International Journal for Modern Trends in Science and Technology, 7(05): 44-48, 2021

ISSN: 2455-3778 online
DOI:https://doi.org/10.46501/IJMTST0705006
Available online at:http://www.ijmtst.com/vol7issue05.html

Real Time Data Pipeline for Twitter Trends Analysis

Mukesh Kumar Sah[1] | Rishabh Sharma[2] | Amritpal Singh[3]

[1]Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab.

To Cite this Article
Mukesh Kumar Sah., Rishabh Sharma & Amritpal Singh. Real Time Data Pipeline for Twitter Trends Analysis. *International Journal for Modern Trends in Science and Technology* 7, 44-48 (2021).

Article Info
Received on 15-April-2021, Revised on 27-April-2021, Accepted on 01-May-2021, Published on 05-May-2021.

ABSTRACT

In social media, Information is present in enormous amount. Extracting data from processed information from social media gives us diverse usages in various fields. In the field of Business Analytics, HealthCare, Technologies and Trending Topics in Social Media posted by the user. Extracting information from social media is providing number of benefits such as knowledge about the latest Technology, Medical field, Business Decisions, etc. Twitter is solitary of the social media which allows the user post tweets of limited number of characters and share the tweet to their followers. Twitter allows application developer to access the tweets for their motive. In the implemented methodology, Tweets are collected, and sentiment analysis is performed on them. Based on the results of sentimental analysis of Trending Topics in Twitter, suggestions can be provided to the user. In this way, the implemented system can help in improving the growth of business, healthcare, technologies and alsoNegative or Positive mentions of a product or service can be determined.

KEYWORDS:*Twitter App API,Apache Flume, Apache Kafka, Apache Spark, Spark Streaming, DStreams, MySQL, Tableau Data Visualization.*

I. INTRODUCTION

In today's generation, the analysis of real-time data is setting off critical for SMEs & Large Corporations alike. Industries like Financial resource, Legal services, IT operation management resource, Marketing and Advertising all require the analysis of massive amounts of real-time data as well as historical data in order to make business decisions.

Big data is determined by velocity, volume, and variety of the data. These characteristics makes Big data unique from regular traditional data.Contrasting regular big data applications, real-time data processing applications is essential for building a distributed data pipeline for capturing, processing, storing, and analyzing the data efficiently.

This project is a means for us to apply the theory of big data distributed data processing in parallel, to build a real-time data processing pipeline using big data frameworks, big data programming tools, open-source tools that can capture large amounts of data from numerous data sources, and process, then store, and then analyze the large-scale data in an efficient manner.

Nowadays, Data is generating in enormous amount. Most of world population are eager to use new technologies and it is increasing day to day, we

can have examples like Social Media i.e., Facebook, Instagram, Twitter, etc., eCommerce websites, OTT platforms like Netflix, Amazon Prime, we can have numerous examples, all these platforms are attracting people in very huge amount. Most of the data generated by these platforms are Semi-Structured and Unstructured data. So, to analyze these types of data, there must be some data processing and analytics tool, data handling tools or visualization tools so that, data can be analyzed and visualize in proper manner which can help many business and organizations today. Here, in our project, we have fetched the data from Twitter Social Media, basically the tweets by the people. From those tweets, we can analyze the behavior of the people, their likes, dislikes, what they are taking about, what products are attracted by them, where they are visiting, and much more. Overall trending topics can be aggregated and the same we have made in our project.1st we fetch data from Twitter App API then, we will send that data to apache flume and again we send it to Apache Kafka, or we can also store it in local file system. After we get the data, we apply our sentiment analysis on those tweets using Apache spark RDD and DataFrames and the results are stored in MySQL Database. MySQL is connected to Tableau and Finally visualization is done through Tableau Data Visualization and the results are shown in Tableau Dashboard.

II. PROJECT DESCRIPTION

Twitter streaming trends is too popular these days and sentiment analysis is a superb choice for building a distributed data pipeline. Every day around 500 million tweets (as of Jan 2, 2021) are produced from all over the world, and around 1% of them are publicly available, that is 5 million tweets. The data pipeline uses Apache Flume and Kafka as a data ingestion system, Apache Spark as a real-time data processing engine, MySQL Database for storing enormous data and retrival, and MySQL with Tableau for real-time analytics.

The Twitter data is obtained by using Twitter App API and is streamed to Apache Flume, furtherApache Kafka which makes it available for Apache Spark that performs sentiment and data processing and and stores the results into MySQLDatabase. The popular sentiments of the trending topics from twitter are explored through a Tableau dashboard.

Tools and IDEs Used

- Twitter App API
- Apache Flume
- Apache Kafka
- Apache Spark
- Spark Streaming, DStreams
- Java 8
- Scala Programming Language
- Scala SBT
- MySQL Database
- Tableau Data Visualization

III. METHODOLOGY

Proposed methodology is implemented creating a Data Pipeline which uses the following architecture:

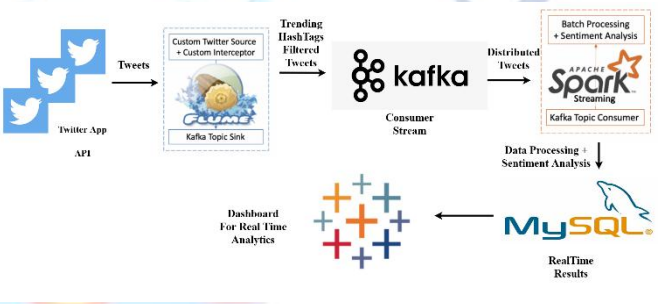

Fig 3.1: Data Architecture of the Project

- Apache Flume and Twitter Streaming API publishes streaming tweets to the 'twitterdata1' topic in an Apache Kafka broker.
- The Apache Spark Streaming Context is subscribed to read the tweets from the 'twitterdata1' topic.
- The Spark processing engine leverages Spark Streaming to perform batch processing on incoming tweets and performs sentiment classification before storing the processed results in the MySQL.
- Tableau connects MySQL and the real-time data is used to create a live dashboard to analyze popularity and sentiment of trending topics on Twitter.

IV. SYSTEM DESIGN

The different components of the data pipeline, Apache Flume, Twitter App API Streaming, Apache Kafka, Apache Spark Streaming, MySQL and Tableau are all run locally for development.

a. Twitter App API Streaming with Apache Flume

Firstly, we will be creating Twitter App API with our Twitter Developer Account. After configuring Twitter App, we will get API keys, and those keys will be configured in Apache Flume conf file.

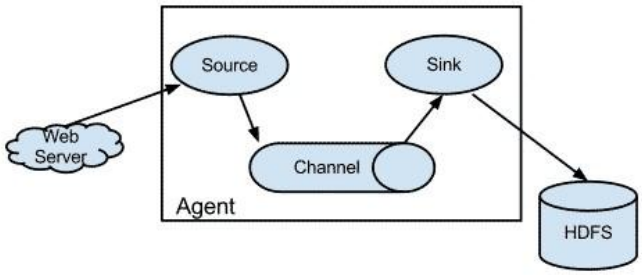

Fig 4.1: Data Flow Model in Apache Flume

In conf file of apache flume, we will be configuring as, twitter will be our source, channel will be memory channel, and apache kafka will be our sink or we can take local file system i.e., HDFS as sink.

b. Apache Kafka

We know Apache Kafka is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables us to pass messages from one end point to another.

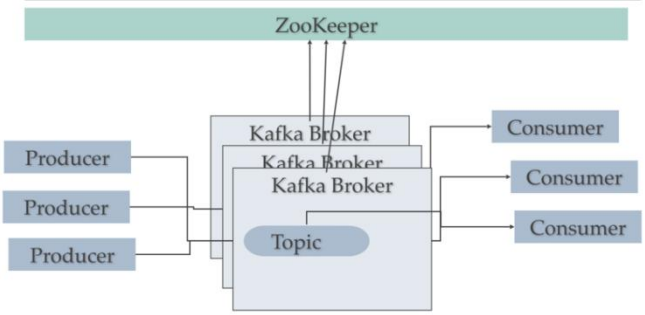

Fig 4.2: Kafka Architecture

Similarly, all the incoming tweets from apache flume will be ingested here in apache kafka. We will be also creating 'twitterdata1' topic in kafka. This created topic is subscribed to read the tweets from apache flume and will be fetched in the topic. After getting the tweets in kafka topic, we will be consuming all the raw tweets whatever we get from Twitter Streaming API.

c. Apache Spark

Apache Spark is a lightning fast anddistributed cluster computing framework. Spark core is the foundation of this overall project.

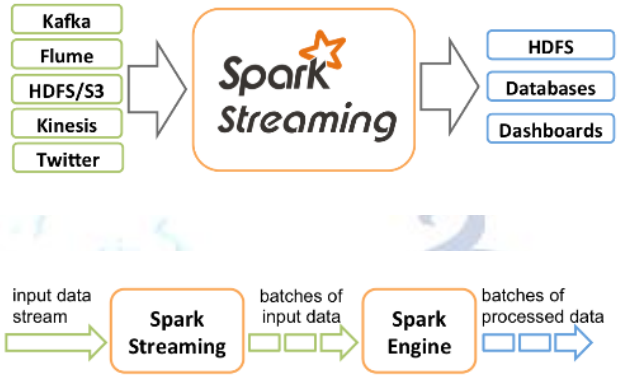

Fig 4.3: Apache Spark Architecture

Whatever raw data we get from the Twitter Streaming API after consuming from Kafka, it is forwarded to apache spark for analyzing the trending topics in twitter. We apply spark RDD and spark DataFrames, to filter out the trending HashTags and topics in the twitter.

d. MySQL Database

MySQL is a fast, easy to use. It is a RDBMS tools which is being used by a number of small and big businesses, organizations and much more. In our project, after applying sentiment analysis on tweets in apache spark, the output results are stored in MySQL Database.

e. Tableau

Tableau is a Business Intelligence tool in order to analyze the data visually. We can design and distribute an interactive and shareable dashboard, which represents the trends, variations, and density of the data in the shape of graphs and charts. Here, in our project, after storing the results in MySQL Database, Tableau is connected to MySQL to get the live results of the trending topics and trending HashTags of the twitter social media.

V. FLOWCHART

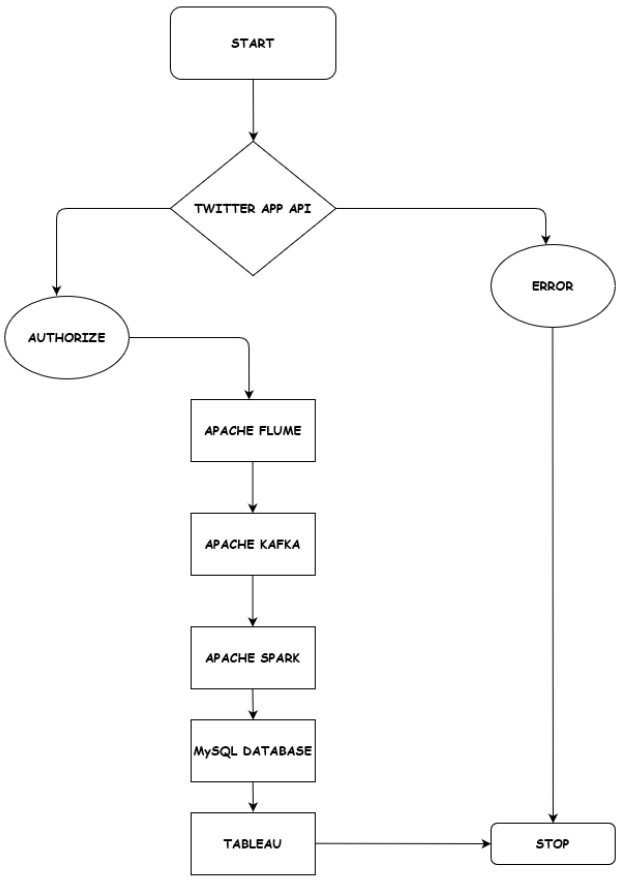

Fig 5.1: FlowChart

VI. WORKING OF PROPOSED APPROACH(PSEUDO CODE)

Creating Twitter Account:

- Applying for Twitter Developer Access
- After Verification, Getting Twitter App API Keys

Setting Up Apache Flume-

- Configuring Our Flume Source as Twitter Source
- Configuring Our Flume Channel as Ch-memory Channel
- Configuring Our Flume Sink as Kafka Sink

Setting Up Apache Zookeeper and Apache Kafka -

- Initializing Zookeeper
- Initializing Kafka
- Creating Kafka Topic
- Executing Apache Flume Conf File
- Consuming Real-Time Tweets in Kafka Consumer

Setting Up Apache Spark and SBT Tool-

- Initializing SBT Tool
- Initializing Spark Streaming
- Executing Our Scala Code in Spark Streaming
- Applying Our Sentiment Analysis on Real-Time raw Tweets to Get Final Output.
- Initializing MySQL Database Service
- Final Output is Stored in MySQL Database

Evaluate Results-

- MySQL is Connected is Tableau Visualization Tool
- And Final Results are Visualized through Tableau.

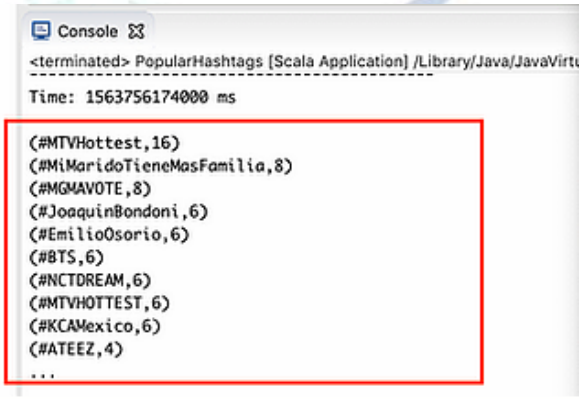

Fig 5.1: Trending HashTags Results

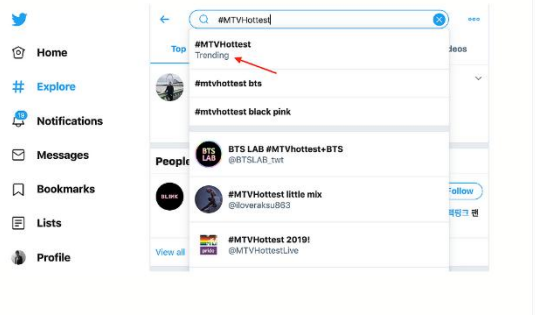

Fig 5.2: Trending HashTags in Twitter

VII. CONCLUSION

Sentiment Analysis is one of the most significant areas of Text Analysis. This study is a distributed approach to process real-time streaming of unstructured data inmassive amounts, in order to make decisions by visualizing positive or negative mentions of a product or service, healthcare problems, trending technologies and business

decisions.This framework was proposed to gather, filter, and mine streams of data in three main phases of Ingestion, Processing, and Visualization.The significant improvement is the speed of processing tweets and implementation of big data frameworks, such as Apache Flume, Apache kafka, Spark Streaming on DStreams.

REFERENCES

[1] "Getting Started for Twitter App API," [Online]. Available: https://developer.twitter.com/en/docs/twitter-ads-api/getting-started. [Accessed 17 06 2020].

[2] "Installing SBT on Linux," [Online]. Available: https://www.scala-sbt.org/1.x/docs/Installing-sbt-on-Linux.html. [Accessed 16 03 2020].

[3] "Introduction to Apache Flume, Flume User Guide," [Online]. Available: https://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html. [Accessed 28 03 2021].

[4] "Introduction, Key Concepts, APIs," [Online]. Available: https://kafka.apache.org/documentation/#introduction. [Accessed 28 03 2021].

[5] "MySQL, MySQL Introduction, Installation," [Online]. Available: https://www.tutorialspoint.com/mysql/mysql-introduction.htm. [Accessed 08 05 2020].

[6] "Spark Streaming," [Online]. Available: https://databricks.com/glossary/what-is-spark-streaming. [Accessed 28 03 2021].

[7] "Spark Streaming on DStreams," [Online]. Available: https://spark.apache.org/docs/latest/streaming-programming-guide.html. [Accessed 28 03 2021].

[8] "Tableau Overview," [Online]. Available: https://www.tutorialspoint.com/tableau/tableau_overview.htm. [Accessed 31 03 2021].

[9] S. DAS, R. K. BEHERA, M. KUMAR and S. K. RATH, *Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction,* 2018.

[10] M. K. SAH and R. SHARMA, PHAGWARA, PUNJAB, 2020.

[11] A. SINGH, "Mr," PHAGWARA, 2020.