Cloud Assisted Trend Analysis of Twitter Data using Hadoop

Aman Gupta
vasugupta42@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Ishaan Bhasin
ishaanbhasin96@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

Narinder Kaur
er.nkaur1@gmail.com
Maharaja Agrasen Institute of Technology, Rohini, Delhi

ABSTRACT

In today's highly developed world, every minute, people around the globe express themselves via various platforms on the Web and in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites, such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets, and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses.

Here we propose to analyze the sentiments of Twitter users through their tweets in order to extract what they think and find out the most trending events happening around the globe.

Hence we are using hadoop and amazon web services for sentiment analysis which will process the huge amount of data on an amazon ec2 hadoop cluster faster.

Keywords: Opinion Mining, Sentiment analysis, Hadoop Cluster, Twitter, Unstructured data, Amazon Web Services, Hive, Flume, Putty, WinSCP, Amazon EC2 computing, Portability.

1. INTRODUCTION

Today we are living in the world which is surrounded by 99% of data. There are different micro blogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource. Hence, there is a need to develop a product which can analyze opinions of people. This product will be useful in increasing market value of industries as well as satisfy needs of customers.

Performing Sentiment Analysis on Twitter is trickier than doing it for large reviews. This is because the tweets are very short (only about 140 characters) and usually contain slangs, emoticons, hash tags and other twitter specific jargon. For the development purpose twitter provides streaming API which allows the developer an access to 1% of tweets tweeted at that time bases on the particular keyword. The object about which we want to perform sentiment analysis is submitted to the twitter API’s which does further mining and provides the tweets related to only that object. Twitter data is generally the unstructured i.e. use of abbreviations is very high. Also, it allows the use of emoticons which are direct indicators of the author’s view on the subject. Tweet messages also consist of a timestamp and the user name. This timestamp is useful for guessing the future trend application of our project. User location if available can also help to gauge the trends in different geographical regions.
HADOOP - The Apache Hadoop project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is a good choice for Twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large-scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes.

APACHE FLUME - Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS.

APACHE PIG - Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

AMAZON WEB SERVICES (AWS) - Amazon Web Services (AWS) is a subsidiary of Amazon.com that provides on-demand cloud computing platforms to individuals, companies, and governments, on a paid subscription basis. The technology allows subscribers to have at their disposal a full-fledged virtual cluster of computers, available all the time, through the Internet.

AMAZON ELASTIC COMPUTE CLOUD (EC2) - Amazon Elastic Compute Cloud (EC2) forms a central part of Amazon.com's cloud-computing platform, Amazon Web Services (AWS), by allowing users to rent virtual computers on which to run their own computer applications. EC2 encourages scalable deployment of applications by providing a web service through which a user can boot an Amazon Machine Image (AMI) to configure a virtual machine, which Amazon calls an "instance", containing any software desired. A user can create, launch, and terminate server-instances as needed, paying by the second for active servers – hence the term "elastic".

2. OBSERVATION

Hadoop and its Ecosystems, for getting raw data from Social Network preferably Twitter, we may use Hadoop online streaming using Apache Flume, this tool configures everything from fetching of Data to a storage location. All Data will be stored in our HDFS in JSON format which we have installed on our Amazon EC2 micro cluster.

3. PROBLEM DEFINITION

Social media is one of the popular media right now to share opinions or variety of topics and twitter is a very popular social site to share everything related to opinions on a variety of topics and discussions on current issues. These tweets generate the huge information related to a different area like government, election, etc. millions of tweets are generated every day and which is very useful in decision making because everyone is sharing their view and opinions on issues or variety of topics. Twitter sites receive petabytes of data every day and these data is nothing but a collection of tweets so these data is very important in real life to analyze different scenario through which it helps us in decision making. The analysis of Twitter data gives a real view or different user opinions regarding what they think and to analysis these data provide a better way for making any decision.

The second main issue is the portability and availability of the system, it needs a running desktop or laptop and a power source to access the database.
4. PROPOSED WORK

For analyzing the large and complex data set a powerful tool is required, we are using hadoop which is an open source implementation of MapReduce algorithm designed by Google for deep analysis and transformation of large data set. This paper we design algorithm for handling the problems raised by larger data volume and dynamic characteristics for finding and performing an operation on social media data sets.

To take in account the portability and availability of system, we used a EC2 micro cluster with standard specifications and ubuntu server and installed hadoop in single node mode which works 24*7 to solve the challenges of big data through Map-Reduce framework where the complete data is mapped to frequent data sets and reduced to smaller sizable data to ease of handling, after this we integrate hadoop ecosystem e.g. Flume and Hive on top of Hadoop running server machine.

5. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow:

- In the first step, we will create an Amazon ec2 cluster with the micro instance and ubuntu 64-bit server which will be running 24*7 for readily access by any virtual terminal example PUTTY.
- We install hadoop, Apache Flume, Apache Hive and configure them to fetch, process and store data.
- We create a twitter app using twitter streaming API to fetch real-time twitter data.
- For doing twitter data analysis first data is uploaded using FLUME in local HDFS. The twitter API used in Flume, through which all the tweets are directly fetch from the twitter site and stored it in the HDFS. Data comes from the twitter site is in a un-structure form called JSON data.
- After storing all Twitter data into the HDFS we are performing the analysis part for these we use hive through which we can convert the un-structure complex data into readable or understandable structure form.

6. EXPERIMENTAL AND RESULT ANALYSIS

All the experiments were performed on an AWS AMAZON EC2 micro cluster running ubuntu server 16.04 LTS and hadoop version 2.8 with 10GB hard disk and 1 GB RAM.

To achieve the proposed methodology we are going to follow the following methods:

- Creating Twitter Application

First of all if we want to do opinion analysis on Twitter data we want to get Twitter data first so to get it we want to create an account on 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.

- 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. All the details we have to fill in the flume-twitter.conf file is shown in the figure.

- Analyze using apache HIVE

After running the Flume with configuration, the Twitter Data will automatically save into HDFS sink where we have set the path storage to save Twitter data that was fetched by Flume. From this data, 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 we have converted the unstructured data into a structured format. For this, we want to use some custom Serde concepts.

Core Analysis Scheme:

a. Tokenization

All the words in a tweet are broken down into tokens. This is the tokenization process. For example, '@Jack That is an awesome car!' is broken down into individual tokens such as '@Jack', 'That', 'is', 'an', 'awesome', 'car'. Emoticons, abbreviations, hash tags, and URLs are recognized as individual tokens. Each word in a tweet is separated by a space. Therefore, on encountering a space, a token is identified.

b. Normalization

The normalization process verifies each token and performs some computing based on what kind of token it is. If the token is an emoticon, its corresponding polarity is taken into account by searching the emoticon dictionary. If the token is an acronym, it is checked in the acronym dictionary and the full form is stored as individual tokens.
c. Sentiment Classifier

The tweets are broken down into tokens where each token is assigned polarity which is a floating point number ranging from 1 to -1.

A. Positive Tweets

Positive tweets are the tweets which show a good or positive response towards something. For example tweets such as —It was an inspiring movie!!!! or —Best movie ever.

B. Negative Tweets

Negative tweets can be classified as the tweets which show a negative response or oppose towards something. For example tweets such as —Waste of time or —Worst movie ever.

C. Neutral Tweets

Neutral tweets can be classified as the tweets which neither show a support or appreciate anything nor oppose or depreciate it. It also includes tweets which are facts or theories. For example tweets such as —Earth is round.

7. TIME EFFICIENCY

Time efficiency is an important aspect where our project scores well. Lower response time has achieved by use of data structures as local variables. This reduces the access time from a hard-disk. Also, the use of Hadoop ensures the distributed processing and it also lowers the access time. Hence overall the time efficiency increases owing to the above-mentioned factors.

8. FUTURE SCOPE

At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also, the use of native language combined with English usage is difficult to interpret.

9. CONCLUSION

Sentiment analysis is a very wide branch of research. We have covered some of the important aspects. We plan ahead to improve our algorithm used for determining the sentiment value. Also, the project as of now can also be expanded to other social media platform usages like movie reviews (IMDB reviews), personal blogs. The accuracy achieved is also mentioned below, emoticons and the use of hashtags for the sentiment evaluation is a very important inference related to sentiment analysis of social media data.
Also the use of hashtags to determine the most trending tweets and filter the results to sentiment analysis the tweets related to the most tweeted topic only.

By running the hadoop cluster on EC2 cluster the computation power and results showed a dramatic change to that of running on a domestic laptop or desktop.

Hence with the current limitations, the accuracy is found to be 72.27 %.

9. REFERENCES

[1] Apoorv Agarwal, Jasneet Singh Sabarwal, “End to End Sentiment Analysis of Twitter Data”
[2] Theresa Wilson, Joanna Moore, Efthymios Kouloumpis, “Twitter Sentiment Analysis – The Good, the Bad and the OMG”