PROPOSAL

USING FLUME TO COLLECT/GATHER DATA FROM TWITTER THEN, USING HIVE TO PROCESS THE DATA. AND FINALLY STORING THE DATA IN HADOOP FILE SYSTEM

23rd January, 2019

By

Kelechi Nzeh, Jermaine Felder, Olson JD Dimanche

Proposed System

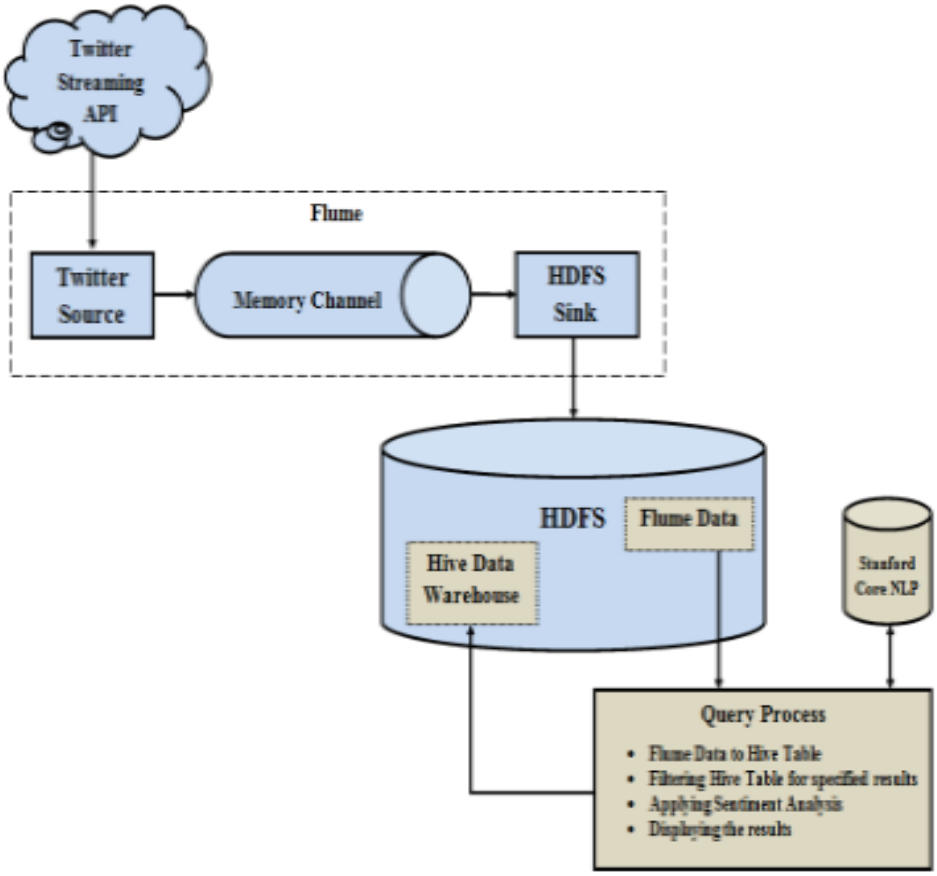
To overcome the drawbacks of existing system we are using Big Data problem statements. By using Hadoop and its Ecosystems, for getting raw data from Twitter by using Hadoop online streaming tool called Apache Flume.

By using this tool we are going to configure everything that we want to gather from Twitter. For this we want to set the configuration and also want to define what information that we want to get from Twitter.

All these will be saved into HDFS (Hadoop Distributed File System) in our prescribed format. From this raw data we are going to filter the information that is needed for us.

And from that we are going to perform the Twitter Analysis by using some UDF's (User Defined Functions) by which we can perform Twitter analysis by taking a data dictionary so that by using that we can decide the frequency of "Trump" and "Toupee" or "Orange" in a negative context.

The subsequent figure shows clearly the architecture view for the proposed system by this we can understand how our project is effective using the Hadoop ecosystems and how the data is going to store from the Flume. How we are going to create tables using Hive, along with how the frequency analysis is going to be performed.



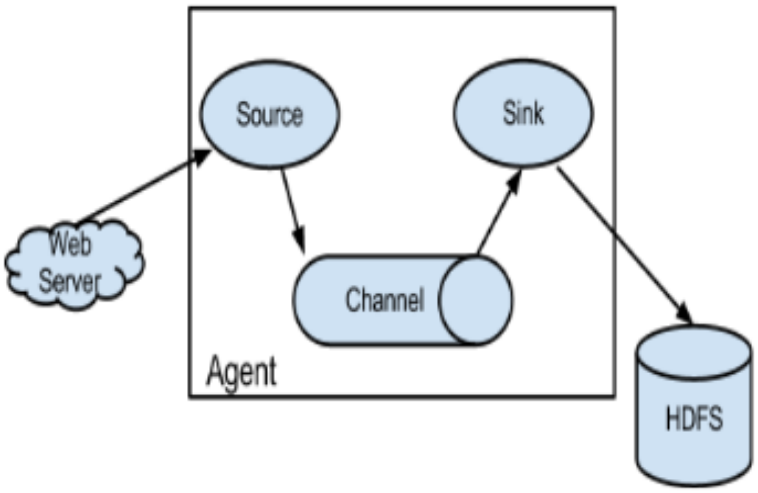
Extracting Twitter Data with Flume

Flume is one of the distribution frameworks which have the capability to use HDFS, HBase, Mongo DB etc., as Sink. As, the Twitter Streaming API gives a constant stream of tweet data coming from the application, it must reside in HDFS securely. The security can be ensured by the generation of keys at the time of creating an application in twitter.

PLAN OF WORK

This project involves the following modules:

- Collecting the required input data [Module I]
- Setting the cluster to store the data [Module II]
- Transfers the data in to hive dataware house [Module III]
-



FLUME

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.

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 kind of interface to process data
stored in HDP. Due its SQL kind of interface, Hive is increasingly becoming the technology of choice for using Hadoop.

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.

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