ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Social Media Analytics Using Big Data

Ankitha B S [1], K Varsha [2], Sathya Narayana N [3]

U.G. Student, Department of Information Science and Engineering, Vemana Institute of Technology, Karnataka, India[1]

U.G. Student, Department of Information Science and Engineering, Vemana Institute of Technology, Karnataka, India[2]

Assistant Professor, Department of Information Science and Engineering, Vemana Institute of Technology, India[3]

ABSTRACT: “Big Data” is data whose scale, diversity and complexity require new architecture, techniques, algorithms and analytics to manage it and extract value and knowledge from it. Social Media Analytics as a part of social analytics is the process of gathering data from stakeholder conversations on digital media and processing into structured insights leading to more information-driven business decisions and increased customer centrality for brands and businesses. "Social Media Analytics is the art and science of extracting insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making. It is a science, as it involves systematically identifying, extracting, and analyzing social media data (such as tweets, shares, likes, and hyperlinks) using sophisticated tools and techniques". Our proposed data processing system uses Apache Kafka for stream processing, Apache Spark for in-memory data processing, Apache Kudu for real-time storage, Apache Superset for retrieving, analyzing and reporting data for business intelligence.

KEYWORDS: Big Data, Social media, Cartography, Spatiotemporal.

I. INTRODUCTION

Every second of every day, Big data gets bigger. Social media alone generates endless streams of data, flowing in from Facebook, Twitter and other social sites. Big Data is changing social media marketing in some pretty big ways.

A. Disadvantage of existing system

1) As social media is growing bigger day by day, data is increasing rapidly, so processing becomes bottleneck and data customization is very difficult.

2) Processing larger social media data becomes harder with current legacy hardware and software system which makes the process slow.

3) Dockers are not used. Auto-scaling of appropriate component based on the workload is not enabled and also existing systems have storage and display analysis issues.

Information increases rapidly at a rate of 10x every five years. During each stage of the Data lifecycle, the management of Big Data is the most demanding issue. In our proposed system the following features are adopted

B. Proposed System

1) We are using Apache Kafka [1] and Apache Spark [2][3] Technologies for streaming and real-time data processing which makes data processing faster. We also propose Sentiment analysis and Geo coding.

2) In our proposed system, the usage of Dockers helps in deploying and automation of different technologies such as Apache Kafka, Apache Spark and Apache Kudu along with enabling of Auto scaling without any computational changes. As we are using all apache open source technologies we have zero costing other than hardware/processing costing.

II. RELATED WORK

There are already some work to leverage big data technologies on Analytics. Some of the examples are as given:

Real Time Text Analytics [4] evaluates the proposed real-time text processing pipeline using open-source big data tools which minimize the latency to process data streams. Sentiment analysis is one of the most interesting techniques to find out the users opinion against a particular discussion, debate, or product. Sentiment analysis requires large-scale processing of data on multiple machines using big data tools. This system doesn't improve the performance of the system through enabling auto-scaling of appropriate component based on the workload.

Big Data Survey technologies [5] the fundamental concepts of Big Data, which include the increase in data, the progressive demand for HDDs, and the role of Big Data in the current environment of enterprise and technology has been detailed. Big Data is promising for business application and is rapidly increasing as a segment of the IT industry. It has significant interest in various fields, including the manufacture of healthcare machines, banking transactions, and social media. This system doesn't provide the efficiency of integrity online, as well as the display, analysis, and storage of Big Data.

Developing Real time data analytics framework [6] a framework for real-time analysis of Twitter data. This framework is designed to collect, filter, and analyse streams of data and gives a chance to sense what is popular during a specific time and condition. The framework consists of three main steps: data ingestion, stream processing, and data visualization. Data ingestion is done by Kafka, a powerful message brokering system to import tweets, and to distribute it based on Topics that it defines, and to make it available over consumers' nodes to be used by analytical tools. Apache Spark is used to access these consumers directly and analyse data by Spark Streaming. This only processes the data that is popular at that time and sentiment analysis is not implemented.

III. PROPOSED SYSTEM

Fig. 1 shows the architecture of our approach. It contains three main components (1) Data sources which are the social media sites. (2) Data streaming and pre-processing which contains pre-processing engine Apache Spark, Real time data pipeline Apache Kafka. (3) Display which displays the dashboard that contains the analytics in it.

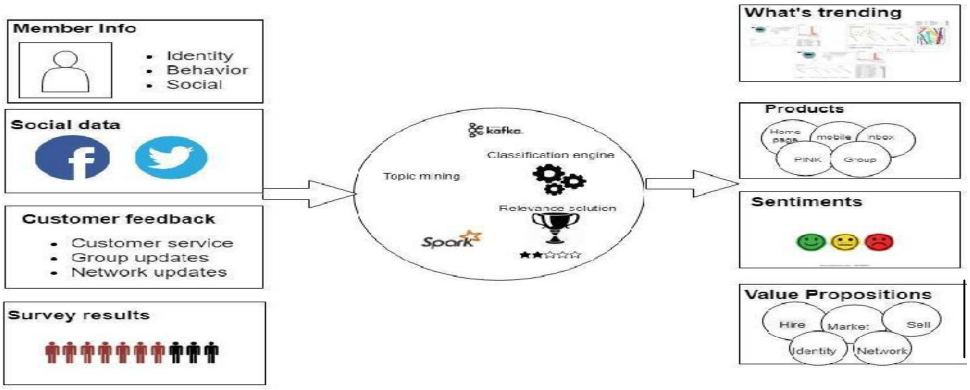


Fig. 1 Architecture Diagram

Description of the Proposed Algorithm:

Aim of the proposed algorithm is to develop a dashboard to show analytics in different ways for the data collected from twitter. The proposed algorithm consists of four main steps.

Step 1: Data Collection and Filtration:

Data is collected from Twitter API using python by connecting to twitter using secret key and access key. Data filtration is done using hashtags based on keywords. Data obtained after filtration is consumed by Big Data messaging queue system, Kafka and will be sent to Big Data pre-processing engine Apache Spark for further processing.

Fig. 2 shows the Flow diagram of our approach to process Big Data using Apache Spark.

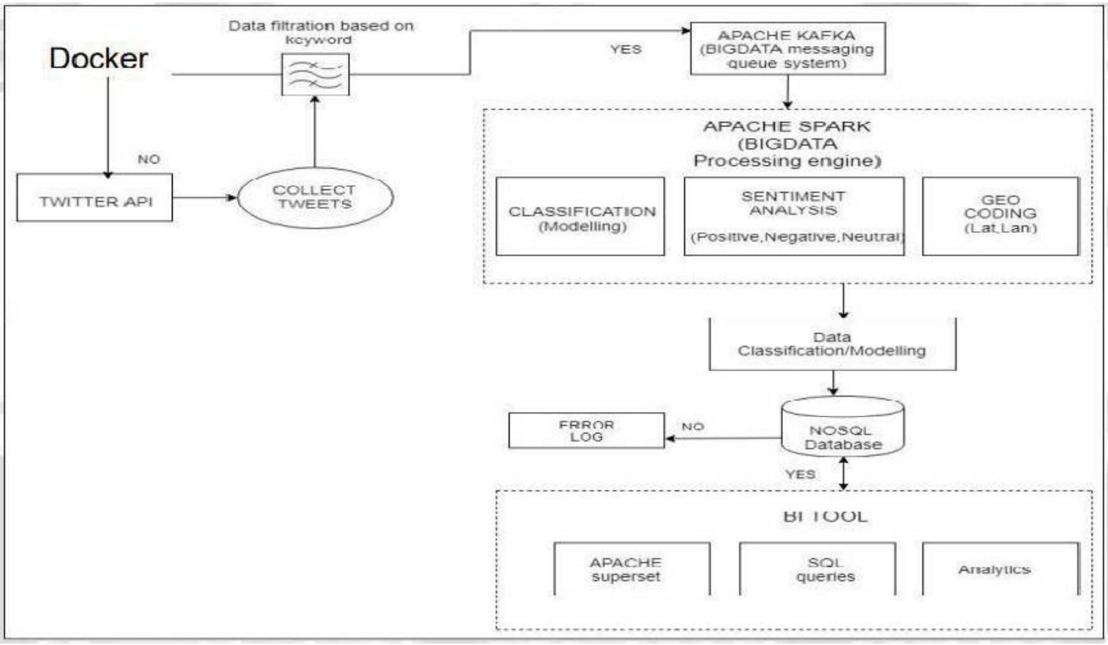

Docker
Data filtration based on keyword
YES
APACHE KAFKA (BIGDATA messaging queue system)
NO
TWITTER API
COLLECT TWEETS
APACHE SPARK (BIGDATA Processing engine)
CLASSIFICATION (Modelling)
SENTIMENT ANALYSIS
(Positive,Negative,Neutral)
GEO CODING (Lat,Lan)
Data Classification/Modelling
ERROR LOG
NO
NOSQL Database
YES
BI TOOL
APACHE superset
SQL queries
Analytics

Fig. 2 Flow Diagram

Step 2: Pre-processing:

Big data pre-processing engine used in this approach is Apache spark which performs classification, sentiment Analysis and geo-coding. Classification is done based on different fields available and user requirements. Sentiment analysis is the process of showing human opinions (positive, negative and neutral) to a particular situation or product. Geo coding is the process of showing analytics based on locations using latitudes and longitudes.

Step 3: Data classification:

This classification is done based on user requirements. If user requires data related to sentiment analysis then only that data will be fetched. The fetched data will be stored in NoSQL database and will be used further to show analytics. NoSQL database is relational database in which data is placed in tables and data schema is carefully designed before the database is built.

Step 4: Analytics:

Analytics in this project are shown using BI tool (Apache superset). SQL queries are used to fetch data from NoSQL database. Final output will be shown in a dashboard which is analytics.

IV. SIMULATION RESULTS

The method introduced above in the paper is demonstrated on a big dataset. And the output analytics is as shown in the below dashboards.

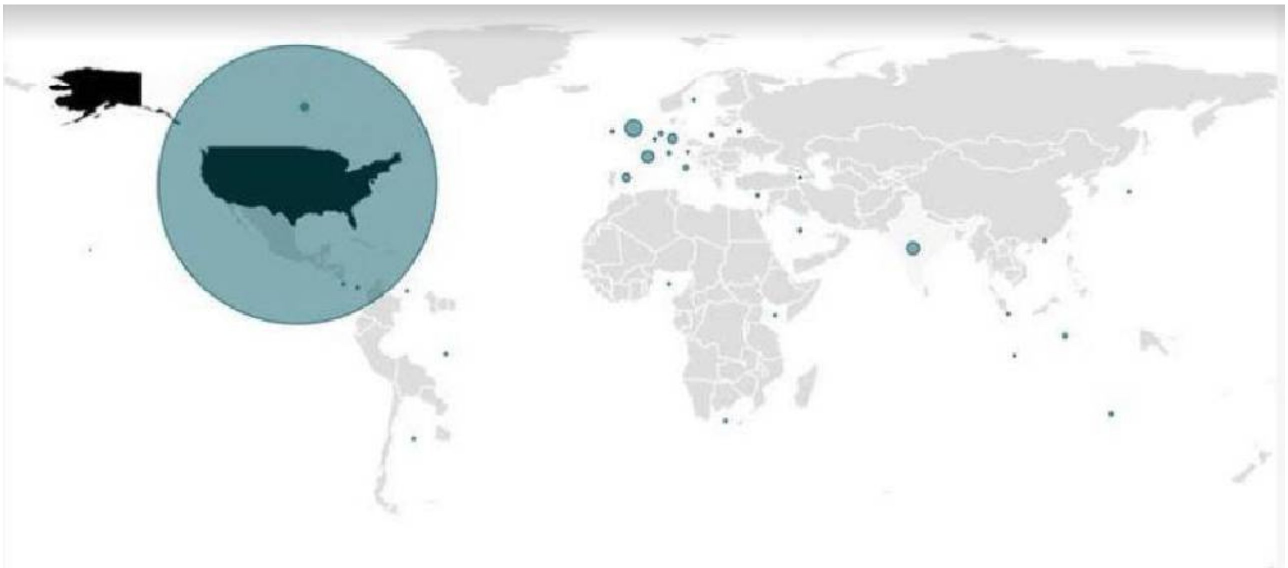
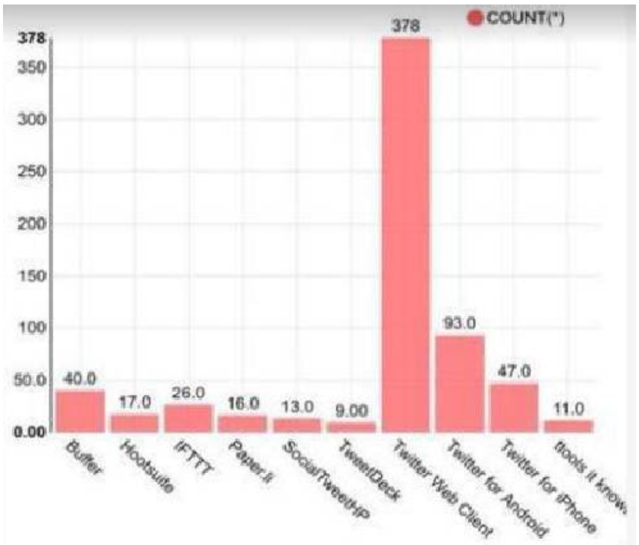


Fig. 3 Analytics based on Geo Coding



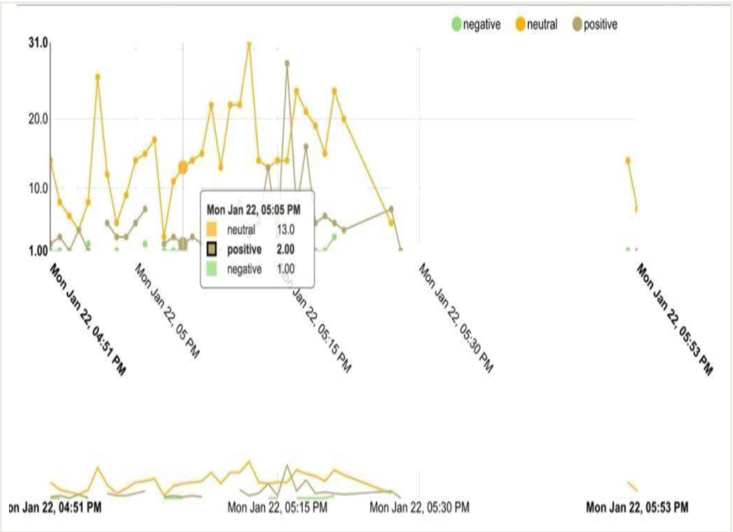


Fig. 4 Analytics based on Devices

Fig. 5 Analytics based on Sentiment Analysis

V. CONCLUSION AND FUTURE WORK

This paper describe a novel approach for loading, managing, computing, and querying big data in the distributed computer cluster system based on Apache Spark. With the large number and rapid growth of social media systems and applications, social big data has become an important topic in a broad array of research areas. The aim of this study has been to provide a holistic view and insights for potentially helping to find the most relevant solutions that are currently available for managing knowledge in social media. One of the current main challenges in data mining related to big data problems is to find adequate approaches to analysing massive amounts of online data (or data streams). Because classification methods require previous labelling, these methods also require great effort for real-time analysis.

REFERENCES

[1]. Kafka, "A high-throughput, distributed messaging system," URL: kafka. apache. org as of, vol. 5, no. 1, 2014.

[2]. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster Computing with Working Sets," Hot Cloud, vol. 10, p. 95, 2010.

[3]. Apache Spark. Available: http://spark.apache.org/

[4]. Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari, and Faisal Bukhari, "Real-time Text Analytics Pipeline Using Open-source Big Data Tools".

[5]. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali,1 Muhammad Alam, Muhammad Shiraz, and Abdullah Gani "Big Data: Survey, Technologies, Opportunities, and Challenges".

[6]. Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabrizi, "Developing a Real-time Data Analytics Framework for Twitter Streamsing Data".