SCHOLEDGE International Journal of Multidisciplinary & Allied Studies
ISSN 2394-336X, Vol.07, Issue 05 (2020), Pg 123-130.
DOI: 10.19085/sijmas070501

www.thescholedge.org |Email: editorial@thescholedge.org

Design & Framework of Real Time Twitter Analysis Using Apache Flume and Spark for Trending Technology

Amit Pimpalkar[1], Arti Zade[2], Devashree Jaronde[3], Gouravi Bajpai[4], Kimaya Bahe[5]
[1]Professor and Head, G H Raisoni Academy of Engineering and Technology, Nagpur, India.
[2,3,4,5] Student, G H Raisoni Academy of Engineering and Technology, Nagpur, India.

Abstract
The Twitter investigation is a dynamic field. It is an approach to recognize the demeanor, perspective, or feelings of the individual towards an item, administration, film, and so forth by examining the assessments and surveys shared through online systems administration media, writes, etc. Diverse electronic long range interpersonal communication stages like twitter grant people to confer their points of view to other people. Twitter becomes the most famous online life stage that permits clients to share data by method for the short messages called tweets consistently. A huge number of individuals cooperate with one another simultaneously and an enormous measure of information is delivered in a moment or two. This examination will try to develop a logical framework with the limit of in-memory getting ready to separate and separate composed and unstructured Twitter data. We have coordinated a relevant examination on tweets about the slanting advances in India as indicated by the city. Our trial results show the itemized order pretty much all the ongoing inclining innovations accessible.

Keywords: Twitter, Apache Spark, Apache Flume, Real-time

I. Introduction

The uptrend graph for the use of social media or micro blogging platforms has become an integral part of the user's routine. The increase in Big Data has reached an extreme level. It is known that the capacity of data will reach 45 trillion GB until 2020. All the notices, photographs, videos activity, and area posted by clients on their interpersonal organization contain valuable data about their socioeconomics, sees, likes loathes, and so on. Organizations are teaching this data for dissecting to get a serious edge. Spatial breaking down has been concentrated well and all extraordinary aspect approach has been talked about. This research paper is totally focused on twitter analysis.

A segment of the employments of the progressing data examination are observation, condition, restorative administrations, business Intelligence promoting, discernment, computerized security, maneuvers, and online informal communication. This paper presents a constant information system for breaking down twitter information. The real-time twitter data is used to analyze the upsurge technology in the cities. When the person or an individual tweets there are 150 attributes are being shared and they are being processed in spark using Scala. Scala language is used to processing of data. In this apache, the flume is used to collect the data from the generator and it is used to collect the streaming data from the generator and the apache spark is used to analyze that raw data and gives it the structure of rows and columns. This examination will assist a person in finding a way to manufacture a bearer in innovation in a proper city. There will be no disarray no puzzlement for picking a city or innovation.

II. Related Work

This assessment will attempt to build up an interpretive system with the constraint of in-memory preparing to expel and dismantle created and unstructured Twitter information. The proposed structure right now information ingestion, stream dealing with, and information depiction deserts the Apache Kafka lighting up a framework that is utilized to perform information ingestion task. Moreover, Spark makes it conceivable to perform current information dealing with and AI figuring powerfully. We have facilitated a coherent assessment of tweets about the tremor in Japan and the responses of individuals the world over with evaluation on the time and inception of the tweets. [1]

In this paper, auspicious examination over enormous information is a key factor for achievement in numerous business and administration spaces. A few instances of these spaces incorporate fund, transportation, vitality, security, military, and crisis reaction. A few major information applications in these spaces depend on quick and convenient investigation dependent on accessible information to settle on quality choices. This paper reviews continuous large information examination applications and their specialized difficulties. [2]

This paper hopes to address this issue by working up a progressing data assessment framework fit for dealing with the persistent treatment of sorted out and unstructured data required for performing particular symptomatic assignments, running from data ingestion and planning to data examination, and portrayal. The proposed designing subject to the Storm/YARN adventures for data ingestion, dealing with examination and view of spilling sorted out and unstructured data. We have completed the proposed building using Apache Storm related APIs for both of a local mode and a scattered mode. All portions had the choice to manage their own functionalities suitably. [3]

This research paper reviews various methodologies continuously examination of Big Data or close to constant in the particular fields of utilization, just as devices and systems being utilized. The overview results show what advancements have been utilized in every one of the fields of utilization and what the explanation behind the decision was. [4]

This research draws consideration from clients confirming by the number of important reports and correspondence cooperation toward that point. In any case, customary theme recognition approaches are not intended to identify the sort of occasion effectively continuously, especially if the information sources are impacted by clamor information and containing differing subjects. To conquer the right now proposed a model for removing and following genuine get-togethers on Social Data Stream which can function admirably progressively by utilizing dispersing calculation and information collection strategy on the discrete signals as another portrayal of the first information. [5]

This paper exhibits a philosophy and foundation conveying such abilities. Not at all like different methodologies that have been taking things down a notch, this work abuses huge scope Cloud offices and a lot bigger assortments of information. In particular, we gathered and broke down more than 46 million tweets from the three most colonize urban communities in Australia to discover designs identified with wellbeing occasions. [6]

This paper proposes conventional engineering for huge information human services expository by utilizing open sources. The mix of high throughput convey purchase in advising for streams, appropriated steady figuring, and flowed amassing system can sufficiently separate a colossal proportion of human administration data going with a brisk rate.[7]

Twitter plays an important role in an individual's life. It helps in various ways for analyzing how much people are incorporate in social media, how many number of people are tweeting in an hour about the most trending topic, surveillances, environment, health care, business intelligence, marketing, visualization, cyber security. Twitter is also used by industries to do campaigning in cities according to the interest of people. Used by politicians to promote their parties and social services done by them. The streaming data is very useful because it gives an accurate analysis about the opinion of people. But in previous papers there no survey for trending technologies in cities of India.

III. Proposed Work

The outline is expected to think, channel, and separate spouting data and gives us sees about which development is slanting in India. The outline comprises of following advances for example information ingestion, Stream preparation, and information perception. Information ingestion is performed by Apache flume, a viable device to gather, total, and move the huge number of log information. Apache Spark is utilized to get to information, channel the information, and afterward break down the information by flash gushing. This permits general preparing errands as well as progressively confounded and significant level information examination calculations. The contextual analysis shows the quality and the significance of continuous information investigation via web networking media spilling data. It requires a proper schema which does not depend upon storing the data on the hard disk but can process the data in memory during the arrival. Analyzing views on social media may prove quite useful for draining conclusion and predicting the activities that occur in specific areas managing web-based life information, including a wide range of information types, for example, instant messages, photographs, and recordings which are showing up in an enormous volume inconsistently, needs a legitimate system which doesn't depend after putting away information on hard plates and can process information in memory, as it Analyzing posts on locales, may demonstrate very helpful for drawing.

a) Apache Flume

Apache Flume is a prorated, consistent feasible system for effective aggregating and propelling large data from various points to a centralized data store. The use of Apache Flume is not only bounded to data collection but the source of the data can be customized. Flume can be used to evacuate huge quantities of data which is not restricted to network traffic but media-generated data, email me pretty much source possible.

b) HDFS

Apache Flume stores the information into any of the unified stores HBase or HDFS. At the point when the pace of approaching information violates the rate at which information can be written to the goal, Flume carries on as a middle person between information originator and the brought together center and give after putting away information on hard plates and can process information in memory, as it shows up. Analyzing posts on goals, for instance, Twitter may show truly significant for causing judgments and making conjectures about activities that occur in the express world at explicit events.

The use of Apache Flume is not only bounded to data collection but the source of Flume can be used f data which is to network traffic but socially generated data, email messages, and Apache Flume stores the data into any of the centralized stores HBase or HDFS. When the ep the rate at which data can be penned to the destination, Flume behaves as a mediator between data centralized hub and provide a Flume provides the feature of routing. The transactions in Flume are channel-based where one sends recipient is maintained for each message. Flume is also used to import a huge density of event data generated by social media sites.

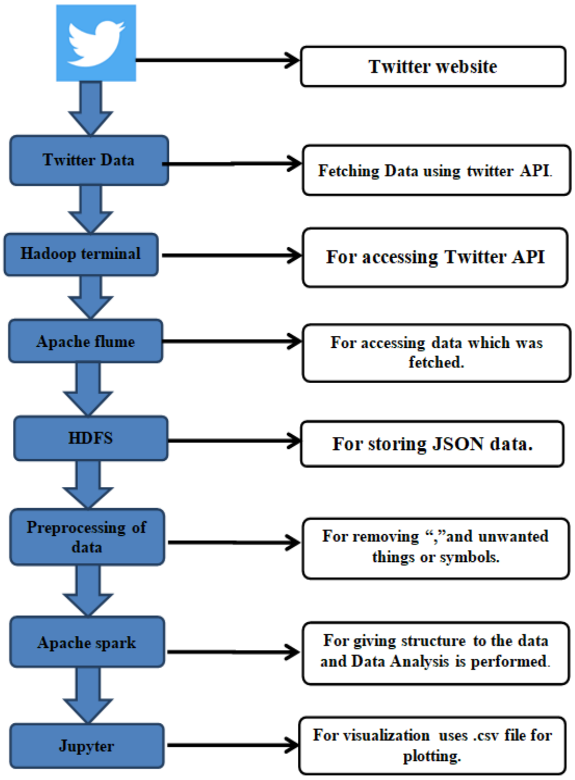

Twitter website
Twitter Data
Fetching Data using twitter API
Hadoop terminal
For accessing Twitter API
Apache flume
For accessing data which was fetched.
HDFS
For storing JSON data.
Preprocessing of data
For removing ","and unwanted things or symbols.
Apache spark
For giving structure to the data and Data Analysis is performed.
Jupyter
For visualization uses .csv file for plotting.

Fig 1 Block diagram of design Framework

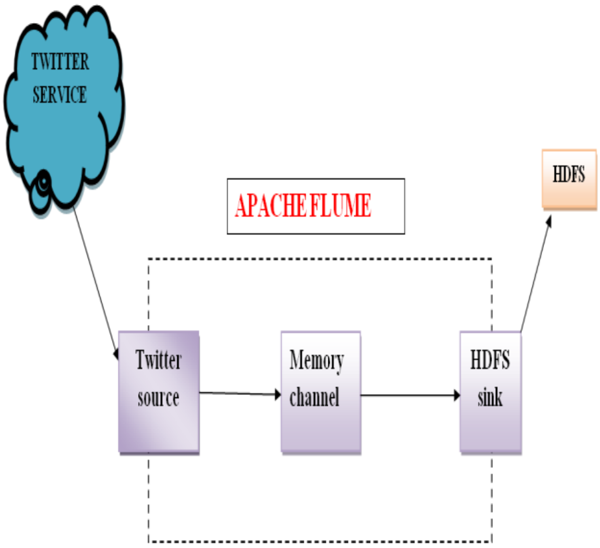


Fig2. Components of Apache Flume

c) Apache spark

Apache Spark is a source which is flexible in memory data processing that allows MIlib, and advance analysis on the Hadoop platform. Apache Spark is well positioned to replace Map Reduce as the default data processing engine for Hadoop.

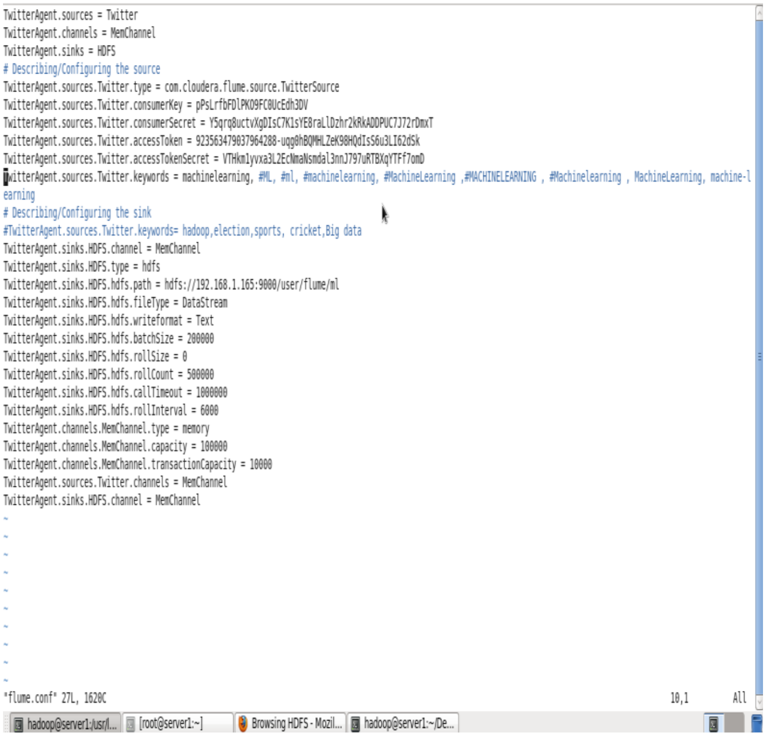
d) Visualization through Jupyter

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

IV. Implementation

Twitter Data set

The unprocessed dataset contains tweet text exactly as posted on Twitter along with metrics such as: Tweet ID, Date and Time of posting the tweet, Name, and username of the account posting the tweet and many other value metadata present in the raw datasheet. Flume is also used to import a huge density of event data generated by social media sites. It uses data for opinions. Twitter, a famous smaller scale blogging webpage, has a lot of APIs that permit us to get the tweets and can control the tweets. Right now, are allotted a key and a mystery token that our application uses to verify for our sake. When the application is verified, and we would then be able to utilize the Twitter APIs to get tweets.



```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.consumerKey = pPsLrfbFDlPKO9FC0UcEdh3DV
TwitterAgent.sources.Twitter.consumerSecret = Y5qrq8uctvXgDIsC7K1sYE8raLlDzhr2kRkADDPUC7J72rDmxT
TwitterAgent.sources.Twitter.accessToken = 923563479037964288-uqg0hBQMHLZeK98HQdIsS6u3LI62dSk
TwitterAgent.sources.Twitter.accessTokenSecret = VTHkm1yvxa3L2EcNmaNsmdal3nnJ797uRTBXqYTFf7omD
TwitterAgent.sources.Twitter.keywords = machinelearning, #ML, #ml, #machinelearning, #MachineLearning ,#MACHINELEARNING , #Machinelearning , MachineLearning, machine-l
earning
# Describing/Configuring the sink
#TwitterAgent.sources.Twitter.keywords= hadoop,election,sports, cricket,Big data
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://192.168.1.165:9000/user/flume/ml
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 200000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 500000
TwitterAgent.sinks.HDFS.hdfs.callTimeout = 1000000
TwitterAgent.sinks.HDFS.hdfs.rollInterval = 6000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 100000
TwitterAgent.channels.MemChannel.transactionCapacity = 10000
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
~
~
~
~
~
~
~
~
~
~
"flume.conf" 27L, 1620C                                                  10,1          All
```

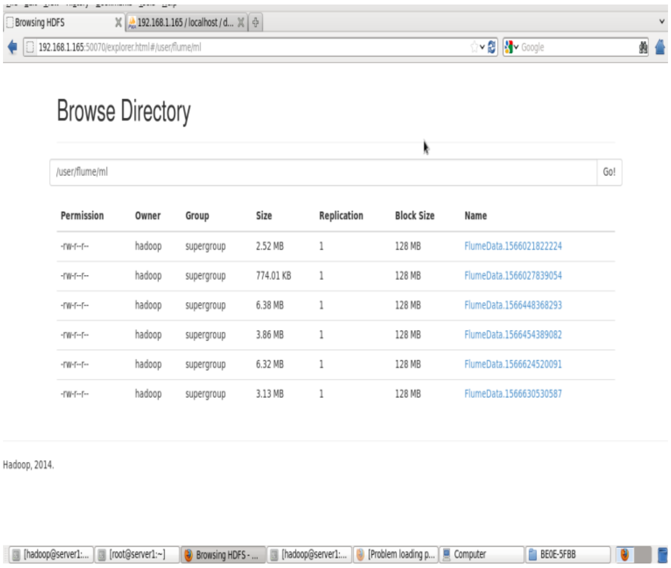
Fig3. Flume configuration

Hadoop Distributed File System

The twitter file gets synchronized with HDFS by command

“Flume-ng agent -n Twitter Agent-f/usr/local/flume/conf/flume.cof.“

Once the synchronization is established then the data will get extracted by the commandHadoopfs-tail–fhe's://192.168.1.165:9000/user/flume/ml/Flu meData.1566630530530587.tmp



Browse Directory

/user/flume/ml

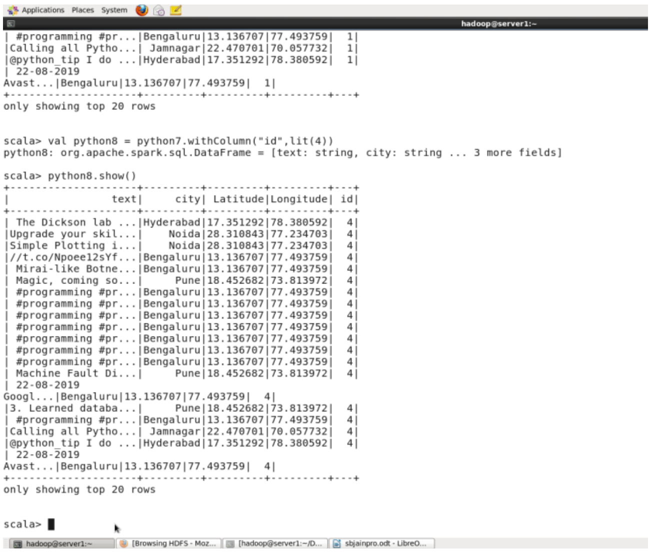
Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	2.52 MB	1	128 MB	FlumeData.1566021822224
-rw-r--r--	hadoop	supergroup	774.01 KB	1	128 MB	FlumeData.1566027839054
-rw-r--r--	hadoop	supergroup	6.38 MB	1	128 MB	FlumeData.1566448368293
-rw-r--r--	hadoop	supergroup	3.86 MB	1	128 MB	FlumeData.1566454389082
-rw-r--r--	hadoop	supergroup	6.32 MB	1	128 MB	FlumeData.1566624520091
-rw-r--r--	hadoop	supergroup	3.13 MB	1	128 MB	FlumeData.1566630530587

Hadoop, 2014.

Fig4. Real time twitter data

Apache spark

A solitary tweet consists of 150 fields on which analysis can be performed. Apache Spark helps to convert unstructured data into structured data.



```
| #programming #pr...|Bengaluru|13.136707|77.493759|  1|
|Calling all Pytho...| Jamnagar|22.470701|70.057732|  1|
|@python_tip I do ...|Hyderabad|17.351292|78.380592|  1|
| 22-08-2019
Avast...|Bengaluru|13.136707|77.493759|  1|
+--------------------+---------+---------+---------+---+
only showing top 20 rows


scala> val python8 = python7.withColumn("id",lit(4))
python8: org.apache.spark.sql.DataFrame = [text: string, city: string ... 3 more fields]

scala> python8.show()
+--------------------+---------+---------+---------+---+
|                text|     city| Latitude|Longitude| id|
+--------------------+---------+---------+---------+---+
| The Dickson lab ...|Hyderabad|17.351292|78.380592|  4|
|Upgrade your skil...|    Noida|28.310843|77.234703|  4|
|Simple Plotting i...|    Noida|28.310843|77.234703|  4|
|//t.co/Npoee12sYf...|Bengaluru|13.136707|77.493759|  4|
| Mirai-like Botne...|Bengaluru|13.136707|77.493759|  4|
| Magic, coming so...|     Pune|18.452682|73.813972|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
| Machine Fault Di...|     Pune|18.452682|73.813972|  4|
| 22-08-2019
Googl...|Bengaluru|13.136707|77.493759|  4|
|3. Learned databa...|     Pune|18.452682|73.813972|  4|
| #programming #pr...|Bengaluru|13.136707|77.493759|  4|
|Calling all Pytho...| Jamnagar|22.470701|70.057732|  4|
|@python_tip I do ...|Hyderabad|17.351292|78.380592|  4|
| 22-08-2019
Avast...|Bengaluru|13.136707|77.493759|  4|
+--------------------+---------+---------+---------+---+
only showing top 20 rows


scala> 
```

Fig5. Structured data in spark

The above figure shows the filtered data from the unstructured downloaded data from twitter. It consists of ID's, text, city, latitude, and longitude for spatial analysis.

V. Result

The prognosticative analysis showed the result of which technology is trending in which city in India. It is about focusing on which technology is trending and build their strategies based on the real public opinions or tweets. We have studied five IT technology that is Hadoop, Artificial intelligence, machine learning, cloud computing, and python. We have considered the tweets of India only and we got 108 tweets of every technology so the number of tweets that were analyzed was 540. We came to know that Hadoop, Artificial Intelligence and Machine Learning is trending over Chennai (India), Python is trending over Bengaluru (India) and Cloud computing is trending over Hyderabad (India). The analysis of all technology concludes that Artificial Intelligence is trending all over India with the largest number of views by the user over India.

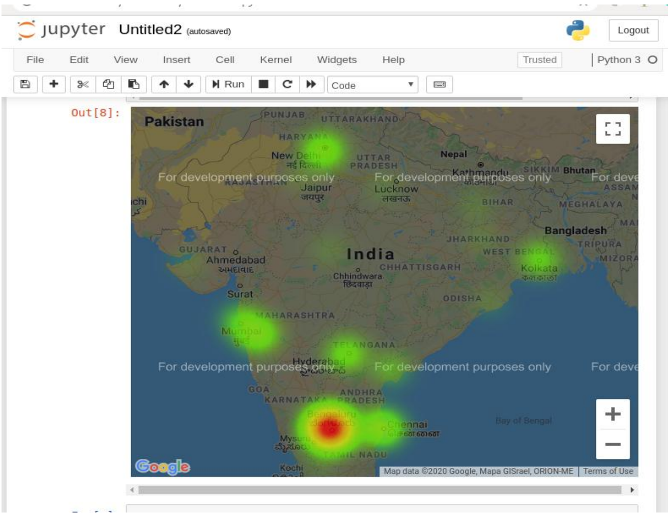


Fig6.Heat map showing cities with trending technologies

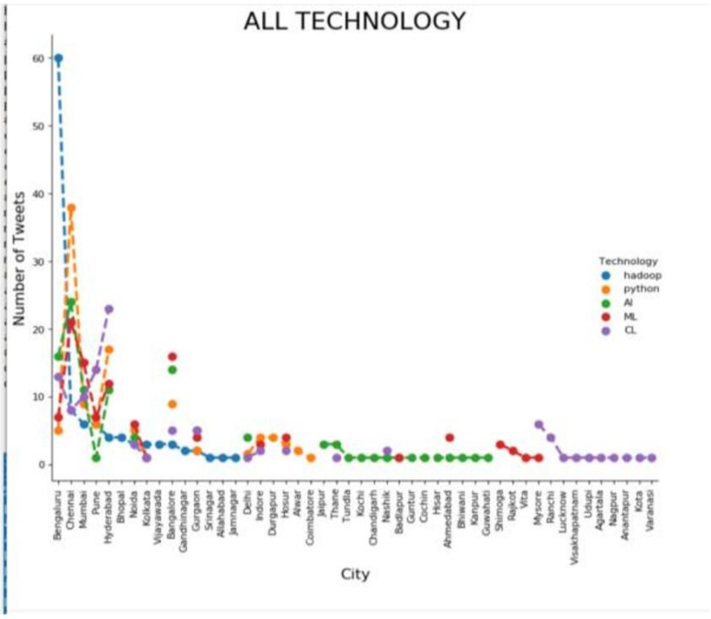


Fig7.Analysis of twitter data

VI. Conclusion

The contextual investigation right now to show the quality and the significance of continuous information examination via web-based networking media spilling data. Spatial investigation is helpful in online life observation as it permits us to increase an outline of the more extensive popular supposition behind a certain point. In this paper, we have analyzed public opinion about which IT technology is trending in India and analyze the opinion about technology through tweeter data. The prediction can be analyzed by seeing the graph of the top 5 trending technologies in the country. With Predictive Analytics we can conclude various things like which technology is trending in which city, which technology has higher carrier scope in IT technology etc.

This project can be used in decision making for an individual who wants to nurture a specific technology belonging to a particular city. It also gives the classification of the most trending technology in India. The accuracy of the project is so good that in IT cities it gives the perfect result as shown above in Bengaluru, it is the highest as we all know.

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.

[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, 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]. D. T. Nguyen and J. E. Jung. "Real-time event detection for online behavioral analysis of big social data" Future Generation Computer Systems, 2016.

[6]. J. Zaldumbide, R. O. Sinnott, "Identification and Validation of Real Time Health Events through Social Media" 2015 IEEE International Conference on Data Science and Data Intensive Systems.

[7]. 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.

[8]. M. T. Jones. Process real-time Big Data with twitter Storm. IBM Technical Library, 2013.

[9]. G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin. Fast data in the era of Big Data: Twitter's real-time related query suggestion architecture. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data.

[10]. A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, JM. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, "Storm@ twitter", InProceedings of the 2014 ACM SIGMOD international conference on Management of data 2014 Jun 18.

[11]. www.twitter.com