int,total_retour_a_domicile int,total_deces int);
```

On calcule le nombre de cas d'hospitalisation,le nombre de personnes en soins intensifs,le nombre de personnes guéries et le nombre de personnes décédées en fonction du sexe ,puis ces résultats sont stockés dans la table **covid_stat_per_sexe**.

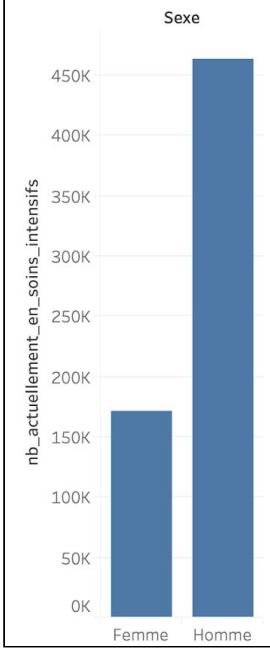
```
insert overwrite table covid_stat_per_sexe SELECT sexe AS sexe ,sum(nb_actuellement_hospitalises) AS nb_actuellement_hospitalises, sum(nb_actuellement_en_soins_intensifs)AS nb_actuellement_en_soins_intensifs,sum(total_retour_a_domicile)AS total_retour_a_domicile,sum(total_deces) AS total_deces FROM covid_per_dep where sexe='Homme' OR sexe='Femme' GROUP BY sexe ;
```

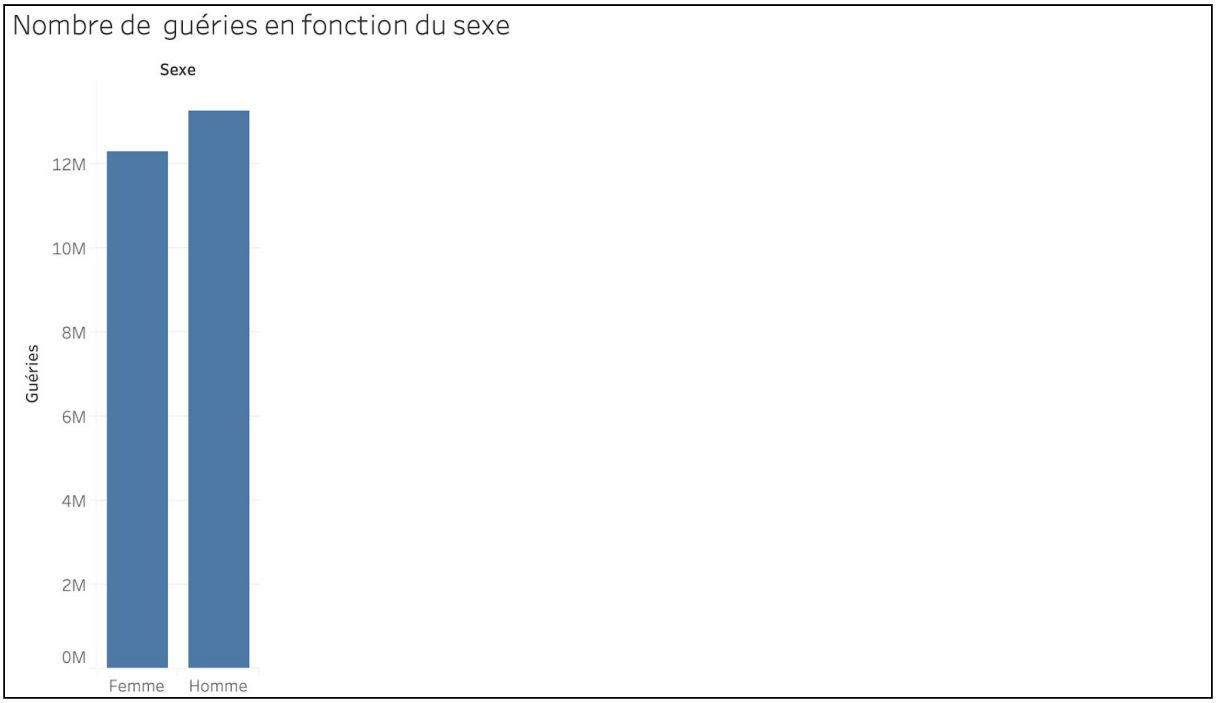
```
hive> select * from covid_stat_per_sexe;
OK
Femme   2142828 171652  12304080        2504202
Homme   2360671 464138  13280462        3598857
Time taken: 0.292 seconds, Fetched: 2 row(s)
hive>
```

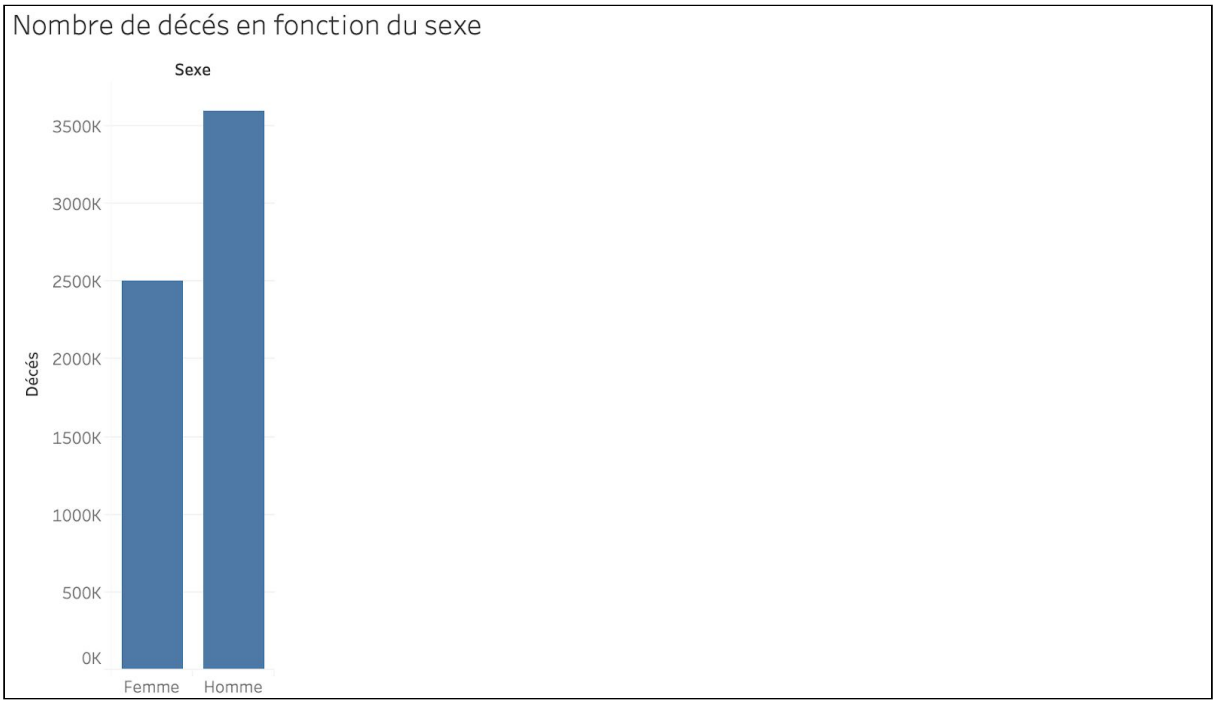
Nombre de cas d'hospitalisation par sexe



Nombre de personnes en soins intensifs par sexe







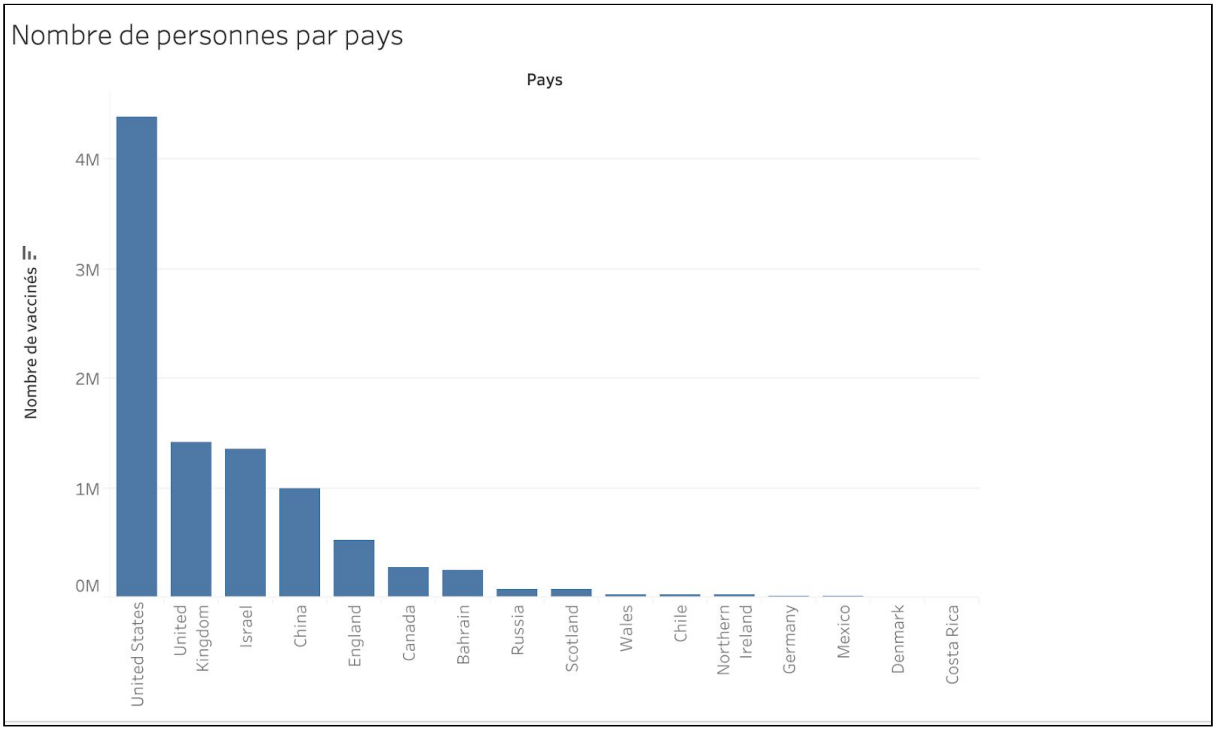
On crée une table **nombre_vaccine.**

```
CREATE TABLE nombre_vaccine(pays varchar(32),nombre_vaccine_par_pays INT)
```

On calcule le nombre de personnes vaccinées pour chaque pays ,puis ces résultats sont stockés dans la table **nombre_vaccine**.

```
insert   overwrite  table  nombre_vaccine   SELECT  pays  p,  sum(total_vaccinations)AS
nombre_vaccine_par_pays   FROM vaccination GROUP BY pays ;
```

```
hive> select * from nombre_vaccine;
OK
Bahrain 244749
Canada  278585
Chile   22574
China   1000000
Costa Rica      55
Denmark 4788
England 521594
Germany 18454
Israel  1361000
Mexico  9748
Northern Ireland        19691
Russia  80500
Scotland        75476
United Kingdom  1416933
United States   4394936
Wales   30447
World   26140449
Time taken: 0.163 seconds, Fetched: 17 row(s)
```



- **Analyse des données locales:**

On récupère les données depuis notre local vers hdfs.
On créer une table appelée **covid**.

```
CREATE EXTERNAL TABLE covid(
   `Date` DATE,
   Pays STRING,
   Infections INT,
   Deces INT,
   Guerisons INT,
   TauxDeces FLOAT,
   TauxGuerison FLOAT,
   TauxInfection FLOAT
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ';'
LOCATION '/user/hive/warehouse/covidtable';
```

Chargement des données sur la table.

```
LOAD DATA INPATH '/user/hive/warehouse/coronavirus.politologue.com_pays_2021-01-08.csv' INTO TABLE covid;
```

On fait la jointure entre les tables **covid** et **tweets** par rapport au nom du pays.

```
SELECT
   cov.*,
   t.`user`.location
    FROM
   (SELECT * from tweets) AS t
     LEFT JOIN covid AS cov ON (cov.Pays =   t.`user`.location);
```

NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	日本 東京
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Galicia, España
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Ahwatukee, AZ
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Strasbourg, France
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Mars, France
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Jakarta,Indonesia.
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Toronto, Ontario
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Finland
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Kaduna, Nigeria
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Paris
2021-01-08	Panama	269091	4321	212656	1.61	79.03	19.37	Panama
2021-01-07	Panama	269091	4321	212656	1.61	79.03	19.37	Panama
2021-01-06	Panama	264956	4283	208620	1.62	78.74	19.65	Panama
2021-01-05	Panama	259770	4238	206087	1.63	79.33	19.03	Panama
2021-01-04	Panama	256230	4197	203688	1.64	79.49	18.87	Panama
2021-01-03	Panama	253736	4140	201816	1.63	79.54	18.83	Panama
2021-01-02	Panama	251764	4103	199694	1.63	79.32	19.05	Panama
2021-01-01	Panama	249733	4064	197928	1.63	79.26	19.12	Panama
2020-12-31	Panama	246790	4022	195138	1.63	79.07	19.3	Panama
2020-12-30	Panama	242744	3975	192601	1.64	79.34	19.02	Panama
2020-12-29	Panama	238279	3933	189764	1.65	79.64	18.71	Panama
2020-12-28	Panama	233705	3892	187552	1.67	80.25	18.08	Panama
2020-12-27	Panama	231357	3840	185966	1.66	80.38	17.96	Panama
2020-12-26	Panama	228724	3799	183522	1.66	80.24	18.1	Panama
2020-12-25	Panama	226660	3756	181749	1.66	80.19	18.16	Panama
2020-12-24	Panama	223674	3715	180045	1.66	80.49	17.84	Panama
2020-12-23	Panama	220261	3664	178140	1.66	80.88	17.46	Panama
2020-12-22	Panama	217202	3632	176428	1.67	81.23	17.1	Panama
2020-12-21	Panama	214038	3597	174951	1.68	81.74	16.58	Panama
2020-12-20	Panama	212339	3566	173508	1.68	81.71	16.61	Panama
2020-12-19	Panama	209584	3527	171745	1.68	81.95	16.37	Panama
2020-12-18	Panama	206310	3504	170170	1.7	82.48	15.82	Panama
2020-12-17	Panama	203295	3481	168238	1.71	82.76	15.53	Panama
2020-12-16	Panama	199947	3439	166600	1.72	83.32	14.96	Panama
2020-12-15	Panama	196987	3411	164855	1.73	83.69	14.58	Panama
2020-12-14	Panama	194619	3382	163334	1.74	83.93	14.34	Panama
2020-12-13	Panama	193007	3356	162105	1.74	83.99	14.27	Panama

NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Andoain
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Somewhere in the World.
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	OWERRI
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Ubatuba, Brasil
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Leicester
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	San Bernardo - Chile
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	أتمشى على أقدامي
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Laguna Beach, Ca
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	England, United Kingdom
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	50.984736,-114.080062
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Saint John, New Brunswick
2021-01-08	France	2701235	66197	174767	2.45	6.47	91.08	France
2021-01-07	France	2701235	66197	174767	2.45	6.47	91.08	France
2021-01-06	France	2701235	66197	174767	2.45	6.47	91.08	France
2021-01-05	France	2676215	65915	173513	2.46	6.48	91.05	France
2021-01-04	France	2655935	65049	172047	2.45	6.48	91.07	France
2021-01-03	France	2651913	64671	171396	2.44	6.46	91.1	France
2021-01-02	France	2639420	64555	171183	2.45	6.49	91.07	France
2021-01-01	France	2636061	64399	170910	2.44	6.48	91.07	France
2020-12-31	France	2721907	64632	194221	2.37	7.14	90.49	France
2020-12-30	France	2701980	64381	193045	2.38	7.14	90.47	France
2020-12-29	France	2675523	64078	191806	2.39	7.17	90.44	France
2020-12-28	France	2664128	63109	190722	2.37	7.16	90.47	France
2020-12-27	France	2661168	62746	189941	2.36	7.14	90.5	France
2020-12-26	France	2652346	62573	189718	2.36	7.15	90.49	France
2020-12-25	France	2649253	62427	189445	2.36	7.15	90.49	France
2020-12-24	France	2628991	62268	188639	2.37	7.18	90.46	France
2020-12-23	France	2607357	61978	187272	2.38	7.18	90.44	France
2020-12-22	France	2592428	61702	186058	2.38	7.18	90.44	France
2020-12-21	France	2580633	60900	184464	2.36	7.15	90.49	France
2020-12-20	France	2574836	60549	183806	2.35	7.14	90.51	France
2020-12-19	France	2562037	60418	183571	2.36	7.17	90.48	France
2020-12-18	France	2544472	60229	182656	2.37	7.18	90.45	France
2020-12-17	France	2528798	59619	181506	2.36	7.18	90.46	France
2020-12-16	France	2510544	59361	180311	2.36	7.18	90.45	France
2020-12-15	France	2492929	59072	179087	2.37	7.18	90.45	France
2020-12-14	France	2481397	58282	177647	2.35	7.16	90.49	France
2020-12-13	France	2478334	57911	176995	2.34	7.14	90.52	France
2020-12-12	France	2466801	57761	176743	2.34	7.16	90.49	France
2020-12-11	France	2452854	57567	175891	2.35	7.17	90.48	France
2020-12-10	France	2439448	56940	174658	2.33	7.16	90.51	France
2020-12-09	France	2425698	56648	173247	2.34	7.14	90.52	France
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Amsterdam
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Valencia, España
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Wahran
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	朝霞市と新座市と志木市
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Northwest Arkansas

8) Conclusion

Nous avons mis en place une architecture d'analyse de données avec Hadoop avec les briques Flume hive,sqoop.Ceci nous a permis d'analyser les données covid provenant de diverses sources(tweeter en temps réel, mysql et en local).
Avec les données tweeter ,on a calculé le top 10 des meilleurs utilisateurs qui ont retweeté sur la covid 19(covid 19, confinement,masques,vaccin) et pour chaque pays le nombre de retweets.

Avec les données d'evolution de cas recuperées depuis Datagouv et opendata on a calculé le nombre de cas d'hospitalisation,le nombre de personnes en soins intensifs,le nombre de personnes guéries et le nombre de personnes décédées en fonction de la région et du sexe.

Pour les données concernant les vaccins,on a calculé le nombre de personnes vaccinées dans chaque pays.
Enfin on a fait la jointure des données de tweets avec celles d'évolution de cas en fonction des pays.