Big Data Analytics over Twitter Real Time Data using Kafka Streaming and KSQL with Confluent Bundled Platform

Mohana Durga Raja[2], Konda Sreenu[2]
[1]*M.Tech Student,* [2]*Assistant Professor*
Sir C R Reddy College of Engineering, Eluru, West Godavari Dt, AP, India

Abstract - As of now dealing with big data is a big deal. If we see Twitter, the data is continuous flow which is really huge. My Paper mainly deals with gathering real time twitter tweets, either those are of reply/retweeted/normal tweets from TwitterAPI and calculating the sentiment analysis on that particular tweet using Kafka Streaming and finally sending all those sentiment analytical data into the kafka topic, hence by making use of this data with the help of KSQL, we could able to calculate the overall sentiment analysis on a particular keyword, like how people reacted (positive/negative/neutral) about particular keyword. And here the keyword might be a person or an organization or anything else, where we wanted to find how people are speaking about that particular keyword.

I. INTRODUTION

Overview and Problem statement - People makes use of Twitter data for all types of business purposes, like brand awareness and monitoring. Twitter provides this data free of cost. Hence by making use of this data, we could able to get the sentiment analysis on a particular keyword.

A. About Big Data - Big Data differs from regular data in few characteristics known as the 3 V's: Big Data volume, velocity, and variety. Also, some other characteristics of Big Data are recently introduced as new V's such as value and veracity.

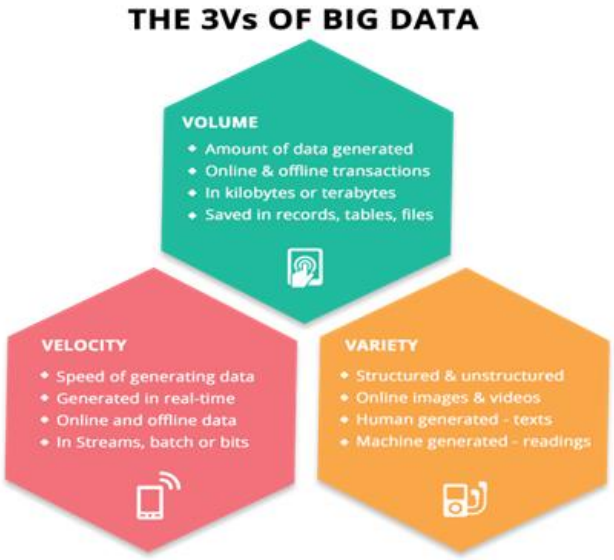

Figure 1: 3Vs of Big Data

So why Big Data is needed anyway?

As We Know that big data is nothing but vast data. In order to perform several data analytics on chunks of data, it is not possible. Hence big data came into the picture, with the ease of big data, we could able to perform several marketing analysis, as well as product analysis. So it would be easier for any organization by making use of big data.

B. Data Analytics - Once big data is given to any database or to any file, the data can be analysed by using several services or any other online tools. From the big data extracting the desired analysis on any product or any keyword simply called as data analytics.

C. Data analytics benefits upon Big Data - Rather than getting the big data and storing at some place, if we perform the analytics over big data we could able to extract the desired output of organization, rather than sending all those bulk data into the database. Storing the bulk data is waste of space and of no use in some cases. Hence storing the analytics results is much preferable than keeping the big data.

1. Big data analytics applications enable data scientists, predictive modelers, statisticians and other analytics professionals to analyse growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That contains semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensors connected to the internet of things.
2. On a broad scale, data analytics technologies and techniques provide means of analysing data and drawing conclusions about them to help companies make informed business decisions. Big data analytics is a form of advanced analytics, which involves complex applications with elements such as models, statistical algorithms and what-if analyses powered by high-performance analytics.

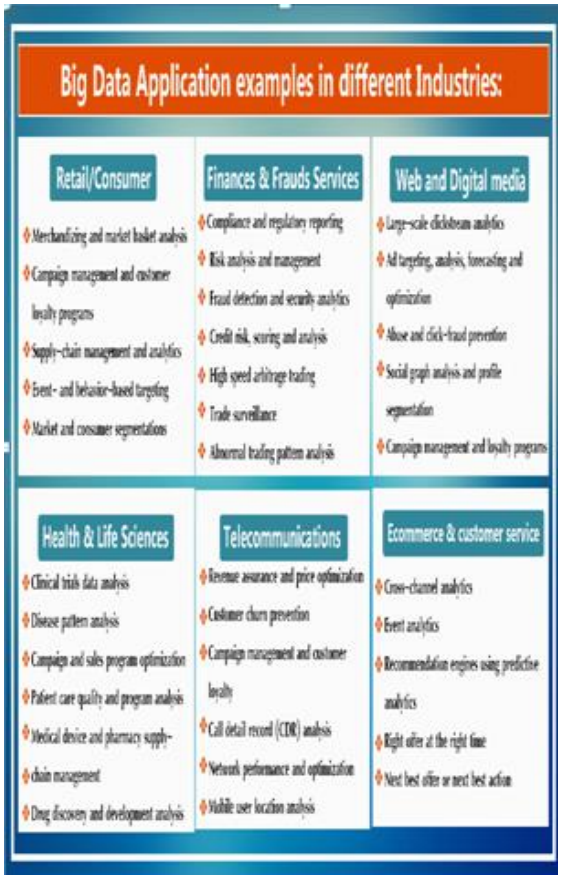

Figure 2

D. Applications of big data analytics - The below diagram shows the several applications for big data.

II. LITERATURE SURVEY

The existing system works on datasets(static), works on static data rather than dynamic data. The existing system is limited to particular data. The procedure includes taking the data sets and sending it to kafka. Later performing analytics by use of strom service.

A. Existing system also uses the storm service for analysing the data analytics coming from the kafka.

Existing project flow looks like given architecture

Data Sets→Collect Data Into

Perform Data Analytics using Strom Service
Store it in data base and hence used to send to UI

The above diagram represents the existing system to get the sentiment out of data set and hence performing the sentiment analysis on that particular static data set and sending it to the database or else sending it to the UI.

B. Data Set - Data sets taken as static.

C. Storm Service - Used to perform sentiment analytics on the data set

III. PROPOSED METHOD

Before going to explain the complete scenario let me give a brief introduction regarding all the services that I have been using in my project to complete the scenario.

A. Services that I used for the scenario -

1. Fetch Data from Twitter Streaming API
2. Confluent
3. Kafka Streaming

1. Fetching Data from Twitter Streaming API - As we know api meant for application programming interface. There are many web services which provides APIs to developers to communicate with services and accessing data in a programmatic way.

Step 1 - Fetch keys for Twitter API

Step 2 - Connect to Twitter Streaming API through java producer, fetching the data to Kafka.

2. Confluent - Confluent bundled platform is a new one that helps developer to start several services just on a fly with a single command. Previously, if we wanted to start kafka we need to start zookeeper alone and then kafka alone. Confluent removed that hectic of starting all these services several times.

Confluent Bundle could able to all these services at a time with a single command "Confluent start"

- Zookeeper\
- Kafka
- Schema-registry
- Kafka-rest
- Connect
- KSQL

Zookeeper - Zookeeper is a monitoring service for kafka. We can say zookeeper maintains as supervisor for below all services and acts as coordinator.

Kafka - Kafka is simply a messaging pipeline, allows continuous flow of data.

Schema Registry - One of the service that confluent bundle starts up, Used to store metadata.

Kafka Rest - Kafka rest acts as restful interface for the kafka cluster.

Connect - It is the super service, that can connect kafka continuous data with the desired data base. Confluent itself comes up with several default connect drivers.

KSQL Server - Could able to Query and visualize the big data resided in kafka topic.

3. Kafka Streaming - Kafka provides one library (Streaming library) to perform data analytics over the continuous flow of data resided in kafka topic. And stores the analytical data back to a kafka topic.

Architectural flow of our proposed System

Fetching the data from the Twitter
Sending the Twitter Json data to Kafka
Performing sentiment Analytics on Tweet Text using Kafka Streaming (Confluent platform)
Visualize and Query the big data using KSQL (Confluent Platform)

B. Differences between existing and proposed system -

1. In the existing system, user could able to perform analytics over only static data sets. In proposed system, user had a choice to opt for any keyword and could able to get live twitter data, which is dynamic.
2. Existing system follows older service compared to the new service, that we have been in the proposed system.
3. Existing system does not allow user to choose theirown keyword but proposed system allows us to query our own keyword.
4. Existing uses storm. Our proposed system uses no extra service other than confluent kafka.

IV. CONCLUSIONS

Data analytics(Sentiment Analysis) became a major part while taking major decisions in ay organization, hence in this paper we discussed how to calculate sentiment by using advanced technologies such as kafka streaming and ksql with confluent platform. On top of that, we are not using extra service called storm, that increases the speed of overall analysis.

V. REFERENCES

[1]. Geetika Gautam, Divakar Yadav. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. IEEE 2014.
[2]. Neethu M S, Rajasree R. Sentiment Analysis in Twitter using Machine Learning Techniques. IEEE 2013.
[3]. W. Yang, X. Liu, L. Zhang, and L. T. Yang. Big Data real-time processing based on Storm. In 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pages 1784–1787. IEEE, 2013.
[4]. Datadog Engineering Blog. Monitoring Kafka performance metrics. 23 May 2016.
[5]. A. R. Baig and H. Jabeen. Big Data analytics for behavior monitoring of students. Procedia Computer Science, 82:43–48, 2016.
[6]. M. T. Jones. Process real-time Big Data with twitter Storm. IBM Technical Library, 2013.
[7]. V. Ta, C. Liu, G. Wandile. Big Data Stream Computing in Healthcare Real-Time Analytics. IEEE International
[8]. Shahrivari. Beyond batch processing: towards real-time and streaming Big Data. Computers, 3(4):117–129, 2014.
[9]. A. B. Patel, M. Birla, and U. Nair. Addressing Big Data problem using Hadoop and map reduce. In 2012 Nirma University International Conference on Engineering (NUiCONE), pages 1–5. IEEE, 2012.
[10]. Carpenter, T. Way, "Tracking Sentiment Analysis through Twitter" in ACM computer survey, Villanova:Villanova University, 2010.
[11]. A. Pak, P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" in Special Issue of International Journal of Computer Application, France:Universitede Paris-Sud, 2010.
[12]. M. Rambocas, J. Gama, "MarketingResearch: The Role of Sentiment. Analysis", The 5 th SNA-KDD Workshop'11 , 2013.
[13]. X. Chen, M. Vorvoreanu, K. Madhavan, "Mining Social Media Data to Understand Students' Learning Experiences", IEEE Transaction, vol. 7, no. 3, pp. 246-259, 2014.
[14]. Twitter Engineering, "200 million Tweets per day", Twitter Official Blog., June 2011, [online] Available: https://blog.twitter.com/2011/200-million-tweets-per-day.