

Twitter data analysis using hadoop ecosystems and apache zeppelin

Stanly Wilson, Sivakumar R

Department of Computer Science, Christ (Deemed to be University), India

Article Info

Article history:

Received Dec 20, 2018

Revised Mar 17, 2019

Accepted Apr 11, 2019

Keywords:

Flume

Hadoop

HDFS

JSON

Pig

Twitter

Tweets

Zeppelin

ABSTRACT

The day-to-day life of the people doesn't depend only on what they think, but it is affected and influenced by what others think. The advertisements and campaigns of the favourite celebrities and mesmerizing personalities influence the way people think and see the world. People get the news and information at lightning speed than ever before. The growth of textual data on the internet is very fast. People express themselves in various ways on the web every minute. They make use of various platforms to share their views and opinions. A huge amount of data is being generated at every moment on this process. Being one of the most important and well-known social media of the present time, millions of tweets are posted on Twitter every day. These tweets are a source of very important information and it can be made use for business, small industries, creating government policies, and various studies can be performed by using it. This paper focuses on the location from where the tweets are posted and the language in which the tweets are written. These details can be effectively analysed by using Hadoop. Hadoop is a tool that is used to analyze distributed big data, streaming data, timestamp data and text data. With the help of Apache Flume, the tweets can be collected from Twitter and then sink in the HDFS (Hadoop Distributed File System). These raw data then analyzed using Apache Pig and the information available can be made use for social and commercial purposes. The result will be visualized using Apache Zeppelin.

*Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Stanly Wilson,

Department of Computer Science,

Christ (Deemed to be University), India.

Email: stanly.wilson@mca.christuniversity.in

1. INTRODUCTION

The world that we live in has been influenced by the social media. In this modern era, the growth of mobile devices has made it easy to access the internet from anywhere, and internet connect people with one another. They express their views and emotions, on the platforms that are given by the social media. Social media users are increasing day by day. They share their personal feelings, reviews of things they use and the feedback they provide are sources for decision making [1]. They generate huge unimaginable amount of data that can be accessed over the net. If it is used in an efficient manner, then it can bring results of various kinds. For the companies and enterprises, these information render help to design their products for the future, since the customer reviews are a sure means of improvement. Governments can use this information for making policies and to understand the problems faced by people. It is useful during the election time for exit polls and understanding the mindset of the people. Being one of the most important social media of our time, Twitter contains a variety of information. Millions of tweets are sent and received by Twitter on a daily basis. The analysis of tweets is a bit tricky because of its limited character length, and they are very often abbreviations, emoticons, argots etc. Twitter provides the possibility to extract 1% of the tweets that are tweeted at that particular time. Based on the keywords given to the Twitter API the tweets are downloaded

and the tweets are sources of various information [2]. Big data is a term that is used to refer a large amount of data which can't be processed using traditional database techniques. Tweets extracted from Twitter are big data considering the huge amount of data. At this juncture, the Hadoop ecosystem comes in as a handy tool since it provides the possibility to extract the tweets, store and analyse it.

1.1. Hadoop

Hadoop is an Apache open-source software for big data analysis. It is scalable, reliable and works in a distributed environment. It is capable of processing a large number of data sets using many clusters of computers [3]. Hadoop is a very good choice to analyse the Twitter data since the range in which Hadoop work is very high. Hadoop has its origin in the Google File System (GFS) and Google's MapReduce. There are two main components in Hadoop [4]. They are Hadoop Distributed File System, known as HDFS which is specific to Hadoop, and an algorithm called MapReduce [5]. MapReduce has two different tasks, and they are the map task and the reduce task. The Map task collects the data and divides it into tuples which form key pair value for the given task [6]. The output of the Map task is then given as input to the Reduce task and the key pair values are combined to form smaller tuples [7]. For this paper the Hadoop is installed in a single cluster mode, means Hadoop is deployed on one system alone. There are several nodes that perform the map and reduce. Hadoop has a master-slave architecture. The NameNode is the master server that regulates the access, JobTracker coordinates the parallel processing of data using MapReduce and the DataNode does the computational works and storing of data. Multiple nodes can easily be attached to Hadoop that can handle huge clusters of data [3]. Hadoop is very reliable since it has features of data duplication. Hadoop keeps multiple copies of the same data as it tries to overcome the hardware failure, and the data is never lost. Hadoop performs the processes in a fault-tolerant manner and it is very much reliable [8, 9].

1.2. Apache Flume

Apache Flume is designed to push a large amount of data from various resources to the HDFS. It is built on Hadoop itself. There are situations in Hadoop to deal with a large number of data and a large number of servers may be employed to it. Issues may arise when multiple systems try to write or access HDFS at the same time. Flume provides a solution to these issues. It is designed to be verily customizable and it can be scaled out very easily. The simplest unit of Flume is an agent, and one Flume agent can be connected to multiple agents. Flume agent has three components and they are the source, channel and sink. The source collects the events to the agent, the channel acts as a buffer that store data from source before writing into the HDFS, and sinks is responsible to store the data to HDFS [10]. It dumps data in HDFS in different formats. In this paper, it is configured to extract and store the data in HDFS in JSON (JavaScript Object Notation) format.

1.3. Apache Pig

Apache Pig is created by Yahoo for writing programs for Hadoop to deal with large sets of data present in the Hadoop cluster. Pig is a scripting language known as Pig Latin. In Hadoop, MapReduce functions are written in Java. Pig Latin comes as an alternative to this and gives the developers the possibility to spend comparatively less time with programming in Java and invest more time in analysing the same. Pig can deal with a problem which would take only 5% of the actual MapReduce program of Hadoop [11]. Pig can handle any type of data with its powerful programming language. Pig scripts can be written in any text editor and could be executed in Hadoop frameworks like MapReduce or Apache Tez. In Pig, extract transform model (ETL) is employed to collect data from various sources, and then stores in HDFS [12]. Pig can run in two modes. One is the local mode and the other in the MapReduce mode. For the implementation, the pig script would be executed in the MapReduce mode.

1.4. Apache Zeppelin

Visualization of data is of similar importance as that of analysing the data. Visualization of data provides for better understanding. Apache Zeppelin is one of that kind which is quite helpful and powerful. It is an open-source, web-based 'notebook' that enables interactive data analytics and collaborative documents. It is used for data ingestion, discovery, analysis and visualization. It has an interpreter concept that allows many languages to be plugged in and processed in its environment like Python, Spark, Hive, Flink, R and others [13]. In this paper, Apache Zeppelin is used only for the data visualization and it runs the Pig Latin Script in its environment.

2. LITERATURE REVIEW

The paper [8] explore into the world of big data. It tries to understand the properties of big data like velocity, volume and variety. They discuss the difficulties in the area of big data such as analysis of customer behaviour, customer recommended system etc. The paper speaks about two important tools. First, it discusses Hadoop and its major components along with the technologies built upon Hadoop. The second technology is NoSQL. It concludes by saying the benefits of Hadoop like the ability for load balancing, scalability and cost efficiency. The paper [14] places its focus on how to perform the sentimental analysis on Twitter using Apache Pig. The tweets are extracted from the Twitter using the Twitter API using Flume, and then they are loaded into the HDFS. It makes use of some dictionary which groups the tweets into three categories like positive, negative and neutral based on their polarity with the words in the dictionary.

The paper [15] brings out a framework for the sentimental analysis of social media. They propose Hadoop as a very efficient and effective tool with the ability to work on both unstructured and semi-structured data. It elaborates on the HDFS and MapReduce aspects of Hadoop. It analyses various existing frameworks that were proposed earlier with Hadoop as their platform and bring out their merits and demerits.

The paper [16] looks at the big data aspect of the tweets. The paper deal in detail with the terminology 'big data' and its main characteristics. The main issues with the big data are the difficulty to store the data, retrieve the data and process it. It proposes Hadoop as a solution to deal with the issues brought by the big data with its HDFS and MapReduce algorithms.

The paper [17] shows how the social media has changed the way the world thinks and decides, and how these data could be used to make predictions on various areas. The paper demonstrates the Twitter data analysis using Hive and Flume of the Hadoop ecosystem. The implementation of the data analysis with its code on Flume and Hive provide a glimpse of the actual working. The results show how often a word is repeated in the tweets and how popular one word from the other.

The paper [18] designed a method to perform the sentimental analysis using the machine learning techniques. Here the data is stored in MongoDB, a NoSQL database. The stored data is pre-processed to remove the unwanted symbols and stop words. Different clustering algorithms (K Means, Birch and Cure) are used, to which the pre-processed tweets are given as input. The paper makes an observation that Cure performs better than K Mean and Birch on a large number of tweets.

The paper [19] concentrates on the Twitter data to study the popularity of Flipkart. The proposed methodology is getting raw data to HDFS, using the streaming API available in Flume and store this raw data in HDFS. This data is tabularized and structured into tables by using Hive. Hive, with its queries, analyse raw data stored in HDFS. As a data dictionary for the analysis, it uses Stanford Core NLP. It also categorizes the sentences into three categories, and they are positive tweets, negative tweets and neutral tweets.

In this paper [20] the emphasis is on the execution time of the analysis. It studies the existing systems and its drawbacks. The tweets are extracted and stored in Hadoop and pre-processed. It makes a comparative study in 4 stages. The tweets are considered in different sizes and number of tweets. It is done with and without Hadoop. When the size of data is less than 5MB the execution is faster without Hadoop. As the size of the data increase, with Hadoop the execution time gets reduced very much.

The paper [10] elaborately looks into the emerging technologies in the field of big data. It takes Twitter as a source for big data and discusses various analysing possibilities with Hadoop ecosystems. Hadoop infrastructure is dealt in detail with a brief explanation to the Hadoop ecosystem technologies like Apache Flume, Apache Zookeeper, Apache Hive, Apache Pig, Apache HBase and Apache Sqoop. The paper uses Hive for the analysis. The system is deployed in the Cloud services provided by Microsoft Azure. The paper brings out a fact that as the number of nodes increases, the execution time for hive query also increases.

In [21] focuses on the analysis of the data that are available especially the Twitter data. The traditional methods of storing and analyzing data is not a good option to do the sentimental analysis since there are very large unstructured data sets that are under consideration. The RDBMS cannot handle unstructured data. So, the Hadoop architecture is used to get the work done. The methodology used in the paper is to get the tweets into HDFS using the API available with Flume and sink it in HDFS. Then hive queries are used to get unstructured data into structured form so that various operations can be performed in it.

3. IMPLEMENTATION

All the analysis is performed on a system having i5-6200U CPU @ 2.30 GHz processor and RAM of 16 GB running in the Ubuntu OS. There are some steps to be followed for Twitter data analysis. They are creating a Twitter application which provides the necessary permissions required for extracting the tweets from Twitter. After obtaining the permissions, Flume is used to extract the tweets from Twitter and they are

processed and analysed using Pig [22]. There are a lot of dependency issues pops up while building on Hadoop. Hadoop is the base and Flume and Pig are tools built on Hadoop. One must be very careful while choosing the version of these tools because all the versions of Pig and Flume do not work on the corresponding versions of Hadoop. For this paper the version of Hadoop is 2.7.1, Flume is 1.6.0, Pig is 0.16.0 and that of Zeppelin is 0.8.0. Moreover, there need to be proper library files for both tools. The detailed procedure is given below.

3.1. Creation of Twitter Application

It is necessary to have a Twitter developer account to create a Twitter application. Those who are registered users of Twitter can use the same login information to create a Twitter developer account by visiting the webpage <https://apps.twitter.com>. Here the users can create, manage and delete the twitter apps. By entering to the site, users can see the option to create a new app and then provide the required information as shown in the Figure 1.

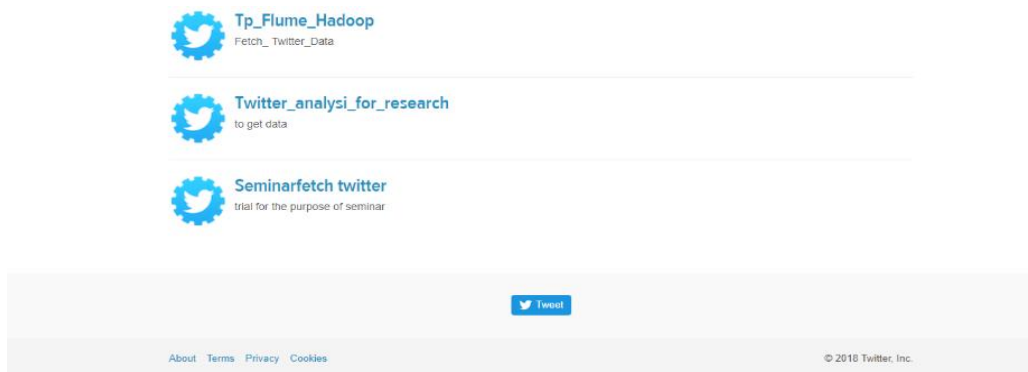


Figure 1. Creation of Twitter Application

Once the application is created, the four required keys will be generated with it. The keys are Consumer key, Consumer secret, Access token and Access token secret. Consumer key is basically the API key which recognizes the user. A user here refers to an entity, either a service or website that tries to use the resources of the application (Twitter). Consumer secret is the authentication password used by the server to identify the user. Access token tells the privileges the user has. Access token secret is the information sent along with the Access token to the application as a password when a user tries to get its resources. Consumer key and Consumer secret are used for authentication while Access token and Access token secret are meant for the authorization [23]. These keys provide users access to the Twitter and extract the tweets. The generation of the keys is given in Figure 2.

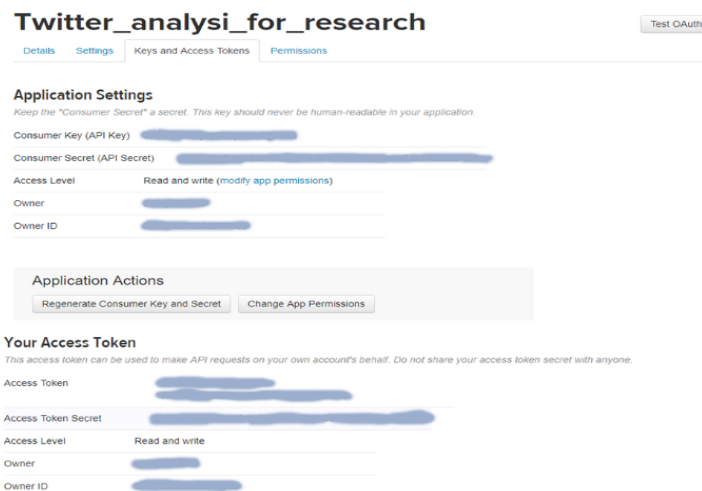


Figure 2. Generating the Consumer and Access keys

3.2. Tweets Extraction using Flume

The keys obtained from Twitter using the developer's account are to be placed in the configuration file located in the Flume installation folder. In this specific configuration file, there are various parameters that can be set like the keywords, language filter, the block size of the extracted files etc. The keywords are the words on which the tweets will be extracted [24]. The tweets that are extracted for in this analysis of this paper were live tweets that were posted at the particular time of downloading and not any tweets prior to it. For the implementation, a custom file is included in the Flume which extracts the tweets in the JSON. The configuration file containing the details for tweets extraction is given in Figure 3.

```

TwitterAgent.sources= Twitter
TwitterAgent.channels= MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type =
    com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.consumerKey = xxxxxxxxxxxx
TwitterAgent.sources.Twitter.consumerSecret = xxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessToken = xxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessTokenSecret = xxxxxxxxxxxx
TwitterAgent.sources.Twitter.keywords= @FIFAWorldCup

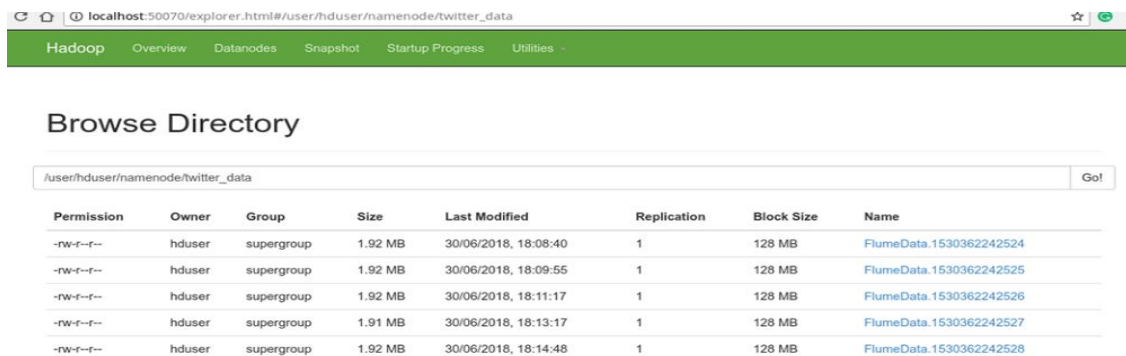
TwitterAgent.sources.Twitter.maxBatchSize = 25000
TwitterAgent.sources.Twitter.maxBatchDurationMillis = 60000
TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=namenode/twitter_data
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 10000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 2000000
TwitterAgent.sinks.HDFS.hdfs.rollCount = 1500000
TwitterAgent.sinks.HDFS.hdfs.rollInterval=6000
TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=5000
TwitterAgent.channels.MemChannel.transactionCapacity=5000
TwitterAgent.sources.Twitter.channels=MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel

```

Figure 3. Flume configuration file

Flume with its three agents mentioned earlier extract the tweets and then sinks it in the HDFS location configured in the configuration file. The tweets downloaded for this work is around 23 GB of data and more than 4.4 million records. All the blocks are of similar size of 1.91 MB each. The tweets are taken from the official page of the 'FIFA World Cup 2018' [25], and all the tweets used in this work are extracted during a match of FIFA World Cup 2018. The HDFS files can be viewed in the browser as shown in Figure 4.



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	1.92 MB	30/06/2018, 18:08:40	1	128 MB	FlumeData.1530362242524
-rw-r--r--	hduser	supergroup	1.92 MB	30/06/2018, 18:09:55	1	128 MB	FlumeData.1530362242525
-rw-r--r--	hduser	supergroup	1.92 MB	30/06/2018, 18:11:17	1	128 MB	FlumeData.1530362242526
-rw-r--r--	hduser	supergroup	1.91 MB	30/06/2018, 18:13:17	1	128 MB	FlumeData.1530362242527
-rw-r--r--	hduser	supergroup	1.92 MB	30/06/2018, 18:14:48	1	128 MB	FlumeData.1530362242528

Figure 4. Tweets in HDFS

There are many attributes in a single tweet, and the downloaded tweets are in unstructured format. A sample of downloaded unstructured tweet is shown in Figure 5.

```

1 {"extended_tweet":{"entities":{"urls":{},"hashtags":{},"user_mentions":{"indices":[0,9],"screen_name":"KingFut","id_str":"630897139","name":"KingFut.com","id":1
2 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:18 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
3 {"in_reply_to_status_id_str":"1009479635204235264","in_reply_to_status_id":"1009479635204235264","created_at":"Wed Jun 20 16:58:18 +0000 2018","in_reply_to_user_id
4 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:20 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
5 {"extended_tweet":{"extended_entities":{"media":{"display_url":"pic.twitter.com/YkjsNdrROM","indices":[175,198],"sizes":{"small":{"w":680,"h":680,"resize":"fit"
6 {"in_reply_to_status_id_str":"1009479748031057920","in_reply_to_status_id":"1009479748031057920","created_at":"Wed Jun 20 16:58:21 +0000 2018","in_reply_to_user_id
7 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:22 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
8 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:23 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
9 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:23 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
10 {"extended_entities":{"media":{"display_url":"pic.twitter.com/Ya4WUdA3le","source_user_id":"3112199881","type":"video","media_url":"https://pbs.twimg.com/ext_tw_vi
11 {"in_reply_to_status_id_str":"100933979215962112","in_reply_to_status_id":"100933979215962112","created_at":"Wed Jun 20 16:58:24 +0000 2018","in_reply_to_user_id
12 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:26 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
13 {"in_reply_to_status_id_str":"1009479860388093954","in_reply_to_status_id":"1009479860388093954","created_at":"Wed Jun 20 16:58:28 +0000 2018","in_reply_to_user_id
14 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:29 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
15 {"in_reply_to_status_id_str":"100947920944691969","in_reply_to_status_id":"100947920944691969","created_at":"Wed Jun 20 16:58:30 +0000 2018","in_reply_to_user_id
16 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:31 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
17 {"in_reply_to_status_id_str":"1009479635204235264","in_reply_to_status_id":"1009479635204235264","created_at":"Wed Jun 20 16:58:31 +0000 2018","in_reply_to_user_id
18 {"extended_tweet":{"entities":{"urls":{},"hashtags":{},"user_mentions":{"indices":[0,13],"screen_name":"FIFAWorldCup","id_str":"138372303","name":"FIFA World Cup
19 {"extended_entities":{"media":{"display_url":"pic.twitter.com/0qzq2Fec0A","indices":[100,123],"sizes":{"small":{"w":680,"h":453,"resize":"fit"},"large":{"w":204
20 {"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Jun 20 16:58:35 +0000 2018","in_reply_to_user_id_str":null,"source":"<a href='\"
21 {"extended_tweet":{"extended_entities":{"media":{"display_url":"pic.twitter.com/17Pzms2M0","indices":[250,273],"sizes":{"small":{"w":680,"h":611,"resize":"fit"

```

Figure 5. Twitter sample file stored in HDFS

3.3. Structure of the Tweet

In Twitter, the users post the opinions, add media, links and hashtags. It is collectively called as a tweet. A tweet may or may not contain user opinions. It can be just an image or a URL link. A specific tweet contains more than 150 attributes. There are four common objects in a tweet like user, place, entities and extended entities. The 'user' object provides the overview of the author of the tweet. Twitter gives an option to enable or disable the location. So, if the location is enabled then that information could be seen in the 'place' object. Tweets contain a collection of hashtags, URL, media, the user opinions etc, and they are in the 'entities' object. The native media of the tweets are in the 'extended' objects [26]. A brief understanding on each of these objects helps to understand the tweets better. Under each object, some important attributes are elaborated. The details of the attributes are taken from the official Twitter developer documents.

The 'user' object has the general information of the Twitter account and the author of the tweet. The attributes in this object are usually constant, means they remain the same for a specific user which do not change tweet to tweet. It has an 'id' which is unique for every user and with which every user can be uniquely identified. The 'screen_name' attribute is the name of the user as provided, and it need not be a name of a person. It acts like a username of the specific Twitter account. Twitter users can follow other Twitter accounts. The 'followers_count' gives the number of followers a particular account has, while 'friends_count' tells the number of accounts the specific user follows. The 'favourites_count' provides the number of tweets that are liked by the user in the timeline. The 'created_at' provide date and time of the creation of the specific Twitter account. The user's language information is provided by the attribute 'lang'. It tells the language known to the user and self-declared by the user. It does not tell anything about the content of the tweet or the language in which the tweets are written [27].

The 'place' object will be present in those tweets in which the geo-tagged is enabled. It provides the location with respect to the coordinates. The 'full_name' tells the name of the place. The 'country_code' tells the shortened name of the country where the location is, while 'country' gives the proper name of the country. The 'bounding_box' provide the geographical coordinates which enclose the particular place and it has four pairs of values. Coordinates are a pair of values consisting of the longitude and latitude that uniquely represent a particular geographical location. The 'type' tells the shape that is formed by these coordinates [28].

The 'entities' object has the following attributes like URL, hashtags, symbols, media and user opinions. They provide information on what the user tweets like the description in the tweet, the media that is added, the URL inserted, the hashtags mentioned and the symbols used. The 'hashtag' object contains the hashtags that are used in the tweet. This object will remain empty if no hashtags are present in the tweet. In this, the text field provides the name of the hashtag. The 'media' field gives the overview on the media, if present any, like the type of media, its dimensions and the URL tells if the media is taken from another link. The 'url' object contains the URL in the tweet if any. The 'display_url' gives the URL pasted in the tweet body. The 'user_mentions' field gives the user opinion mentioned in the tweets. If it is more than 140 characters then this will be present in the 'entities' field of extended tweets [29].

The 'extended' objects became a part of tweets in November 2017, and it comes into the picture when the limit of tweet extends between 140 characters to 280 characters. The 'extended' objects contain those longer messages. It could also contain 'entities' with a collection of hashtags, media and user opinions. There are retweet and quote tweet. A retweet is just resending the same tweets to the followers, while quote

tweet is sending the tweet with certain comments added to it. They have their fields similar to what has mentioned above like user, place and entities [26]. Among different objects that are present, the paper concentrates on the 'lang' and 'country' attributes. There are two 'lang' attributes. One is in the general section, and the other in the 'user' object. The paper focuses on the 'lang' in the general section which tells the language in which the tweets are written. The 'country' attribute is the attribute seen in the 'place' object.

3.4. Analysing the tweets using Apache Pig

The tweets that are downloaded and stored in HDFS are in raw form and in JSON format. Apache Pig uses a shell environment called 'grunt' and it has two modes as mentioned earlier. All the scripts are executed in the MapReduce mode. Apache Pig needs to be configured to process the JSON files. There are some jar files that need to be loaded into Pig [14] and they are

```
REGISTER '/usr/local/pig/elephant-bird-hadoop-compat-4.5.jar';
REGISTER '/usr/local/pig/json-simple-1.1.1.jar';
REGISTER '/usr/local/pig/elephant-bird-pig-4.5.jar';
REGISTER '/usr/local/pig/elephant-bird-core-4.5.jar';
```

These files are not the part of Pig and they need to be loaded every time when these scripts are executed. They also could be kept in the configuration files such that they would be loaded each time when the 'grunt' shell is launched. After loading these *jar* files, the tweets that are stored in the HDFS need to be loaded using the script

```
getTweets = LOAD hdfs://localhost:54310/user/hduser/namenode/twitter_data/FlumeData.*'
USING com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap;
```

All the blocks in HDFS start with 'FlumeData' and the above script loads all of them for the analysis. Here there are two things that need to be considered in particular. They are the 'id' and the 'lang'. The 'id' tells the unique id of the user and makes sure that the tweets are not duplicated in the analysis. The 'lang' part contains the language in which the tweets are written. The scripts that perform these actions are

```
split_tweets = FOREACH getTweets GENERATE myMap#'id' as id,myMap#'lang' as lang;
grouped= GROUP split_tweets BY lang;
wordcount = FOREACH grouped GENERATE group, COUNT(split_tweets);
dump wordcount;
```

The above scripts perform three actions. First, it maps the specific 'id' and 'lang' attributes of the tweets, and store it in the *split_tweets*. Secondly, *split_tweets* are grouped according to the 'lang' and are kept in *grouped*. Thirdly, the attributes are counted and saved in *wordcount*. Here *wordcount* contains the desired output. The 'dump' command is used to display the result. All the scripts mentioned here could be kept in a single file with the extension *.pig* and could be executed as a single script with the command *pig*. For example, *pig filename.pig*.

3.5. Results and Discussions

There are many languages in which the tweets are written. The result is the total number of tweets written in each language. The output of the execution is given the Figure 6.

(bn, 3)	(vi, 6163)
(eu, 2642)	(el, 1764)
(sl, 4)	(lt, 880)
(iw, 13)	(cy, 880)
(th, 16739)	(ht, 15888)
(de, 7045)	(in, 58168)
(fr, 33568)	(ca, 882)
(ru, 882)	(en, 3498958)
(ko, 24645)	(hu, 1)
(hi, 10548)	(lv, 1)
(tl, 7963)	(fi, 1762)
(pl, 13200)	(mr, 878)
(ja, 9695)	(nl, 5269)
(sv, 880)	(is, 881)
(cs, 5)	(ro, 881)
(fa, 4)	(ar, 8826)
(uk, 879)	(es, 357556)
(tr, 9668)	(it, 6172)
(ne, 1)	(no, 886)
(zh, 1761)	(da, 3)
(ml, 1)	(et, 4406)
(pt, 16775)	(und, 275665)

Figure 6. Total Count of Languages

Figure 6 shows the individual count of the languages in which the tweets are written. The languages are represented in *ISO 639-1* format. This two letter words format is used to identify a language uniquely. From the above, it can be inferred that 34,98,958 tweets are written in English (*en*) which takes the major portion of the whole tweet language. The second largest is Spanish (*es*) whose count it 3,57,556. The last entry is *und* which indicates the 'undetermined', means the mixture of languages may be used [30]. The visualization of the result using Apache Zeppelin is shown in Figure 7.

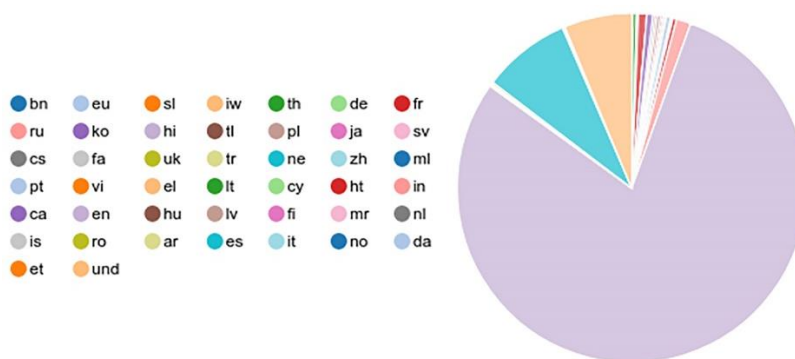


Figure 7. Visualization of the result using Zeppelin

The above analysis is made on the languages in which the tweets are written. The same could be done to analyse the locations or the countries from where tweets are written. There are times people do not provide the permissions to share or access their locations. In such situations, the tweets don't give the location or country information. The country information of tweets is in the 'places' object from where it needs to be fetched. The country from where these tweets are tweeted is shown in Figure 8, and it can be inferred that most of the tweets are written from United States (19344) and the next is from Ecuador (14048). Here only part of the result is shown.



Figure 8. Total Count of Countries

4. CONCLUSION

The analysis of tweets is not something new. It was there for quite some time. One of the most analysed aspects of tweets is the sentiment analysis. This research paper focused on two other attributes provided by the tweets like the language in which the tweets are written and the country details. These information present in tweets could be of great importance for the businesses and industries. They help to understand which language group, or people from which country respond to their products that are in the market or that will be launched soon. These days companies use location-based marketing, in order to promote and enhance their business. The details obtained from the tweets helps to figure out where to concentrate or advertise more in order to offer better service. An enhanced language and country analysis help to have sentiment analysis of users from a particular language group and location that could be used to provide a personalized experience for the users.

REFERENCES

- [1] P. Barnaghi, et al., "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets," *Conference ACM SIGKDD, ACM*, vol. 15, pp. 10-13, 2015.
- [2] A. Barskar and A. Phulre, "Opinion Mining of Social Data Using Hadoop," *International Journal of Engineering Science and Computing*, vol. 6, pp. 3849-3851, 2016.
- [3] K. Shvachko, et al., "The Hadoop Distributed File System," *IEEE 26th Symposium on Mass Storage Systems and Technologies*, pp. 1-10, 2010.
- [4] C. Kaushal and D. Koundal, "Recent Trends in Big Data using Hadoop," *International Journal of Informatics and Communication Technology*, vol. 8, pp. 39-49, 2019.
- [5] M. Wankhede, et al., "Analysis of Social Data Using Hadoop Ecosystem," *International Journal of Computer Science and Information Technologies*, vol. 7, pp. 2402-2404, 2016.
- [6] P. Ganesh, et al., "Performance Evaluation of Cloud service with Hadoop for Twitter Data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, pp. 392-404, 2019.
- [7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, pp. 107-113, 2008.
- [8] J. Singh and V. Singla, "Big Data: Tools and Technologies in Big Data," *International Journal of Computer Applications*, vol. 112, pp. 6-10, 2015.
- [9] D. C. Vinutha and G. T. Raju, "An Accurate and Efficient Scheduler for Hadoop MapReduce Framework," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, pp. 1132-1142, 2018.
- [10] A. Bhardwaj, et al., "Big Data Emerging Technologies: A Case Study with Analyzing Twitter Data using Apache Hive," *Proceedings of 2015 RA ECS UIET Panjab University Chandigarh*, pp. 1-6, 2015.
- [11] C. Olston, et al., "Pig Latin: A Not-So-Foreign Language for Data Processing," *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM*, pp. 1099-1110, 2008.
- [12] R. Singh and P. J. Kaur, "Analyzing Performance of Apache Tez and MapReduce with Hadoop Multinode Cluster on Amazon Cloud," *Journal of Big Data*, vol. 3, pp. 19, 2016.
- [13] A. MadhaviLatha and G. V. Kumar, "Streaming Data Analysis using Apache Cassandra and Zeppelin," *International Journal of Innovative Science, Engineering & Technology*, vol. 3, pp. 8-15, 2016.
- [14] A. Barskar and A. Phulre, "Opinion Mining of Twitter Data using Hadoop and Apache Pig," *International Journal of Computer Applications*, vol. 158, pp. 1-6, 2017.
- [15] A. P. Rodrigues, et al., "Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey," *International Journal of Computer Applications*, vol. 151, 2016.
- [16] M. K. Danthala, "Tweet Analysis: Twitter Data Processing using Apache Hadoop," *International Journal of Core Engineering and Management*, vol. 1, pp. 94-102, 2015.
- [17] Sangeeta, "Twitter Data Analysis using Flume & Hive on Hadoop Frame Work," *International Journal of Research in Advanced Engineering Technologies*, pp. 119-123, 2016.
- [18] G. S. Supraja, et al., "A Big Data Methodology for Sentiment Analysis of Twitter Data," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, pp. 7324-7329, 2015.
- [19] S. Nadagoud and K. Naik D., "Market Sentiment Analysis for Popularity of Flipkart," *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 4, pp. 2117-2123, 2015.
- [20] N. Patil, et al., "Twitter Sentiment Analysis using Hadoop," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, pp. 8230-8236, 2016.
- [21] C. Penchalaiah, et al., "Effective Sentiment Analysis on Twitter Data using Apache Flume and Hive," *International Journal of Innovative Science Engineering and Technology*, vol. 1, pp. 101-105, 2014.
- [22] P. Kirar, et al., "Opinion Mining of Twitter Data using Hive," *International Journal of Computer Applications*, vol. 156, pp. 44-49, 2016.
- [23] <https://stackoverflow.com/questions/28057430/what-is-the-access-token-vs-access-token-secret-and-consumer-key-vs-consumer-s>, 2018.
- [24] M. Wankhede, et al., "Location based Analysis of Twitter Data using Apache Hive," *International Journal of Computer Applications*, vol. 153, pp. 21-26, 2016.
- [25] <https://twitter.com/FIFAWorldCup>, 2018.
- [26] <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json>, 2018.
- [27] <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object>, 2018.
- [28] <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects>, 2018.
- [29] <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/entities-object>, 2018.
- [30] https://www.loc.gov/standards/iso639-2/php/code_list.php, 2018.