Social Network Sentiment Analysis for security uses using: Apache Flume and Hive

OUERHANI Marouane

Business intelligence engineering student at ESPRIT, TUNISIA
marwen.werheni@esprit.tn

Abstract

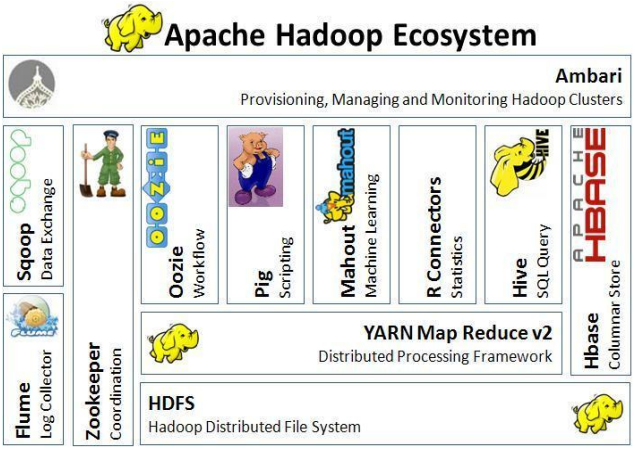
We can do analysis on data collected from social Network for many uses . In this paper, we are going to talk how effectively sentiment analysisis done on Twitter data using Flume. Twitter is an online web application which contains rich amount of data that can be a structured, semi-structured and un-structured data. We can collect the data from the twitter by using BIGDATA eco-system using online streaming tool Flume. And doing analysis on Twitter is also difficult due to language that is used for comments. And, coming to analysis there are different types of analysis that can be done on the collected data. So here we are taking sentiment analysis, for this we are using Hive and its queries to give the sentiment data based up on the groups that we have defined in the HQL (Hive Query Language).

Keywords: *Analysis, BIGDATA, Comment, Flume, Hive, HQL, Sentiment Analysis, Structured, Semi-Structured, Twitter, Tweets.*

1. Introduction

From 20th century onwards this WWW has completely changed the way of expressing their views. Present situation is completely they are expressing their thoughts through online blogs, discussion forms and also some online applications like Facebook, Twitter, etc. If we take Twitter as our example nearly 1TB of text data is generating within a week in the form of tweets. So, by this it is understand clearly how this Internet is changing the way of living and style of people. Among these tweets can be categorized by the hash value tags for which they are commenting and posting their tweets. So, now many companies and also the survey companies are using this for doing some analytics such that they can predict the success rate of their product or also they can show the different view from the data that they have collected for analysis. But, to calculate their views is very difficult in a normal way by taking these heavy data that are going to generate day by day.

Fig. 1: Describes clearly Apache Hadoop Ecosystem.



The above figure shows clearly the different types of ecosystems that are available on Hadoop so, this problem is taking now and can be solved by using BIGDATA [15]Problem as a solution. And if we consider getting the data from Twitter [1] one should use any one programming language to crawl the data from their database or from their web pages. Coming to this problem here we are collecting this data by using BIGDATA online streaming Eco System Tool known as Flume and also the shuffling of data and generating them into structured data in the form of tables can be done by using Apache Hive[9].

2. Problem Statement

2.1 Existing System

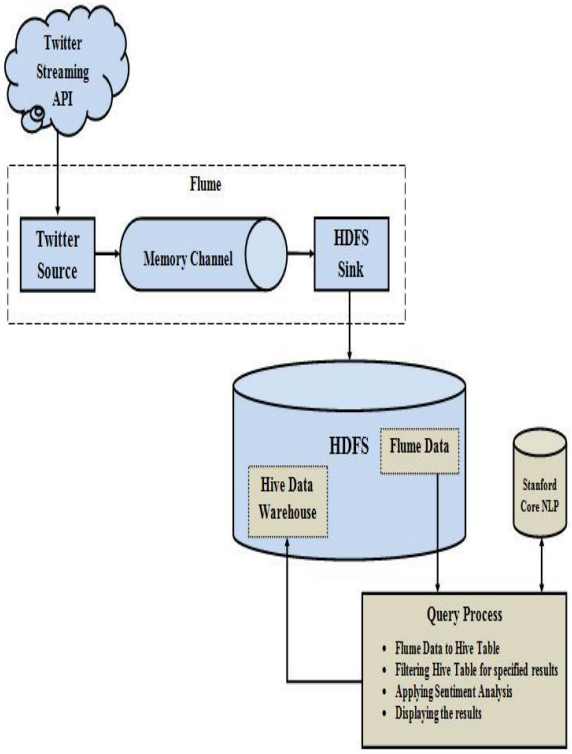
As we have already discussed about the older way of getting data and also performing the sentiment analysis on those data. Here they are going to use some coding techniques for crawling the data from the twitter where they can extract the data from the Twitter web pages by using some code that may be written either in JAVA, Python etc. For those they are going to download the libraries that are provided by the twitter guys by using thisthey are crawling the data that we want particularly.[1] After getting raw data they will filter by using some old techniques and also they will find out the positive, negative and moderate words from the list of collected words in a text file. All these words should be collected by us to filter out or do some sentiment analysis on the filtered data.[2],[7]. These words can be called as a dictionary set by which they will perform sentiment analysis. Also, after performing all these things and they want to store these in a database and coming to here they can use RDBMS[14] where they are having limitations in creating tables and also accessing the tables effectively.

2.2Proposed System

As it can have seen existing system drawbacks, here we are going to overcome them by solving this issue using Big Data problem statement. So here we are going to use Hadoop and its Ecosystems, for getting raw data from the Twitter we are using Hadoop online streaming tool using Apache Flume[13]. In this tool only we are going to configure everything that we want to get data from the Twitter.[3] For this we want to set the configuration and also want to define what information that we want to get form Twitter. All these will be saved into our HDFS (Hadoop Distributed File System)[12] in our prescribed format. From this raw data we are going to create the table and filter the information that is needed for us and sort them into the Hive Table. And form that we are going to perform the Sentiment Analysis by using some UDF's (User Defined Functions).

The following figure shows clearly the architecture view for the proposed system by this we can understand how our project is effective using the Hadoop ecosystems and how the data is going to store form the Flume, also how it is going to create tables using Hive also how the sentiment analysis is going to perform[8].

Fig. 2: Architecture diagram for proposed system.



3. Methodology

As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

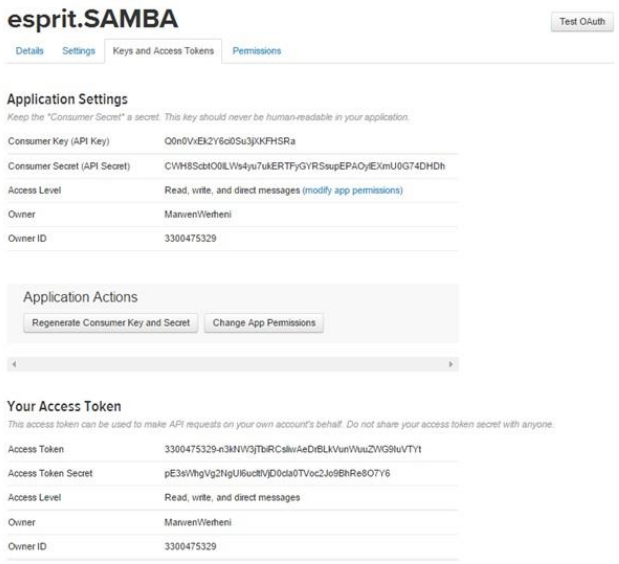
Creating Twitter Application.
Getting data using Flume.
Querying using Hive Query Language (HQL)

3.1 Creating Twitter Application

First of all if we want to do sentiment analysis on Twitter data we want to get Twitter data first so to get it we want to create an account in Twitter developer and create an application by clicking on the new application button provided by them.[3] After creating a new application just create the access tokens so that we no need to provide our authentication details there and also after creating application it will be having one consumer keys to access that application for getting Twitter data. The following is the figure that show clearly how the application data looks

after creating the application and here it's self we can see the consumer details and also the access token details. We want to take this keys and token details and want to set in the Flume configuration file such that we can get the required data from the Twitter in the form of twits.

Fig. 3: Creating Twitter application from Twitter Developer.



The figure show clearly the application keys that are generated after creating application and in this keys we can see the top two keys are the API key and API secret. And coming to the reaming two keys it is nothing but know as the access tokens that we want to generate it by ourselves by clicking the generate access token. After clicking that we can get the two keys that are our account access token and coming to that one is Access token and the other one is the Access token secret.

3.2 Getting data using Flume

After creating an application in the Twitter developer site we want to use the consumer key and secret along with the access token and secret values. By which we can access the Twitter and we can get the information that what we want exactly here we will get everything in JSON format and this is stored in the HDFS that we have given the location where to save all the data that comes from the Twitter. The following is the configuration file that we want to use to get the Twitter data from the Twitter.

Fig. 4: Flume configuration files for Twitter data.

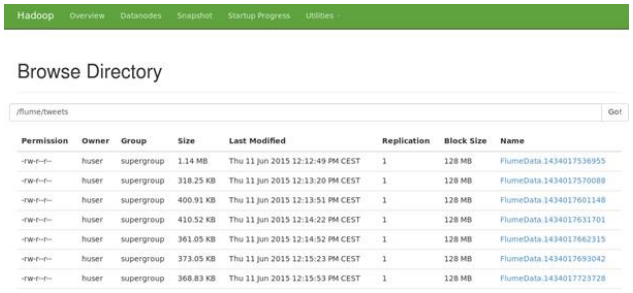
```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = Q0n0VxEk2Y6ci0Su3jXKFHSRa
TwitterAgent.sources.Twitter.consumerSecret = CWH8ScbtO0lLWs4yu7uKERTFyGYRSsupEPAOylEXmU0G74DHDh
TwitterAgent.sources.Twitter.accessToken = 3300475329-n3kNW3jTbiRCsliwAeDrBLkVunWuuZWG9IuVTYt
TwitterAgent.sources.Twitter.accessTokenSecret = pE3sWhgVg2NgUl6ucltlVjD0cla0TVoc2Jo9BhRe8O7Y6

TwitterAgent.sources.Twitter.keywords = linux hack,linux vulnerability,linux vuln,linux bug,linux exploit,linux patch,linux fix,linux CVE
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/flume/tweets/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 10
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

3.3 Querying using Hive Query Language (HQL)

After running the Flume by setting the above configuration then the Twitter data will automatically will save into HDFS[6] where we have the set the path storage to save the Twitter data that was taken by using Flume. The following is the figure that shows clearly how the data is stored in the HDFS in a documented format and the raw data that we got form the Twitter is also in the JSON format that is shown clearly below:

Fig. 5: Twitter data in HDFS (Hadoop Distributed File System).



From these data first we want to create a table where the filtered data want to set into a formatted structured such that by which we can say clearly that we have converted the unstructured data into structured format. For this we want to use some custom serde concepts. These concepts are nothing but how we are going to read the data that is in the form of JSON format for that we are using the custom serde for JSON so that our hive can read the JSONdata[10]and can create a table in our prescribed format.

Fig. 6: HQL Query for creating Tweets table.

```
CREATE EXTERNAL TABLE tweets (
 username STRING,
 lang STRING,
 screen_name STRING,
 id BIGINT,
 created_at STRING,
 text STRING,
 post_id BIGINT,
 post_created_at STRING,
 hashtags STRING,
 retweet BOOLEAN,
 favorited BOOLEAN,
 retweet_count BIGINT,
 friends_count INT,
 followers_count INT,
 statuses_count INT,
 verified BOOLEAN,
 utc_offset INT,
 time_zone STRING,
 retweeted_username STRING,
 retweeted_screen_name STRING,
 retweeted_id BIGINT,
 retweeted_text STRING,
 retweeted_retweet_count BIGINT
)
PARTITIONED BY (datehour INT, rating STRING)
LOCATION '/user/flume/tweets';
```

Also we are using another UDF's (User Defined Functions) for performing the sentiment analysis on the tales that are created by using Hive. From that we can perform the sentiment analysis. And acquire the results where a new table is created by partition concept such patch, exploits, cve, zero-day and finally the table others that contain the reset of results all those table are stored also in HDFS the next figure show that clearly.

Fig. 7: Query of the table Result.

```
select * from tweets where instr(text,"PATCH")<>0 or instr(text,"patch")<>0' | sed 's/[\t]/,/g' > /home/huser/analyse-twitter/patch/patch.txt
select * from tweets where instr(text,"CVE")<>0 or instr(text,"cve")<>0' | sed 's/[\t]/,/g' > /home/huser/analyse-twitter/cve/cve.txt
select * from tweets where instr(text,"Zero day")<>0 or instr(text,"zero day ")<>0' | sed 's/[\t]/,/g' > /home/huser/analyse-twitter/zeroday
select * from tweets where instr(text,"exploit")<>0 or instr(text,"EXPLOIT")<>0' | sed 's/[\t]/,/g' > /home/huser/analyse-twitter/exploit.txt
```

Fig. 8: Result after performing the sentiment analysis.

4. Conclusions

There are different ways to get Twitter data or any other online streaming data where they want to code lines of coding to achieve this. And, also they want to perform the sentiment analysis on the stored data where it makes some complex to perform those operations. Coming to this paper we have achieved by this problem statement and solving it in BIGDATA by using Hadoop and its Eco Systems. And finally we have done sentiment analysis on the Twitter data that is stored in HDFS[6]. So, here the processing time taken is also very less compared to the previous methods because HadoopMapReduce and Hive are the best methods to process large amount of data in a small time.

5. Future Work

In this paper it has shown the way for doing sentiment analysis for Twitter data. Also, we can do this by using R language to implement clustering, scoring and profiling algorithm that help us to know and identify wish is the most populated tweet and user.

Acknowledgments

We are grateful to express sincere thanks to our faculties who gave support and special thanks to our department for providing facilities that were offered to us for carrying out this project.

References

[1] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

[2] Tang, H., Tan, S., Cheng, X., A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal, v.36 n.7, p.10760-10773, September, 2009.

[3] A. Pak and P. Parouek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010.

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, Vol. 51, Iss. 1, pp. 107-113, January 2008.

[5] S. Ghemawat, H. Gobioff and S-T. Leung, "The Google File System," ACM SIGOPS Operating System Review, Vol. 37, Iss. 5, pp. 29-43, December 2003.

[6] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[7] Bahrainian, S.A., Dengel, A., Sentiment Analysis using Sentiment Features, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013.

[8]"Sentimental Analysis", Inc. [Online]. Available: [Accessed 23 March 2013].

[9] (Online Resource) Hive (Available on:http://hive.apache.org/).

[10] (Online Resource)http://jsonlint.com/

[11] (Online Resource)http://nlp.stanford.edu/software/corenlp.shtml.

[12]T. White, "The Hadoop Distributed Filesystem," Hadoop: The Definitive Guide, pp. 41-73, GravensteinHighwaNorth, Sebastopol: O'Reilly Media, Inc., 2010.

[13] (Online Resource) http://flume.apache.org/

[14]S. W. Ambler. Relational databases 101: Looking at the whole picture.www.AgileData.org, 2009.

[15] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier For Innovation, Competition, And Productivity", May 2011.

Author: OUERHANI Marouane Business intelligence engineering student at ESPRIT he was born in Tunis at 24/12/1991.