Twitter Data Analysis Using FLUME & HIVE on Hadoop FrameWork

Sangeeta

Asst. Professor, Department of Computer Science & Engineering , Gurgaon Institute of Technology & Management
Gurgaon, Haryana, India
Email: Sangeeta.yogi@gmail.com

Abstract : Twitter, one of the largest social media site receives millions of tweets every day on variety of important issues. This huge amount of raw data can be used for industrial , Social, Economic, Government policies or business purpose by organizing according to our requirement and processing. Hadoop is one of the best tool options for twitter data analysis as it works for distributed Big data , Streaming data , Time Stamped data , text data etc. This paper discuss how to use FLUME and HIVE tool for twitter post analysis. FLUME is used to extract real time twitter data into HDFS. Hive which is SQL like query language is used for some extraction and analysis.

Keywords : FLUME , HIVE , Hadoop, MapReduce, HDFS.

I. INTRODUCTION

Micro blogging today has become a very popular communication tool among Internet users. Twitter, one of the largest social media site receives millions of tweets every day on variety of important issues. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. These posts analysis can be used for decision making in different areas like government , Elections, Business, Product review etc. Also sentiment analysis is one of the important area of analysis of twitter posts that can be very helpful in decision making.

Performing Sentiment Analysis on Twitter is trickier than doing it for large reviews. This is because the tweets are very short (only about **140 characters**) and usually contain slangs, emoticons, hash tags and other twitter specific jargon. For the development purpose twitter provides streaming API which allows the developer an access to 1% of tweets tweeted at that time bases on the particular keyword. The object about which we want to perform sentiment analysis is submitted to the twitter API's which does further mining and provides the tweets related to only that object. Twitter data is generally unstructured i.e use of abbreviations is very high. Also it allows the use of **emoticons** which are direct indicators of the author's view on the subject. Tweet messages also consist of a **timestamp** and the **user name**. This timestamp is useful for guessing the future trend application of our project. **User location** if available can also help to gauge the trends in different geographical regions.

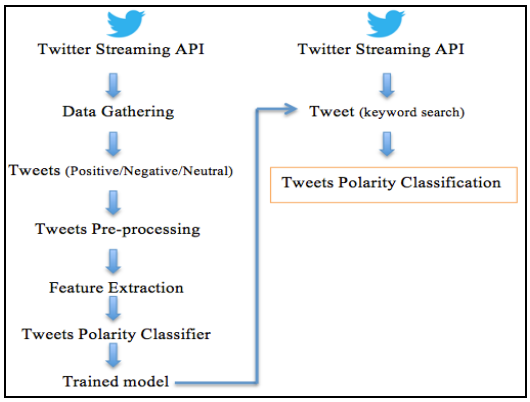

Fig 1 : Twitter Analysis Steps

For doing twitter data analysis first data is collected using FLUME in local HDFS . Tweets are preprocesses for removing noise and meaningless symbols. Feature vector is extracted using Unigram or N-Gram . After that HIVE can be used for twitter posts analysis.

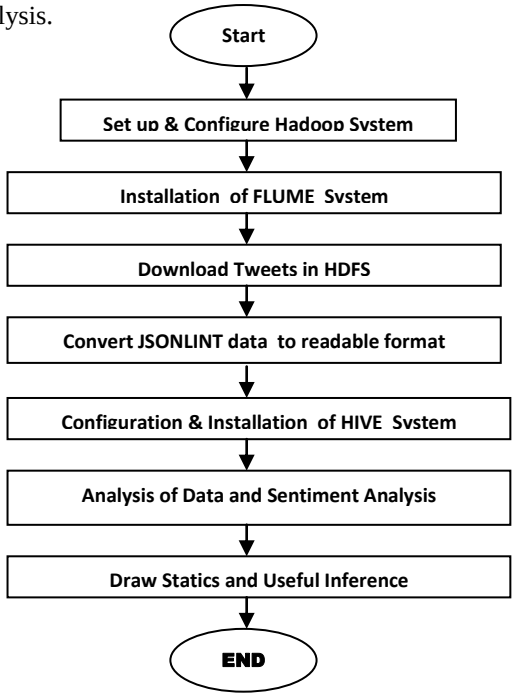

Fig 2 : Installation & Work Flow

The above diagram shows the complete step wise working of twitter posts analysis.

II. INTRODUCTION TO HADOOP

Apache Hadoop is good choice for twitter analysis as it works for distributed big data. Apache Hadoop is an open source software framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop **MapReduce** is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Hadoop framework includes different modules like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase for different functionality as shown in below diagram. I will be using FLUME and HIVE for twitter analysis.

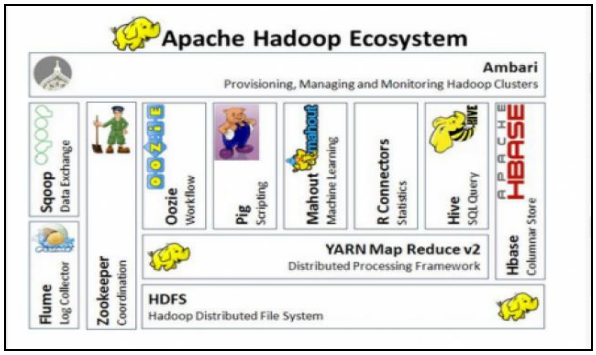

Fig 3 : Apache Hadoop Framework

Hadoop use HDFS (Hadoop Distributed File System) file system**.** The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data. Benefit of using Hadoop is distributed storage , Distributed Processing , Security, Reliability, Speed , Efficiency, Availability, Scalability and lots more. This is the reason of using Hadoop for tweet processing.

III. FLUME

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS. After the installation of VMWRE and Hadoop for single node next step come the installation of FLUME. For this you need to log in to twitter. After that go to apps on twitter and create an new application. After you agree with all terms and conditions you will got new application. Then set Consumer Key , Consumer Secret , Owner Key and Owner Secret ID . Now access token need to be created. After the creation of access token and refresh you will get all the 4 information. Now you Go to flume home and download Apache Flume.

Download the flume-sources-1.0-SNAPSHOT.jar and add it to the flume class path as shown below in the conf/flume-env.sh file

FLUME_CLASSPATH="/home/training/Installations/apache-flume-1.3.1-bin/flume-sources-1.0-SNAPSHOT.jar"

This will automatically be dumped in downloads. Store in desired library. You need to go to apace flume , then go to downloads and extract here and place it In bin and lib. After this you need to configure the file Flume-twitter.conf .

The flume.conf should have all the agents defined as below:

1 TwitterAgent.sources = Twitter

2 TwitterAgent.channels = MemChannel

3 TwitterAgent.sinks = HDFS

4 TwitterAgent.sources.Twitter.type= com.cloudera.flume.source.TwitterSource

5 TwitterAgent.sources.Twitter.channels = MemChannel

6 TwitterAgent.sources.Twitter.consumerKey = <consumerKey>

7 TwitterAgent.sources.Twitter.consumerSecret = <consumerSecret>

8 TwitterAgent.sources.Twitter.accessToken = <accessToken>

9 TwitterAgent.sources.Twitter.accessTokenSecret = <accessTokenSecret>

10 TwitterAgent.sources.Twitter.keywords = Narender Modi, BJP, Election

11 TwitterAgent.sinks.HDFS.channel = MemChannel

12 TwitterAgent.sinks.HDFS.type = hdfs

13 TwitterAgent.sinks.HDFS.hdfs.path =

14 hdfs://localhost:9000/user/flume/tweets/

15 TwitterAgent.sinks.HDFS.hdfs.fileType

= DataStream

16 TwitterAgent.sinks.HDFS.hdfs.writeFormat

= Text

17 TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

18 TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

19 TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

20 TwitterAgent.channels.MemChannel.type

= memory

21 TwitterAgent.channels.MemChannel.capacity

= 10000

22 TwitterAgent.channels.MemChannel.transaction Capacity = **100**

While configuration of this file configure sink as HDFS and Set path for storing tweets in HDFS. Run the configuration file and tweets start downloading in HDFS in specified path. To do this execute flume comments. Start flume using the below command

bin/flume-ng agent --conf ./conf/ -f conf/flume.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent

After a couple of minutes the Tweets should appear in HDFS. If no tweet downloaded in the specified path then refresh. Temporarily data remain in container / Channel and In few seconds tweets start dumping in HDFS. The data downloaded in HDFS is in JSON format. That need to be converted into readable format. Add jsonserde.jar File to convert Json data in readable format.

IV. 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. Apache Hive (HiveQL) with Hadoop Distributed file System is used for Analysis of data. Hive provides a SQL-like interface to process data stored in HDP. Due its SQL-like interface, Hive is increasingly becoming the technology of choice for using Hadoop. To set up HIVE in Hadoop :

Build or Download the JSON SerDe

Before we can query the data, we need to ensure that the Hive table can properly interpret the JSON data. By default, Hive expects that input files use a delimited row format, but our Twitter data is in a JSON format, which will not work with the defaults. And we can use the Hive SerDe interface to specify how to interpret what we've loaded. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process.

To build the hive-serdes JAR, from the root of the git repository:

$ cd hive-serdes

$ mvn package

$ cd ..

This will generate a file called hive-serdes-1.0-SNAPSHOT.jar in the target directory.

After that Create the Hive directory hierarchy using command below:

$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse

$ sudo -u hdfs hadoop fs -chown -R hive:hive /user/hive

$ sudo -u hdfs hadoop fs -chmod 750 /user/hive

$ sudo -u hdfs hadoop fs -chmod 770 /user/hive/warehouse

You'll also want to add whatever user you plan on executing Hive scripts with to the hive Unix group:

$ sudo usermod -a -G hive <username>

After that Configure the Hive metastore. The Hive metastore should be configured to use MySQL. Follow these instructions to configure the metastore. Make sure to install the MySQL JDBC driver in /var/lib/hive/lib.

Now you need to create tweet table. Run hive, and execute the following commands:

ADD JAR <path-to-hive-serdes-jar>;

CREATE EXTERNAL TABLE tweets (

id BIGINT,created_at STRING,

source STRING,

favorited BOOLEAN,

retweeted_status STRUCT<text:STRING,

user:STRUCT<screen_name:STRING,name:STRING>,

retweet_count:INT>,

entities STRUCT<

urls:ARRAY<STRUCT<expanded_url:STRING>>,user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,

hashtags:ARRAY<STRUCT<text:STRING>>>,

text STRING,

user STRUCT<

screen_name:STRING,

name:STRING,

friends_count:INT,

followers_count:INT,

statuses_count:INT,

verified:BOOLEAN,

utc_offset:INT,

time_zone:STRING>,

in_reply_to_screen_name STRING

)

PARTITIONED BY (datehour INT)

ROW FORMAT SERDE

'com.cloudera.hive.serde.JSONSerDe'

LOCATION '/user/flume/tweets';

Now you have you data in relational form which can be easily analyzed. Actually this data looks in relational form bur is not. Data is analyzed using Map-Reduce form. Now Queries can be fired on this data for analysis.

I collected data for Narender Modi , BJP and election using tweets in tweets table. I want to access 12 most common hashtags on the data. For this I fired the query :

SELECT

LOWER(hashtags.text),

COUNT(*) AS total_count

FROM tweets

LATERAL VIEW EXPLODE(entities.hashtags) t1 AS hashtags

GROUP BY LOWER(hashtags.text)

ORDER BY total_count DESC

LIMIT 12;

Results in:

Narender Modi 38890

BJP 22122

Election 21232

Campaigning 18176

Congress 17656

Votes 18111

Centre 15034

Delhi 11390

Sonia Gandhi 9034

Win 8202

Shiv Sena 1090

Result shows the tags in decreasing order of no of occurrences.

V. CONCLUSION

As twitter post are very important source of opinion on different issues and topics. It can give a keen insight about a topic and can be a good source of analysis. Analysis can help in decision making in various areas. Apache Hadoop is one of the best options for twitter post analysis. Once the system is set up using FLUME and HIVE , it helps in analysis of diversity of topics by just changing the keywords in query. Also it do the analysis on real time data, so is more useful. The analysis what I did could be helpful in finding people mood for election voting. And can be helpful in strategy planning. Also opinion mining can also be done on that data for finding polarity(Positive, Negative, Neutral) of tweets collected.

REFERENCES

[1] Sunil B. Mane , Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde , "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, Suseela S , "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 : 2278-1021.

[3] Matthew Koehler, Spencer Greenhalgh, Andrea Zellner, Michigan State University, United States , "Potential Applications of Sentiment Analysis in Educational Research and Practice – Is SITE the Friendliest Conference?", Mar 02, 2015 in Las Vegas, NV, United States ISBN 978-1-939797-13-1 Publisher: Association for the Advancement of Computing in Education (AACE).

[4] 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.

[5] Peiman Barnaghi, Parsa Ghaffari, John G. Breslin , "Text Analysis and Sentiment Polarity

on FIFA World Cup 2014 Tweets" , Conference ACM SIGKDD'15, August 10-13, 2015, Sydney, Australia. Copyright 2015 ACM 1-58113-000-0/08/2015.

[6] "Mining Data from Twitter" from AbhishangaUpadhyay, Luis Mao, Malavika Goda Krishna (PDF)

[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] 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/

[9] "Install Apache Hadoop 2.6.0 in Ubuntu (Multi node/Cluster setup)", [online], available at http://pingax.com/install-apache-hadoop-ubuntu-cluster-setup/

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

[11] https://blog.cloudera.com/blog/ 2012/11/analyzing-twitter-data-with-hadoop-part-3-querying-semi-structured-data-with-hive/