Sentiment Analysis on Twitter Data Using Apache Flume and Hive

Ms. Pooja S. Patil[1], Ms. Pranali B. sable[2], Ms. Reshma J. Fasale[3], Mr. P. A. Chougule[4]

[123] Department of Computer Science Engg., Dr.J.J.Magdum College Of Engg., Jaysingpur,India.
[4] Assistant Professor, Department of Computer Science Engg., Dr.J.J.Magdum College Of Engg., Jaysingpur,India.

---***---

Abstract -*'BIG DATA' has been getting much importance in different industries over the last year or two, on a scale that has generated lots of data every day. Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a reasonable amount of time. It has tremendous potential to transform business and power in several ways. Here the challenge is not only storing the data, but also accessing and analyzing the required data in specified amount of time. One of the popular implementation to solve the above challenges of big data is using Hadoop. Hadoop is well-known open-source implementation of the MapReduce programming model for processing big data. It is highly scalable compute platform. Hadoop enables users to store and process bulk amount which is not possible while using less scalable techniques. As of now we know present industries and some survey companies are mainly taking decisions by data obtained from web. As we see WWW is a rich collection of data that is mainly in the form of unstructured data from which we can do analysis on those data which is collected on some situation or on a particular thing. In this paper, we are going to talk how effectively sentiment analysis done on the data which is collected from the Twitter 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). Here we have categorized this sentiment analysis into 3 groups like tweets that are having positive, moderate and negative comments.*

Keywords: **Analysis, BIGDATA, Comment, Flume, Hive, HQL, Sentiment Analysis, Structured, Semi-Structured, Twitter, Tweets, Un-Structured, WWW (Word Wide Web).**

1. Introduction

Over past ten years, industries and organizations doesn't need to store and perform much operations and analytics on data of the customers. But around from 2005, the need to transform everything into data is much entertained to satisfy the requirements of the people. So Big data came into picture in the real time business analysis of processing data. 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.

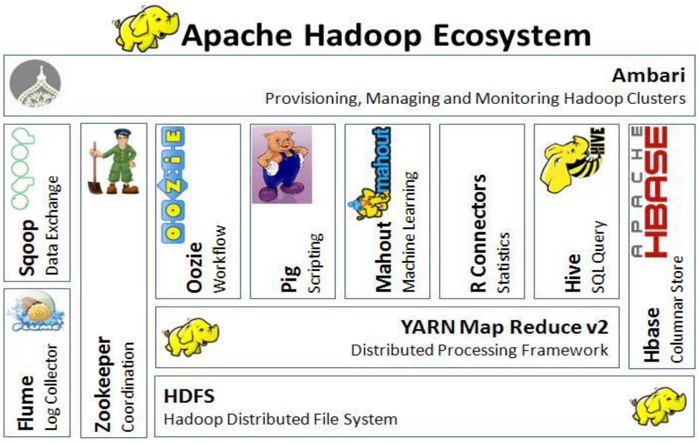


Fig. 1: Describes clearly Apache Hadoop Ecosystem.

The above figure shows clearly the different types of ecosystems that are available on Hadoop. If we consider getting the data from Twitter 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.

2. State of Art

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:

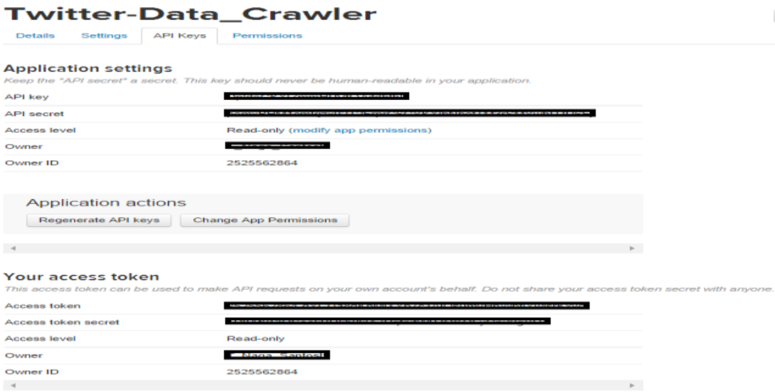
1. **Creating Twitter Application.**
2. **Getting data using Flume.**
3. **Querying using Hive Query Language (HQL)**

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. In this tool only we are going to configure everything that we want to get data from the Twitter. 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) 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) by which we can perform sentiment analysis by taking Stanford Core NLP as the data dictionary so that by using that we can decide the list of words that coming under positive, moderate and negative. 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.

3. Experimental Setup

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. 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.



Application settings

Keep the "API secret" a secret. This key should never be human-readable in your application.

API key [redacted]
API secret [redacted]
Access level Read-only (modify app permissions)
Owner [redacted]
Owner ID 2525562864

Application actions

Your access token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access token [redacted]
Access token secret [redacted]
Access level Read-only
Owner [redacted]
Owner ID 2525562864

Fig. 2: 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.

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 where we have the set the path storage to save the Twitter data that was taken by using Flume. 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 and can create a table in our prescribed format. 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 that all the comments that are having positive will go into the positive partition and all the comments that are having moderate will go into moderate partition and finally all the comments that are having negative will go into negative partition. The following figure shows clearly how the rating is done and how the data is partitioned into 3 types.

4. Sentiment Analysis

In short, Sentiment Analysis is the process of detecting the contextual polarity of text. In other words, it determines whether a piece of writing is positive, negative or neutral. An alternative term is opinion mining, as it derives the opinion, or the attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic. For example, do people on Twitter think that Chinese food in San Francisco is good or bad? Analyzing tweets for sentiment will answer this question for you. You can also learn why people think the food is good or bad, by extracting the exact word indicating why people did or didn't like the food. **"I love the summer in New York, but I hate the winter."** The individual scores would show "love the summer" as positive and "hate the winter" as negative. However, the sentiment for the entire comment would be neutral, because the positive sentiment for the word love would cancel out the negative sentiment for the word hate. Because Sentiment Analysis can track a particular topic, many companies use it to track or monitor their products, services or reputation in general.

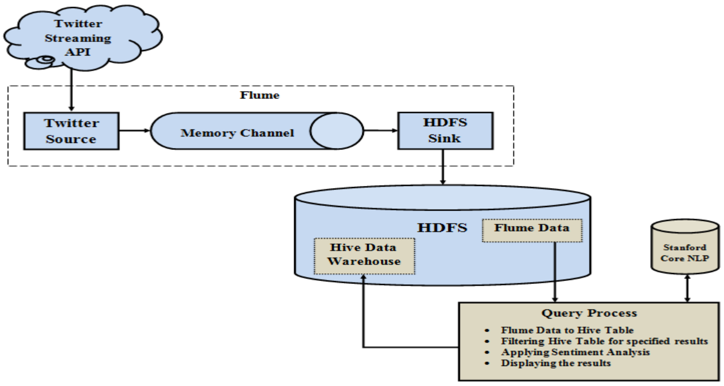
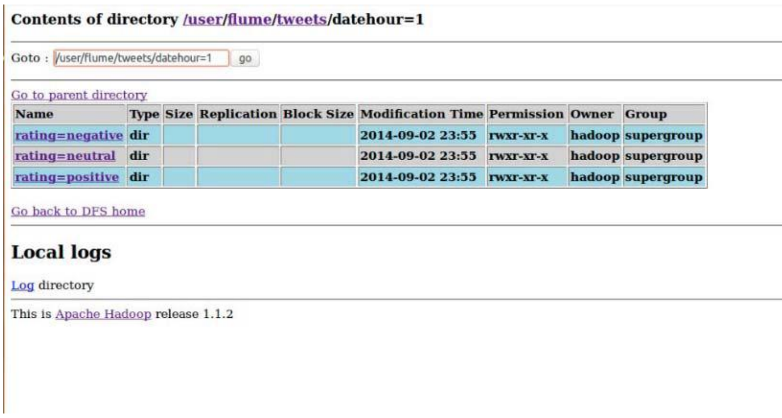


Fig. 3: Architecture diagram for proposed system.

5. Result

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 JSON data and can create a table in our prescribed format.

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 that all the comments that are having positive will go into the positive partition and all the comments that are having moderate will go into moderate partition and finally all the comments that are having negative will go into negative partition. The following figure shows clearly how the rating is done and how the data is partitioned into 3 types.



Contents of directory /user/flume/tweets/datehour=1

Goto : /user/flume/tweets/datehour=1 go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
rating=negative	dir				2014-09-02 23:55	rwxr-xr-x	hadoop	supergroup
rating=neutral	dir				2014-09-02 23:55	rwxr-xr-x	hadoop	supergroup
rating=positive	dir				2014-09-02 23:55	rwxr-xr-x	hadoop	supergroup

Go back to DFS home

Local logs

Log directory

This is Apache Hadoop release 1.1.2

Fig. 4: Result after performing the sentiment analysis.

6. Application and scope

The major applications of the big data are :

1. Sentiment Analysis: Sentiment data is unstructured data that represents opinions, emotions, and attitudes contained in sources such as social media posts, blogs, online product reviews, and customer support interactions.

Different companies and Organizations use social media analysis to understand how the public feels about something at a particular moment in time, and also to track how those opinions change over time.

2. Text Analytics: It is the process of deriving the high quality information from the raw data such as unstructured data and predicting the analysis.

3. Volume Trending: Here volume is estimated in terms of amount of data to process a job. Volume trending is a big issue nowadays. Day by day it has been increasing in a much higher rate in the organizations and social media sites etc.

4. Predictive Analytics: Predictive analysis gives the predictive scores to the organizations to help in making the smart decisions and improve business solutions. It optimizes marketing campaigns and website behaviour to increase customer responses in business, conversions and meetings, and to

decrease the burden on the people. Each customer's predictive score informs actions to be taken with that customer.

5. Massively Scalable Architectures

6. Social Media Data: With Hadoop, we can mine Twitter, Facebook and other social media conversations for sentiment data about people and used it to make targeted, real time decisions that increase market share.

7. Web Click stream data: Hadoop makes easy to track customers and their activities in different issues like products purchasing and viewing etc. It makes analyzers to know the behaviour and interest of the customers and can able to visualize similar type of products to the customers.

Scope: The Proposed system can finds the most popular information about the people, organizations and can be used in the field of analytics.

Applications :

- ✓ Finds the most number of follows in the social networking sites.
- ✓ This system can be useful to track the business analysis of the organizations.
- ✓ Allows researchers to retrieve and analyze the data easily from large datasets.

7. Future work

In this paper it has shown the way for doing sentiment analysis for Twitter data. Also, we can do this by using Oozie by creating a work flow so that we can give a time slang such that it will work based upon that time we allocated for performing a particular work. Also at last we can also visualize the word map i.e., the most frequent words that are used in positive, moderate and negative fields by using R language to visualize.

8. Conclusion

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. So, here the processing time taken is also very less compared to the previous methods because Hadoop Map Reduce and Hive are the best methods to process large amount of data in a small time.

9. References

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

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

3. 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.

4. 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.

5. 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.

6."Sentimental Analysis", Inc. [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis [Accessed 23 March 2013].

7.(Online Resource) Hive (Available on :http://hive.apache.org/).