Twitter Data Analysis for Live Streaming by Using Flume Technology

Mohan Manoj Vissamsetti, Yalamandala Prasanth and
T. Prem Jacob

March 11, 2020

TWITTER DATA ANALYSIS FOR LIVE STREAMING BY USING FLUME TECHNOLOGY

Mr. V. Mohan Manoj[1], Mr. Y. Prasanth[2], Student, Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Dr. T. Prem Jacob[3], Associate Professor, Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
vmohanmanoj.avg.9@gmail.com[1], prasanthyalamandala99@gmail.com[2], premjac@yahoo.com[3]

Abstract:Nowadays people all over the world express themselves bynumerous platforms on the internet every minute. Twitter is a type of social media that is being used by millions of people in order to express their opinions or expressions in the form of tweets. The assets of the streaming process generate an enormous amount of real-time data mentionedas big data. In order to understand the instance, we are going to analyze the tweets to find out the trending word among the given word using big data. We are going to make use of flume technology to extractreal-time dataon twitteras well asstore inthe Hadoop distributed file system. PIG, HIVE as well as itsinquiries to provide the sentiment data depending on the groups demarcated in HIVE query language. Data will be visualized with the help of a word cloud.

Index terms: Hadoop, Flume, Pig, HIVE, HDFS, Word cloud, JSON

1.INTRODUCTION

Twitter is among the largest social media sites, tweets in millions on different topicsevery day.Social media is quite popular as well as an exciting platform to express your viewpoints as well as different kinds of ideas globally just by sitting at one place. Being a passive application, we can follow the tweets of a person and will reach us through a notification. Twitter will generate nearly 1TB of text data in the form of tweets. There is an option on twitter to follow the tweets of persons that you admire. It is beingessential for business purposes while clients, as well as customers, are recurrentTwitter users as well as expected to follow the feeds. We can easily reach out the updates of government, sportsmen, actors, businessmen, and celebrities regarding their updates corresponds to information, their decisions correspond to their fields will be constantly posted,following your admired persons in something very close to real-time. Twitter is been one of the trending social media, theirupdates will be instantly posted to the twitter feed. The facility to Tweet using mobile phones alsopermits tocontinue this real-time linking does not matter where you are present. The requirement is to have an account in a device that has internet connectivity.

In this project, we are going to extract the data (which will be a form of tweets in JSON format) of live streaming directly from the server using the flume technology. All this process will be under the extraction module. In the transformation module, the data which we extracted will be transformed into the normal text usinghive which is constructed on the top of Hadoop. Based on the analyzed data it will be visualized using a word cloud based on repeated words. The most repeated word will be displayed in higher font size, based on the totalnumber of times the word is repeated in the analyzed data.Several times the wordrecurs in tweets will have a higher font size. Least times the word has recurred in tweets will have the lower font size. The most repeated words will be displayed in the central position of the word cloud. Visualizing data in word cloud will be easy to differentiate the most repeated words among the other words.

Apache flume is an open-source service for collecting data which is meant for shifting the data from the fount to the final destination. And it is available, distributed, scalable, customizable as well as reliablesystem to collect as well as transmitting a huge quantity of data logs in an efficient manner from a number of sources to thecentralized data store. The huge amount of streaming data from various fount to the centralized storage unit makes data loading easy and efficient. There runsa number of services of a particular company on several servers. Flume can aggregate the data from multiple servers in real-time and can also transfer them to several destinations. Flume supports fan-in fan-out and Multi-hop flows as well as contextual routing.

An open-source framework called Hadoop can process the large data clusters in the distributed computing atmosphere. The main components of Hadoop are MapReduce, distributed file system as well asvarious related projects like HBase, Sqoop, apacheHIVE, pig. Among them,Hadoop distributed file system andmap-reduce are called the primary components.Usually,a set of linked computer systems that work collectively asa system is known as the cluster. In basic language, a computer cluster that is utilized for Hadoop is known as the Hadoopcluster. In distributed computing, the large amount of unorganized data is stored as well as examined by the specially designed Hadoop cluster for such kinds of computations. A low-cost commodity computer system is required for running these particular clusters. A company has loads of services running on the multiple servers that generate a lot of data (logs). So,we have to examine these logs all in all.In order to get these logs processed, we require anextensible and scalable,reliable, as well as controllable data collection service in a distributed environment that can flow semi-structured from one location to the other for their processing purposes.

Recurrently used keywords show up better in a cloudof words. These clouds provide clarity at the time of text examinationin pursuance of effective communication of your data outcomes. They are mainly known for analyzing the text because the frequently occurring word can be spotted effortlessly.A word that is used more frequently is shown in a bold and comparatively larger size, these are the efficient methods to recognize the relevant portions of the text. They also assist the business users for contrasting as well as comparing 2 unique pieces of text sample to evaluate the similarities in the wording of both.

2.LITERATURE REVIEW

Anjali Barskar, Ajay phulre [2017]: Opinion mining of twitter data using Hadoop and apache pig,it is mainly used for knowing the opinions for a particular situation. To extract the data directly from the server they are using flume technology to serve the purpose. And the data will be transferred to the Hadoop environment. They are using HIVE, pigtools which are built over the Hadoop environment which will be used for converting the unstructured format to a structured format.

A.S. Nagdive, Tugnayat, G. BRegulwar, D. Petkar [2019]:reported in “Web Server log analysis for unstructured data using apache flume and pig”, that a large number of the log files will be produced in a web server and those data will be in an unstructured format. Analyzing the logs of the website will give the usage of a website, can be taken out by analyzation. Log files will be taken out through web mining. Such log files are generated at a higher level and will be analyzed with the help of parallel processing and will be stored in this manner to handle a large amount of data. That data will be analyzed through a HIVE, which will be like SQL it works over the Hadoop environment.

Soumysharma, SandeepKumar [2019]: Sentiment analysis on twitter posts using Hadoop,everyone around the world is updated with the usage of twitter to express their opinions which can be seen by other people who are using the twitter with an account. Twitter API will retrieve these data which includes tweets and creates a database from those tweets, which will be used for analyzation further to predict the supporting regards to a situation. Using Ni-Fi technology we can transfer the data from local storage to the Hadoop clusters. The data will be in an unstructured format, here HIVE used for classifying the unstructured data to a usable data termed as structured data. In order to express whether the tweet is a positive or negative sentiment, we are using Natural Language Processing.

Manish Wankhede, Vijay Trivei, VineethRichhariya[2016]: Location-based analysis of twitter data using apache HIVE, In this paper the frequency of tweets supporting the situation or vice versa that are posted regards to the location to find the maximum and a minimum number of the tweets. The data will be extracted from the server by using flume technology and analyzation techniques will be performed by HIVE and pig in a Hadoop environment. To transform unstructured data to structured data we can use jsonserde. we can find a number of tweets that are posted location wise.

Ms. Pooja S. Patil, Ms. Pranali B. Sable, Ms. Reshma J. Fasale, Mr. P. A. Chougule [2016]: Sentiment Analysis on Twitter Data Using Apache Flume and HIVE, Sentiment examination is similar to opinion mining, to find out the response for an incident. The data which we extracted from twitter will be stored in the Hadoop distributed file system in the format of JSON. By using the user-defined functions, we are going to convert the other format to a structured format, and analysis also was done using that function.

Mr.SagarNadagoud,[2015]: Market Sentiment Analysis for Popularity of Flipkart, Here, they discussed, One should get the data from the web pages or the database rather than that we can utilize anykind of programming language to sneakinto the data from their database. As we have to analyze the market analysis in the initial step, we should collect the data to be analyzed. The adjectives which qualify the tweet has to be mentioned in the dictionary for a reference to the data. We can collect the data of online streaming data by the Ecosystem tool known as a flume. After that step, we can transform the data into a structured format. And by performing the sentimental examinationit could find out whether they expressed in a positive note or negative comments referencing from the dictionary.

G.Kavitha,NomaanImtiaz[2018]: Discovery Public Opinion by Executionof Sentimental Analysis on Real-Time Twitter Data, In this paper, they are visualizing the data through a bar chart classifying the positive, negative and neutral tweets. The data will be compared with the dictionary defined and will consider the nature of the tweet. The data will be extracted, analyzed using flume and Hadoop technologies. Through this, we can find the trend, viral topics of society.

3.PROPOSED SYSTEM

As we have seen in the earlier papers the analyzation of twitter data has done to the log files. Inorder to overcome the drawbacks, we are using Hadoop as well as its ecosystems. For getting raw data from twitter we are using apache flume. We have utilized the standard platform as Hadoop on a single node Ubuntu machine for solving the problems related to the big databythe map-reduce framework. The Flume has been utilized to fetchthe real-time dataon twitteras well as store it in hdfs. The condition for flume as well as HIVE is that Hadoopmust be installed before its use.The format used to transfer the data between server and web application is JSON. HIVE is used to transform the data into text.

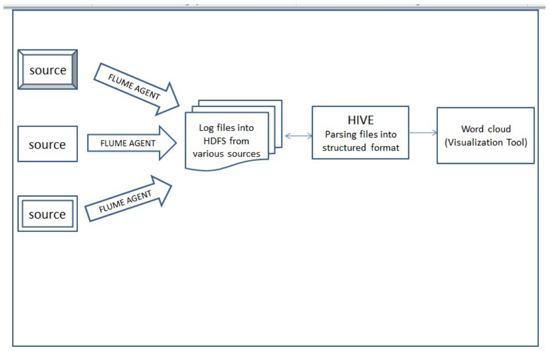


Figure 1. System Architecture

The workflow of the project is shown above, the data which is stored in sources will be extracted using the flume agent and will be stored in a Hadoop environment. The extracted data will be in the format of JSON, as many tools have been developed over the Hadoop to manage the data. Using the tools of Hadoop, we can manage the data to transform from semi-structured format to structured format. Each technology used in the above process is explained below.

4.TECHNOLOGIES AND TOOLS

4.1APACHE FLUME

It is a tool to collect, aggregate as well as transport the huge quantities of streaming data like events, log files from different sources to the centralized data storage unit. It is provided with basic as well as supple architecture depending on the flow of streaming data. It makes use of a basic extensible data model that permits online analytical applications. It is a system utilized for transmittinghuge amounts of streaming data within the hdfs. Accumulating log data existingin thelog files from the web servers as well as combining it in hdfsforits analysisis the main usage of this system. There are few flume components lists event, client, agent, source, channel, sink. An event is a fundamental unit of data transported like single log entry or a tweet by a flume from its originating points to the destination. The client produces data in the form of events. The agent is referred to as a self-governing JVM process that canhost components offlume like sinks, channelsas well as sources. Therefore,provided with the capability of collecting, storing as well asforwarding events from one place to another.The source is an active component that receives the event and places it in the channel. Channel is a passive component that buffers the event and sends it to the sink. The method flume agents use to transfer events from their sources to destination. Sink removes the event from a channel as well as move them to the upcoming agent to the event's final destination orin the flow Flume pushes data to the sink because of which writes to sink can overwhelm data read from sinks.

The configuration file to access twitter server will be in presence of the consumer and consumer secret key and access token along with the access secret token has to be mentioned regarding the account

Code for flume configuration file

```
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=SU4mEtHIQ
SOI5OJIwZ2KawWVb
TwitterAgent.sources.Twitter.consumerSecret=DfRTZOK
pUZGkcKOKBkn1NtUskRi78fOjCVEq465iHn8I38lb
TwitterAgent.sources.Twitter.accessToken=12119549056
58712064V48cc8JqQ3oSJQ4FKbbg7phcWjXyee
TwitterAgent.sources.Twitter.accessTokenSecret=PEfDzr
AzTuO7yg8W3PnC703buqB9233bCvphUiAzbLfae

TwitterAgent.sources.Twitter.keywords=Hadoop,machin
elearning,spark,informatica,unix
TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=/user/sairavi/twittert
oday
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
```

TWITTER APPLICATION CREATION

In order to extract the tweets, we have tomake a twitter application. Creation of twitter application is as follows.

Step1:

Initially, we have to sign in to the twitter account as well as dolittle work with the twitter application window in which we canmanage, delete and createtwitter apps by navigating to the following link https://apps.twitter.com/

Step2:

From the above link click on thebutton written with create New App. Then you will be going to a window and has to fill your detail information in order to get an application.

Step3:

A new app will be created which is used to create Consumer and Consumer secret key, Access token and Access token secret used to edit in flume.conf. file. In order to fetch twitter data which is lively tweeting in the account, the above-mentioned keys are used.

Step4:

In the details of the app, we can find the keys under keys and tokens. Out of which consumer keys will be shown and access tokens will be generated for every instance.

Step5.

Consumer Key, Consumer Secret, Access tokens are used to configure the flume agent.

4.2 HADOOP

Hadoop gives the agenda for handling the large clusters of data and it is also termed as the eco-system for all the open-source projects. The large data sets also processed by the assistance of the Hadoop in distributed computing. HadoopcontainsHadoop distributed file system (hdfs) as well as connected to handling huge data. Map-reduce as well as (hdfs) are the primary components of Hadoop. In order to handle more amount of data different computers connected as a single system. Such computer clusters meant for Hadoopare known as Hadoop clusters. These clusters will execute in low-cost commodity computers. The Hadoop architecture is composed of the hugeHadoop clusters that are organized in various racks. In each record, there will be master machines and slave machines. About 40 slave machines are contained in each and every rack. Each rack and the connection between the racks will be managed with rack switch. Job trackers are referred to as Masters as well as some machines work as name node. They are favoring with more ram and CPU and less storage. Task Trackers andData Node are mentioned as slaves, these slaves have huge storage at the local disk as well as modesttotals of CPU along with RAM. Job trackers will be under the map to reduce components. Secondary name, as well as name node, will be under the hdfs component. The client neither plays the role of a slave nor a master for loading the data within the cluster, acquiesce map-reduce jobs, retrieve the data to observe the response afterward the job is done. The master is made up of three main parts name and secondary name node, job tracker. Name node holds the metadata for hdfs like which part ofthe file is stored in which portionof the cluster, block information, user permissions. When in use all this information is stored in ram, but this information also stored in disk for persistence storage. The map reduces assists the parallel processing of data and it is organized by the job tracker. Slaves are responsible for store data, process computation.

A data block is directly written toadata node by the client. Then, data nodes duplicate this particular blockto some other nodes. When all the 3 data nodes are written with a block only then cycle reappearsfor the upcoming block.

4.3 HIVE

It is the infrastructure instrument of data warehouse for processing the organized data in Hadoop. It stores its data in his system. Hive stores its metastore in one RDBMS database. It exists at the top of Hadoop to recapitulate the bigdata,as well as generates queries as well as analyze it effortlessly. In a database, the schema is stored by the HIVE as well as it also processes the datain HDFS. An SQL query language is provided by HIVE and is called as HQLor HIVEQL. It owns data in the HIVE table means it is called a HIVE internal table. It doesn't own the data. This means it's a part of HIVE, HDFS only owns the external HIVE tables. To create the external table, it is required to use an external keyword. An external table is a way to protect data against accidentally drop commands. It provides a bucketing concept, another technique for decomposing table sets into manageable parts. Joins can also be performed using it.

The word cloud is a visualization toolto act in aneasy way. The more exact a word seems in the source of text, thebolderand bigger it looks in the cloud of words. It is the group or cluster of the words shown in unique sizes. If the word appears in a bigger size as well as bolder, the more frequently it is declared in the text provided. These are the ideal methods to highlights the relevant portions of the data in the textual form from blogs to databases. Significant textual data points can be highlighted using a word cloud.

4.4 JAVASCRIPT OBJECT NOTATION

It isalight-weight anddata-interchange notation. It is primarily utilized for transmitting databetween the server as well as the web-application. JSON is used for storing as well as organizing the contentgenerated with the help of CMS at the site. It is a self-governing data format and is the best substitute for XML.

4.5 WORD CLOUD

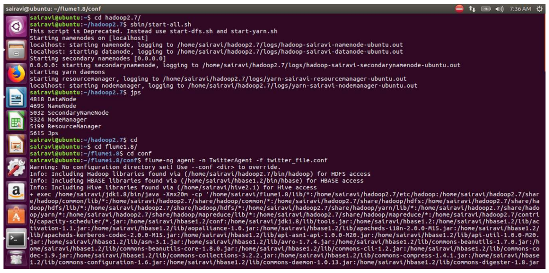
Word clouds show you what is emphasized in your text. It is a data visualization practice for representing thedata in the form of textand every word specifies its frequency or importance. It is widely used for analyzing data from social network websites. We can differentiate words in visualization. We are implementing a word cloud using python.

5. METHODOLOGY

The following procedure is used to achieve the purpose of the proposed system. The steps that include are

1.Initially, we have to create a twitter app that includes the twitter tokens which will be used for fetching real-time data.
2. The data will be extracted fromtwitterusing flume technology and it will be stored in local HDFS. The twitterData comes from the site will be in an unstructured format known as JSON.
3. The analysis part will be done after storing twitter data into HDFS,the analysis part will be done using HIVE. We can transform intoa structured format using HIVE.
4. The structured data will be in the form of text and will be visualized through word cloud using python.

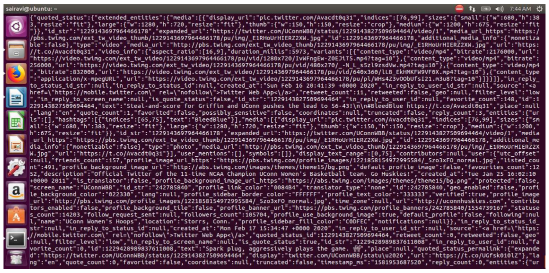
6. RESULTS



```
sairavi@ubuntu:~$ cd hadoop2.7/
sairavi@ubuntu:~/hadoop2.7$ sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/sairavi/hadoop2.7/logs/hadoop-sairavi-namenode-ubuntu.out
localhost: starting datanode, logging to /home/sairavi/hadoop2.7/logs/hadoop-sairavi-datanode-ubuntu.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/sairavi/hadoop2.7/logs/hadoop-sairavi-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /home/sairavi/hadoop2.7/logs/yarn-sairavi-resourcemanager-ubuntu.out
localhost: starting nodemanager, logging to /home/sairavi/hadoop2.7/logs/yarn-sairavi-nodemanager-ubuntu.out
sairavi@ubuntu:~/hadoop2.7$ jps
4818 DataNode
4695 NameNode
5032 SecondaryNameNode
5324 NodeManager
5199 ResourceManager
5615 Jps
sairavi@ubuntu:~/hadoop2.7$ cd
sairavi@ubuntu:~$ cd flume1.8/
sairavi@ubuntu:~/flume1.8$ cd conf
sairavi@ubuntu:~/flume1.8/conf$ flume-ng agent -n TwitterAgent -f twitter_file.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/sairavi/hadoop2.7/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/home/sairavi/hbase1.2/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/home/sairavi/hive2.1) for Hive access
+ exec /home/sairavi/jdk1.8/bin/java -Xmx20m -cp '/home/sairavi/flume1.8/lib/*:/home/sairavi/hadoop2.7/etc/hadoop:/home/sairavi/hadoop2.7/shar
e/hadoop/common/lib/*:/home/sairavi/hadoop2.7/share/hadoop/common/*:/home/sairavi/hadoop2.7/share/hadoop/hdfs:/home/sairavi/hadoop2.7/share/ha
doop/hdfs/lib/*:/home/sairavi/hadoop2.7/share/hadoop/hdfs/*:/home/sairavi/hadoop2.7/share/hadoop/yarn/lib/*:/home/sairavi/hadoop2.7/share/hado
op/yarn/*:/home/sairavi/hadoop2.7/share/hadoop/mapreduce/lib/*:/home/sairavi/hadoop2.7/share/hadoop/mapreduce/*:/home/sairavi/hadoop2.7/contri
b/capacity-scheduler/*.jar:/home/sairavi/hbase1.2/conf:/home/sairavi/jdk1.8/lib/tools.jar:/home/sairavi/hbase1.2:/home/sairavi/hbase1.2/lib/ac
tivation-1.1.jar:/home/sairavi/hbase1.2/lib/aopalliance-1.0.jar:/home/sairavi/hbase1.2/lib/apacheds-i18n-2.0.0-M15.jar:/home/sairavi/hbase1.2/
lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/sairavi/hbase1.2/lib/api-asn1-api-1.0.0-M20.jar:/home/sairavi/hbase1.2/lib/api-util-1.0.0-M20.
jar:/home/sairavi/hbase1.2/lib/asm-3.1.jar:/home/sairavi/hbase1.2/lib/avro-1.7.4.jar:/home/sairavi/hbase1.2/lib/commons-beanutils-1.7.0.jar:/h
ome/sairavi/hbase1.2/lib/commons-beanutils-core-1.8.0.jar:/home/sairavi/hbase1.2/lib/commons-cli-1.2.jar:/home/sairavi/hbase1.2/lib/commons-co
dec-1.9.jar:/home/sairavi/hbase1.2/lib/commons-collections-3.2.2.jar:/home/sairavi/hbase1.2/lib/commons-compress-1.4.1.jar:/home/sairavi/hbase
1.2/lib/commons-configuration-1.6.jar:/home/sairavi/hbase1.2/lib/commons-daemon-1.0.13.jar:/home/sairavi/hbase1.2/lib/commons-digester-1.8.jar
```

Figure 2.Commands to extract twitter data

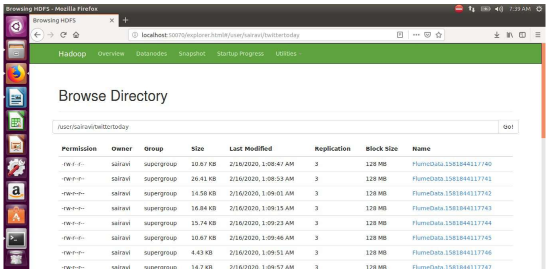
These are the commands required to extract twitter data from the server of live streaming. Here, we are using flume to collect data from server.



```
{"quoted_status":{"extended_entities":{"media":[{"display_url":"pic.twitter.com/Avacdt0q31","indices":[76,99],"sizes":{"small":{"w":680,"h":38
3,"resize":"fit"},"large":{"w":1280,"h":720,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"medium":{"w":1200,"h":675,"resize":"fit
"}},"id_str":"1229143697964466178","expanded_url":"https://twitter.com/UConnWBB/status/1229143827509694464/video/1","media_url_https":"https:/
/pbs.twimg.com/ext_tw_video_thumb/1229143697964466178/pu/img/_E1RHoUrHIERZ2XW.jpg","id":1229143697964466178,"additional_media_info":{"monetiza
ble":false},"type":"video","media_url":"http://pbs.twimg.com/ext_tw_video_thumb/1229143697964466178/pu/img/_E1RHoUrHIERZ2XW.jpg","url":"https:
//t.co/Avacdt0q31","video_info":{"aspect_ratio":[16,9],"duration_millis":5973,"variants":[{"content_type":"video/mp4","bitrate":2176000,"url":
"https://video.twimg.com/ext_tw_video/1229143697964466178/pu/vid/1280x720/IvWFngEw-Z0EJiT5.mp4?tag=10"},{"content_type":"video/mp4","bitrate":
256000,"url":"https://video.twimg.com/ext_tw_video/1229143697964466178/pu/vid/480x270/_-N_L_s5zl9zsdVw.mp4?tag=10"},{"content_type":"video/mp4
","bitrate":832000,"url":"https://video.twimg.com/ext_tw_video/1229143697964466178/pu/vid/640x360/lLB_EkHMKFW9VF0X.mp4?tag=10"},{"content_type
":"application/x-mpegURL","url":"https://video.twimg.com/ext_tw_video/1229143697964466178/pu/pl/WHs4Z3vOObUfsI21.m3u8?tag=10"}]}}]},"in_reply_
to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sun Feb 16 20:41:39 +0000 2020","in_reply_to_user_id_str":null,"source":"<a
href=\"https://mobile.twitter.com\" rel=\"nofollow\">Twitter Web App<\/a>","retweet_count":11,"retweeted":false,"geo":null,"filter_level":"low
","in_reply_to_screen_name":null,"is_quote_status":false,"id_str":"1229143827509694464","in_reply_to_user_id":null,"favorite_count":148,"id":1
229143827509694464,"text":"Steal-and-score for Griffin and UConn pushes the lead to 56-43!\n\n#BleedBlue https://t.co/Avacdt0q31","place":null
,"lang":"en","quote_count":1,"favorited":false,"possibly_sensitive":false,"coordinates":null,"truncated":false,"reply_count":3,"entities":{"ur
ls":[],"hashtags":[{"indices":[65,75],"text":"BleedBlue"}],"media":[{"display_url":"pic.twitter.com/Avacdt0q31","indices":[76,99],"sizes":{"sm
all":{"w":680,"h":383,"resize":"fit"},"large":{"w":1280,"h":720,"resize":"fit"},"thumb":{"w":150,"h":150,"resize":"crop"},"medium":{"w":1200,"
h":675,"resize":"fit"}},"id_str":"1229143697964466178","expanded_url":"https://twitter.com/UConnWBB/status/1229143827509694464/video/1","media
_url_https":"https://pbs.twimg.com/ext_tw_video_thumb/1229143697964466178/pu/img/_E1RHoUrHIERZ2XW.jpg","id":1229143697964466178,"additional_me
dia_info":{"monetizable":false},"type":"photo","media_url":"http://pbs.twimg.com/ext_tw_video_thumb/1229143697964466178/pu/img/_E1RHoUrHIERZ2X
W.jpg","url":"https://t.co/Avacdt0q31"}],"user_mentions":[],"symbols":[]},"display_text_range":[0,75],"contributors":null,"user":{"utc_offset"
:null,"friends_count":157,"profile_image_url_https":"https://pbs.twimg.com/profile_images/1221858154972995584/_5zo3xFO_normal.jpg","listed_cou
nt":491,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme15/bg.png","default_profile_image":false,"favourites_count":12
52,"description":"Official Twitter of the 11-time NCAA Champion UConn Women's Basketball team. Go Huskies!","created_at":"Tue Jan 25 16:02:10
+0000 2011","is_translator":false,"profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme15/bg.png","protected":false,
"screen_name":"UConnWBB","id_str":"242785840","profile_link_color":"0084B4","translator_type":"none","id":242785840,"geo_enabled":false,"profi
le_background_color":"022330","lang":null,"profile_sidebar_border_color":"FFFFFF","profile_text_color":"333333","verified":true,"profile_image
_url":"http://pbs.twimg.com/profile_images/1221858154972995584/_5zo3xFO_normal.jpg","time_zone":null,"url":"http://uconnhuskies.com","contribu
tors_enabled":false,"profile_background_tile":false,"profile_banner_url":"https://pbs.twimg.com/profile_banners/242785840/1554739167","statuse
s_count":14203,"follow_request_sent":null,"followers_count":105704,"profile_use_background_image":true,"default_profile":false,"following":nul
l,"name":"UConn Women's Hoops","location":"Storrs, Conn.","profile_sidebar_fill_color":"C0DFEC","notifications":null}},"in_reply_to_status_id_
str":null,"in_reply_to_status_id":null,"created_at":"Mon Feb 17 15:34:47 +0000 2020","in_reply_to_user_id_str":null,"source":"<a href=\"https:
//mobile.twitter.com\" rel=\"nofollow\">Twitter Web App<\/a>","quoted_status_id":1229143827509694464,"retweet_count":0,"retweeted":false,"geo"
:null,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":true,"id_str":"1229428989837611008","in_reply_to_user_id":null,"fa
vorite_count":0,"id":1229428989837611008,"text":"Spark plug, aggressively plays the game. 👯","place":null,"quoted_status_permalink":{"expande
d":"https://twitter.com/UConnWBB/status/1229143827509694464","display":"twitter.com/UConnWBB/statu\u2026","url":"https://t.co/UGfsk010lZ"},"la
ng":"en","quote_count":0,"favorited":false,"coordinates":null,"truncated":false,"timestamp_ms":"1581953687520","reply_count":0,"entities":{"ur
```

Figure 3. Twitter data in the format of JSON

The data extracted from twitter will be in the format of JSON as shown above. JSON is a semi structured format. It is a subset of java script programming language.


Browsing HDFS - Mozilla Firefox
7:39 AM
Browsing HDFS
localhost:50070/explorer.html#/user/sairavi/twittertoday
Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities
Go!

Browse Directory

/user/sairavi/twittertoday

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	sairavi	supergroup	10.67 KB	2/16/2020, 1:08:47 AM	3	128 MB	FlumeData.1581844117740
-rw-r--r--	sairavi	supergroup	26.41 KB	2/16/2020, 1:08:53 AM	3	128 MB	FlumeData.1581844117741
-rw-r--r--	sairavi	supergroup	14.58 KB	2/16/2020, 1:09:01 AM	3	128 MB	FlumeData.1581844117742
-rw-r--r--	sairavi	supergroup	16.84 KB	2/16/2020, 1:09:15 AM	3	128 MB	FlumeData.1581844117743
-rw-r--r--	sairavi	supergroup	15.74 KB	2/16/2020, 1:09:23 AM	3	128 MB	FlumeData.1581844117744
-rw-r--r--	sairavi	supergroup	10.67 KB	2/16/2020, 1:09:46 AM	3	128 MB	FlumeData.1581844117745
-rw-r--r--	sairavi	supergroup	4.43 KB	2/16/2020, 1:09:51 AM	3	128 MB	FlumeData.1581844117746
-rw-r--r--	sairavi	supergroup	14.7 KB	2/16/2020, 1:09:57 AM	3	128 MB	FlumeData.1581844117747

Figure 4.Data fetched from twitter and sank into HDFS

The twitter data will be stored in the browsed directory(UNIX) as shown above. Data is stored in local of the unix environment.

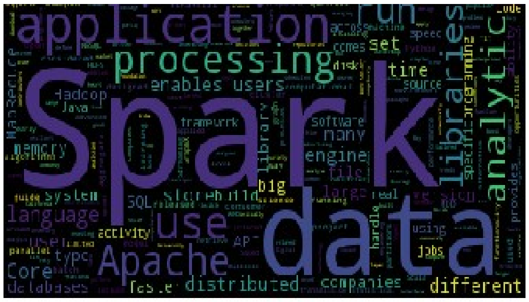


Figure 5:VisualizingWord Cloud

The output will be shown using a word cloud which is used to emphasize data.Easy to differentiate between words.

7.CONCLUSION ANDFUTURE ENHANCEMENT

Previously the data of twitter in which we analyzing was a log file. Inorder to achieve this, we have to perform a code of lines coding to achieve this. To perform sentimental analysis on stored data we have to perform complex operations. Here we can extract twitter data of live streaming by using flume. The extracted twitter data file has to be transformed from JSON format to text format using HIVE. From this approach, we can perform a task in less time. Visualization can be done using a word cloud.

The way of extracting and analyzing can be done with technologies(spark) which will take less time to process and easy to code. We can also visualize the data in other approaches according to their requirements.

REFERENCES

[1]. Anjali Baskar, Ajay Phurle, “Opinion Mining of Twitter Data using Hadoop and Apache Pig”, International Journal of computer Applications(IJCA) Volume 158-No 9,January 2017.

[2].A.S. Nagdive, R.M Tugnayat, G.B Regulwar, D.Petkar, ”Web Server log Analysis for Unstructured data Using Apache Flume and Pig”, International Journal of Computer Sciences and Engineering(JCSE)Vol-7,Issue-3,March 2019 E-ISSN:2347-2693.

[3].Soumy Sharma, Sandeep Kumar, “Sentimental Analysis on Twitter posts Using Hadoop”,International Research Journal of Engineering and Technology(IRJET) Volume:06 Issue:04, April 2019, e-ISSN:2395-0056, p-ISSN:2395-0072.

[4].Manish Wankhede, Vijay Trivei, Dr.VineethRichhariya, “Location based Analysis of Twitter Data using Apache HIVE”, International Journal of computer Applications(IJCA), Volume:153-N0 10, November 2016

[5].Dr.poojaS.patil, Ms. PranaliB.sable, Ms.ReshmaJ.Fasale, Dr.P.A.chougule, “Sentimental Analysis on Twitter Data using Apache Flume and HIVE” Volume:03 Issue:02, Feb-2016, e-ISSN:2395-0056, p-ISSN:2395-0072.

[6].Mr.SagarNadagoud, Mr.KotreshNaik.D, “Market Sentimental Analysis for Popularity of Flipkart”, International Journal of Advanced Research in Computer Engineering and Technology(IJARCET), Volume:04, Issue:05, May 2015

[7].G.Kavitha, B.Saveen, NomaanImtiaz[2018], “Discovery Public Opinion by Performing Sentimental Analysis on Real time Twitter Data”, International Conference on Circuits and Systems in Digital Enterprise Technology(ICCSDET) Publisher:IEEE, 2018.

[8] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

[9] Vishal A. Kharde and S.S. Sonawane, “ Sentiment Analysis ofTwitter Data: A Survey of Techniques”, International Journal of Computer Applications, Vol. 139, No.11,2016, pp:5-15.

[10] Srishti Sharma, ShampaChakraverty and AkhilSharma.A context-based algorithm for sentiment analysis. In International Journal of Computational Vision and Robotics, Vol. 7, No. 5, 2017, NetajiSubhas Institute of Technology, Dwarka, New Delhi, India, 2017.

[11] http://flume.apache.org/ (online resource).

[12] http://www.Hadoopadmin.co.in/sources-of-bigdata/.