SENTIMENT ANALYSIS ON TWITTER DATA OF THE DEMONETIZATION OF 500 AND 1000 RUPEE NOTES USING THE FLUME & HIVE ON HADOOP FRAMEWORK

[1]Pallavi Kakde, [2]Prof. V. A. Losarwar
[1]M.E Student, [2]Associate Professor
[1, 2]Department of Computer Science and Engineering,
[1, 2]PES College of Engineering Aurangabad, Maharashtra, India.

Abstract:In recent years, because of the popularity of social networking has drastically expanded and the tremendous information being delivered by the Social media like Facebook, Twitter. Twitter one of the biggest social media website gets a huge number of tweets each day on a variety of important issue. In this paper for handling the huge data, we have used the Hadoop and Apache Flume and Hive to analyze the sentiment. We take the real-time tweets on demonetization using tweeter streaming API and store the tweets in Hadoop using tweet agent. Hive is query language to the transformation of semi-structured and with the help of dictionary we classified tweets into 3 constraints positive, negative and neutral

Keywords: **Hadoop,HIVE, Semi Structured, Flume, Twitter, Social Media**

I.INTRODUCTION:

In todays developed world, consistently, people around the world express their opinion, feeling via different stages Online. What's more, in every moment, a tremendous measure of unstructured information is produced. Twitter, one of the biggest web-based social networking webpage gets a many tweets each day on the important issues. This colossal measure of crude information can be utilized for modern, social, financial, government approaches or business reason by sorting out as indicated by our necessity and preparing.[1]

1.1 Background of Demonetization:

Demonetization is the process of banning or ending the currency in the form of coins and notes as the decision of a country. The real objects of demonetization were formalization (of the economy), attack on black money, less-cash economy, bigger tax base, digitization, a blow to terrorism.

On November 8, Prime minister of India Narendra Modi took a historic decision by announcing that the high-denomination notes (Rs 500 and Rs 1,000) than in circulation would cease to be legal tender. Sudden demonetization is not new in India this is the third demonetization since 1946 and 1978. Be that as it may, the flow of the higher category banknotes amid that period was exceptionally restricted and a large portion of the higher section banknotes was held with banks as it were. Agreeing to Save Bank of India (RBI) records 2016, Indian rupee banknotes worth 16,664 billion are being flowed among the open. Of these 86% (14.180 billion) are in Rs 500 and Rs 1000 banknotes [2].

1.2Objective:

Twitter has over a billion clients and ordinary individuals produce billions of tweets more than 100 hours out of each moment and this number is regularly expanding. To analyze and understand the movement happening on such a massive scale, a social SQL database isn't sufficient. Such sort of information is handled by distributed framework like Hadoop.The main objective of this paper is to gather efficient real-time data on a demonetizationtopic from Twitter.

- Tweets Retrieval: A large number demonetization of tweets is collected through the Streaming API and Hadoop online streaming tool using Apache Flume
- Storage: This data is stored in Hadoop (HDFS: Hadoop Distributed Filesystem) so as to form a key-value pair that is JSON format which is needed to feed to mapper in map-reduce programming approach. The data is stored in the flume sink.
- Data Processing: real-time data is collected in JSON format, process the data using Apache Hadoop and using map-reduce programming model and Apache hive framework.
- Data Analysis: The output obtained from the reducer stage in the format of the table

- Sentiment Representation: Representation of sentiments classified data in the form of pie charts and map and the table. the final output we will get the output of tweets in the sorted form of classified tweets that is Positive, Negative and Neutral tweets.[3]

II. RELATED WORK

There is growing interest in sentiment analysis from the social media and related content:
Sangeeta, theywere to analyze the effect of the demonetization policy implemented by using the concept of sentiment analysis. Twitter data is collected from API and they distinguished the data positive negative and neural by using Meaning cloud based on collected data they calculated the net score of twitter and result shown in different maps[4**] Monika Sharma,**they proposed system for twitter Data Analysis Using FLUME & PIG on Hadoop Frameworks achieved sentiment analysis of twitter that data gathered from twitter API related to BJP and Election so it will be beneficial to all people for voting[5]**KomalSutar,**they have tracked all the tweets from November 11th to November 12th, 2016, concerning the terminationof 500 and 1000 rupees notes and analyzed a total of 6268 tweets and the classification is based on the polarity[6]**Mahalakshmi R,**Twitter Sentiment Analysis of Demonetization on Citizens of INDIA using R A polarity score been assigned to each of thetoken. In order to determine the sentiment behind the text the aggregated sum of the score is beencalculated. Depending on the calculated score the text is been classified as positive, negative and neutral.[7]**Ajinkya Ingle,**The analysis of twitter data is processed to check performance of the proposed methodology. The following tests were carried out i,e Polarity test and sentiment score which result's sentiment analysis of the given data set . With the use of R language they visualized the result. Twitter volume separated by positive tweets and negative tweets and the polarity score [8]**SnehaKindare,**experience of handling and parallel processing of huge amount of data. Data collection process introduces us to Java twitter streaming API. They have exposure work with prominent Sentiment Analysis of Twitter Data Using Hadoop .[9]**Jyoti Yadav,**they gathered twitter data from twitter4j API ,then data is tokenized through the TFIDF and Porter Stemming Algorithm root of word found that compared to affine dictionary and compared the weight and the final result calculated with the k means algorithm[10]

III. Proposed System:

3.1System Architecture:

We have overcome few drawbacks from the existing systems by using Hadoop and its services. For getting raw data from Twitter we have used tweeter streaming API, we have used Hadoop's online streaming tool Apache Flume. In this tool, we have added the keys and tokens from Tweeter developer account, we defined what information or keywords we want to retrieve from Twitter. The retrieved data is stored into HDFS (Hadoop Distributed File System) in JSON format. From this raw data, we create tables and filter the information that is needed for us using HIVE. We have performed the data analysis using some UDF's (User Defined Functions) along with AFINN dictionary. The following figure shows an architectural view of the proposed system.

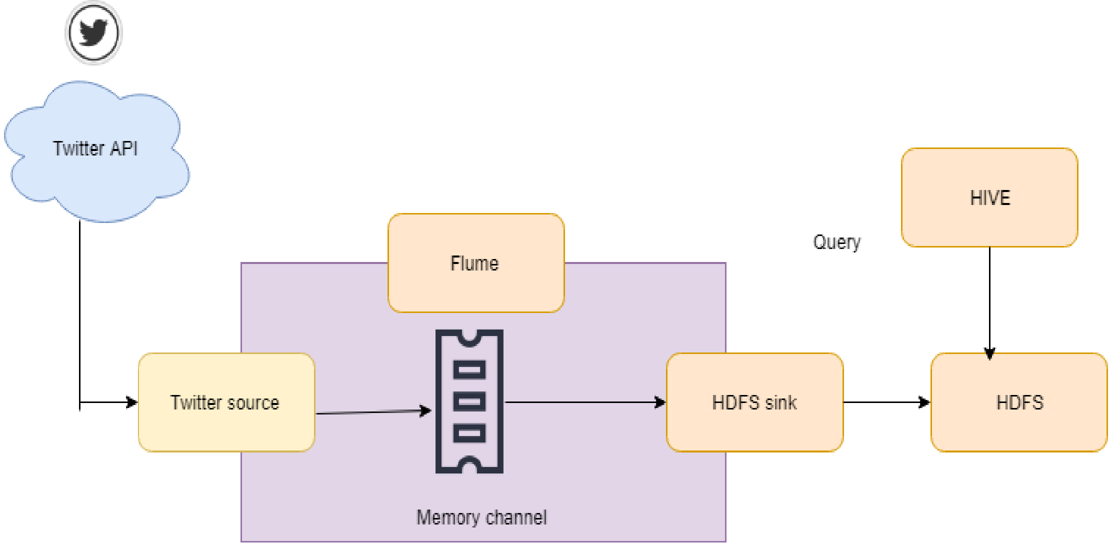


Fig: 3.1 proposed system

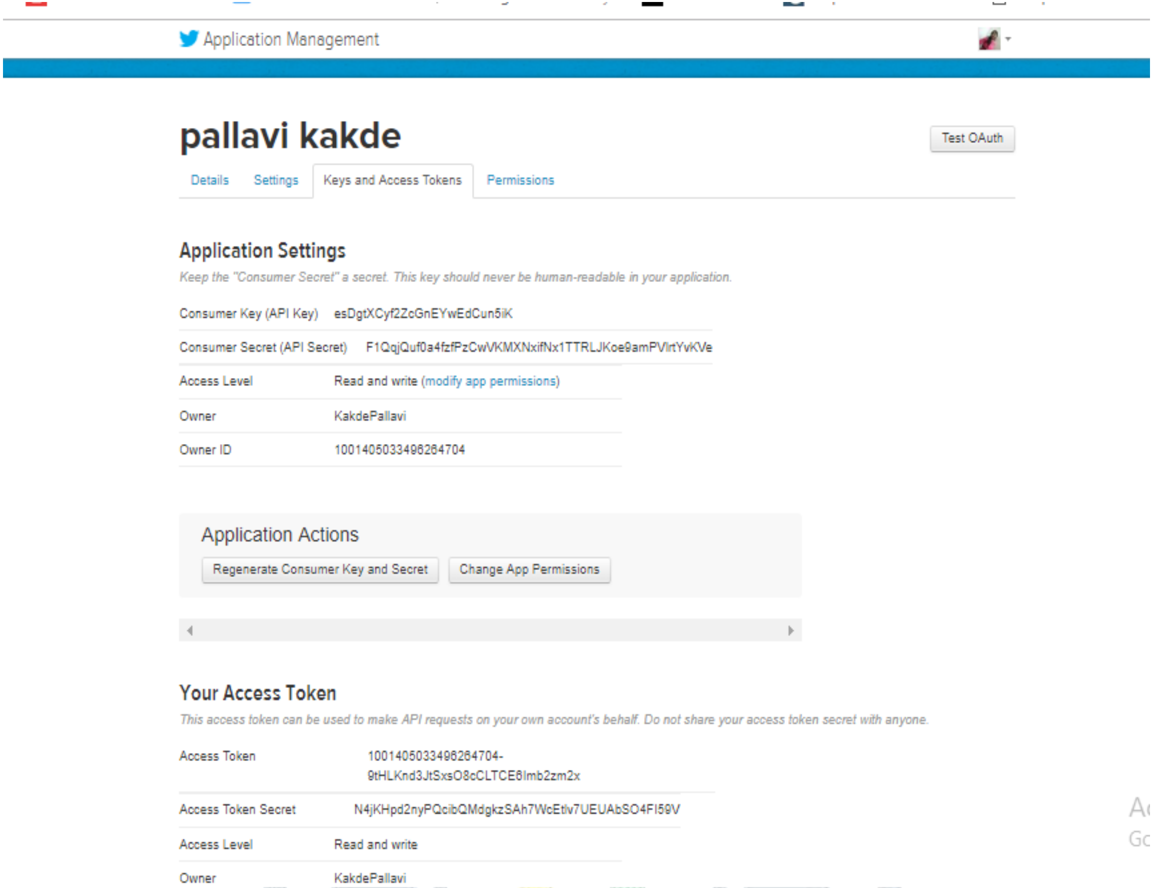
IV.Methodology:

- Creating Twitter Application.
- Retrieving data using Flume.
- Querying using Hive Query Language (HQL)

4.1 Creating twitter application:

First of all, for performing analysis on Twitter data, we need to create an account as a Twitter developer, create an application on http://apps.twitter.com. After login with developer account creating a new application, after successfully created application we generate unique access token along with consumer key. Now we got one consumer key to access the application for getting Twitter data.[4]

The following is the figure that shows how the application looks after creating the application andthus we can see the consumer details and also the access token details. We took these keys and token details and set in the Flume configuration file such that we got the required data from the Twitter in the form of tweets.[10]



pallavi kakde

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) esDgtXCyf2ZcGnEYwEdCun5iK

Consumer Secret (API Secret) F1QqjQuf0a4fzfPzCwVKMXNxifNx1TTRLJKoe9amPVlrtYvKVe

Access Level Read and write (modify app permissions)

Owner KakdePallavi

Owner ID 1001405033498264704

Application Actions

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token 1001405033498264704-9tHLKnd3JtSxsO8cCLTCE6Imb2zm2x

Access Token Secret N4jKHpd2nyPQcibQMdgkzSAh7WcEtlv7UEUAbSO4FI59V

Access Level Read and write

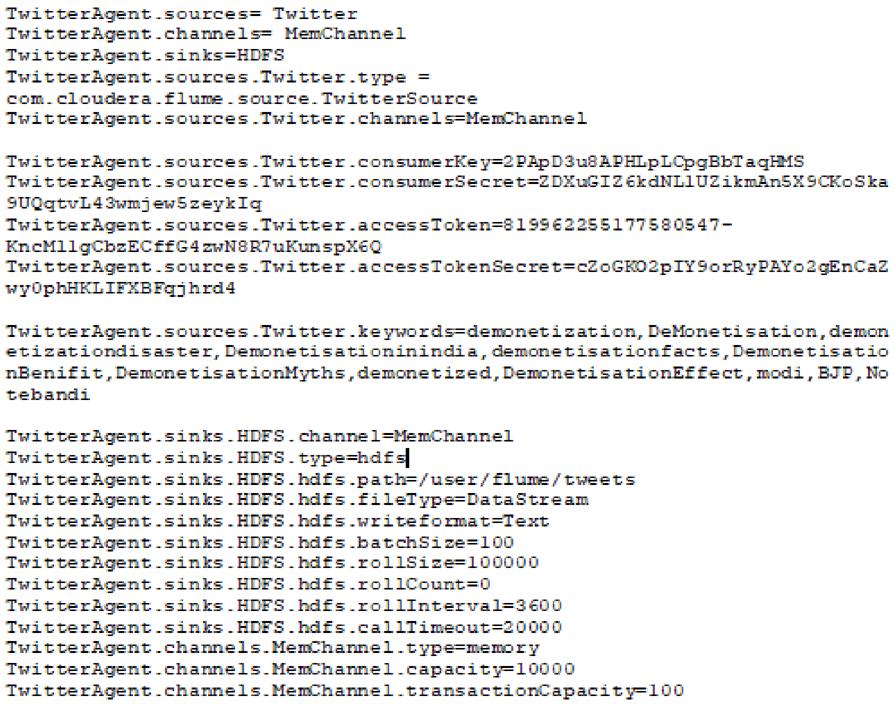
Owner KakdePallavi

Fig. 4.1 account credential

4.2 Retrieving data using Flume.

After creating twitter developer account on twitter site we got the keys and access token that we put on the flume's configuration file that is flume.conf file

In Our paper we took #demonetization, #demonetizationbenifits,#demonetizationinindia, #demonetized as keyword to fetch the tweeter data.



```
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=2PApD3u8APHLpLCpgBbTaqHMS
TwitterAgent.sources.Twitter.consumerSecret=ZDXuGIZ6kdNLlUZikmAn5X9CKoSka
9UQqtvL43wmjew5zeykIq
TwitterAgent.sources.Twitter.accessToken=819962255177580547-
KncMllgCbzECffG4zwN8R7uKunspX6Q
TwitterAgent.sources.Twitter.accessTokenSecret=cZoGKO2pIY9orRyPAYo2gEnCaZ
wy0phHKLIFXBFqjhrd4

TwitterAgent.sources.Twitter.keywords=demonetization,DeMonetisation,demon
etizationdisaster,Demonetisationinindia,demonetisationfacts,Demonetisatio
nBenifit,DemonetisationMyths,demonetized,DemonetisationEffect,modi,BJP,No
tebandi

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=/user/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=100
TwitterAgent.sinks.HDFS.hdfs.rollSize=100000
TwitterAgent.sinks.HDFS.hdfs.rollCount=0
TwitterAgent.sinks.HDFS.hdfs.rollInterval=3600
TwitterAgent.sinks.HDFS.hdfs.callTimeout=20000
TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=100
```

Fig. 4.2 flume.conf file

We can access the tweeter and we got the live streaming data on demonetization.Here we got JSON formatted semi-structured data and its stored in HDFS on provided location that is “user/flume/tweets” in this path we got streaming data in the batch format

4.3 Querying using Hive Query Language (HQL):

After running the flume automatically tweets stored in HDFS. The data we got it's in JSON semi-structured format, that we need to load the tweets into the Hive using JSON format so we create a table where the semi-structured data converted into structure data Through the Hive SerDe. SerDe is inbuilt in Apache Hive and it stands for serializerdeserializer

- First ,we need to add the jar files of SerDe
  ADD jar /usr/lib/apache-flume-1.4.0-bin/lib/hive-serdes-1.0-SNAPSHOT.jar;
- After successfully adding the jar files we creating a table we used below query we stored the tweets id and text for performing the sentiment analysis.

Create external table load_tweets_d(id BIGINT,text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'

LOCATION '/user/flume/tweets';

- **Data-Preprocessing:**

  Tokenization: For analyzing each tweet we need to break it into single words that is uni- grams(n==1)which are called tokens and these tokens are compared with dictionary in order to generate sentiments.
  These are uni-grams(n=1), bi-grams(n=2), tri-grams (n=3),....N-grams.
  The general equation of N-grams is of the next word in a sequence would be; P(wn| w1n-1) ≈ P(wn | wn-N+1n-1) ,where word sequence w1, w2, ... , wn-1 is represented as w1n-1. P(wn| w1n-1) ≈ C(wn−1wn)/C(wn−1).[10]
  For breaking the tweets into individual word we used split **() UDF**function, If we use the split() function to split the text as words, it will return an array of values. So, storing thetweet_id andthe array of words we create another Hive table .

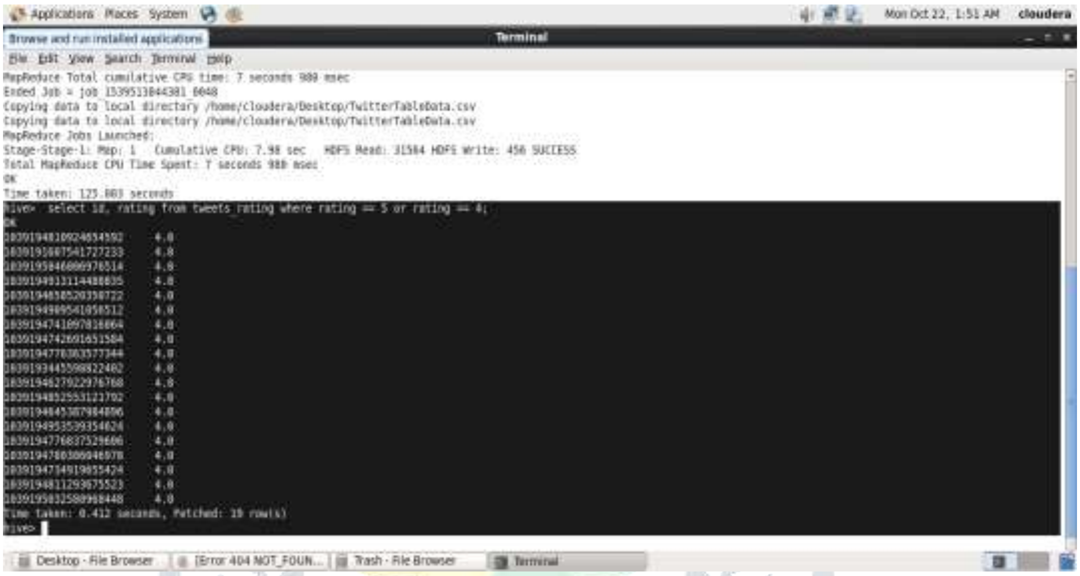
  create table split_words_d as select id as id,split(text,' ') as words from load_tweets_d;
- Now split each word inside the array as a new line. For this we have to use a UDTF(User Defined Table Generating Function). We have used built-in function UDTF called explode which will extract each element from an array and create a new row for each element.
  create table tweet_word as select id as id,word from split_words_d LATERAL VIEW explode(words) was word;

- Now we added the dictionary into HDFS and create the table for dictionary as word and rating. In our paper we have used AFFIN dictionary which consist 2400 word stored as word and its rating ,rated from +5 to -5 depending on their meaning.
  After that we have to join the two tables dictionary table and tweets_word table using below query

  create table word__join_new as select tweet_word.id,tweet_word.word,dictionary.rating from tweet_word LEFT OUTER JOIN dictionary ON(tweet_word.word =dictionary.word);

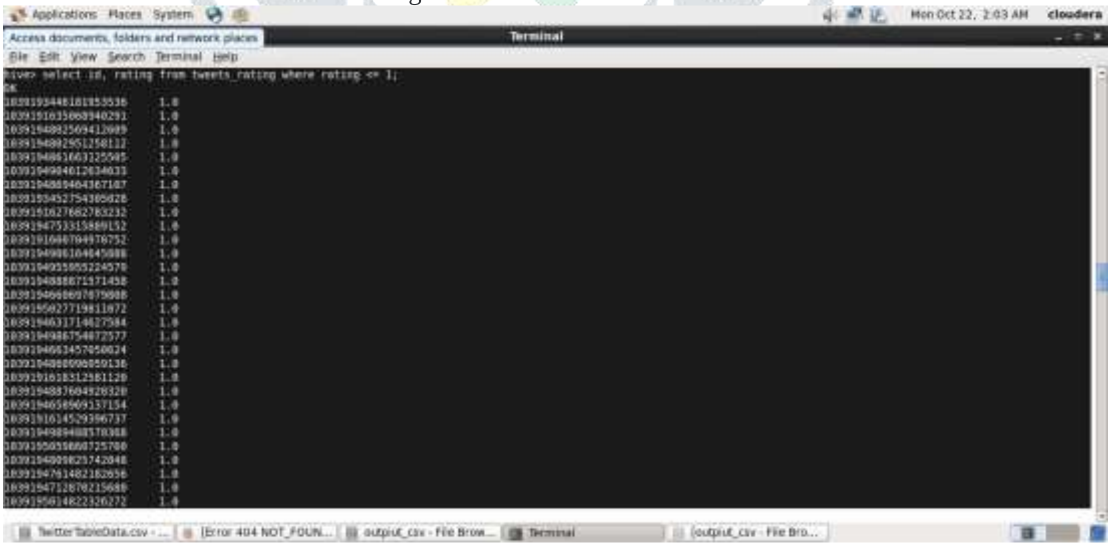
- Now we performed the 'groupby' operation on the tweet_id so that all the words of one tweet came to a single place. And then, we performing an Average operation on the rating of the words of each tweet so that the average rating of each tweet can be found.
  selectid,AVG(rating) as rating from word__join_new GROUP BY word_join_new.id order by rating DESC;

We can filter the data based on rating and decides the Positive, Negative & Neutral tweets (Rating can be from 5,4,3,2,1,-1,-2,-3,-4,-5)



```
MapReduce Total cumulative CPU time: 7 seconds 980 msec
Ended Job = job_1539513844381_0048
Copying data to local directory /home/cloudera/Desktop/TwitterTableData.csv
Copying data to local directory /home/cloudera/Desktop/TwitterTableData.csv
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 7.98 sec   HDFS Read: 31584 HDFS Write: 456 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 980 msec
OK
Time taken: 125.803 seconds
hive>  select id, rating from tweets_rating where rating == 5 or rating == 4;
OK
1039194810924654592     4.0
1039191607541727233     4.0
1039195846006976514     4.0
1039194913114488835     4.0
1039194658520358722     4.0
1039194909541058512     4.0
1039194741097816064     4.0
1039194742691651584     4.0
1039194778363577344     4.0
1039193445598822402     4.0
1039194627922976768     4.0
1039194852553121792     4.0
1039194645387984896     4.0
1039194953539354624     4.0
1039194776837529606     4.0
1039194780300046078     4.0
1039194734919855424     4.0
1039194811293675523     4.0
1039195832580968448     4.0
Time taken: 0.412 seconds, Fetched: 19 row(s)
hive>
```

Fig 4.3 Positive tweets


Applications
Places
System
Mon Oct 22, 2:03 AM
cloudera
Access documents, folders and network places
Terminal
File
Edit
View
Search
Terminal
Help
TwitterTableData.csv - ...
[Error 404 NOT_FOUN...
output_csv - File Brow...
Terminal
[output_csv - File Bro...

```
hive> select id, rating from tweets_rating where rating <= 1;
OK
1039193448181853536     1.0
1039191635068940291     1.0
1039194082569412609     1.0
1039194802951258112     1.0
1039194861663125505     1.0
1039194904612634633     1.0
1039194889464367107     1.0
1039193452754305028     1.0
1039191627682783232     1.0
1039194753315889152     1.0
1039191660784978752     1.0
1039194906164645888     1.0
1039194935585224578     1.0
1039194888871571458     1.0
1039194660697875808     1.0
1039195027719811072     1.0
1039194631714627584     1.0
1039194986754872577     1.0
1039194663457050024     1.0
1039194888096059136     1.0
1039191618312581120     1.0
1039194887604928320     1.0
1039194658909137154     1.0
1039191614529396737     1.0
1039194989488578368     1.0
1039195035866725700     1.0
1039194009825742048     1.0
1039194761482182656     1.0
1039194712870215688     1.0
1039195014822326272     1.0
```

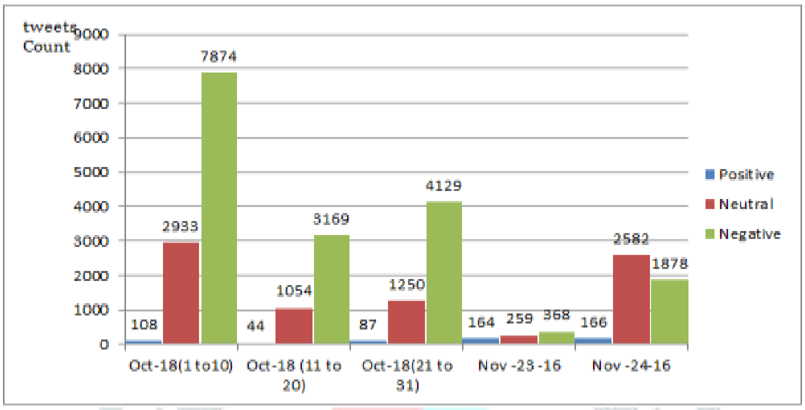
Fig 4.4 Negative Tweets

IV. Results:

We have used dictionary based methodology for the analysis of sentiments describedin the previous section has been used to analyzed sentimentsin a number of ways – sentiment analysis for initial day impactand the current days.. Tweets have beencollected in two phases: first phase after the policy decisionwas announced on November 2016 and the second phase we have collected October 2018 data. Collected data as below:

Table 4.1 Collection of Tweets in Table

Batches	Gathered Tweets	Positive	Neutral	Negative
Nov -23 -16	1495	164	259	368
Nov -24-16	6200	166	2582	1878
Oct-18(1 to10)	14,915	108	2933	7874
Oct-18 (11 to 20)	9490	44	1054	3169
Oct-18(21 to 31)	10,112	87	1250	4129



Graph 4.1 Comparisons of Tweets With Respect to Days in Batches

VII.Performance analysis

6.1 Experimental Setup:

The proposed system was implemented using a 5.4.2 version of Cloudera. Hadoop mostly work in a multimode environment but for research purposes, we used a single node virtual environment is to create an illusion of several nodes. A n Intel Core i5-3210M CPU@2.50GHz processor with 8 GB memory was used to simulate the Hadoop Environment. Data was gathered from Twitter using Apache Flume, a distributed, reliable, and available service for efficiently gathering, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).[11][12]

6.2 Analysis:

Generally, the analysis of sentiment is done by the using the four indexes:
Accuracy with Precision plus Recall and F1-score. These indexes aredepending on the confusion matrix The equations show the indexes:

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F-measure = 2*Precision*recall/(Precision + recall)
- Accuracy = TP + TN /(TP + TN + FP + FN) [13]

We evaluated our experiment results by using following Information Retrieval matrices for that we took 100 tweets

Table 4.5 Performance Evaluation of System on Affine Dictionary

True Positive	True Negative	False Positive	False Negative
7	17	25	5

Precision	Recall	F-measure	Accuracy
0.218	0.5833	0.317	0.73

From above equations and table we got 73 % accuracy in terms of count of tweets gathered from AFFIN Dictionary

VII. Conclusion:

As twitter post are very important source of opinion on different issues and topics .analysis can help decision-making in various areas. It helps in analysis of diversity of topic by just changing the keywords in query. Apache Hadoop is one of best option for tweeter post analysis.

I have collected data in batches on date basis for date of batch 1 October to 10 October our positive result is 2.21 % ,neutral result is 32.80 % negative result is 64.99 % likewise for other batch dates.

I have collected some historic data from 23 November and 24 November 2016 so from that I got positive result 3.59 %, neutral result is 40.60 % and negative result is 55.89%.

Even though demonetization move created adverse short-term policy impact the real impact of demonetization must be assessed in the medium/long term, at this point of juncture we cannot precisely conclude whether demonetization is a failure or a success.

VIII. References:

[1]Dhanya Nary Biju,Yojna Arora Twitter Data Analysis using Hadoop,www.ijariie.com,Vol-4 Issue-5 2018 IJARIIE-ISSN(O)-2395-43969093 www.ijariie.com 306
[2]Prabhsimran Singh, Ravinder Singh Sawhney, Karanjeet Singh Kahlon,SENTIMENT ANALYSIS OF DEMONETIZATION OF 500 & 1000 RUPEEBANKNOTES BY INDIAN GOVERNMENT,ScienceDirect ICT Express,March 2017
[3]Pallavi Kakde, Prof. V. A. Losarwar ,Sentiment analysis of demonetization of 500 & 1000 rupee banknotes Using Apache Flume and Hive,International Journal of Management, Technology And Engineering,Volume 8, Issue IX, SEPTEMBER/2018
[4]Sangeeta,Twitter Data Analysis Using FLUME & PIG on Hadoop Frameworks, Special Issue on International Journal of Recent Advances in Engineering & Technology (IJRAET)Feb 2016
[5]Monika Sharma, Twitter Sentiment Analysis on Demonetization an Initiative Government of India April 2017
[6]KomalSutar, SnehalKasab ,SnehaKindare, Pooja Dhule, Sentiment Analysis: Opinion Mining of Positive, Negative or Neutral Twitter Data Using Hadoop, IJCSN International Journal of Computer Science February 2016
[7]Mahalakshmi R, Suseela,Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data .International Journal of Advanced Research in Computer and Communication Engineering May 2015
[8]Ajinkya Ingle, Anjali Kante, ShriyaSamak, Anita Kumari ,Sentiment Analysis of Twitter Data Using Hadoop,Dec 2015
[9]KomalSutar, SnehalKasab ,SnehaKindare, Pooja Dhule Sentiment Analysis: Opinion Mining of Positive,Negative or Neutral Twitter Data Using Hadoop,IJCSN International Journal of Computer Science February 2016
[10]JyotiYadav,SonalAroraReview Paper on Sentiment Analysis of The Demonetization of Economy ,International Journal on Future Revolution in Computer Science & Communication Engineering,Volume: 4 Issue: 4 April 2018
[11]Pratik B,Shalvi Naresh Raut,AshwiniRatanPatil,AbhayPatil,Public Opinion Analysis Using Hadoop,International Journal on Recent and Innovation Trends in Computing and Communication March 2017
[12]Ramesh R DivyaG,Big Data Sentiment Analysis using Hadoop,IJIRST –International Journal for Innovative Research in Science & Technology| Volume 1 | Issue 11 | April 2015
[13]https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c
[14]https://medium.com/digital-trends-index/sentiment-analysis-a-better-way-to-measure-success-3b40a7d5d89
[15] https://cubalytictalks.blogspot.com/2018/08/confusion-matrix.html