

Université Hassan 1^{er}
Ecole Nationale des Sciences Appliquées de Berrechid
Département de mathématique et informatique
Filière : Ingénierie des Systèmes d'Information et BIG DATA
Module : Bases de données NoSQL et BIG DATA
Semestre : S8

Compte Rendu : Analyse des données Twitter **(Flume & Hive & HDFS)**



Réalisé Par :

ELANSSARI Yassine

EZ-ZAHAR Zakaria

Encadré Par :

Prof. KARIM Lamia

Année Universitaire : 2021/2022

Table des matières

<u>Introduction</u>	1
<u>Problématique</u>	2
I. <u>Outils et logiciels utilisés</u> :	3
II. <u>Source des données</u> :	4
III. <u>Architecture Generale</u> :	5
IV. <u>Réalisation</u> :	6
<u>Conclusion</u>	21

INTRODUCTION

Le micro blogging est devenu aujourd'hui un outil de communication très apprécié pour les utilisateurs d'internet. Twitter, l'un des plus grands sites de médias sociaux reçoit des millions de tweets chaque jour sur une variété de enjeux importants. Les auteurs de ces messages écrivent sur leur vie, partagent leurs opinions sur une variété de sujets et discuter des problèmes actuels. Ces analyses de messages peuvent être utilisées pour la prise de décision dans différents domaines comme le gouvernement, les élections, les affaires et l'examen des produits, etc. aussi des sentiments

L'analyse des sentiments est l'un des domaines d'analyse importants des messages Twitter qui peut être très utile dans la prise de décision. Effectuer une analyse des sentiments sur Twitter est plus délicat que de le faire pour de grandes critiques. C'est parce que les tweets sont très courts (environ 140 caractères seulement) et contiennent généralement des argots, émoticônes, balises de hash tag et autre jargon spécifique à Twitter. À des fins de développement twitter fournit une API de streaming qui permet au développeur d'accéder à 1 % des sur le mot-clé particulier.

Problématique

Comment peut-on analyser les données de twitter en utilisant les outils de big data ?

Notre analyse consiste à répondre aux questions suivantes :

- ✓ Quels étaient les hashtags utilisés et combien de fois chaque hashtag a été utilisé ?
- ✓ Identifier l'hashtag le plus tendance de la journée, Combien de fois le plus tendance des hashtag a été tweeté ?
- ✓ Déterminer le score de chaque tweet posté ? Identifiez si le tweet avait un sentiment positif ou négatif ?

Outils et logiciels utilisés



Apache Flume est un logiciel de la fondation Apache destiné à la collecte et à l'analyse de fichiers de log.



HDFS est un système de fichier distribué permettant de stocker et de récupérer des fichiers volumineuses en un temps record.



Apache Hive est une infrastructure d'entrepôt de données intégrée sur Hadoop permettant l'analyse, le requêtage via un langage proche syntaxiquement de SQL ainsi que la synthèse de données.

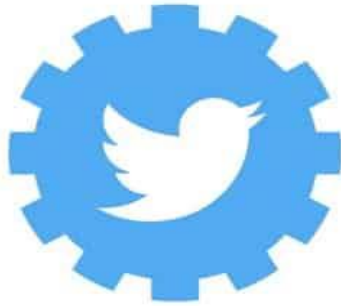


OS : UBUNTU 18.04



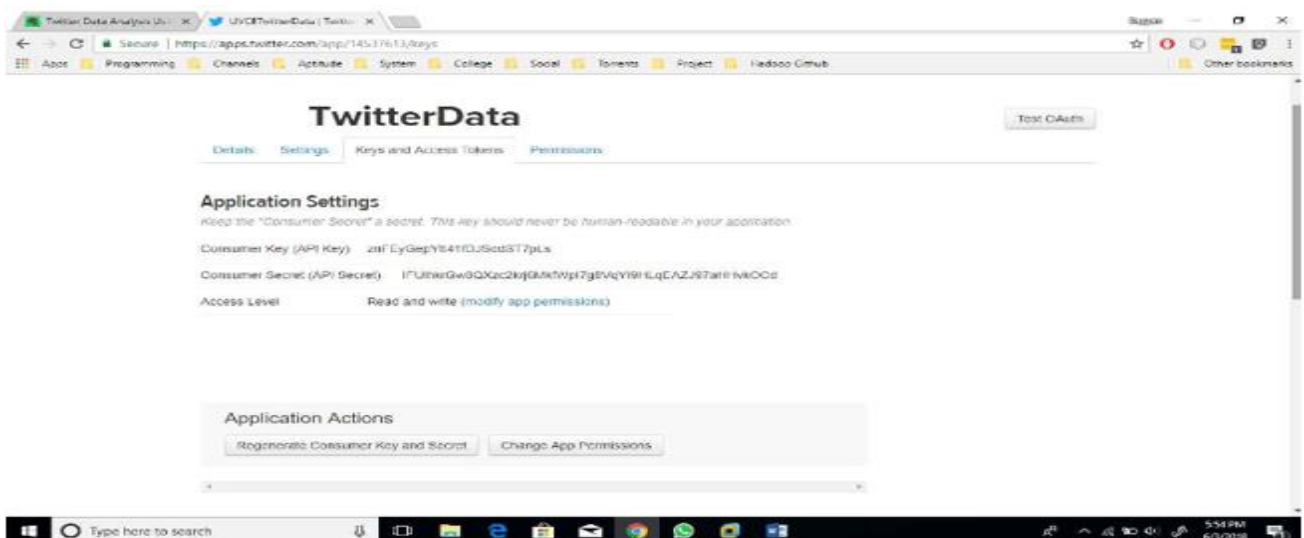
JDK : 8

Source des Données

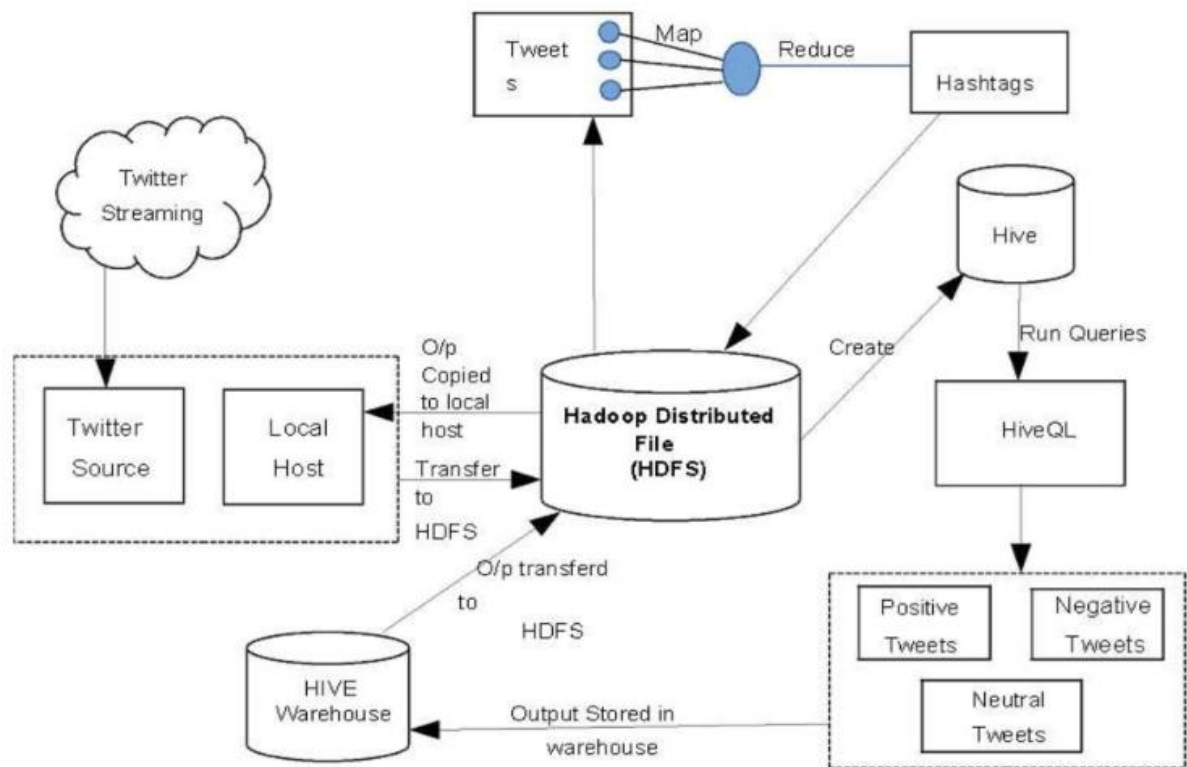


Twitter API

Le compte de développeur Twitter peut être créé sur la page des applications des développeurs Twitter. Dans cette page nous devons fournir une page de compte Twitter valide dans le champ du site Web à partir duquel nous devons obtenir données en continu. Si nous fournissons des détails valides sur cette page, notre application sera créée comme indiqué dans les captures d'écran ci-dessous.



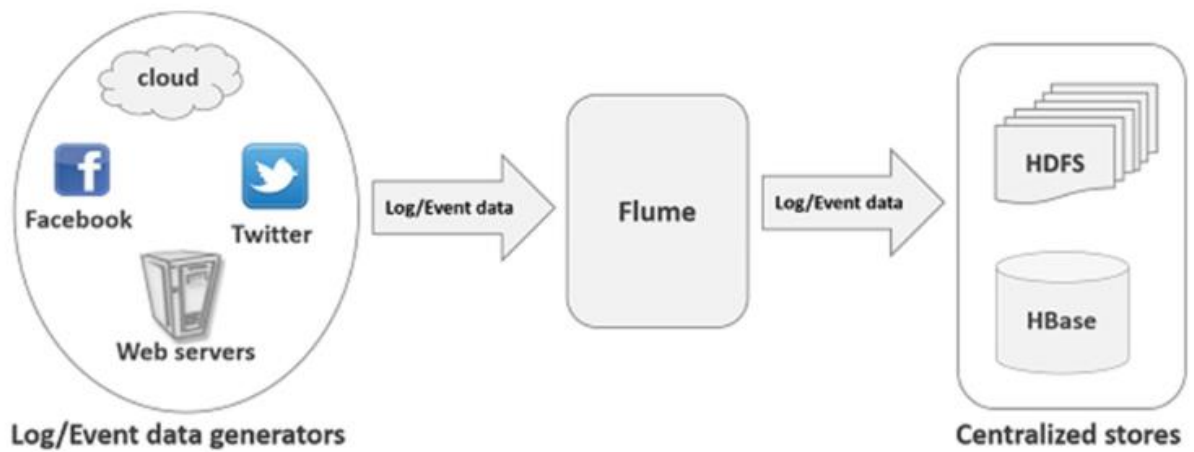
Architecture Générale



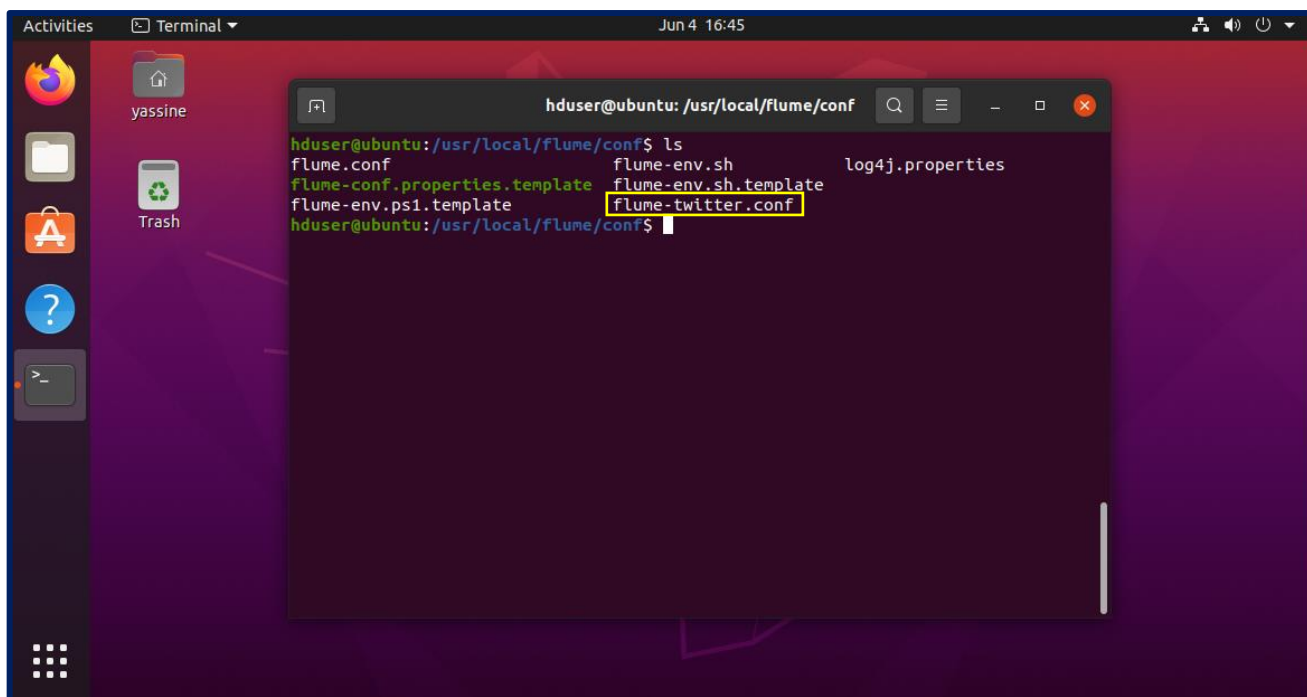
Réalisation

✚ Generation des données avec Apache Flume :

Pour faire l'analyse des données Twitter, les données sont collectées en utilisant FLUME dans HDFS local.



Pour cela il faut crée un fichier de configuration de flume dans le répertoire conf :



Flume-twitter.conf :

```
GNU nano 4.8 flume-twitter.conf
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = znFEyGepYtt41fDJSodST7pLs
TwitterAgent.sources.Twitter.consumerSecret = IFULhkrGw8QXzc2krj6MkfWpI7g8VqYI9HLqEAZJ97aHHvkOCd
TwitterAgent.sources.Twitter.accessToken = 1497149762-w4Dcg1lOEef8dk3mHNBHCaqEbVP3ewW2Jtdg8xC
TwitterAgent.sources.Twitter.accessTokenSecret = LnT7MTLiPM69ZAV0qM7CINNg3kqgURdC9SmxZlLYfFSVm
TwitterAgent.sources.Twitter.keywords =narendramodi,aap,bjp,Arvindkejrival,shivsena,congress

TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/user/hduser/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Après que hdfs est bien démarré on exécute ce fichier dans le répertoire bin de flume :

- flume-ng agent -n TwitterAgent -c ./conf/ -f /usr/local/flume/conf/flume-twitter.conf

```
Activities Terminal Jun 4 17:21
hduser@ubuntu: /usr/local/flume/bin

hduser@ubuntu: /usr/local/flume/conf
hduser@ubuntu: /usr/local/flume/bin$ flume-ng agent -n TwitterAgent -c ./conf/ -f /usr/local/flume/conf/flume-twitter.conf
Info: Including Hadoop libraries found via (/usr/local/hadoop/bin/hadoop) for HDFS access
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
+ exec /usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java -Xmx20m -cp './conf:/usr/local/flume/lib/*:/usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/local/hadoop/share/hadoop/common/*:/usr/local/hadoop/share/hadoop/hdfs:/usr/local/hadoop/share/hadoop/hdfs/lib/*:/usr/local/hadoop/share/hadoop/hdfs/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*:/usr/local/hadoop/share/hadoop/mapreduce/lib/*:/usr/local/hadoop/share/hadoop/mapreduce/*:/contrib/capacity-scheduler/*.jar:/usr/local/hive/lib/*' -Djava.library.path=:/usr/local/hadoop/lib org.apache.flume.node.Application -n TwitterAgent -f /usr/local/flume/conf/flume-twitter.conf
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/flume/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
22/06/04 17:20:54 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
22/06/04 17:20:54 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/usr/local/flume/conf/flume-twitter.conf
22/06/04 17:20:54 INFO conf.FlumeConfiguration: Processing:HDFS
22/06/04 17:20:54 INFO conf.FlumeConfiguration: Processing:HDFS
```

Un flux de données json sera stocker dans le répertoire de hdfs indiquée dans fichier de flume :

Browsing HDFS

localhost:50070/explorer.html#/user/hduser/flume/tweets

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/hduser/flume/tweets Go!

Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	hduser	supergroup	432.85 KB	Jun 04 17:01	1	128 MB	Twitter.json

Showing 1 to 1 of 1 entries Previous 1 Next

Contenu du fichier :

Activities Text Editor Jun 4 17:46

Twitter.json /home/hduser

```
1 [{"retweet_count": 7, "created_at": "Fri Jul 29 12:59:31 +0000 2016", "text": "It is being reported by virtually everyone, and is a fact, that the media pile on against me is the worst in American political history!", "id": 641766061380228000, "source": "href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>", "in_reply_to_screen_name": "Ethel C. Adams", "user": {"location": "Michigan", "id": 4020825913, "id_str": 4020825913, "name": "Ethel", "screen_name": "Ethel_", "geo_enabled": "FALSE", "lang": "ru", "protected": "FALSE", "verified": "FALSE", "followers_count": 1037, "friends_count": 27, "listed_count": 4, "favourites_count": 990, "statuses_count": 188, "profile_background_color": "0", "contributors": "", "is_quote_status": "FALSE", "entities": {"user_mentions": [{"screen_name": "Ethel C. Adams", "name": "try", "id": 4447341869 }]}]}]
2 [{"retweet_count": 2, "created_at": "Fri Jul 29 12:59:31 +0000 2016", "text": "I am now in Texas doing a big fundraiser for the Republican Party and a @FoxNews Special on the BORDER and with victims of border crime!", "id": 872283722010901000, "source": "href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>", "in_reply_to_screen_name": "Barry C. Bowles", "user": {"location": "Michigan", "id": 2703669882, "id_str": 2703669882, "name": "Barry", "screen_name": "Barry_", "geo_enabled": "FALSE", "lang": "en", "protected": "FALSE", "verified": "FALSE", "followers_count": 487, "friends_count": 490, "listed_count": 8, "favourites_count": 5106, "statuses_count": 2790, "profile_background_color": "0", "contributors": "", "is_quote_status": "FALSE", "entities": {"user_mentions": [{"screen_name": "Barry C. Bowles", "name": "try", "id": 5398181484 }]}]}]
3 [{"retweet_count": 8, "created_at": "Fri Jul 29 12:59:51 +0000 2016", "text": "The @WashingtonPost quickly put together a hit job book on me comprised of copies of some of their inaccurate stories. Don't buy, boring!", "id": 534262624541602000, "source": "href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>", "in_reply_to_screen_name": "Dan C. Snyder", "user": {"location": "Oregon", "id": 4533220285, "id_str": 4533220285, "name": "Dan", "screen_name": "Dan_", "geo_enabled": "FALSE", "lang": "en", "protected": "FALSE", "verified": "FALSE", "followers_count": 1430, "friends_count": 47, "listed_count": 9, "favourites_count": 3111, "statuses_count": 3614, "profile_background_color": "0", "contributors": "", "is_quote_status": "TRUE", "entities": {"user_mentions": [{"screen_name": "Dan C. Snyder", "name": "try", "id": 2319468009 }]}]}]
4 [{"retweet_count": 6, "created_at": "Fri Jul 29 12:59:31 +0000 2016", "text": ".@AnnCoulter's new book, 'In Trump We Trust, comes out tomorrow. People are saying it's terrific - knowing Ann I am sure it is!", "id": 791731982444207000, "source": "href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>", "in_reply_to_screen_name": "Nina C. Buchanan", "user": {"location": "New York", "id": 3558058220, "id_str": 3558058220, "name": "Nina C. Buchanan", "screen_name": "NinaCBuchanan", "geo_enabled": "FALSE", "lang": "en", "protected": "FALSE", "verified": "FALSE", "followers_count": 1430, "friends_count": 47, "listed_count": 9, "favourites_count": 3111, "statuses_count": 3614, "profile_background_color": "0", "contributors": "", "is_quote_status": "TRUE", "entities": {"user_mentions": [{"screen_name": "Dan C. Snyder", "name": "try", "id": 2319468009 }]}]}]

Bracket match found on line: 1 JSON Tab Width: 8 Ln 1, Col 1 INS
```

🔗 Analyse des données :

Pour analyser les tweets nous sommes besoins d'un dictionnaire avec notation pour les mots individuels.

Dictionary.txt :

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/hduser/data Go!

Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	hduser	supergroup	27.44 KB	Jun 04 17:00	1	128 MB	Dictionary.txt

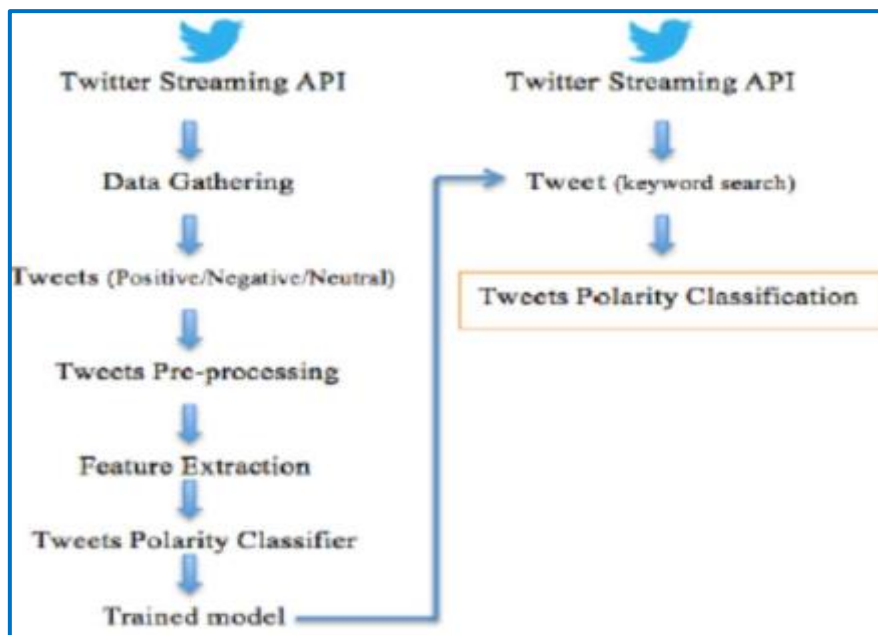
Showing 1 to 1 of 1 entries

Previous 1 Next

Dictionary.txt		
/home/hduser		
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2
13	aboard	1
14	absentee	-1
15	absentees	-1
16	absolve	2
17	absolved	2
18	absolves	2
19	absolving	2
20	absorbed	1
21	abuse	-3
22	abused	-3
23	abuses	-3
24	abusive	-3
25	accept	1
26	accepted	1
27	accepting	1
28	accepts	1
29	accident	-2

Classification :

À la fin, le système classera les données Twitter en positif, négatif, neutre à l'aide de Dictionnaire de données.



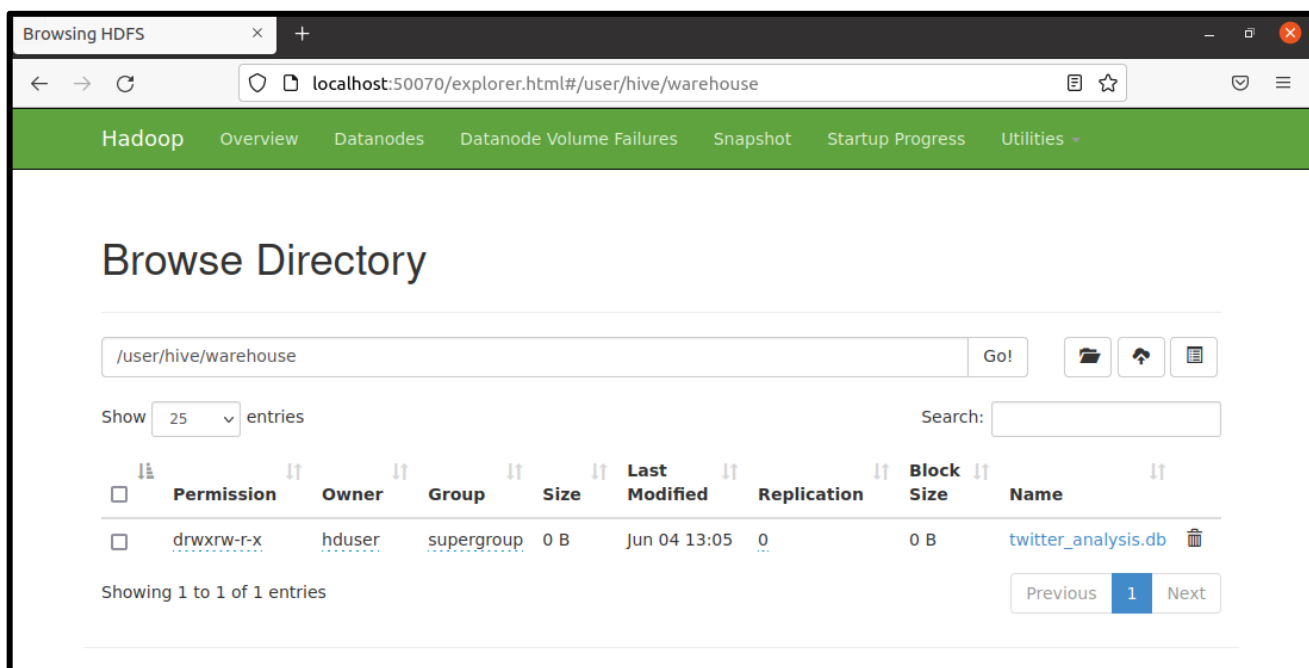
1- Lancement de Hive :

```
Activities Terminal Jun 4 18:27
hduser@ubuntu: /usr/local/hive/bin

hduser@ubuntu: /usr/local/hive/bin$ ls
beeline derby.log hive hive-config.cmd hiveserver2 hplsql.cmd metatool
beeline.cmd ext hive.cmd hive-config.sh hplsql metastore_db schematool

hduser@ubuntu: /usr/local/hive/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-2.1.0.jar!/hive-log4j2.properties
A sync: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show databases ;
OK
default
twitter_analysis
Time taken: 1.237 seconds, Fetched: 2 row(s)
hive> use twitter_analysis;
OK
Time taken: 0.029 seconds
hive>
```



2- Chargement de fichier Dictionary.txt dans Hive :

Pour cela il faut d'abord crée un tableau hive de mêmes champs que celle de fichier :

```
hive> CREATE TABLE dictionary (word string, rating int)
> row format delimited
> fields terminated by '\t';
OK
Time taken: 0.864 seconds
hive> █
```

Ensuite en charge le dictionnaire depuis hdfs dans le tableau hive :

```
hive> LOAD DATA INPATH '/user/hduser/data/Dictionary.txt' INTO TABLE dictionary;
Loading data to table twitter_analysis.dictionary
OK
Time taken: 0.71 seconds
hive>
```

3- chargement de fichier json de twitter :

```
hive> CREATE TABLE twitter_json ( json string );
OK
Time taken: 0.072 seconds
hive> LOAD DATA INPATH '/user/hduser/flume/tweets/Twitter.json' INTO TABLE twitter_json;
Loading data to table twitter_analysis.twitter_json
OK
Time taken: 0.39 seconds
hive> █
```

Le premier fichier entier est chargé dans la table twitter_json en tant que chaîne, puis analysé à l'aide de get_json_object :


```

hive> CREATE TABLE twitter as
> select get_json_object(twitter_json.json, '$.retweet_count') as retweet_count,
>        unix_timestamp(get_json_object(twitter_json.json, '$.created_at'),
>        "EEE MMM d HH:mm:ss Z yyyy") as created_at,
>        get_json_object(twitter_json.json, '$.text') as text,
>        get_json_object(twitter_json.json, '$.id') as id,
>        get_json_object(twitter_json.json, '$.source') as source,
>        get_json_object(twitter_json.json, '$.in_reply_to_screen_name') as in_reply_to_screen_name,
>        get_json_object(twitter_json.json, '$.user.location') as location,
>        get_json_object(twitter_json.json, '$.user.id') as u_id,
>        get_json_object(twitter_json.json, '$.user.id_str') as id_str,
>        get_json_object(twitter_json.json, '$.user.name') as name,
>        get_json_object(twitter_json.json, '$.user.screen_name') as screen_name,
>        get_json_object(twitter_json.json, '$.user.geo_enabled') as geo_enabled,
>        get_json_object(twitter_json.json, '$.user.lang') as lang,
>        get_json_object(twitter_json.json, '$.user.protected') as protected,
>        get_json_object(twitter_json.json, '$.user.verified') as verified,
>        get_json_object(twitter_json.json, '$.user.followers_count') as followers_count,
>        get_json_object(twitter_json.json, '$.user.friends_count') as friends_count,
>        get_json_object(twitter_json.json, '$.user.listed_count') as listed_count,
>        get_json_object(twitter_json.json, '$.user.favourites_count') as favourites_count,
>        get_json_object(twitter_json.json, '$.user.statuses_count') as statuses_count,
>        get_json_object(twitter_json.json, '$.user.profile_background_color') as
>        profile_background_color,
>        get_json_object(twitter_json.json, '$.contributors') as contributors,
>        get_json_object(twitter_json.json, '$.is_quote_status') as is_quote_status,
>        get_json_object(twitter_json.json, '$.entities.user_mentions.screen_name') as
>        user_mention_screen_name,
>        get_json_object(twitter_json.json, '$.entities.user_mentions.name') as user_mention_name,
>        get_json_object(twitter_json.json, '$.entities.user_mentions.id') as user_mention_id
> from twitter_json;

```

Dans la table twitter les données sont stockées sous forme de chaîne afin de convertir la colonne created_at en colonne de date requise. unix_timestamp est utilisé.

La fonction unix_timestamp() :

S'il est appelé sans argument, renvoie un horodatage Unix (secondes depuis '1970-01-01 00:00:00' UTC) sous la forme d'un entier non signé. Si UNIX_TIMESTAMP() est appelé avec un argument de date, il renvoie la valeur de l'argument en secondes depuis '1970-01-01 00:00:00' UTC. date peut être une DATEchaîne, une DATETIMEchaîne, un TIMESTAMPou un nombre au format AAAAMMMJJ ou AAAAMMMJJ. Le serveur interprète la date comme une valeur dans le fuseau horaire actuel et la convertit en une valeur interne en UTC. Les clients peuvent définir leur fuseau horaire comme décrit dans Fuseaux horaires.

4- Quels sont les hashtags utilisés dans le fichier et combien de fois chaque hashtag a été utilisé ?

```

hive> CREATE TABLE hashtag as
> select hashtag from twitter
> LATERAL VIEW explode(split(text, ' ')) text_tweet as hashtag
> where hashtag like '#%';

```

```
hive> Select * from hashtag;
OK
#tcdisrupt
#tcdisrupt
#TeamTrump
#TrumpPence16
#NotoTrump
#MAGA
#TrumpPence16
#TrumpPence16
#MAGA
#AlwaysTrump
#WheresHillary?
#TrumpPence16
#ImWithYou
#TrumpTrain
#CrookedHillary
#ThrowbackThursday
#ImWithYou
#Hillary
#Obamacare,
#ImWithYouhttps://t.co/Eph6qy7zyB
#MakeAmericaSafeAgain
#ImWithYou
#MakeAmericaGreatAgain
#ImWithYou
#ImWithYou
#LawandOrder
#ImWithYouVideo:
#LawandOrder
#ImWithYouTranscript:
#MakeAmericaGreatAgain
#Obamacare
#AmericaMasked
```

- ✓ La deuxième partie de la question est le nombre de hashtags, afin d'obtenir que la table hashtag_count soit créée pour stocker les hashtags et leur nombre.

```
hive> CREATE TABLE hashtag_count as
> select hashtag, count(hashtag) as hashtag_count from hashtag
> GROUP BY hashtag
> ORDER BY hashtag_count;
```

La requête est utilisée pour valider la création de la table. La table Hashtag_count contient tous les hashtags et leur nombre.

Lors de la création de la table ORDER BY hashtag_count DESC ; peut être utilisé pour imprimer les résultats dans l'ordre décroissant au lieu de l'utiliser pendant l'instruction select.

Les hashtags les plus couramment utilisés sont :


```
hduser@ubuntu: /usr/local/hive/bin
hive> Select * from hashtag_count ORDER BY hashtag_count DESC;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hduser_20220604191850_6983e7ce-488a-4c42-bc84-0f90d2d80d1a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=<number>
Job running in-process (local Hadoop)
2022-06-04 19:18:52,455 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local164110658_0005
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 1417330 HDFS Write: 394148 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
#InWithYou      23
#MAGA            18
#TrumpPence16   16
#CrookedHillary 16
#MakeAmericaGreatAgain 14
#AmericaFirst   13
#Trump2016       12
#RNCinCLE        10
#NeverHillary    5
#MakeAmericaSafeAgain 4
#Trump           4
#MakeAmericaGreatAgain! 4
#LESM            3
#Hillary         3
#2A              2
#GOPConvention#AmericaFirst 2
```

5- Identifiez le hashtag le plus tendance de la journée. Combien de fois le hashtag le plus tendance a-t-il été tweeté

```
hive> CREATE TABLE date_ as
> select id, name as user_name, text, from_unixtime(created_at,'yyyy-MM-dd') as date_
> from twitter;
```

```
hive> Select * from date_;
OK
641766061380228000 Ethel It is being reported by virtually everyone, and is a fact,that the media pile on against me is the worst in American
ical history! 2016-07-29
872283722010901000 Barry I am now in Texas doing a big fundraiser for the Republican Party and a @FoxNews Special on the BORDER and with vict
border crime! 2016-07-29
534262624541602000 Dan The @WashingtonPost quickly put together a hit job book on me comprised of copies of some of their inaccurate storie
't buy, boring! 2016-07-29
791731982444207000 Nina .@AnnCoulter's new book, 'In Trump We Trust, comes out tomorrow. People are saying it's terrific - knowing Ann I am
t is! 2016-07-29
516680352480685000 Erica Just leaving Akron, Ohio, after a packed rally. Amazing people! Going now to Texas. 2011-09-10
744884451423099000 Cameron Great meeting with active & retired law enforcement officers- at the Fraternal Order of Police lodge in Akron, O
https://t.co/EUwhDC644R 2009-03-18
781654794440300000 Max Statement on Clinton Foundation 2009-03-18
488845101475781000 Marion Will be interviewed on @foxandfriends at 8:30 A.M. Eastern. ENJOY! 2009-03-18
366912068027820000 Heather Some day, when things calm down, I'll tell the real story of @JoeNBC and his very insecure long-time girlfriend, @mc
ika. Two clowns! 2009-03-18
481044586874523000 Ellen Tried watching low-rated @Morning_Joe this morning, unwatchable! @morningmika is off the wall, a neurotic and not v
ight mess! 2016-08-23
268653953230593000 George @realbill2016: @realDonaldTrump @Brainykid2010 @shl Trump leading LA Times poll 2016-08-23
240253756777193000 Kathryn @twitter meets @seepicturely accepted #tcdisrupt cc.@boscomonkey @episod 2016-08-23
614111043815355000 Sheryl @twitter meets @seepicturely accepted #tcdisrupt cc.@boscomonkey @episod 2016-08-23
211014694431747000 Craig I am self funding my campaign and only work for YOU, the American people!#Trump2016 2016-01-25
522735212763929000 Joan @Brainykid2010: @shl @realDonaldTrump The ad was actually very good! 2016-07-29
591002608868294000 Anne @55Lidsville: #TeamTrump @KellyannePolls You need to show the crowds at the rallies use Periscope! Show HC's 139 YT
rs vs DT 38K 2016-07-29
```

Création d'un tableau avec hashtags et date :

Ici, deux colonnes sont utilisées dans la colonne de date de la table date_ et le texte est divisé à l'aide de fonctions de split pour obtenir uniquement les hashtags du texte.

```
hive> CREATE TABLE hashtag_date as
> select hashtag, date_ from date_
> LATERAL VIEW explode(split(text,' ')) text_tweet as hashtag
> where hashtag like '#%';
```

```
hive> Select * from hashtag_date;
OK
#tcdisrupt      2016-08-23
#tcdisrupt      2016-08-23
#TeamTrump      2016-07-29
#TrumpPence16   2016-07-29
#NotoTrump      2011-09-10
#MAGA           2016-07-29
#TrumpPence16   2016-07-29
#TrumpPence16   2011-09-10
#MAGA           2011-09-10
#AlwaysTrump    2011-09-10
#WheresHillary? 2011-09-10
#TrumpPence16   2016-07-11
#ImWithYou      2016-07-11
#TrumpTrain     2016-07-11
#CrookedHillary 2016-07-11
#ThrowbackThursday 2016-07-11
#ImWithYou      2011-09-10
#Hillary        2011-09-10
#Obamacare,     2011-09-10
#ImWithYouhttps://t.co/Eph6qy7zy8 2011-09-10
#MakeAmericaSafeAgain 2011-09-10
#ImWithYou      2011-09-10
#MakeAmericaGreatAgain 2011-09-10
#ImWithYou      2011-09-10
#ImWithYou      2011-09-10
#LawandOrder    2011-09-10
#ImWithYouVideo: 2011-09-10
#LawandOrder    2011-09-10
#ImWithYouTranscript: 2011-09-10
#MakeAmericaGreatAgain 2011-09-10
```

Création d'un tableau avec les hashtags tendances par date :

```
hive>
hive> CREATE TABLE trending_hashtag_byday as
> select date_, hashtag,count(hashtag) as hashtag_count from hashtag_date
> GROUP BY date_, hashtag
> ORDER BY hashtag_count DESC;
```

```
hive> Select * from trending_hashtag_byday;
OK
2011-09-10      #MakeAmericaGreatAgain 13
2011-09-10      #AmericaFirst 12
2011-09-10      #ImWithYou 12
2011-09-10      #MAGA 12
2011-09-10      #TrumpPence16 10
2011-09-10      #Trump2016 9
2011-09-10      #RNCinCLE 8
2011-09-10      #CrookedHillary 4
2011-09-10      #LESM 3
2011-09-10      #MakeAmericaSafeAgain 3
2011-09-10      #Trump 3
2015-12-22      #ImWithYou 3
2015-12-18      #ImWithYou 2
2016-07-08      #NeverHillary 2
2016-07-24      #CrookedHillary 2
2011-09-10      #MakeAmericaGreatAgain! 2
2011-09-10      #NeverHillary 2
2016-03-20      #CrookedHillary 2
2016-05-24      #ImWithYou 2
2015-12-18      #CrookedHillary 2
2009-03-18      #CrookedHillary 2
2016-01-29      #ImWithYou 2
2016-01-19      #RNCinCLE 2
2016-08-23      #tcdisrupt 2
2011-09-10      #MAGATickets 2
2016-05-21      #MAGA 2
2016-05-24      #Trump2016 2
2011-09-10      #GOPConvention#AmericaFirst 2
2016-07-29      #TrumpPence16 2
2011-09-10      #LawandOrder 2
```

Les hashtags les plus populaires sont # MakeAmericaGreatAgain, #ImWithYou, #MAGA, #AmericaFirst.

6- Déterminer le score de chaque tweet posté ? Identifiez si le tweet avait un sentiment positif ou négatif ?

Approche pour réaliser l'analyse des sentiments ci-dessus, c'est nécessaire :

1. Pour avoir un tableau avec toutes les colonnes requises de twitter
2. Pour créer une jointure entre la table Twitter et le dictionnaire, faites correspondre tous les mots du dictionnaire avec le texte des tweets.
3. Trouvez la note de tous les mots correspondants et fournissez les résultats de l'analyse des sentiments.

Création d'un tableau avec toutes les colonnes requises à partir des données Twitter :

```
hive> CREATE TABLE split_text as
> select id as tweet_id, user_name, tweet_words, date_ from date_
> LATERAL VIEW explode(split(text, ' ')) text_tweet as tweet_words;
```

```
hive> select * from split_text ;
OK
641766061380228000      Ethel      It      2016-07-29
641766061380228000      Ethel      is      2016-07-29
641766061380228000      Ethel      being   2016-07-29
641766061380228000      Ethel      reported 2016-07-29
641766061380228000      Ethel      by      2016-07-29
641766061380228000      Ethel      virtually 2016-07-29
641766061380228000      Ethel      everyone, 2016-07-29
641766061380228000      Ethel      and      2016-07-29
641766061380228000      Ethel      is      2016-07-29
641766061380228000      Ethel      a      2016-07-29
641766061380228000      Ethel      fact,that 2016-07-29
641766061380228000      Ethel      the      2016-07-29
641766061380228000      Ethel      media   2016-07-29
641766061380228000      Ethel      pile    2016-07-29
641766061380228000      Ethel      on      2016-07-29
```

Jointure entre la table split_text et la table dictionnaire :

```
hive> CREATE TABLE join_tweet_dict as
> select st.tweet_id, st.user_name, st.date_, st.tweet_words, d.rating
> FROM split_text st LEFT OUTER JOIN dictionary d on (st.tweet_words = d.word);
```

```
664259398231896000      Edith      2011-09-10      JOBS,      NULL
664259398231896000      Edith      2011-09-10      JOBS,      NULL
664259398231896000      Edith      2011-09-10      JOBS!      NULL
664259398231896000      Edith      2011-09-10      Crooked    NULL
664259398231896000      Edith      2011-09-10      Hillary    NULL
664259398231896000      Edith      2011-09-10      will       NULL
664259398231896000      Edith      2011-09-10      sell       NULL
664259398231896000      Edith      2011-09-10      us         NULL
664259398231896000      Edith      2011-09-10      out,       NULL
664259398231896000      Edith      2011-09-10      just       NULL
664259398231896000      Edith      2011-09-10      like      2
664259398231896000      Edith      2011-09-10      her        NULL
664259398231896000      Edith      2011-09-10      husband    NULL
664259398231896000      Edith      2011-09-10      did        NULL
664259398231896000      Edith      2011-09-10      with       NULL
664259398231896000      Edith      2011-09-10      NAFTA.     NULL
624787394042523000      Ricky      2011-09-10      Another    NULL
624787394042523000      Ricky      2011-09-10      new        NULL
624787394042523000      Ricky      2011-09-10      poll.      NULL
624787394042523000      Ricky      2011-09-10      Thank      NULL
624787394042523000      Ricky      2011-09-10      you        NULL
624787394042523000      Ricky      2011-09-10      for        NULL
624787394042523000      Ricky      2011-09-10      your       NULL
624787394042523000      Ricky      2011-09-10      support!   NULL      NULL
624787394042523000      Ricky      2011-09-10      Join       NULL
624787394042523000      Ricky      2011-09-10      the        NULL
624787394042523000      Ricky      2011-09-10      MOVEMENT  NULL      NULL
624787394042523000      Ricky      2011-09-10      today!     NULL
624787394042523000      Ricky      2011-09-10      #ImWithYou NULL      NULL
624787394042523000      Ricky      2011-09-10      NULL
680459189437637000      Diane      2016-04-04      Great      NULL
680459189437637000      Diane      2016-04-04      new        NULL
680459189437637000      Diane      2016-04-04      poll-      NULL
680459189437637000      Diane      2016-04-04      thank      2
```

Création d'un tableau qui a un score - sur toutes les notes et fournissant des résultats d'analyse des sentiments :

```
hive> CREATE TABLE score as
> select tweet_id, user_name, date_, sum(rating) as score
> FROM join_tweet_dict
> GROUP BY tweet_id, user_name, date_;
```

Le tableau des scores contient la somme des notes des données de tweet.

103386304747066000	Marion	2011-09-10	I	NULL	
103386304747066000	Marion	2011-09-10	am	NULL	
103386304747066000	Marion	2011-09-10	truly	NULL	
103386304747066000	Marion	2011-09-10	enjoying		2
103386304747066000	Marion	2011-09-10	myself	NULL	
103386304747066000	Marion	2011-09-10	while	NULL	
103386304747066000	Marion	2011-09-10	running	NULL	
103386304747066000	Marion	2011-09-10	for	NULL	
103386304747066000	Marion	2011-09-10	president.		NULL
103386304747066000	Marion	2011-09-10	The	NULL	
103386304747066000	Marion	2011-09-10	people	NULL	
103386304747066000	Marion	2011-09-10	of	NULL	
103386304747066000	Marion	2011-09-10	our	NULL	
103386304747066000	Marion	2011-09-10	country	NULL	
103386304747066000	Marion	2011-09-10	are	NULL	
103386304747066000	Marion	2011-09-10	amazing	4	
103386304747066000	Marion	2011-09-10	-	NULL	
103386304747066000	Marion	2011-09-10	great	3	
103386304747066000	Marion	2011-09-10	numbers	NULL	
103386304747066000	Marion	2011-09-10	on	NULL	
103386304747066000	Marion	2011-09-10	November		NULL
103386304747066000	Marion	2011-09-10	8th!	NULL	

```
hive> CREATE table sentiment_analysis as
> SELECT tweet_id, user_name, date_, score,
> CASE WHEN score > 0 THEN 'Positive'
> WHEN score < 0 THEN 'Negative'
> ELSE 'Neutral'
> END as tweets_sentiment_analysis FROM score;
```

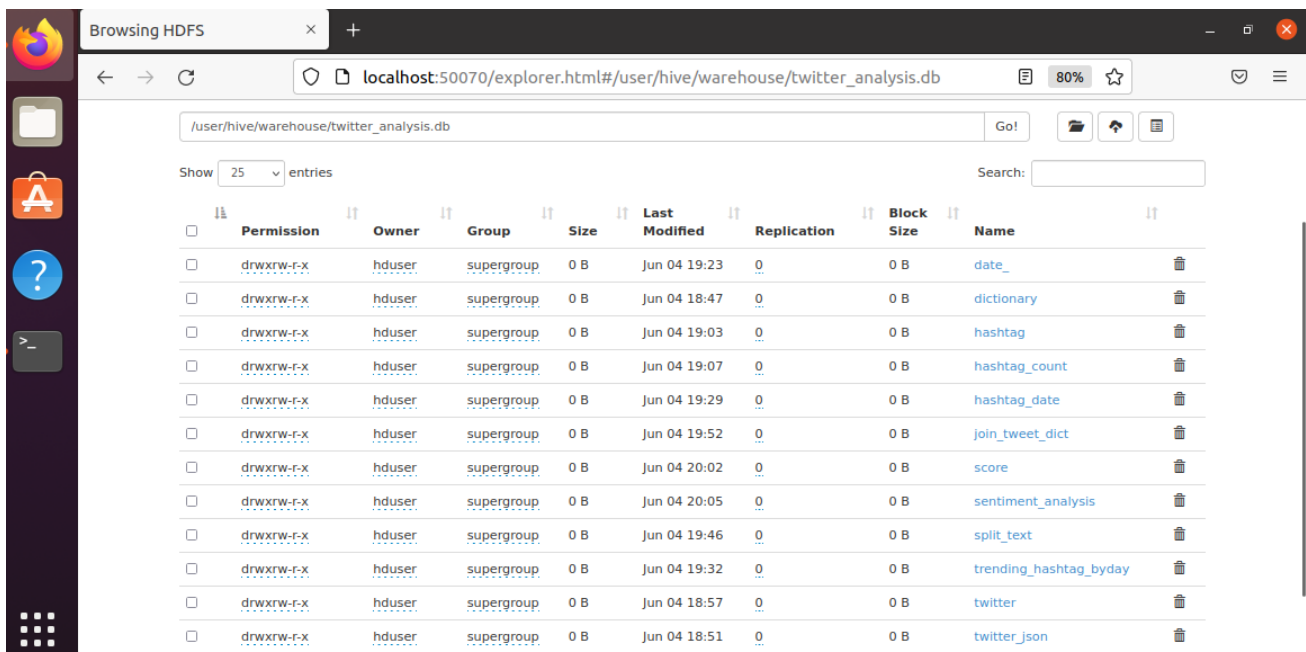
La table d'analyse des sentiments est créée et donne un résultat positif si la somme des notes est supérieure à 0, négative si la somme est inférieure à 0 et neutre si la somme est 0. La requête suivante est utilisée pour valider la table.


```

Time taken: 1.708 seconds
hive> select * from sentiment_analysis;
OK
100696141719273000    Michelle    2016-06-07    -5    Negative
101270445959461000    Frances    2016-04-04    2    Positive
103386304747066000    Marion    2011-09-10    9    Positive
105743242835042000    Alex    2011-09-10    NULL    Neutral
105859228650774000    Judith    2011-09-10    NULL    Neutral
106128175972515000    Nicholas    2011-09-10    -2    Negative
106841024367780000    Annie    2016-07-24    NULL    Neutral
107921004221536000    Gene    2011-09-10    NULL    Neutral
108397478405756000    Rhonda    2011-09-10    NULL    Neutral
109813990267358000    Jennifer    2011-09-10    -1    Negative
110624316674377000    Valerie    2011-09-10    -2    Negative
111442643526292000    Daniel    2015-12-15    NULL    Neutral
111565792997795000    Amy    2011-09-10    NULL    Neutral
113523162983355000    Marc    2015-12-22    NULL    Neutral
119416231902478000    Shirley    2011-09-10    2    Positive
121033230143694000    Dan    2011-09-10    -2    Negative
121873790298340000    Joshua    2011-09-10    2    Positive
124202382045715000    Lester    2009-03-18    -4    Negative
127363447354159000    Jackie    2011-09-10    NULL    Neutral
128176348280256000    Johnny    2016-04-04    NULL    Neutral
128390143224214000    Sue    2016-06-07    2    Positive
128919251286755000    Dianne    2015-12-22    1    Positive
132527313979710000    Judith    2016-01-25    -2    Negative
132658794295731000    Valerie    2011-09-10    NULL    Neutral
134237176213536000    Jonathan    2011-09-10    NULL    Neutral
136262215022259000    Lynn    2011-09-10    4    Positive
137400163920091000    Alexandra    2015-12-15    0    Neutral
138226334124559000    Kristin    2011-09-10    NULL    Neutral
138496495922264000    Rita    2011-09-10    1    Positive
144151617886614000    Scott    2015-12-15    NULL    Neutral

```

Notre entrepôt de données :



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:23	0	0 B	date_
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 18:47	0	0 B	dictionary
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:03	0	0 B	hashtag
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:07	0	0 B	hashtag_count
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:29	0	0 B	hashtag_date
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:52	0	0 B	join_tweet_dict
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 20:02	0	0 B	score
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 20:05	0	0 B	sentiment_analysis
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:46	0	0 B	split_text
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 19:32	0	0 B	trending_hashtag_byday
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 18:57	0	0 B	twitter
drwxrw-r-x	hduser	supergroup	0 B	Jun 04 18:51	0	0 B	twitter_json

CNCLUSION

Dans notre projet, nous avons pris les données de twitter à l'aide d'Apache Flume, puis nous avons traité ces données dans Hive, afin de stocker dans HDFS.

Nous avons une idée complémentaire pour notre projet sur laquelle on travaille actuellement, l'idée est que nous allons stocker ces données de HDFS dans MySQL en utilisant Sqoop.