Semantic Analysis of Twitter Data Using Hadoop

Sadiya Nazrana Syed Asad[1*], Ms. Vaishali Bhagat[2]
[1, 2] Dept of Computer Science and Engineering
[1, 2] P.R.Pote (Patil) Education & Welfare Trust's Group of Institutions's
College of Engineering and Management, Amravati, Sant Gadge
Baba Amravati University, Amravati, India.

Abstract- *Blogging and networking platforms like Facebook, Reddit, Twitter and LinkedIn are social media channels where users can share their thoughts and opinions.These opinions can be mined using various technologies and are of most importance to make predictions since they directly convey the viewpoint of the masses.Twitter contains a huge volume of data, storing and processing this data is a complex problem. Hadoop is a big data storage and processing tool for analyzing data with 3Vs, i.e. data with huge volume, variety and velocity.This paper discuss how to use FLUME and HIVE tool for twitter post analysis.*

I. INTRODUCTION

In this paper we are going to study system which will help us to analyze data semantically for different purposes by using Hadoop. let first we will understand why semantic analysis is important.

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as **opinion mining**, deriving the opinion or attitude of a speaker.

Business: In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.

Politics: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!

Public Actions: Sentiment analysis also is used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. On **Twitter**, each year the number of tweets flowing in is increasing rapidly. The data that is tremendous which is Big data is very hard to process due to its enormity so we use Hadoop technology.**Hadoop** is undoubtedly the preferred choice for such a requirement due to its key characteristics of being reliable, flexible, economical, and a scalable solution. While Hadoop provides the ability to store this large scale data on HDFS (Hadoop Distributed File System), there are multiple solutions available in the market for analyzing this huge data like MapReduce, Pig and Hive. With the advancements of these different data analysis technologies to analyze the big data, there are many different school of thoughts about which Hadoop data analysis technology should be used when and which could be efficient.For doing twitter data analysis first data is collected using FLUME in local HDFS after that HIVE can be used for twitter posts analysis.

II. PROBLEM STATEMENT

Existing System: Previously they used Lexicon-based system, There are some steps of lexicon-based that is used in this research, such as determining the polarity of words, negation handling, and also giving score to every each entity in the tweet For those they are going to download the libraries that are provided by the twitter guys by using this they are crowded the data that we want predominantly. After getting raw data they will filter it and they will find out the constructive, 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 twitter analysis on the filtered data. These words can be called as a dictionary set by which they will perform twitter analysis. There was so many problems and limitations in creating tables and also accessing table effectively from RDBMS.

Proposed System: To overcome these problem of large scale data we will use Hadoop for analyzing data semantically. All these will be saved into HDFS (Hadoop Distributed File System) in our prescribed format. project is effective using the Hadoop ecosystems and how the data is going to store from the Flume, also how it is going to create tables using Hive also how the sentiment analysis is going to perform.

Given figure shows us how data will flow:

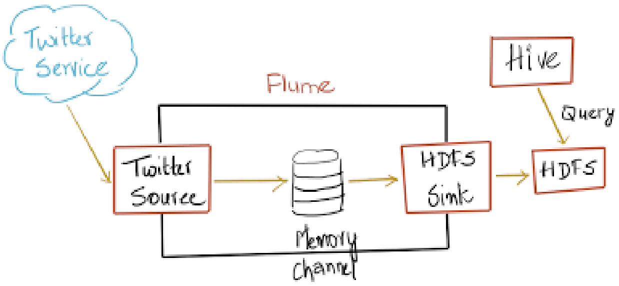

Fig2.2: Twitter data flow

III. METHODOLOGY

Extracting Twitter Data with Apache Flume:

By using flume data will be extracted from twitter. It is an open source component which is designed to locate and store the data in a distributed environment and collects the data as per the specified input key.

Querying Data with Hive: It will analyze the data.

IV. PLAN OF WORK

This project involves the following mod:

- Collecting the required input data from twitter which has to be dumped into HDFS file.
- Setting the cluster to store the data we have to make cluster of data and then it data will store in nodes.
- Transfers the data into hive data warehouse Dumping the chunks of data from module I to the cluster [18] file system in all data nodes (hdfs, Hadoop distributed file system) using flume from the name node by forming the data pipeline between them.

V. INTRODUCTION TO HADOOP

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop framework includes different modules like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase. I will be using FLUME and HIVE for twitter analysis.

HDFS (Hadoop File System) was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

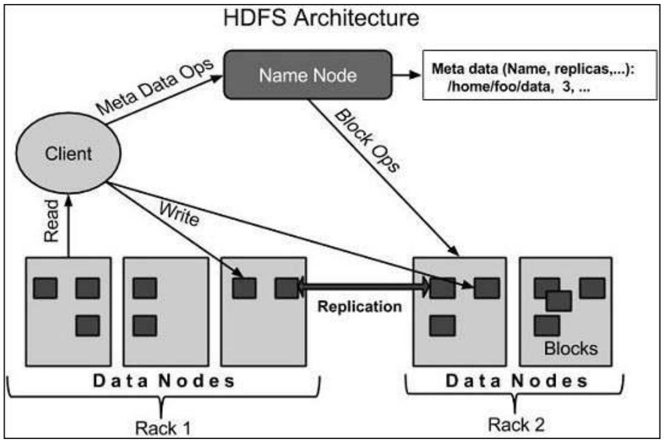

Fig2.1: HDFS Architecture

HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data. Benefit of using Hadoop is distributed storage, Distributed Processing,Security, Reliability, Speed, Efficiency, Availability, Scalability and lots more. This is the reason of using Hadoop for tweet processing.

VI. FLUME and HIVE

Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data from various web servers to HDFS.The data in these agents will be collected by an intermediate node known as Collector. Just like agents, there can be multiple collectors in Flume.

Finally, the data from all these collectors will be aggregated and pushed to a centralized store such as HBase or HDFS. The following diagram explains the data flow in Flume.

- In the flume configuration file we need to:
- Name the components of the current agent.
- Describe/Configure the source.

- Describe/Configure the sink.
- Describe/Configure the channel.
- Bind the source and the sink to the channel.

For naming components:

- agent_name.sources = source_name
- agent_name.sinks = sink_name
- agent_name.channels = channel_name

Now we are going to extract data from twitter:
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

Now we will see the function of **source**, sinkand channel :

Each source will have a separate list of properties. The property named “type” is common to every source, and it is used to specify the type of the source we are using.

Along with the property “type”, it is needed to provide the values of all the required properties of a particular source to configure it,

- agent_name.sources. source_name.type = value
- agent_name.sources. source_name.property2 = value
- agent_name.sources. source_name.property3 = value

For twitter we can write:

- TwitterAgent.sources.Twitter.type = Twitter (type name)
- TwitterAgent.sources.Twitter.consumerKey =
- TwitterAgent.sources.Twitter.consumerSecret =
- TwitterAgent.sources.Twitter.accessToken =
- TwitterAgent.sources.Twitter.accessTokenSecret =

Just like the source, each **sink** will have a separate list of properties. The property named “type” is common to every sink, and it is used to specify the type of the sink we are using. Along with the property “type”, it is needed to provide values to all the required properties of a particular sink to configure it, as shown below.

- agent_name.sinks. sink_name.type = value
- agent_name.sinks. sink_name.property2 = value
- agent_name.sinks. sink_name.property3 = value

For twitter we can write:

- TwitterAgent.sinks.HDFS.type = hdfs (type name)
- TwitterAgent.sinks.HDFS.hdfs.path=HDFS directory’s Path to store the data.

Flume provides various channels to transfer data between sources and sinks. Therefore, along with the sources and the channels, it is needed to describe the channel used in the agent.

- agent_name.channels.channel_name.type = value

For twitter we can write:

- TwitterAgent.channels.MemChannel.type= memory (type name).

So in this way we will configure our flume and then we can extract data easily from twitter then by binding channel with source and sink we can start extracting.

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

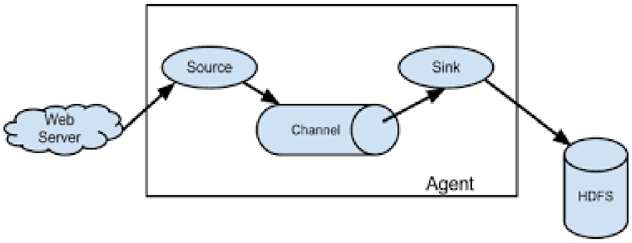

Fig3.1 Flume flow chart

VII. CONCLUSION

This project will give us hands on experience of handling and parallel processing of huge amount of data. Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly. This project will help us not only to gain knowledge about installation and configuration of Hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyze sentiment, or sentiment analysis.

The future of this data analysis field is vast. This project not only analyses the sentiments of the user but also computes other results like the user with maximum friends/followers, top tweets etc. hence Hadoop can also be effectively used to compute such results in order to determine the current trends with respect to particular topics. This can be very useful in the marketing sector.

VIII. FUTURE WORK

In this paper it has shown the way for responsibility sentiment t analysis for Twitter data. Also, we can do this by using re-tweeted by creating work flow so that it can givea 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] Tarun R R1 ; Sahana J S2 ; Sadvik B S3 ; Shashank S4 ; Prof. Mahesh T R5”CONTEXT BASED SENTIMENT ANALYSIS OF TWITTER USING HADOOP FRAMEWORK”,International Journal of Computer Science and Mobile Computing,IJCSMC, Vol. 8, Issue. 5, May 2019, pg.193 – 202.

[2] Y.Sushmitha Reddy*P*, M.Padma,”Sentiment Analysis of Twitter by using Apache Flume”,IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 9, September 2016.

[3] www.ijiset.com

[4] Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari,”Sentiment Analysis of Twitter Data”, International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015.

[5] Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde,” Real Time Sentiment Analysis of Twitter Data Using Hadoop”, Sunil B. Mane et al, /(IJCSIT) Internationa l Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100.

[6] M. Bouazizi and T. Ohtsuki, “Sentiment Analysis in Twitter: from Classification to Quantification of Sentiments within Tweets,” in Proc. IEEE ICC, May 2016.

[7] Srishti Sharma, Shampa Chakraverty and Akhil Sharma. A context-based algorithm for sentiment analysis. In International Journal of Computational Vision and Robotics, Vol. 7, No. 5, 2017, Netaji Subhas Institute of Technology, Dwarka, New Delhi, India, 2017.

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