Social Media based Sentimental Analysis using Hive and Flume

Rahul Deva
Department of CSE
JEMTEC, Greater Noida, India

Garima Kulshreshtha
Department of ECE
IILM CET, Greater Noida, India

Abstract – Now a day Social media has been so popular among all, receiving more attention, time and people are using this as platform to express their views and other likings and disliking. Everybody post many things like their love, sorrow, anger and opinions about products, places and etc. There are so many social media's platform like Facebook, Twitter, Instagram, Whatsapp, Snapchat, Linkedin, Blogger, Quora and many more. People are spending lots of time with these sites and expressing their personal opinion on many issues.

Twitter is one of the popular social media used for official statements as well as personal views, which allows users to publish micro-blogging short messages called tweets with limited length that are visible to your friends or followers. Twittering is also a less gated method of communication: anyone can share information with people that you wouldn't normally exchange email or IM messages with, opening up your circle of contacts to an ever-growing community of like-minded people. So in short tweets can be named positive, negative, in support or unbiased. In this paper we are analyzing sentiments of Twitters messages.

Keywords: Facebook, Twitter, sentiment, hive, flume, big data, hadoop, MapReduce, unstructured data.

I. INTRODUCTION

Twitter users have found many different uses, including basic communication between friends and family, a way to publicize an event, or as a customer relations tool for companies to communicate with their consumers. With this much of expansion and availability of data, the measure of web based life information being created is increasing very fast. Every minutes seconds there are 11,0000+ tweets, 775,000+ status updates, 1,10,00,000+ instant messages , 698,445 google searches. These all data is unstructured and big in size.

Big data consists everything from the click stream data from the web to genomic data from biological research and medicines. This enormous amount of data is not easy to process as it contains the records of million people that includes everyday massive amount of data from social sites, cell phones GPS signals, videos etc. This data is different from structured data (which is stored in relational database systems) in terms of five parameters –variety, volume, value, veracity and velocity (5V's). The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management are:

1. Volume: Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Storage of such massive quantities of data is one of the issues that needs to be take care of. Data volumes are expected to grow 50 times by 2020.

2. Variety: The sources of Big Data are heterogeneous. The files comes in various formats and of any type, it may be structured, semi-structured or quasi-structured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.

3. Velocity: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. For Some organizations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.

4. Value: Which addresses the need for valuation of enterprise data? It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of data.

5. Veracity: The increase in the range of values typical of a large data set. Since the data is gathered from variety of sources and also it is of different type then, the data cannot be considered 100% correct.

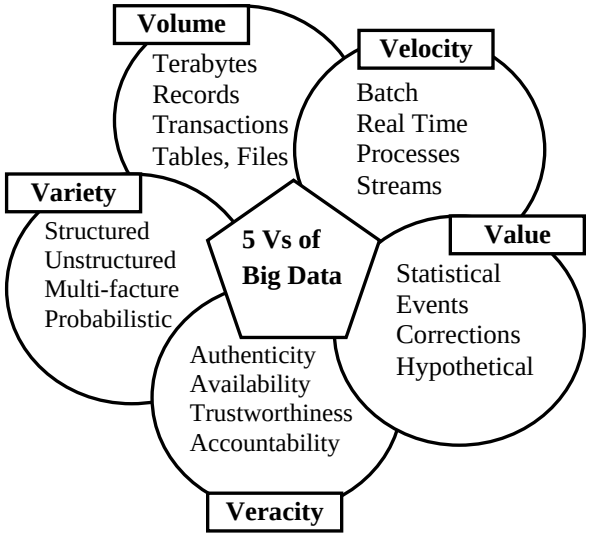

Figure1: Parameters of bigdata

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all values associated with the same intermediate key[2].

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It works on the principle of distributed file system and distributes the file among the nodes and allows the system to continue work in case of a node failure. This approach reduces the risk of catastrophic system failure.

Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System (HDFS) consists of three Components: the Name Node, Secondary Name Node and Data Node.

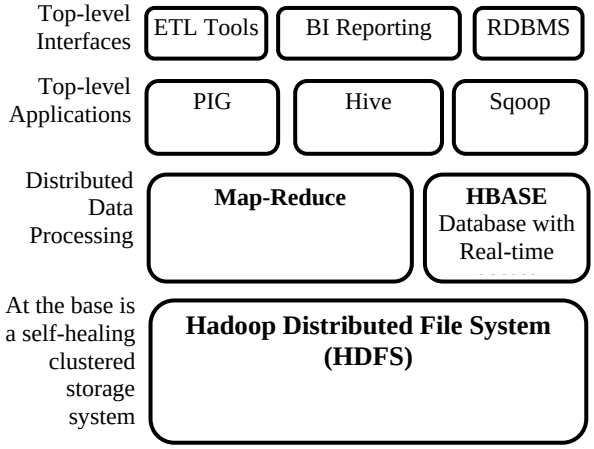

Figure2: Hadoop Architecture

The disadvantage of Hadoop is that it is not suitable for small data. Due to its high capacity design, the HDFS lacks the ability to efficiently support the random reading of small files.

II. RELATED WORK

Sentiment data is basically unstructured data that consists opinions, emotions, and views contained in social media posts, blogs, online product reviews, and customer support interactions. Many companies 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. Sentimental Analysis is also known as Opinion mining. In Today's world it has the title of the most popular trends as it helps in finding the reviews of a product. It basically uses ETL (Extract, transform, load) method.

Many researchers are working to implement sentiment analysis with different different methods and techniques. Many of them had worked on this topic using different techniques and tools. Some of the work related to the Sentimental analysis are:
In Mahalakshmi R, Suseela [3] Social Sentiment Analysis and Data Visualization on Big Data.: This method is composed of a HDFS system based on Hadoop ecosystem and Mapreduce functions for sentiment analysis.

Manoj Kumar Danthala [5] Twitter Data processing with the help of Apache Hadoop : A method of analyzing data (e.g twitter data) with help of Apache Hadoop using map reduce function by generating mapper, combiner, partitioner and reducer that have a function or work of processing and analyzing tweets on Hadoop. Analysis is done on tweets and tweets ids.

Analysis of Market Sentiment Analysis for Flipkart Popularity [6]: In this paper analysis of sentiments on data sets are done with the help of Big data hadoop and some of its eco system components. The output data have 3 categorized group having positive, neutral and negative reviews, just like twitter.

III. METHODOLOGY

Hive provides an SQL dialect, called Hive Query Language (abbreviated HiveQL or just HQL) for querying data stored in a Hadoop cluster.

Hive is most suited for data warehouse applications, where relatively static data is analyzed, fast response times are not required, and when the data is not changing rapidly. Hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights, reports, etc. used for data analysis transformation of very large data. As we have seen many work related to our work so here are the steps to how to fetch data, process that data and stoted in

HDFS and to how to process this work using different technique. So, for this follow the following steps:

1. Create Twitter Application.
2. Data fetching using Flume
3. Query using HQL.

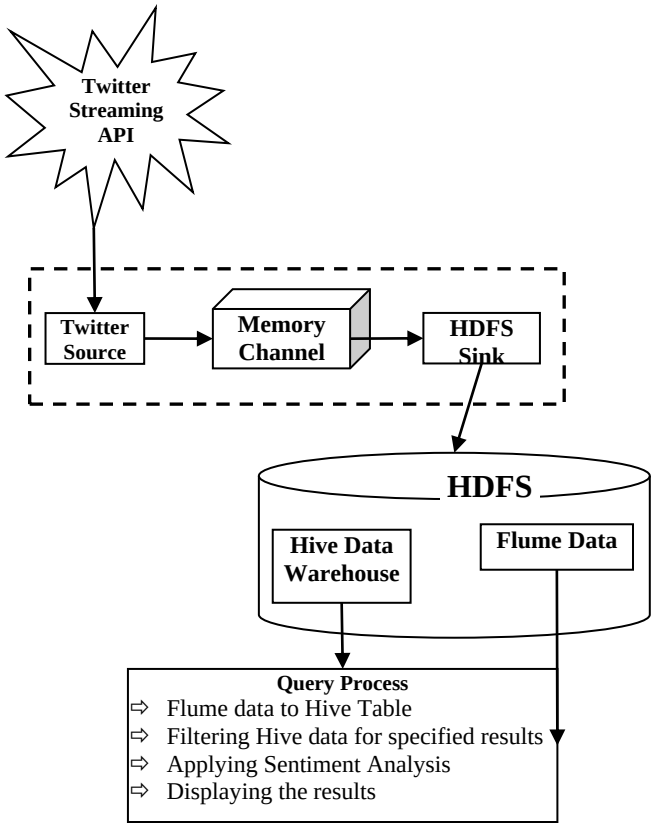

Figure 3: Architecture for proposed system

1) Create Twitter Application

For doing analysis on Twitter data we need twitter information by making a twitter application. Twitter application creation steps are given underneath:

- First open the connection of dev.twitter.com/app in Mozilla Firefox Browser and sign in the twitter account and do some work with twitter Application window where you can make i.e. create, delete, and manage Twitter Apps.
- In the next step click on the Create New App button. At that point you will get an application frame to fill your detail data.
- At that point the new App will be made. New app is utilized to make Consumer Key, Access Key, and Access Token Key. This will be used to edit in the Flume.conf file/record. While getting data's from Twitter these Consumer Key, Access Key, Access Token Key is used to fetch data which is lively tweeting in the account.
- These keys are used to Access Tokens tab and it can observe a button with the name of Create my access token. By clicking this we can produce the access token.
- Consumer keys (API Key), Consumer Secret (API Secret), Access tokens are utilized to arrange the Flume operator.

2) Flume for fetching the data:

After performing above step, now in order to fetch information from Twitter, following steps have to perform:

- We will use the API key and secret API key and the access token and secret values.
- Fetch data that we needed and it will be in JSON format and we will put this data in the HDFS in the location where we have saved all the data that comes from the Twitter.
- Configuration file used to get real time data from the Twitter. All the details or the points of interest needed to filled in the flume-twitter.conf file i.e. configuration file of Flume.

3) Query utilizing HQL

By setting the above design & then running the Flume then the Twitter information will automatically will saved into HDFS in different different directories where the storage path is set by us to save the Twitter data/information that was extracted by using Flume.

- We can also keep the data in local file directory but we haven't because lodaing data in local flie directory is a lengthy process.
- From the collected data we will create a table, table name mytweets_raw where the filtered data will be kept into a formatted structured such that we can clearly show that we have converted the unstructured data into structured data or in organized way.
- After loading the real time data in hive table, more tables are created like dictionary table which stores the polarity and the word and tweets_sentiment table which will contain all the tweets id and its sentiment. Many such more tables are created and different operations are done on data.

IV. DATASETS AND RESULT ANALYSIS

Before we apply queries on the information, we have to make sure that table of Hive can appropriately translate the JSON formatted data using JSON validator. Naturally, Hive takes the input files that uses a format of delimited row, yet fetched data still in JSON format, which is not going to work.

What's more, we can utilize the interface of Hive SerDe in orfer to determine translation of the data. SerDe are the interfaces which guides the Hive about any modify/change of data that Hive can process. For that a jar file is added "hive-serdes-1.0-SNAPSHOT.jar" into the directory /usr/local/hive/lib. This will be used by hive shell to extract the clean data from the downloaded data into the hive table.

Figure 1: Input data Stored in Hive

By using jar file of hive and custom serde files unstructured data can be stored into hive table name mytweets_raw in the structured format. The figure demonstrate's the structured information stored in table named mytweets_raw. And this is also our input data in which sentiment analysis is done.

These fetched datasets (tweets) are stored in HDFS. The following figure shows the tweets data in flume directory and this is the list of twitter data extracted which contains the keyword as specified in the configuration file. We can check the files by downloading them and seeing the tweets relating to the keyword. Here our keyword or the data fetched from twitter is of Virat.

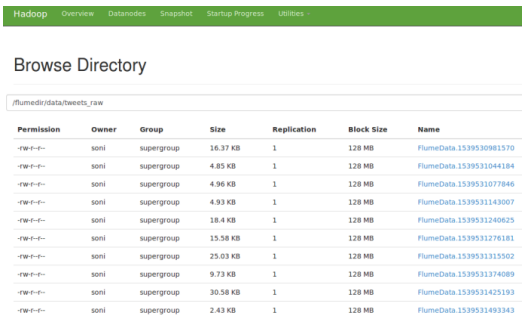

Figure 2: Data in Flume directory

The sentiment of the tweet is calculated by using polarity. Our output shows the sentiment of the tweets ie. Whether the tweet has positive nature, negative nature or it is neutral in nature. The output table consists of the tweet id and the sentiment. As every tweet is having its unique id so it is easy to analyze the sentiment of every tweet.

Tweet ID	Sentiments
4042167955965161472	Neutral
4042167960763498497	Neutral
4042168084713455616	Positive
4042168114283413506	Positive
4042168196848287745	Negative
4042169212901654528	Negative

Figure 3: Sample of Output

V. CONCLUSION

There are different ways to analyze Twitter data (unstructured data, Semi Structured) and want to analyze sentiments of twitter data. We have used Hive and Flume. Hive and Flume are the tools of Bigdata Hadoop and are efficient for extracting and loading the data from structured data and unstructured data. Hive is basically SQL like tools used for managing and queering unstructured data where as Flume is reliable and is available for collecting, aggregating and moving large amount of streaming event data. There are different methods real time streaming data by using codes or using Mapreduce etc. In this paper 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.

REFERENCES

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

[2]. Pooja S. Patil1, Pranali B. sable, Sentiment Analysis on Twitter Data Using Apache Flume and Hive, IRJET, Feb-2016.

[3]. Mahalakshmi R, Suseela S, "Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data".

[4]. Sunil B. Mane , Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde , "Real Time Sentiment Analysis of Twitter Data Using Hadoop".

[5]. Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal of Core Engineering & Management (IJCEM).

[6]. Mr. Sagar Nadagoud, Mr. Kotresh Naik.D, "Market Sentiment Analysis for Popularity of Flipkart", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET).

[7]. Online Resource of Hive Available on: http://hive.apache.org/

[8]. Online Resource of Flume Available on: https://flume.apache.org/

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

[10]. "MapReduce: Simplified Data Processing on Large Clusters," by J. Dean and S. Ghemawat.