International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 02 (MAR-APR, 2016)

SOCIAL DATA ANALYSIS USING APACHE FLUME, HDFS, HIVE

Mrs. Sulochana Panigrahi
PG Scholar, Department of Computer Science and Engineering,
New Horizon College of Engineering,
Bangalore, Karnataka, India
sulochanap01@gmail.com

Dr. S. Mohan Kumar
Associate Professor, Department of Computer Science and Engineering,
New Horizon College of Engineering,
Bangalore, Karnataka, India
drsmohankumar@gmail.com

Abstract: - Twitter is one of the most popular micro blogging website in today's globalized world. Twitter messages can be mined to gain valuable information. Although Twitter provides a list of most popular topics people tweet about known as Trending Topics in real time, it is often hard to understand what these trending topics are about. Therefore, various efforts are being made to classify these topics into general categories with high accuracy for better information retrieval. In this paper, we are going to talk how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. Twitter is an online web application which of data that can be a structured, semi-structured and un-structured data. Collect the data from the twitter by using CLOUDERA VM 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) and use Visual Studio to show in User Interface.

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

I. INTRODUCTION

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.

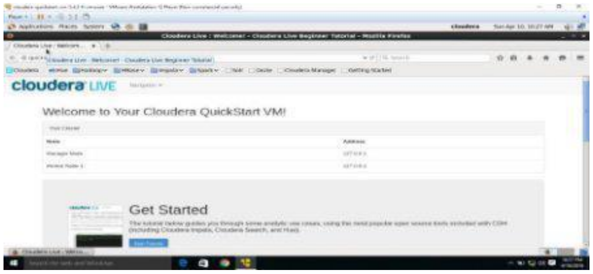

cloudera LIVE
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Welcome to Your Cloudera QuickStart VM!
Get Started

Fig. 1: Describes clearly Cloudera VM

The above figure shows clearly the different types of service that are available on cloudera VM so, this problem is taking now and can be solved by using BIGDATA Problem as a solution. And 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.

II. 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 this they 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]. 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, where they are having limitations in creating tables and also accessing the tables effectively.

2.2 Proposed 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. 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. 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.

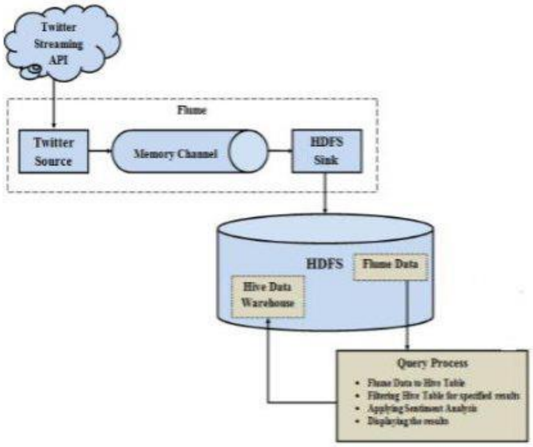


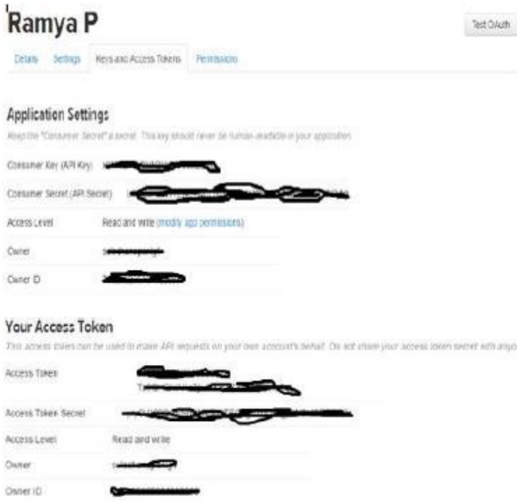
Fig. 2: Architecture diagram for proposed system

III. 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.

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 given figure clearly show that how the application data looks 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. 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 we 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.



Ramya P

Application Settings

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

Consumer Key (API Key) [redacted]

Consumer Secret (API Secret) [redacted]

Access Level Read and write (modify app permissions)

Owner [redacted]

Owner ID [redacted]

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 anyo

Access Token [redacted]

Access Token Secret [redacted]

Access Level Read and write

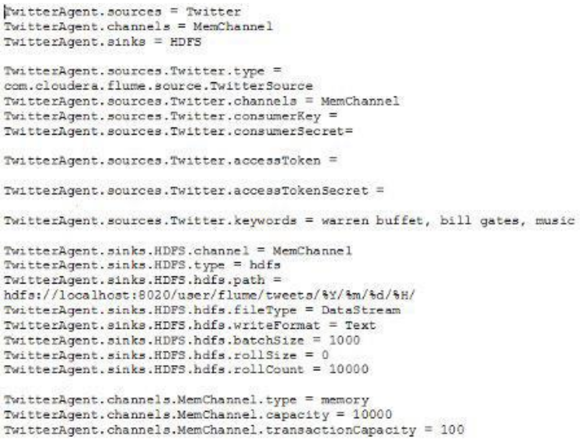
Owner [redacted]

Owner ID [redacted]

Fig. 3: Creating Twitter application from Twitter Developer

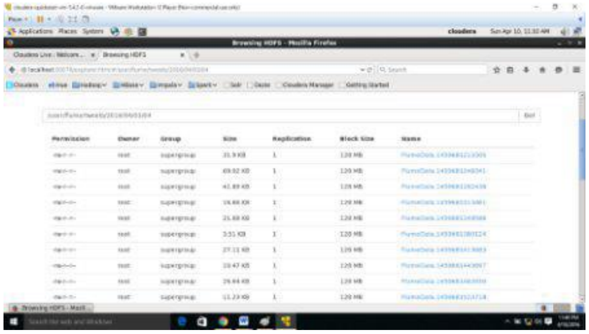
3.2 Getting data using Flume

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.



```
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 =
TwitterAgent.sources.Twitter.consumerSecret=

TwitterAgent.sources.Twitter.accessToken =

TwitterAgent.sources.Twitter.accessTokenSecret =

TwitterAgent.sources.Twitter.keywords = warren buffet, bill gates, music

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:8020/user/flume/tweets/%Y/%m/%d/%H/
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.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Fig. 4: Flume configuration files for Twitter data



Permission	Owner	Group	Size	Replication	Block Size	Name
[illegible]	root	supergroup	31.9 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	69.92 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	41.89 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	[illegible]	1	128 MB	[illegible]
[illegible]	root	supergroup	[illegible]	1	128 MB	[illegible]
[illegible]	root	supergroup	3.51 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	27.11 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	19.47 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	29.64 KB	1	128 MB	[illegible]
[illegible]	root	supergroup	11.23 KB	1	128 MB	[illegible]

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

File: /user/flume/twitter/2014/09/02/FlumeData.1409675778359

Goto: /user/flume/twitter/2014/09/02 go

Go back to dir listing
Advanced view/download options

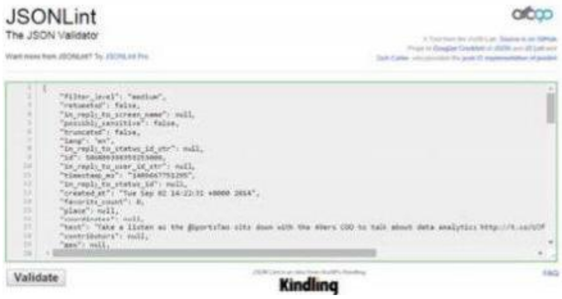
View Next chunk

[illegible]

Fig. 6: Twitter data in JSON format

3.3 Querying using Hive Query Language (HQL)

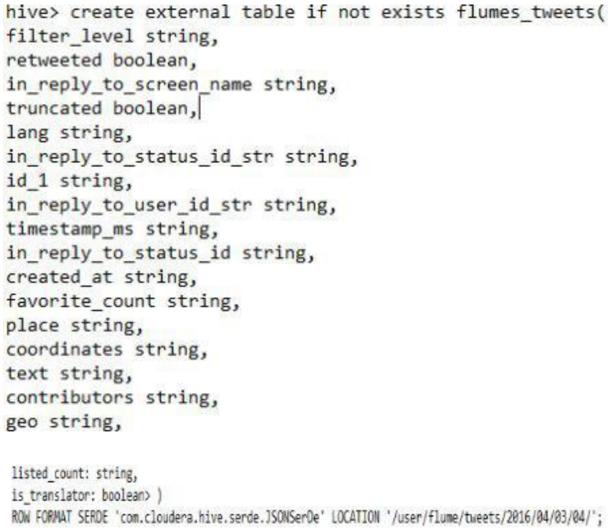
After running the Flume by setting the above configuration then the Twitter data will automatically will save into HDFS where we have the set the path storage to save the Twitter data that was taken by using Flume. The following is the figure that shows clearly how the data is stored in the HDFS in a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



```
{
    "filter_level": "medium",
    "retweeted": false,
    "in_reply_to_screen_name": null,
    "possibly_sensitive": false,
    "truncated": false,
    "lang": "en",
    "in_reply_to_status_id_str": null,
    "id": [illegible],
    "in_reply_to_user_id_str": null,
    "timestamp_ms": "[illegible]",
    "in_reply_to_status_id": null,
    "created_at": "Tue Sep 02 14:22:31 +0000 2014",
    "favorite_count": 0,
    "place": null,
    "coordinates": null,
    "text": "Take a listen as the @sportsfan sits down with the 49ers CIO to talk about data analytics http://t.co/[illegible]
    "contributors": null,
    "geo": null,
```

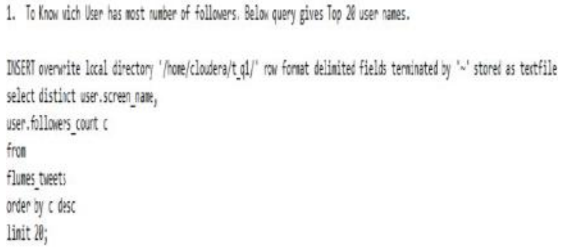
Fig. 7: Validating JSON data for HQL.

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. From that we can perform the sentiment analysis and acquire the results where a new table is created such that all the comments those to know which User has most number of followers in figure 10.



```
hive> create external table if not exists flumes_tweets(
filter_level string,
retweeted boolean,
in_reply_to_screen_name string,
truncated boolean,
lang string,
in_reply_to_status_id_str string,
id_1 string,
in_reply_to_user_id_str string,
timestamp_ms string,
in_reply_to_status_id string,
created_at string,
favorite_count string,
place string,
coordinates string,
text string,
contributors string,
geo string,

listed_count: string,
is_translator: boolean> )
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user/flume/tweets/2016/04/03/04/';
```

Fig. 8: HQL Query for creating Tweets table.



1. To Know wich User has most number of followers. Below query gives Top 20 user names.

```
INSERT overwrite local directory '/home/cloudera/t_q1/' row format delimited fields terminated by '~' stored as textfile
select distinct user.screen_name,
user.followers_count c
from
flumes_tweets
order by c desc
limit 20;
```

Fig. 9: Inserting data by performing sentiment analysis.

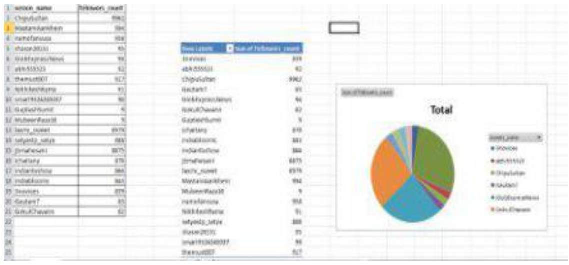


Fig 10: Sentimental Analysis through Excel Sheet

IV. CONCLUSION & FUTURE SCOPE

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. In this paper it has shown the way for doing sentiment analysis for Twitter data. Also, we can do 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.

REFERENCES

[1]. Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde , Real Time Sentiment Analysis of Twitter Data Using Hadoop, International Journal of Computer Science and Information Technologies. Vol 5 ISSUE 3, 2014.

[2]. Penchalaiah C, Murali G , Suresh Babu, Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol.1, ISSUE 8, ISSUE 8, October 2014.

[3]. Manoj Kumar Danthala, Tweet Analysis: Twitter Data processing Using Apache Hadoop, International Journal of Core Engineering & Management (IJCEM), Volume 1, Issue 11, February 2015.

[4]. Manoj Kumar Danthala, Dr. Siddhartha Ghosh, Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using Big Insights, International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 05, May-2015.

[5]. Judith Sherin Tilsha S, Shobha ,A Survey on Twitter Data Analysis Techniques to Extract Public Opinion , International Journal of Advanced Research in Computer Science and Software Engineering Research Paper , Volume 5, Issue 11 , November 2015 .

[6]. Munesh Kataria1, Ms. Pooja Mittal, Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql, International Journal of Computer Science and Mobile Computing a Monthly Journal of Computer Science and Information Technology. IJCSMC, Vol. 3, Issue 7, July 2014.

[7]. Kushal Sharma, Prashant Singh, Sachin Mote, Sudarshan Patil,Twitter Sentimental Analysis using Hadoop, International Journal of Computer Application , Volume 2, Issue 5, January 2015 .

[8]. Mr. Agar Nadagoud , Channa basaveshwara ,Market Sentiment Analysis for Popularity of Flipkart, International Journal of Advanced Research in Computer Engineering &Technology(IJARCET) , Volume 4, Issue 5, may2015.