Comparison of Hadoop Ecosystem on Twitter's Hashtag Dataset

Dr. D. S Rajpoot
UIT, RGPV, Bhopal

Jay Prakash Maurya
Dept. of CSE, LNCT, Bhopal

Abstract. **Intelligence gathering from Social media is a challenging area in data analytics today. Twitter is a popular social media and the primary source of information for real time data sharing. Twitter uses very short and suited symbols for knowledge and these symbols helps for message predictions. Twitter server provides API's by which a twitter user can gather and analyse data of tweets. Twitter 's API provides data in more informative form that is helpful for interesting research work . This paper is based on gathering, mining, and knowledge discovery & visualisation from Twitter data. This paper highlights a process to collect twitter's data, store and analyse the data sets on Hadoop Ecosystem.**

Keywords **:** *Hadoop. Hashtag. Bigdata. HIVE. PIG.*

I. Introduction

Twitter is an online social network, by which users can share short messages, called tweets. Service was launched in 2006 with headquarters in California, USA. During the years of its deployment gained worldwide popularity with 320 million active users. Users signed in to service can share their moments with friends, include links to pictures or videos in tweets, make comments and re-tweet other tweets. Twitter also supports functionality of following other users [1]

Apart from creating personal accounts, many enterprises use for feedback of products, promote company activities, discounts or newsletters. Therefore, Twitter became interesting platform for branding and marketing. Hence, data mining on Twitter can result in interesting results about users and create value for companies. Based on the article, "Twitter data strategy chief Chris Moody" talks about cases when data mining can took place in real-world scenarios on Twitter data. For example, thanks to Twitter, Aircraft Company could surprise one of their customers travelling on board with little present, when they discovered she was travelling to see her new grandchild. Chris arguing that such scenarios can happen on daily basis. Moreover, he says that Twitter gives opportunity to understand people in context like never before.

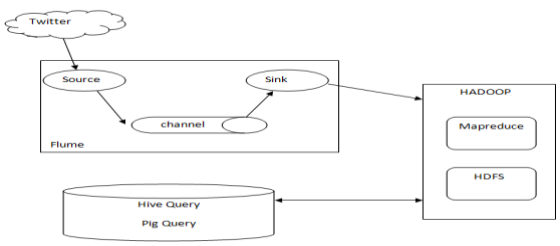

Fig 1. Solution for analysing twitter data

Challenges

There are different challenges in information filtering in micro-blogging environment. They are as follows:

- **Short text**: In Twitter, the text is restricted to 140 characters per post. In terms of text classification, short texts contain sparse data; therefore it is a challenge to classify them.
- **Informal Language:** Another challenge is the informal structure of the language used on Twitter. It contains slangs, abbreviations, stop words etc. It is important to identify keywords and common words useful for text classification.
- **Different Languages:** Twitter is used by users around world in different Languages; therefore it contains tweets in many languages.
- **Identifying topics**: It is necessary to identify relevant topics and identifying relevant tweets against irrelevant topics.
- **Constantly changing vocabulary:** The vocabulary is constantly changing with new words and phrases.

Tools used

- *Hadoop*
- *Apache Flume*
- *Apache Hive*

III. Problem Definition

The work focuses on sentiment analysis of the most popular micro blogging platform, Twitter. The tweets are important for analysis because tweets are in high frequency, real timed. Tweets generates bulk information related to government, election, disasters etc. Millions of tweets is generated & shared every day and useful in decision making about opinions on issues. The algorithms that process tweets are under strict constraints of storage and time. The analysis of tweeted data gives real view of user opinions about their thinking and provide a better way for making any decision. Their Should be a best approach for tweets analytic for betterment of decision.

IV. Proposed Work

This work uses a powerful tool designed for analysis and transformation of Bigdata, Hadoop & map reduce.

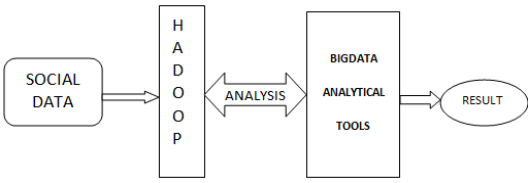

Fig 2. Workflow Diagram

The work uses algorithm for handling Bigdata and the dynamic data characteristics for performing operation on social media data sets. For analysing, work is based on hadoop with single node ubuntu machine to solve the challenges of big data through MapReduce framework [12]. The huge data is mapped in frequent dataset and reduced in smaller size for chuncks of handling, after that big data analytical tools are used to refine such unstructured data and data analysis is applied.

V. Proposed Methodology

Steps:

1. Fetch real time social data by API's [3] and store it HDFS format.(For fetching social data bigdata tools apache flume and kafka),
2. Pre- process on data to provide structure to the Data.
3. Analyzing huge data.

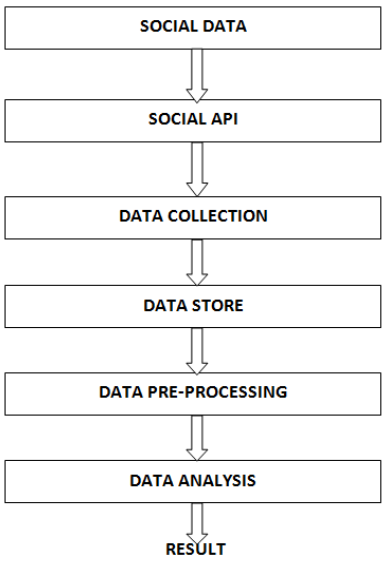

Fig 3. Analysis Steps

VI. Result

Result are taken on configuration of system having an i5-Processor- CPU @ 2.30 GHz clock and 4 GB of RAM running ubuntu14.0 [9]. To achieve the result follow given steps:

- Twitter Application login .
- Start Flume for getting data.
- Run query by Hive for Analysis.
- Optimizing query .

A. *Creating Twitter Application*

Create an account in Twitter developer and create an application shown in fig. 4. Create access tokens to provide authentication and also for creating application. The access token have consumer keys to access application for seeking Twitter data. These keys and token helps to configuration of Flume. The required data from the Twitter is in the form of tweets returned by flume.

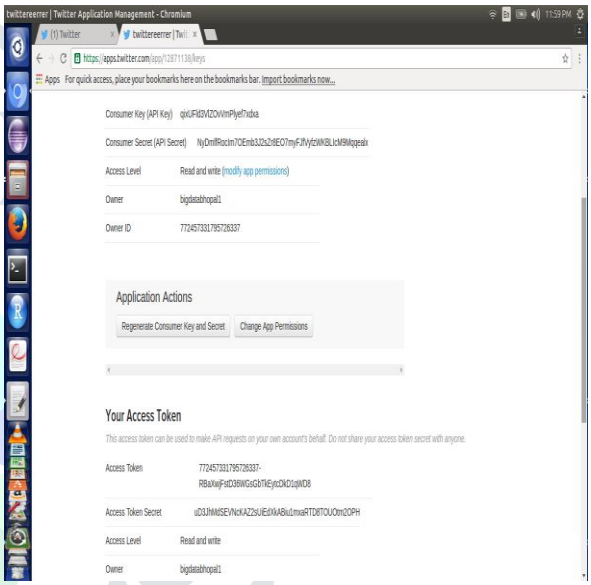

Fig 4. Creating twitter app and Generation access token keys

B. *Getting data with Flume*

Using Twitter developer website the consumer key and secret along with the access token and secret values are used to get twitter data in JSON format and is stored in the HDFS (Configuration file shown in figure-5). The work used GST topic for data from tweet, data are taken into three different slots by it is easy to predict the trends accurately, all configurations are in flume-twitter.

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
Â
TwitterAgent.sources.Twitter.type =
com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey =
Rx7wBGq7jJnQoqynJ7JYeSVMR
TwitterAgent.sources.Twitter.consumerSecret =
mPepi0riR5BYbRV6WP4Th7IxOQdv3ZLbxuepfdCEHMRQkLVZgp
TwitterAgent.sources.Twitter.accessToken = 787561696298622976
634YxHgRECs1EUB463neqVyLyUqamRi
TwitterAgent.sources.Twitter.accessTokenSecret =
X77rhwXqm5VjxzDAcWep5iHMGrZOKxcbjAwbxwgkbswa1
Â
TwitterAgent.sources.Twitter.keywords = GST, gst
Â
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:9000/user/flume/tweets/
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
Â
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Fig 5. Fetching Data using Flume with keyword GST

C. *Querying using Hive*

After setting the path to store data,the Twitter data has been extracted by using Flume which is in HDFS file system shown in figure-6. The raw data those we got form the Twitter is also in the JSON format. Custom *serde* concepts was used before querying, these serde properties were used in work to structure the table so that its was easy to query. These concepts are nothing but how we are going to read the data that is in the form of JSON.

Ensure that the Hive table can properly be interpret by JSON data. Hive accepts the input files in delimited row format, but Twitter data is in a JSON format. SerDe (Serializer and Deserializer) is used, which is an interfaces that tells Hive how it should translate the data into another format (Hive can process). In work, a jar file was added by following command

ADD JAR <path-to-hive-serdes-jar>; (1)

The figure 6 shows the structure data stored in table tweets.

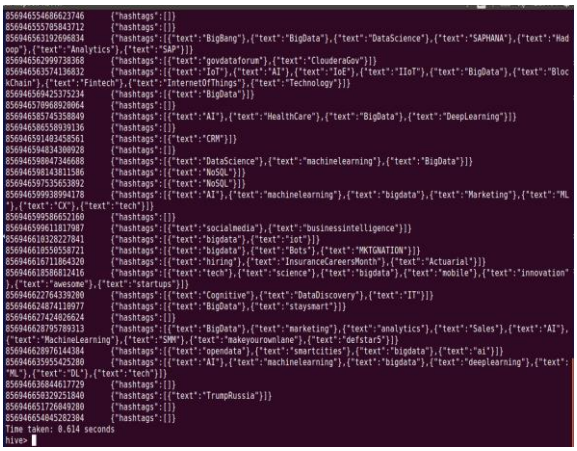

Fig 6. Data store into table tweets

After that from the table tweets the tweets id and the array of hashtag keywords with some unwanted keywords like text , hashtags are pruned. The keywords after pre-processing the tweets table and collect only the hashtag keywords. And finally the top hashtag along with its frequency value for different slots are collected. Figure 7, 8, 9 shows the result hashtag keywords for different slots.

```
hive> select * from top_result;
OK
GST     58
TransformingIndia       6
IndiaForGST     5
Modi    4
photoshop       3
चरचाकाविषय_है  3
gst     3
SonuSong        3
GSTForNewIndia  3
CarolinaWx      3
Time taken: 8.412 seconds
hive>
```

Fig 7. Keywords along with frequency of slot1

```
hive>
    > select * from top_result;
OK
GST     52
Modi    5
CarolinaWx      5
GSTForNewIndia  4
SonuSong        4
photoshop       4
Bihar   3
Brighton        3
3Novices        2
BJP     2
Time taken: 0.192 seconds
hive>
```

Fig 8. Keywords along with frequency of slot2

```
    > select * from top_result;
OK
GST     71
Modi    12
modi    10
parliment       10
Baahubali       6
Brighton        3
GSTForNewIndia  2
GSTMasterClasses        2
NitishKumar     2
150thBirthday   1
Time taken: 0.449 seconds
hive>
```

Fig 9. Keywords along with frequency of slot3

Using hashtag keywords along with its frequency on different slots, the common hashtag keywords with its frequency can be found, Shown in figure 10.

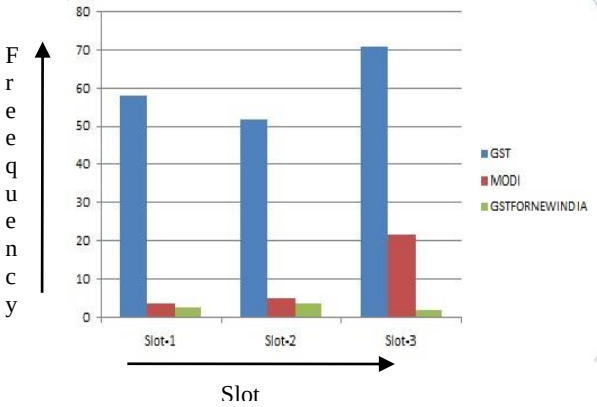

Fig 10. Common keywords of three slots

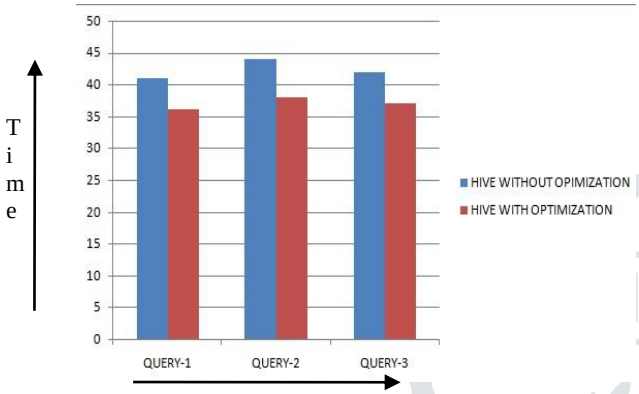

Fig 11. Execution time taken by query on hive tables

D. *Optimizing Query Performance*

For optimize the hive query performance, serialization process is applied on starting table and store the resultant table into new table and then apply the entire query on new table. The result is faster as compared to previous table before serialization on tables. The results are taken after execution of different query on two hive table's first table is with optimization and second is without optimization for which the output of queries with different execution time and the time taken by query is shown in figure 11.

E. *Comparison of Hive & Pig*

After getting the query execution time taken by pig is less then hive for analyzing JSON data is. From these result we can say that pig is best suitable for analyzing JSON data. For these twitter data analysis pig is generating less number of map-reduce job and hive is generating more number of mapreduce job for analyzing twitter data, so pig is better in many parameters as compared to hive.

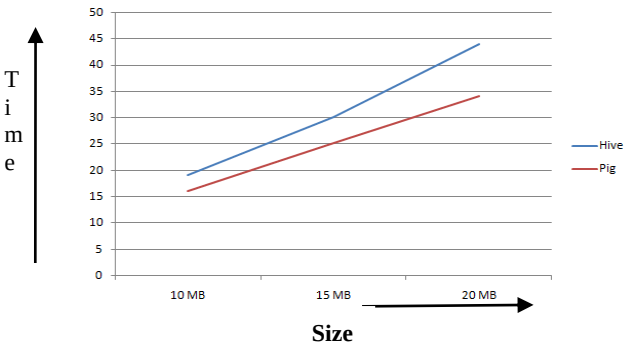

Fig 12. Comparison of HIVE & PIG

VII. Conclusion

Paper Shows the original dataset is in HDFS file system and Hadoop Mapreduce model. This study analyses the data-sets in HDFS, on two ecosystems Hive & Pig respectively, which are more scalable and efficient than traditional Relational Database Systems. Experiment results are the comparative study between Hive and Pig.

The paper work is based on, fetching real time twitter data and store Query DFS and then we develop an trend analysis using hive to analyse the twitter hashtag keywords with its frequency. Using the methodology the trendiest keywords right now on the twitter through hashtag popularity level can be found. The work can be extended towards database query optimization using optimization techniques on hive tables.

References

[1]. Shing H. Doong, "Predicting Twitter Hashtags Popularity Level", in 2016 49th Hawaii International Conference on System Sciences,IEEE,DOI10.1109/HICSS.2016.247

[2]. Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced Analytics Tools", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3]. "Twitter's API HowStuffWorks." HowStuffWorks. N.p., n.d. Web. 24 Oct. 2014.

[4]. Judith Sherin Tilsha S , Shobha M S, "A Survey on Twitter Data Analysis Techniques to Extract Public Opinion", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[5]. Ramesh R, Divya G, Divya D, Merin K Kurian , "Big Data Sentiment Analysis using Hadoop ", (IJIRST)International Journal for Innovative Research in Science & Technology,Volume 1 , Issue 11 , April 2015 ISSN : 2349-6010

[6]. Praveen Kumar, Dr Vijay Singh Rathore," Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014, pp 7123-7126.

[7]. G.Vinodhini , RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

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

[9]. Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[10]. Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, "Big Data Analytics: Hadoop and Tools" in 2016 IEEE Bombay Section Symposium (IBSS), IEEE 2016.

[11]. Can Uzunkayaa, Tolga Ensaria, Yusuf Kavurucu, "Hadoop Ecosystem and Its Analysis on Tweets" in World Conference on Technology, Innovation and Entrepreneurship, Procedia - Social and Behavioral Sciences 195 (2015) 1890 – 1897, Elsevier 2015.

[12]. White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel Corporation.

[13]. https://hadoop-apache.org/docs/r1.2.1/ streaming.html#Hadoop+Streaming