International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Designing Framework for Real Time Twitter Data Analytics using Apache Flume and Pig

Ashlesha S. Nagdive, Rajkishor Tugnayat

Abstract: ***In the world of technology, people prefer social media to express themselves. Record says Twitter has more than 321 million active users with 100 million users posting approximately 340 million tweets a day. Twitter is the largest source of breaking news on social issues specially election-related where people can express their views also suggest their opinion. Twitter is generating unlimited unstructured text data. Hadoop is one of the finest tools accessible for analyzing twitter data because it supports processing of distributed big data, streaming data, time stamped data, text data etc. Whereas Apache Flume is used to extract real time twitter data into HDFS. This study attempts to establish an analytical framework to derive and interpret structured as well as unstructured Twitter data. The proposed framework comprises of real time twitter data insertion, its processing, and data visualization utilizing Apache Flume and pig. In this project we fetch positive and negative tweets on election data from twitter and analyzing the party status and the probability to win the election.***

Keywords : unstructured twitter data, HDFS, Apache flume, Pig, Textblob, Dash.

I. INTRODUCTION

In the modern world, information is readily available through internet and social media has become an indispensable part of people's life. It isn't only interactive platform for creating, distributing and sharing wide range of information. It is effective platform for marketing by various organizations to reach their target audience. With the evolution of big data, social media marketing business has scaled new heights. It is estimated that by 2020 the volume of data will exceed 40 trillion gigabytes. With access to such humongous amounts of data, marketers are able to employ it to get actionable insights for designing efficient marketing strategies. All the updates, photos and videos posted by users provide information about their demographics, likes, dislikes, comments etc. Businesses are, managing and analyzing this information to get a competitive edge. Real-time data analysis requires data ingestion and processing the stream of data prior to Storage of data. Certain applications of the real-time data analytics includes web services, weather forecasting, medical health care, banking sector, retail industry, multimedia, cyber security, and social media. This paper represents designing of framework for analyzing twitter data for prediction of election results based on tweets of people.

II. PROCEDURE RELATED WORK

This Design framework is gathering of data, filter data, and analyzes streaming data which throws light on the trends based on time and condition. Framework comprises in three steps; unstructured data ingestion or insertion, data streaming process, and data visualization for further analysis and prediction. Ingestion of data is achieved by Kafka, a popular and powerful message broker system designed to import tweets, distribute it based on Topics and to make it available over consumers nodes for transformation by analytical tools[3]. Apache Spark provides a direct contact to the users and analyzes data through Spark Streaming.

Sentiment analysis of twitter data from the citizens of the country can provide valuable insight during election campaigns[1]. Such campaign through social media, even makes the party aware of the next step to be done in elections and can focus on necessary action taken for betterment of society.

Social media data that accumulates a huge volume of data every second require a proper framework that processes data as and when it arrives [3]. Identifying and Processing posts on social sites like twitter may prove quite useful for drawing inferences and predicting specific activities that are about to happen in the world in near future [4].

By employing real-time data analytics significant events including emergencies, can be detected. Architecture was developed for analyzing social media text by considering specific predefined keywords and other important related aspects of huge dataset from tweets. These keywords are predefined as positive, negative as well as neutral text words related to particular politician. People generally tweets their sentiment based on current scenario in political issues and problems faced by people which may be positive, negative or neutral. These keywords then helps in generating and prediction the result before elections.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Ashlesha S. Nagdive*, Assistant Professor Information Technology, G. H. Raisoni College of Engineering, Nagpur, India. Email: ashlesha.nagdive@gmail.com
Dr. Rajkishor Tugnayat, Principal, Shri Shankarprasad Agnihotri College of Engineering, Wardha, India. Email: tugnayatrm@rediffmail.com

III. METHODOLOGY

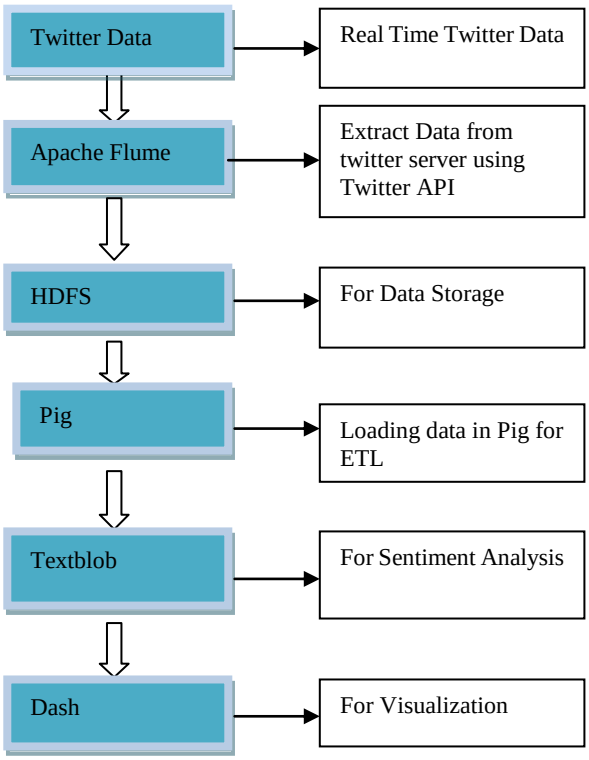

Fig.1 Block Diagram of Design Framework

Methodology for design framework to identify behavioral patterns through the concept of sentiment analysis of text data of twitter is depicted in fig.1.Following are various phases and tools through which data has to get processed for analytics.

A. Apache Flume

Apache Fume is a data ingestion tool that allows collection or gathering data, aggregation and transportation of extensive text data or logs from diversified sources to a centralized data store i.e HDFS. Apache Flume has proved to be an extremely reliable, distributed, and configurable tool. It is particularly created to transcribe streaming data from various web servers to HDFS.

B. Hadoop Distributed File System (HDFS)

Apache Flume gathers and stores the data in one of the two centralized stores HBase or HDFS. The rate of incoming data exceeds, the rate at which data can be written to the destination, Flume serves as a regulator between data producers and the centralized stores maintaining a constant flow of data between them[5]. Flume provides the feature of circumstantial routing. Flume consists of channel-based transactions. It ensures reliable message delivery.

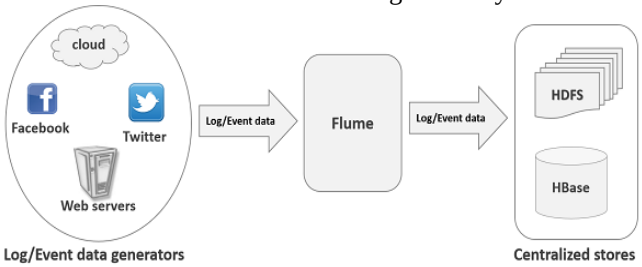

Fig 2. Architecture of Apache Flume

C. Pig

Pig is used as an ETL tool for Hadoop. It makes Hadoop more approachable and usable. It opens an interactive and script based execution environment called Pig Latin, which loads and processes input data using a series of operations and transforms to produce the desired output. MapReduce model is not always convenient and efficient When there is large amount of data which need to processed using Hadoop, it involves more overhead and complex. A solution to the problem is Pig, extension of Hadoop.

D. Textblob

TextBlob is a library in python providing the functionality to preprocess text data. Pattern analyzer and Naïve Bayes analyzer are two sentiment implementation analysis. It return the result as a named tuple form as Sentiment(classification, p_pos, p_neg).When data is fetched and gathered from twitter, the sentiment property returns a named tuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] with 0.0 being interpreted as very objective and 1.0 being specific i.e subjective[6].

E. Visualization Through Dash

Data presented in the form of graphics can be analyzed better than data presented in words. Dash is a Python Framework designed for visualization. Data visualizations convert large and complex data into graphical format so that patterns, trends and correlations can be visualized. Exploratory Data Analysis or EDA is a major part of data visualization.

IV. IMPLEMENTATION

A. Twitter Data Set

Sentiment analysis is defined as the process of mining various data sources for opinions or views using text analysis. Politicians use sentiment analysis to gain an understanding of peoples outlook towards them and their policies. Twitter, a popular social media websites, provides a set of APIs that allows us to fetch and manipulate tweets. In this process, we are given a key and a secret token that the application uses for authentication. Once the application is authenticated, Twitter APIs is been used to fetch tweets. One can fetch data from a Twitter feeds either by use of R language or by using Jaql, as Jaql is designed to handle JSON data and the default data format for tweets. It is flexible and easy to use Jaql.

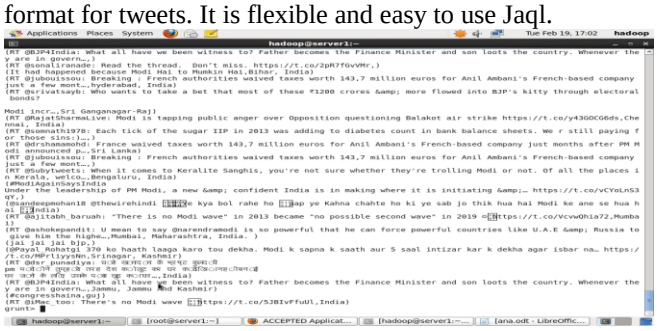

Fig 3. Tweets on Modi

Retrieval Number: F7726038620/2020©BEIESP
DOI: 10.35940/ijrte.F7726.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

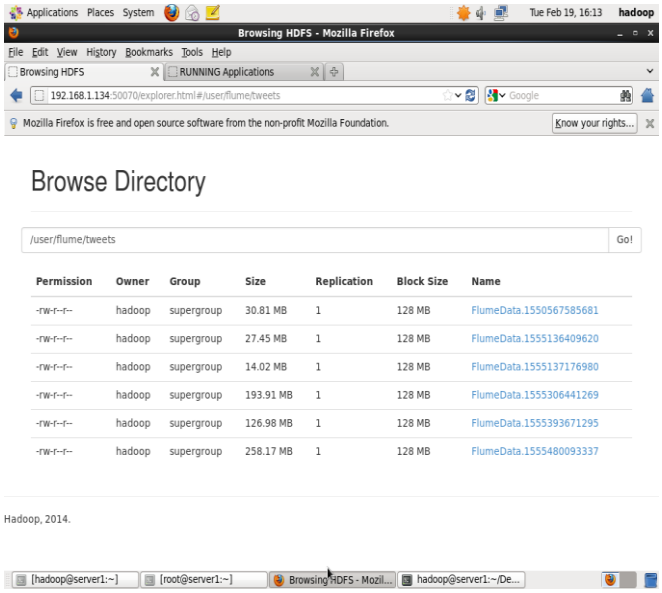

Fig.4. Real Time Twitter Dataset

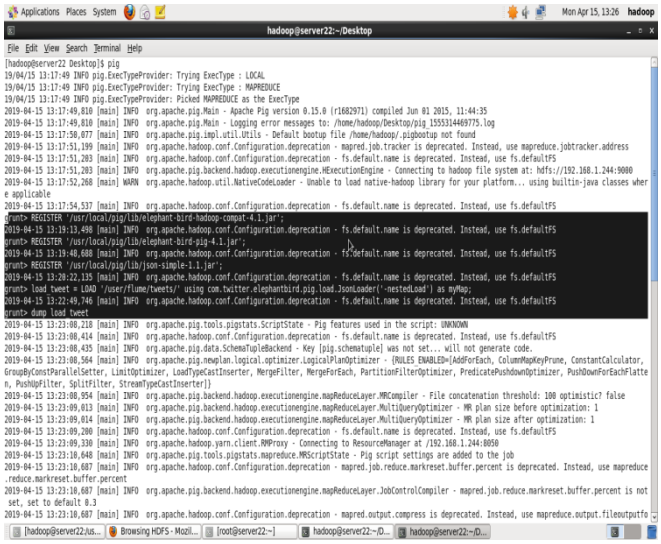

Fig.5. Processing Unstructured Twitter Data

B. Process of ETL

ETL implifies **E**xtract, **T**ransform, **L**oad, three database functions, combined into one tool to move the data from one database to the other.

Extract process reads the data from the dataset. In this stage, the data is gathered, often from various types of data sources.

Transform it is the process of converting the extracted data into a format compatible with another database. Transformation occurs by combining the existing data with other data, following rules or lookup tables.

Load is defined as the process of writing the data into the target database.

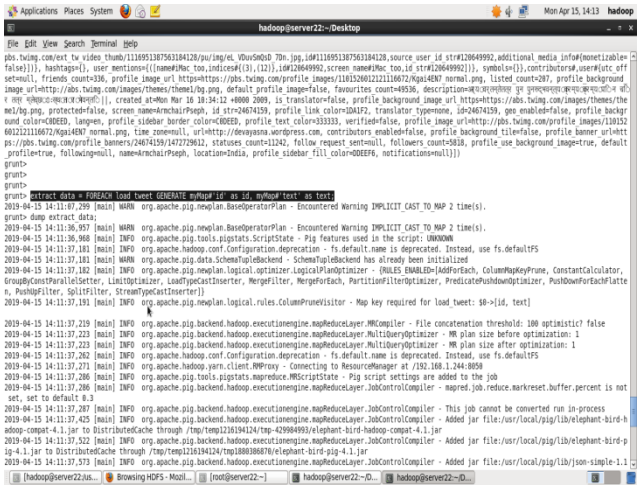

Fig 6. Extracting Twitter Data

The above code will give specific Id and text data of user which makes it structured and also make it easy for analysis.

V. RESULT

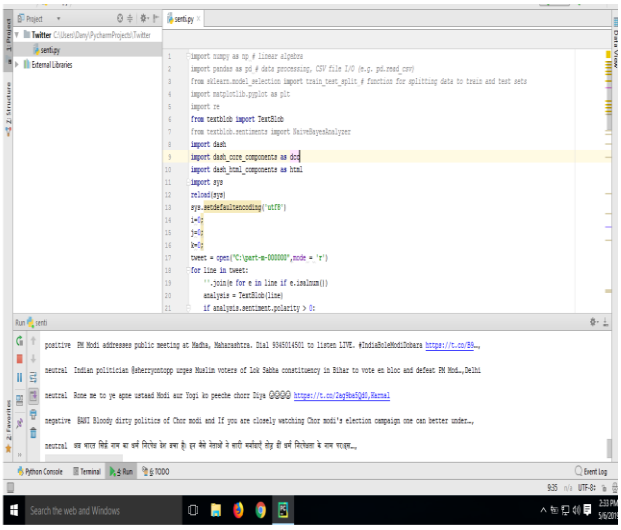

Fig.7.Program in Python

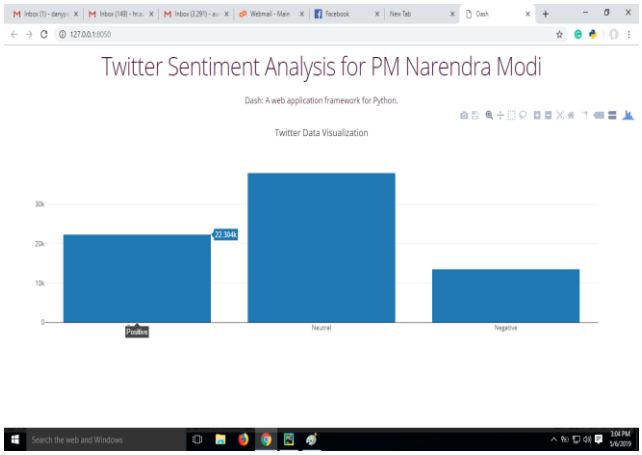

Fig.8. Predictive Analysis of Twitter Data

Predictive analytics is not just about using technological advances to win the electoral battles. But, about focusing political efforts to plan and build their strategies based on real public sentiments. Politicians can now really be part of people's everyday lives. Fig8. Represents prediction of twitter data before elections 2019, analyzing maximum neutral tweets about respected Prime Minister Mr. Narendra Modi.

VI. CONCLUSION

This research case study focuses on the significance of design framework for real-time data analytics using social media data.Sentiment analysis is helpful as it provides access to the wider public opinion on a particular topic/situation. In this paper we have analyze public opinion about elections 2019 and analyze opinion about "Modi" through twitter data. People all over world have expressed their views on election 2019 especially about political leaders and their work towards society. Thus prediction can be analyzed through positive , negative and neutral tweets.With Predictive Analytics, even small campaigns are now able to target the voters they need, talk about the issues voters care about, through their views on social media like twitter.

REFERENCES

1. Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabriz, "Developing a Real-timeData Analytics Framework For Twitter Streaming Data," 2017 IEEE 6th International Congress on Big Data, 978-1-5386-1996-4/17
2. N. Mohamed, J. Al-jaroodi, Real-Time Big Data Analytics: Applications and Challenges. International Conference on High Performance Computing & Simulation (HPCS), 2014
3. S. Cha and M. Wachowicz. Developing a real-time data analytics framework using Hadoop. 2015 IEEE International Congress on Big Data, pages 657–660, June 2015
4. B. Yadranjiaghdam, N. Pool, N. Tabrizi, "A Survey on Real-time Big Data Analytics: Applications and Tools," in progress of International Conference on Computational Science and Computational Intelligence, 2016.
5. A. Bifet, "Mining Big Data in real time," Informatica, 37(1), 2013, Pages 15 -20.
6. D. T. Nguyen and J. E. Jung. Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems, 2016.
7. J. Zaldumbide, R. O. Sinnott, "Identification and Validation of RealTime Health Events through Social Media," 2015 IEEE International Conference on Data Science and Data Intensive Systems, Pages 9 – 16, doi 10.1109/DSDIS.2015.27
8. V. Ta, C. Liu, G.W. Nkabinde, "Big Data Stream Computing in Healthcare Real-Time Analytics", 2016, IEEE International Conference on Cloud Computing and Big Data Analysis, Pages: 37 42, doi: 10.1109/ICCCBDA.2016.7529531
9. M. Wachowicz, M.D. Artega, S. Cha, and Y. Bourgeois, "Developing a streaming data processing workflow for querying space–time activities from geotagged tweets" Computers, Environment and Urban Systems Journal. 2015.
10. M. Wachowicz, M.D. Artega, S. Cha, and Y. Bourgeois, "Developing a streaming data processing workflow for querying space–time activities from geotagged tweets" Computers, Environment and Urban Systems Journal. 2015

AUTHORS PROFILE

Ashlesha S. Nagdive, PhD research scholar has completed Bachelors of Engineering in Information Technology in 2008 from Amravati university and Masters of Engineering in Embedded Systems & Computing from G.H. Raisoni College of Engineering, Nagpur, in 2011.Currently Pursuing PhD from Amravati university. Also working as Assistant Professor in Information Technology department at G.H Raisoni College of Engineering, Nagpur since 2010. Member of IEEE and published various papers in International Journal and conferences. Area of interest is Big Data & Hadoop, Data Analytics, data visualization.

Dr. Rajkishore M. Tugnayat, Principal of Shri Shankarprasad Agnihotri College of Engineering Wardha. He has completed his PhD from Nagpur university. He has more than 20 years of teaching experience and Research Experience. He is a member of IEEE .He has publications in various International Conferences and International Journals. Subject of Expertise is Software Engineering, Big Data, Computer Networks and Image Processing.