iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4807

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Twitter Sentiment Analysis using Hive

Stuti Saraswat[1], Vani Karkare[2], Deepali Garg[3], Sachin Goyal[4], Ratish Agarwal[5]
[1, 2, 3, 4, 5]*Department of Information Technology, UIT RGPV, Bhopal, India UG Student, Department of IT, UIT RGPV, Bhopal, India*

Abstract: One of the most important current research areas is sentiment analysis. Today, micro blogging websites have become a source of a wide variety of information. This is because people now prefer micro blogs to post their opinions about various products and services. Now, companies manufacturing these products have started to analyze these blogs to get a gist of the overall sentiment about a particular product. A big challenge is to build models and technologies to summarize this overall sentiment. In this paper, we take into consideration, a popular micro blogging website, Twitter and provide a methodology for categorizing tweets into positive, negative or neutral sentiment using Apache Hadoop and Hive.
Keywords: Big Data, Apache Hadoop, Sentiment Analysis, Apache Hive, MapReduce

I. INTRODUCTION

There has been a lot of advancement in almost every area. But, many products still fail to satisfy the needs and expectations of their customers. Specific teams in every company work in order to produce a quality product for their customers and therefore rely on survey results. These surveys can be seen on many websites. But people don't prefer to fill up these surveys as they are lengthy and thus consume a lot of time and effort. And, with the evolution of the Internet and social media, people now prefer using social media for posting their reviews. So, the best solution to this problem is opinion mining or sentiment analysis. Twitter is one of the most popular websites and is used by a large population. Therefore, it receives millions of reviews everyday on diverse issues and products. These reviews can then be used for decision making.

II. LITERATURE SURVEY

Bigdata is a collection of large datasets consisting of data in the range of petabytes, zetabytes and its complexity makes it difficult to process using traditional data processing technologies. This paper gives study of few bigdata emerging technologies.[1] The large amount of data generated from micro blogging websites like twitter can be used for decision making. One of the best tools that can be used for twitter data analysis is Hadoop. Hive, which is a query language is used for analysis.[2] Hadoop MapReduce is a framework for writing applications that process large amounts of data in parallel and on large clusters. MapReduce refers to two tasks, a Map task and a Reduce task.[3] The tools presently used for analysis of data are slow, costly and are unable to handle Bigdata. Therefore, these cannot be used for real-time analysis. Hadoop, an Apache open source platform can be used for real time analysis.[4] Social media has become an important platform for sharing information. Sentiment analysis is process of classification whereby opinions are classified as positive, negative or neutral and used for decision making.[5]

III. SENTIMENT ANALYSIS

A sentiment can be defined as the expression or opinion of an author on a particular object or aspect.These sentiments, found within comments and feedbacks prove to be useful indicators for many different purposes.
Sentiments can be categorized as positive or negative or according to an n-point scale ,for example- good, satisfactory, bad, very bad, etc..Sentiment analysis, as the name suggests is a categorization process and each category represents a sentiment. It automates the process of detection of subjective information. In simple words, the purpose of sentiment analysis is the extraction of information about the attitude of the writer or speaker on a particular topic or the polarity of a document.Sentiment analysis is useful for companies because it provides a means to evaluate product acceptance and to build strategies for improving the product quality.

IV. APACHE HADOOP

Since Apache Hadoop works for distributed big data, it is a good choice for twitter data analysis. Hadoop is an open source software framework for distributed storage and distributed processing of large data-sets on clusters. Hadoop MapReduce is a framework for writing applications that process large amounts of data in parallel and on large clusters.
Hadoop MapReduce is a framework for writing applications that process large amounts of data in parallel MapReduce refers to two separate tasks that a Hadoop program performs-
Map task: It takes input data and changes it into a set of data and then the individual elements are separated down into tuples.

Reduce task: The output from the Map task acts s input to the Reduce task. It combines data tuples into a small set of tuples. It always occurs after the Map task.

For every MapReduce there is a JobTracker node and it is accountable for the distribution of mapper and reducer functions and also for monitoring the results. The actual job is done by the TaskTracker and the result is then returned to the JobTracker.

Hadoop is supported by Linux and all its flavors. If we have some other operating system, then we can first install the virtualbox software and then install Linux on it.

Hadoop framework consists of many modules. These are MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper and Hbase. These modules provide different functionalities.We have used Hive for our work.

V. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

Hadoop is capable of working on any mountable distributed file system. But, the most common choice for file system when using Hadoop is Hadoop Distributed File System (HDFS). It is a file system which is used for storing large sets of data in a block with size 64MB .This data is stored in a distributed manner on Hadoop cluster. A cluster is a set of daemons running on different servers of the network. Hadoop Distributed File System follows master slave architecture. Master is single Namenode which is responsible for the management of file system metadata. There is one or more slave nodes called Datanodes which store the actual data. Secondary Namenode acts as a backup for the Namenode. A file is split into a number of blocks and these blocks are stored in various Datanodes. The mapping of blocks to Datanodes is determined by the Namenode. The Datanodes are responsible for read and write operations and also look after block creation, replication and deletion according to the instructions of the Namenode. HDFS has a shell and a number of commands are available to interact with it.

VI. APACHE HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. It stores schema in a database and processed data into HDFS. It provides SQL type language for querying called HiveQL or HQL. It is familiar , fast , scalable and extensible

A. Workflow Between Hive And Hadoop

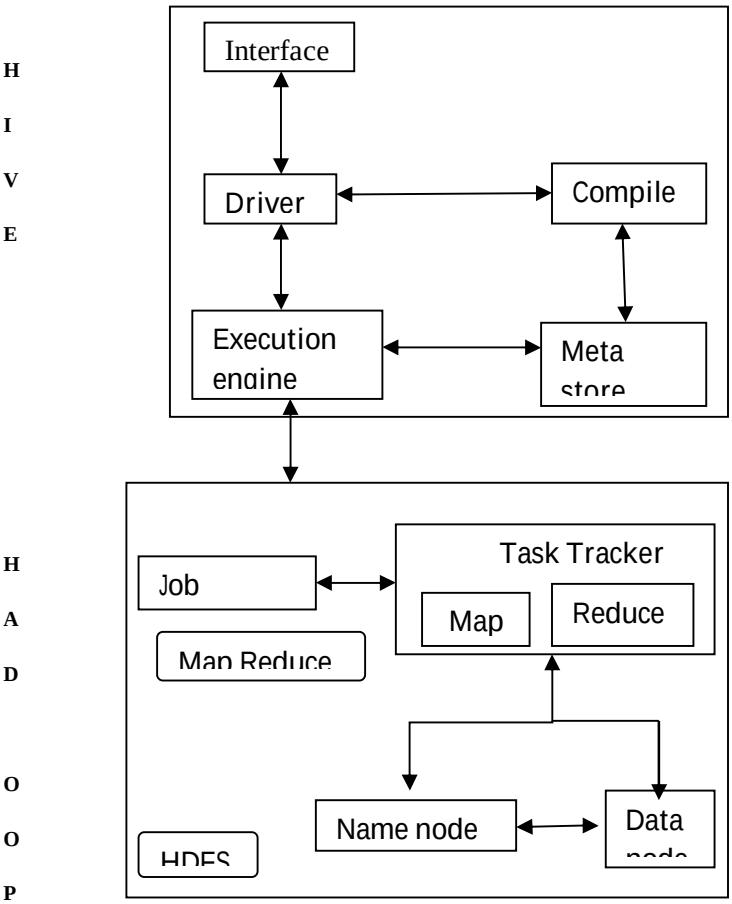


Fig 1 Working of Hadoop and Hive

B. Steps of Working

1) Execute Query: Hive interface sends query to be executed to the driver .
2) Get Plan: The query compiler parses the query to check the syntax and query plan.
3) Get metadata: The compiler sends metadata request to the metastore.
4) Send metadata: The metastore sends metadata to the compiler as a response.
5) Send plan: The compiler checks the requirements and sends the plan to the driver.
6) Execute plan: The driver sends the plan to be executed to the execution engine.
7) Execute job: Execution engine sends the job to the JobTracker (Namenode) and it assigns the job to the TaskTracker (Datanode). Here, the query executes the MapReduce job.
8) Fetch result: The execution engine receives the result from the Datanode.
9) Send result: The execution engine sends the result to the driver.
10) Send result: The driver sends the result to Hive interface.

VII. METHODOLOGY

As the tweets coming in from twitter are in Json format, we need to load the tweets into Hive using json input format. We will use Cloudera Hive jsonserde for this purpose.

After downloading Cloudera Jsonserde, we need to copy the jar file into *lib* directory of your installed Hive folder. We need to ADD the jar file into Hive as shown below:

A. Syntax

1) ADD jar 'path of the jar file: After successfully adding the Jar file, we need to create a Hive table to store the Twitter data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

For performing Sentiment Analysis, we need the tweet_id and tweet_text, so we will create a Hive table that will extract the id and tweet_text from the tweets using ClouderaJsonserde. The tweet is in nested json format. From this tweet we will extract the id, which is the tweet_id and text, which is the tweet_text.

B. Workflow

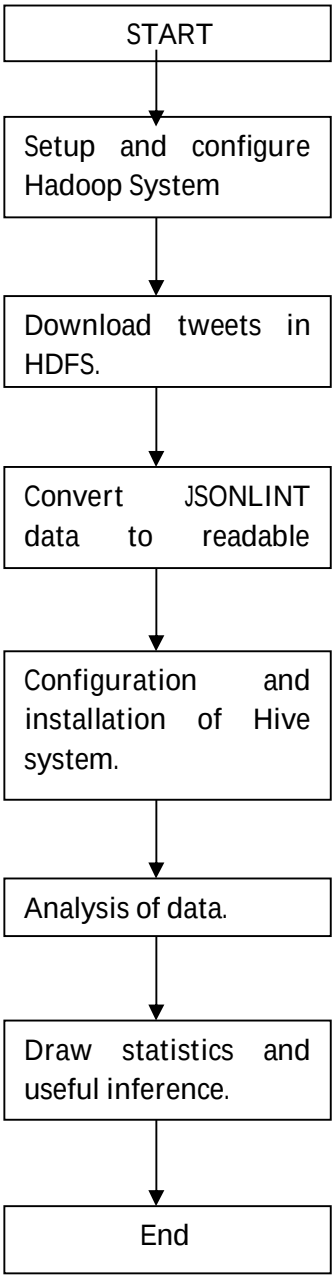


Fig 2 Flow Diagram of proposed methodology

The command for creating a Hive table to store id and text of the tweets is as follows:

create external table load_tweets(id BIGINT,text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user/flume/tweets'
We can check the schema of the table using the below command:
describeload_tweets;

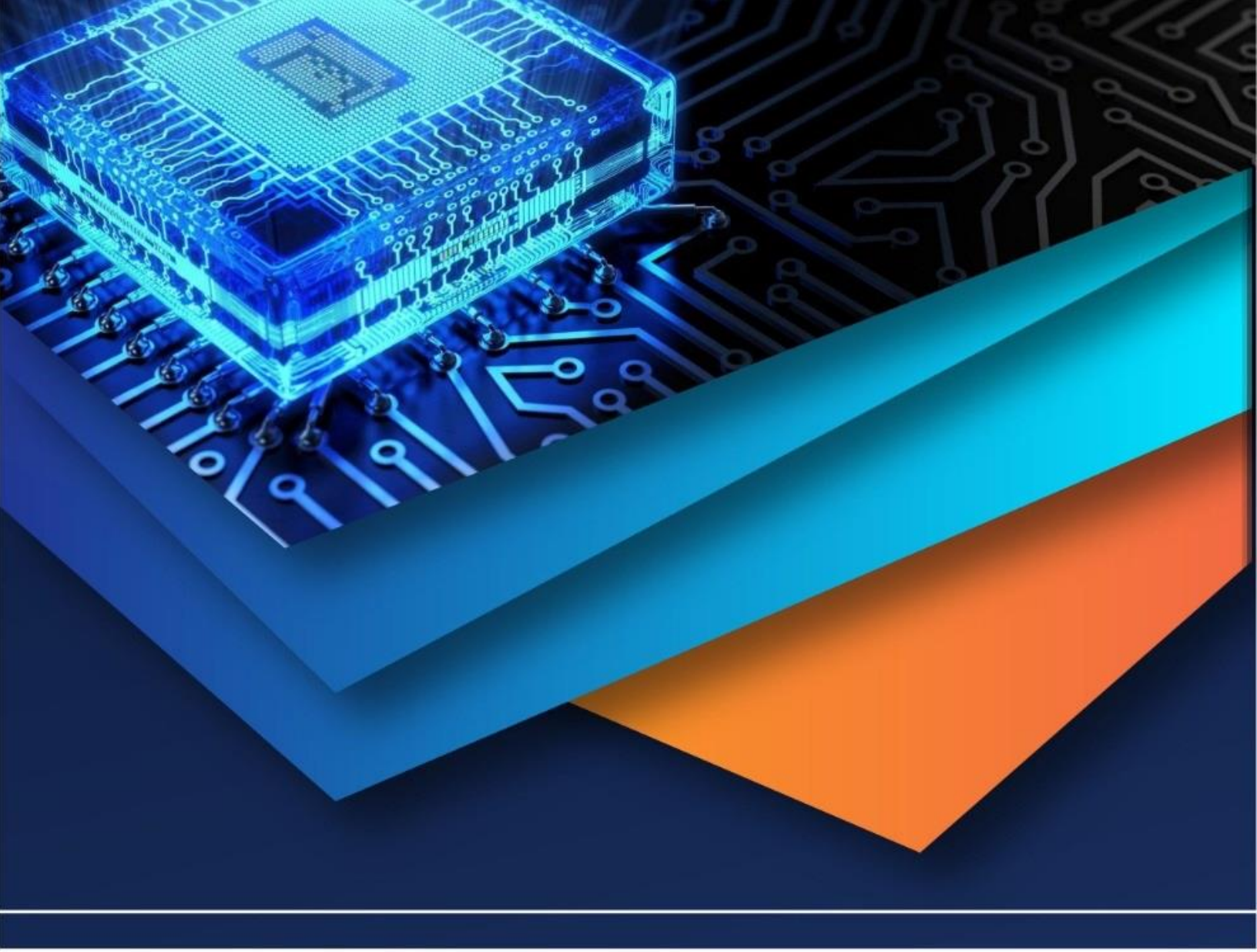
We can view the tweet_id and tweet_text which are present in the table by using the below command:
select * from load_tweets;
Next, we will split the text into words using the split() UDF available in Hive.
If we use by using the following command:
select * from tweet_word;
the split() function to split the text as words, it willreturn an array ofvalues. So, we willcreate another Hive table and store the tweet_id and the array of words.
create table split_words as select id as id,split(text,' ') as words from load_tweets;
We can see the schema of the table by using the 'describe' command. Now, we can view the contents of the table by using the below command:
select * from split_words;
Next, let's split each word inside the array as a new row. For this we need to use a UDTF (User Defined Table Generating Function). We have built-in UDTF called explode which will extract each element from an array and create a new row for each element.
Now, we create another table which can store id and word.
create table tweet_word as select id as id,word from split_words LATERAL VIEW explode(words) w as word;
Syntax for LATERAL VIEW explode UDTF is as follows:
lateralView: LATERAL VIEW udtf(expression) tableAlias AS columnAlias (',' columnAlias)*fromClause: FROM baseTable (lateralView)
In general, explode UDTF has some limitations; explode cannot be used with other columns in the same select statement. So we will add LATERAL VIEW in conjunction with explode so that the explode function can be used in other columns as well.
We can see the schema of the table by using the 'describe' command.
We will use dictionary called AFINN to calculate the sentiments. AFINN is a dictionary which consists of 2500 words rated from +5 to -5 depending on their meaning.
create table dictionary(word string,ratingint) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
Now, let's load the AFINN dictionary into the table by using the following command:
LOAD DATA INPATH '/AFINN.txt' into TABLE dictionary;
We have this AFINN dictionary in the root directory of HDFS. We can view the contents of the dictionary table by using this command:
select * from dictionary;
Now, we will join the tweet_word table and dictionary table so that the rating of the word will be joined with the word.
create table word_join as select tweet_word.id,tweet_word.word,dictionary.rating from tweet_word LEFT OUTER JOIN dictionary ON(tweet_word.word =dictionary.word);
We can view the contents of the table by using the below command:
select * from word_join;
Now we will perform the 'groupby' operation on the tweet_id so that all the words of one tweet will come to a single place. And then, we will be 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 fromword_join GROUP BY word_join.id order by rating DESC;
In the above command, we have calculated the average rating of each tweet by using each word of the tweet and have arranged the tweets in descending order as per their rating.

VII. CONCLUSION

One of the latest areas of interest for research communities around the globe is Bigdata analytics. The data generated from social media websites like twitter is an important source of information. Analysis of this social media data can help in decision making on various topics. Hadoop is one of the best technologies for twitter data analysis. Once the system is set up, analysis can be done using Hive by changing the keywords in the query. Apache Hive provides an easy to use platform for MapReduce programming for people who are comfortable in using SQL. It provides a way to analyze real time, ever-increasing data. Through this analysis businesses can make informed decisions.

REFERENCES

[1] Sunil B. Mane , Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde , "Real Time Sentiment Analysis of Twitter Data Using Hadoop", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 , ISSN:0975-9646
[2] Mahalakshmi R, SuseelaS , "Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2015 , pp 304-306, ISSN : 2278-102
[3] Xing Fang and Justin Zhan "Sentiment analysis using product review data", Journal of Big Data (JBD) Springer, 2015, pp: 1 -14
[4] Qureshi, S. R., & Gupta, A, "Towards efficient Big Data and data analytics: A review", IEEE International Conference on IT in Business, Industry and Government (CSIBIG),March 2014 pp-1-6
[5] Patnaik, L. M, "Big Data Analytics: An Approach using Hadoop Distributed File System.", International Journal of Engineering and Innovative Technology (IJEIT), vol 3, May 2014, pp. 239-243
[6] Singh, J., &Singla, V, "Big Data: Tools and Technologies in Big Data" International Journal of Computer Applications, 2015
[7] Bhandarkar, M, "MapReduce programmingwith apache Hadoop," IEEE International Symposium on Parallel & Distributed Processing (IPDPS), 2010, pp.1-2
[8] Batool, R., Khattak, A. M., Maqbool, J., & Lee, S, "Precise tweet classification and sentiment analysis," IEEE 12th International Conference on Computer and Information Scienc
[9] Wu, X., Zhu, X., Wu, G. Q., & Ding, W, "Data mining with big data." IEEE Transactions on Knowledge and Data Engineering, 2014, pp.97-107
[10] Zhang, F., Cao, J., Khan, S. U., Li, K., & Hwang, K, "A task-level adaptiveMapReduce framework for real-time streaming data in healthcare applications " Future Generation Computer Systems Elsevier,2015, pp.149-160.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY