Live Streaming Data Analysis By Distributed Technology Hive

Thirupathi Gandla
tgandla1@student.gsu.edu

Saivivek Therala
stherala1@student.gsu.edu

Abstract – At present Analytics tools and models that are available in the market are very expensive, unable to handle Big Data and less secure. There is a growing trend of applications that need to handle Big Data as many organizations like Twitter, Facebook collect more and more data from their day to day operations. The traditional analytics systems take a long time to come up with results and the solutions are also not feasible for handling such large volumes of data because of high execution times, so it is not beneficial to use for Real Time Analytics. So, the proposed work resolves all these problems by combining the Apache Open source platform which solves the issues of Real Time Analytics using HADOOP. It also provides scalability and reduced cost over analytics by open source software. Organizations have begun processing Big Data using Map Reduce structure. Hadoop provides an environment for executing Map Reduce over distributed memory clusters thereby supporting the processing of large volumes of data in a distributed environment avoiding single point of failure. Hadoop is flexible and scalable architecture. The proposed work helps the users to analyze the twitter data (android term related) using Hive that supports the queries like SQL type which is called as HiveQL. These queries are compiled as map reduce jobs and are executing using Hadoop.

I. INTRODUCTION

Mobile phones became an essential companion in our day to day lives. They helps us to keep in touch with friends, family, colleagues, access email, browse internet etc. Mobile phones were brought to life with the help of an operating system. In the present world, Android and IOS are having the major mobile operating systems market share in the world. Android holds a market share of 61.9% in the current US market while Apple`s IOS holds only 32.5% of share. Internationally android`s market share is even lot better when compared to Apple`s IOS. So keeping these facts in mind we are inspired to perform big data analytics on tweets related to android operating system.

Big data is the buzzing word in the present software industry. Huge amounts of data is being generated daily from various sources. Companies are trying to perform analytics on big data and get some valuable output which gives an edge over there competitors. In order to achieve this we need to program map reduce jobs in Hadoop ecosystem. It is very difficult to develop the code and reuse it for different business cases. On the other hand, People are very much comfortable to query data using SQL like queries.

A team of developers at Facebook developed a data warehouse tool namely called as HIVE. Hive supports the queries like SQL type which is called as HiveQL. These queries are compiled as map reduce jobs and are executed using Hadoop. Through HiveQL we can plugin custom map reduce scripts into the queries. It is easy to extract, transform and load the data using Hive NoSQL. In Hive, Query execution is done using Map Reduce.

Social Media is an electronic and versatile based web application that will permit the creation, access and exchange of user-generated content that is universally open. Other than person to person communication media like twitter and face-book, the term social media to include posts, blogs, wikis and news, all commonly yielding unstructured content and available through the web. Social media is particularly critical for research into computational sociology that examines questions utilizing quantitative procedures for instance, computational statistics, machine learning and unpredictability, thus called big data for Data mining and simulation modelling. Social media has prompted various information services, tools and analytics platforms. The tools available to researchers are either give superficial access to the raw data or non-superficial access. Analysts require to program analytics in a language such as Java. So the proposed work is much better than the available ones with respect to cost, efficient handling Big Data and scalability. Analysts and organizations feel the need to gain new insights from online networking; they require the analytics tools and ability to change this enormous information data which will have big volume and variety volume and variety into the separate methodologies in order to reach certain

determinations. Social media analytics is useful tool for getting details of customer feelings that are distributed across online sources. The Apache Hadoop software library is a system that takes into account the distributed processing of large information across groups of computers using basic programming models. It give very versatile and adaptable building design for parallel handling. Rather than depending on hardware to deliver high accessibility, the library itself is intended to distinguish and handle failures at the application layer, so delivering a very accessible service on top of a cluster of computers, each of which may be inclined to failures. Social analytics collects and analyzes user opinions and convert them into bits of knowledge and help organizations in identifying areas of user satisfaction or any client grievance for the product.

II. LIMITATIONS OF AVAILABLE SYSTEMS AND TOOLS FOR ANALYTICS

The constraints are as follows:

- The available system like Twitter-Monitor and Real Time twitter trend Mining System require extensive data cleaning, data scraping and integration strategies that will ultimately increase the overhead.
- The available frameworks are wasteful for Real Time Analytics.
- The available strategies and frameworks experience time consuming procedures and the proposed work eliminates all those disadvantages specified previously.

III. TECHNIQUE OF PROPOSED SYSTEM

Social media has procured enormous popularity and interest with advertising groups. Twitter is an effective tool for any organization to get data about how individuals are excited and reacting about its products. Twitters draw in clients and perform correspondence specifically with them and thus, users can give verbal advertising to organizations by talking about the items. With the assistance of limited resources and realizing that one can't target specifically the destination customers, advertising divisions can be more productive in their approach of being so as to promote particular about customers they ought to connect with. In this proposed work, Apache Hive can be utilized to plan direct information pipeline that will empower to break down Twitter data. So as to discover who is prominent in social media one should know the mechanism of twitter which works on tweets and retweets. A retweet is a repost of an update similar to forwarding an email. Querying Twitter data in a traditional RDBMS is inefficient. There are numerous Twitter API which give streaming of twitter data. In the proposed work

Creating Twitter API:

Build up a Twitter API on the Twitter side. The Twitter Programming interface specifically corresponds with the source and sink component by means of network based application. The verification keys and tokens are established that aids in communication over Twitter Server.

Establishing the connection by means of Source and Sink Mechanism:

After making of Twitter API, outline the source and sink mechanism that will help in quick data downloading approach from Twitter Server to HDFS (Hadoop Distributed File System).

The source agent communicates with the Twitter API and Channels the data.

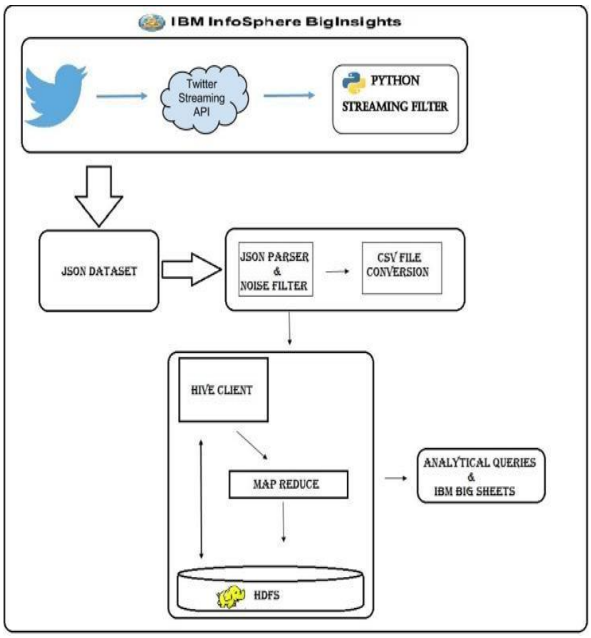
The streaming data is in type of JSON i.e, occasion type of data. The data is lined and directed by means of channel component. At last the information is sink down into HDFS. At that point the tweets are broke down utilizing HIVE.

Analyzing the data on HDFS i.e, Tweets using Hive:

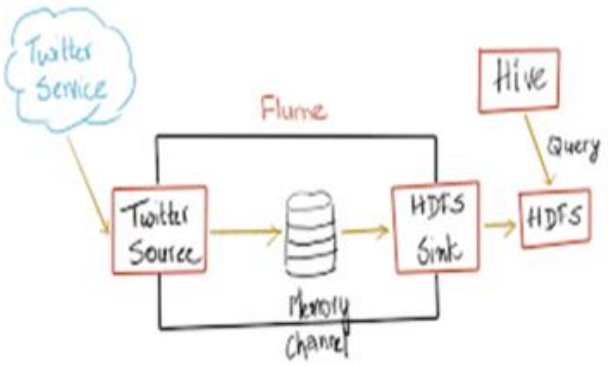
The data now stored on data nodes is analyzed using Hive. Assume we need to perform Twitter Trend Analysis, at that point we need to simply fire the Count query that will tally the particular word count about any keyword.

IV. ARCHITECTURE OF PROPOSED SYSTEM

The proposed system works on the phenomenon of combination of Open Source software along with hardware. The proposed high level architecture is as shown in the figure.



The twitter data downloading rate is expanded with the assistance of system based application and the source and sink mechanism. The reliable connection is set up by making the application on twitter side. The system delays and various latencies are removed for speedy data downloading. The sink, channel and source mechanism is used for the information exchange to HDFS. The Twitter server is present in public domain. In this way we can get to the twitter information by setting up association with the Twitter server. There are many methods available for securing the information from Twitter. One can get to the information by means of Application Programming Interface (API) provided by the twitter. This API sets up the connection between twitter server and the developer's application. One needs to write the program in Python language that will bring the information in the proposed work. Source and sink mechanism is performing the same work effectively. We have implemented the required algorithm for setting up connection with the twitter server and persistently bringing the streamed data. The sink and source mechanism bring the information effectively as well as diminishes the network latency and delay. It source the information from twitter server and channel (line) the information and finally sink the data into Hadoop distributed file system (HDFS). The HDFS is Hadoop Ecosystem daemon. Hadoop is the scalable and flexible architecture that support parallelism. Presently, the data has arrived on the HDFS. The available data is handled with the help of Apache Hive which is a query language designed to handle complex and big data. We can store the data on cloud or in local system. We can also use Apache PIG which is combination of Pig Latin language and compiler that will deliver the map reduce arrangements for parallelism. The Hive and Pig provide simple interface to handle the data. The results are generated with the help of graphs, charts etc. The generated results and reports are analyzed to perform analytics and take decisions in like manner.



V. APPLICATIONS OF PROPOSED WORK

Twitter Trend Analysis:

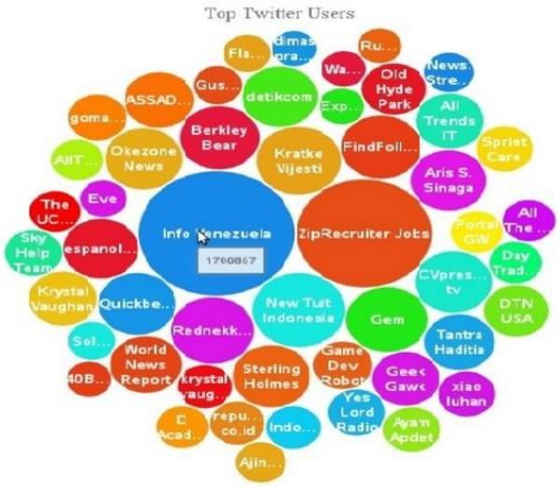
The proposed system compute the pattern on social media that is beneficial for marketing people. The pattern is computed by counting the android keywords. The COUNT algorithm is applied to count all the android keywords which will be viewed as valuable while deciding the pattern. The data visualized will help the marketing people to take certain choices.

VI. EXPERIMENTAL RESULTS

1.

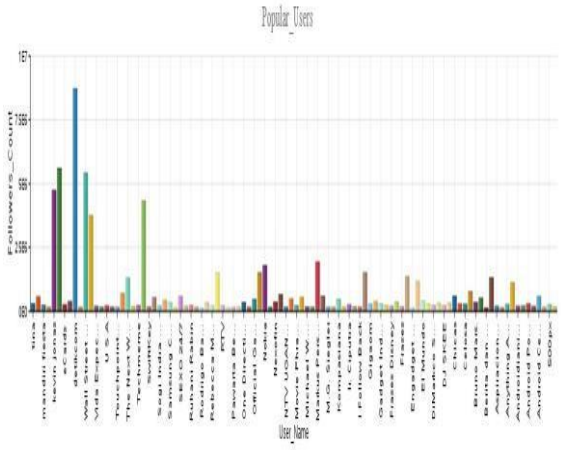
The above figure is the visualization of the result to the query for finding the users contribution towards tweets in twitter dataset. From this data visualization we can observe that certain keywords like 'view', 'android' are highlighted in the final visualization. It implies that the mentioned keywords were more frequently used by the twitter users.

2.



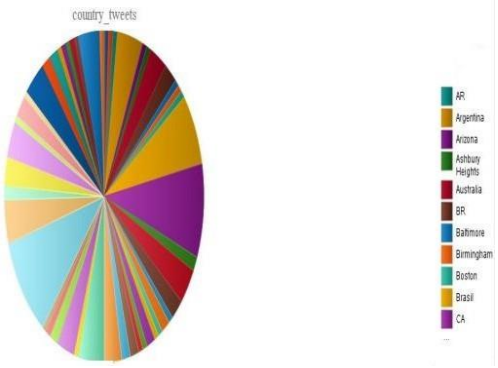
This is the sample visualization of result for the query to identify active users in twitter from the considered data set. From the above visualization we can say that Info Venezuela is the active user in the dataset.

3.



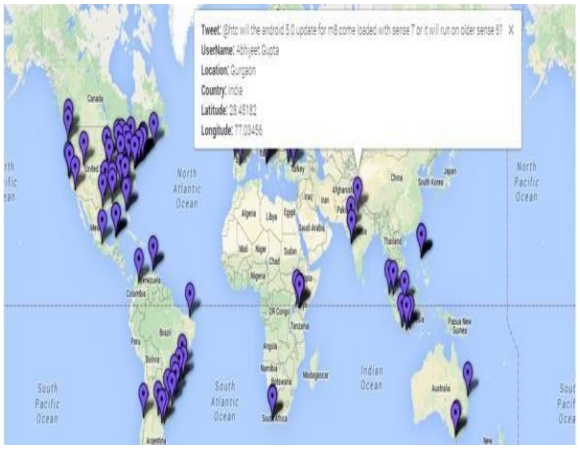
This is the visualization of result obtained for the query to list out the popular users from the considered dataset. From the above visualization we can infer that detikcomis the popular user in the considered dataset.

4.



This is the visualization of result for the query to find the tweet contribution from various countries. Here each sector represents each country tweet contribution. We can say from the above figure that Boston has highest tweet contribution.

5.



This is the visualization of result for the query to plot the geographical locations of tweets from the considered dataset.

VII. CONCLUSION AND FUTURE WORK

The proposed work gives data information about issues with the available tools and systems in the market. There are various systems to get analytics available in the market but are very costly, less efficient and less secure. So the proposed framework utilizes a productive Apache Open Source product which exhibits the model that can have Twitter Pattern Analysis utilizing HADOOP where no additional

work like scraping, cleansing and data security required. It also gives the fast data downloading methodology for effective Twitter Trend Analysis. The proposed work concludes with the phenomenon of Open Source Software along with Commodity Hardware that will expand IT Industry Profit.

Social Media provides valuable datasets, but the challenge is in collecting and analyzing the data quickly. In this project we have analyzed and visualized twitter data on a particular keyword. In future work, we would like to perform domain specific analysis and try to capture valuable insights from data**.**

VIII. REFERENCES

[1] http://infolab.stanford.edu/~ragho/hive icde2010.pdf

[2] https://hive.apache.org/

[3] Claudio Cioffi-Revilla "Computational social science", WILEY Interdisciplinary Reviews: Computational Statistics, Vol. 2, no. 3, May/June 2010:pp.259–271

[4] Gaurav D Rajurkar, Rajeshwari M Goudar "A speedy data downloading approach for Twitter Trend and Sentiment analysis using HADOOP", 2015 International Conference on Computing Communication Control and Automation.

[5] Andreas M. Kaplan , Michael Haenlein "Users of the world, unite! The challenges and opportunities of Social Media", Business Horizons (2010) 53, 59—68 ELSEVIER

[6] Michael Mathioudakis, Nick Koudas, "TwitterMonitor:Trend Detection over the Twitter Stream", SIGMOD'10, June6–11, 2010, Indianapolis, Indiana, USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06

[7] Min Song, Meen Chul Kim, "RT2M : Real-time Twitter Trend Mining System", 978-0-7695-4998-9/13 © 2013 IEEE International Conference on Social Intelligence and Technology

[8] Saeideh Shahheidari, Hai Dong, Md Nor Ridzuan Bin Daud "Twitter sentiment mining: A multi domain analysis" 978-0-7695-4992-7/13 © 2013 IEEE Seventh International Conference on Complex, Intelligent, and Software Intensive Systems