TWITTER ANALYSIS WITH SPARK, HIVE

G.Sai Ram pavan AM.EN.U4AIE20125
Department of Compter Science(Artificial Intelligence)
Amrita Viswa Vidyapeetam
kollam,kerala, India
amenu4aie20125@am.students.amrita.edu

JITHIN JOHN AM.EN.U4AIE20135
Department of Compter Science(Artificial Intelligence)
Amrita Viswa Vidyapeetam
kollam,kerala, India
amenu4aie20135@am.students.amrita.edu

K.Akash Varma AM.EN.U4AIE20141
Department of Compter Science(Artificial Intelligence)
Amrita Viswa Vidyapeetam)
kollam,kerala, India
amenu4aie20141@am.students.amrita.edu

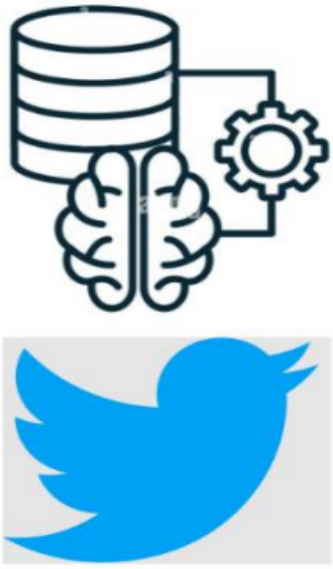
N.Moneesh AM.EN.U4AIE20150
Department of Compter Science(Artificial Intelligence)
Amrita Viswa Vidyapeetam
kollam,kerala, India
amenu4aie20150@am.students.amrita.edu

***Abstract*—'BIG DATA' is getting much significance totally different businesses over the final year or two, on a scale that has produced lots of data each day. Huge Information could be a term connected to information sets of exceptionally huge measure such that the conventional databases are incapable to process their operations in a sensible sum of time. It has heavy power to convert trade and power in a few ways. Here the challenge isn't as it were putting away the information, but moreover getting to and examining the specified information in indicated sum of time. One of the well-known usages to unravel the over challenges of enormous information is utilizing Hadoop. It is exceedingly adaptable compute stage. In this paper, we are progressing to conversation how viably analysis done on the information which is collected from the Twitter utilizing Flume and we did analysis using Apache spark also. Twitter is a web application which contains wealthy sum of information that can be a organized, semi-structured and unstructured information.**

***Index Terms*—BIGDATA, Flume, Hive, Twitter, Tweets, Un-Structured**

I. Introduction

n past, businesses and companies doesn't got to store, do many operations and analytics on data of the clients. But around from 2005, the have to be change everything into data is much engaged to fulfil the prerequisites of the individuals. So Enormous information came into picture within the genuine time commerce examination of preparing information. The investigation of huge information is the most up to date subject of intrigued for analysts all around the world. Big Data is defined as an amount of data that exceeds the capabilities of typical database technologies to store, access, manage, and calculate. Companies can forecast client behaviour, enhance marketing strategies, and gain competitive advantage over its competitors by studying this massive volume of data. From 20th century onwards this WWW has totally changed the way of communicating their sees. Display circumstance is totally they are communicating their considerations through online blogs, talk shapes additionally a few online applications like Facebook, Twitter, etc. Twitter clients have found numerous diverse employments, counting essential communication between companions and family, a way to publicize an occasion, or as a client relations instrument for companies to communicate with their buyers. With this much of extension and accessibility of information, the degree of web based life data being made is expanding exceptionally quick. Each minutes seconds there are 11,0000+ tweets, 775,000+ status upgrades, 1,10,00,000+ moment messages , 698,445 google looks. On the off chance that we take Twitter as our case about 1TB of content information is creating within a week within the frame of tweets. These all information is unstructured and enormous in measure. So, by this it is get it depicts Web is developing the way of living and fashion of population.



Identify applicable funding agency here. If none, delete this.

II. Ease of Use

A. Apache flume:

Apache Flume is a tool for separating and transferring huge volumes of streaming data from multiple sources, such as log files and events, to a centralised data storage. Flume is a tool that is extremely dependable, distributed, and configurable. Its main purpose is to transfer streaming data (log data) from several web servers to HDFS. Flume is a programme that accelerates the transfer of log data from application servers to HDFS.

The Apache Flume architecture :

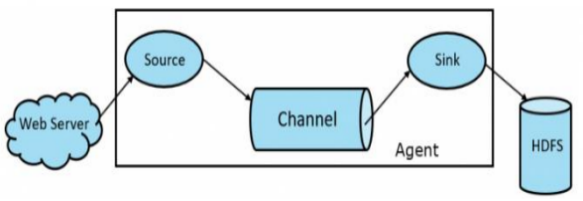
1) Flume Source: it can be found on data producers such as Facebook and Twitter. The source gathers data from the generator and sends it to the Flume Channel within the shape of Flume packets.

2) Flume Channel: it is a Intermediary Store that buffers Events supplied by Flume Source until they are received by Sink. Channel serves as a link between the Source and the Sink.

3) Flume Sink: it is present in data repositories such as HDFS and HBase. Flume sink absorbs Channel events and stores them in Destination stores such as HDFS

4) Flume Agent: it is a Java process that operates on the Source – Channel – Sink Combination. Flume can contain many agents. Flume may be thought of as a network of linked Flume agents scattered across nature.

5) Flume Event: it is the data unit transmitted by Flume. Event is Flume's generic representation of the Data Object.



Flume has several applications:

Consider the case of an e-commerce web application that needs to examine consumer behaviour in a certain region. They'd have to shift the existing log data into Hadoop for analysis in order to do so. Apache Flume saves the day in this situation.

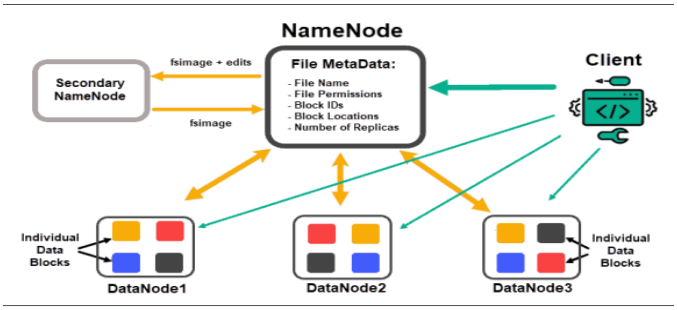
B. Hadoop

Apache Hadoop is a free and open source stage for putting away and preparing gigantic datasets extending in measure from gigabytes to petabytes. Hadoop permits clustering a few computers to examine huge datasets in parallel, instead of requiring a single huge computer to store and dissect the information

1) Hadoop Distributed File System(Hdfs): The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has numerous similitudes with existing dispersed record frameworks. In any case, the contrasts from other dispersed record frameworks are critical. It is profoundly fault-tolerant and is planned to be conveyed on low-cost equipment. It gives tall throughput get to application information and is reasonable for applications having huge datasets. Uses a Master/Slave Design, where a cluster comprises of a single NameNode (master hub) and all the other hubs are DataNodes (Slave hubs). HDFS can be sent on a wide range of machines that bolster Java. In a Hadoop cluster, the NameNode serves as the master, guiding the Datanode (Slaves). Namenode is mostly used to store Metadata, or information about information. Meta Information can be the exchange logs that trace a user's activity in a Hadoop cluster, as well as the title of the record, measure, and data near the location(Block number, Piece ids) of Datanode that Namenode keeps to find the nearest DataNode for faster communication. DataNodes are instrumented by Namenode using operations such as delete, make, imitate.

DataNode:

DataNodes serve as slaves. DataNodes are primarily used to store data in a Hadoop cluster; the number of DataNodes can range from one to 500 or even more. The Hadoop cluster can hold more data as the number of DataNodes increases. As a result, it is recommended that the DataNode have a high storage capacity in order to hold many number of file blocks.

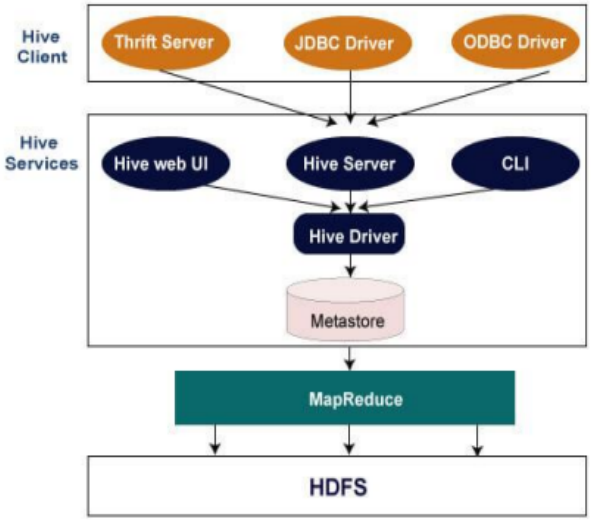


2) Applications of hdfs: Companies use Hadoop for understanding clients prerequisites Diverse companies such as fund, telecom utilize Hadoop for finding out the customer's prerequisite by looking at a huge sum of information and finding valuable data from these tremendous sums of information

C. HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy, it allows users to read, write, and manage petabytes of data using SQL. It enables SQL developers to write Hive Query Language statements similar to standard SQL statements Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive.

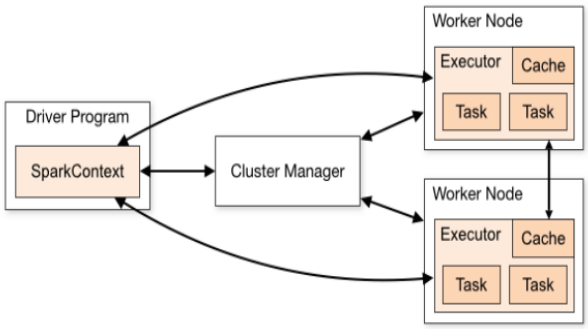
1) Architecture of hive:



2) Hive applications: It is utilized for running end-of-day reports, checking on every day exchanges, making ad-hoc inquiries, and performing information examination.

D. Apache spark

Spark is an Apache project advertised as “lightning fast cluster computing”. It has a thriving open-source community and is the most active Apache project at the moment and it provides a faster and more general data processing platform. Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop. It is used in sentiment analysis, predictive intelligence, customer segmentation, and recommendation engines, among other things. Another mention-worthy application of Spark is network security.



III. LITERATURE REVIEW

Research paper on ”Real-time Twitter data analysis using Hadoop ecosystem” which was published by Anisha P. Rodrigues; Niranjan N. Chiplunkar on 19 October 2018 , in this paper they have discussed about big data introduction, different tools used for big data, hadoop ecosystem, their proposed method to do twitter analysis using hadoop by using apache pig,flume,hdfs and at last they compared the performance of doing this task between apache pig,hive by performance evaluation, execution time.

Research paper on ”Sentiment Analysis on Twitter Data Using Apache Flume and Hive” which was published by Ms. Pooja S. Patil, Ms. Pranali B. sable, Ms. Reshma J. Fasale, Mr. P. A. Chougule on February -2016 , in this paper they have discussed about structured, unstructured data forms, hadoop ecosystem, creation of twitter application, flume, hive , analysed the results after doing the experiment, future applications of big data.

Research paper on ”Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive” which was published by Dr. ADITYA BHARDWAJ on December 2015 , in this paper they have discussed overview of big data, hadoop, apache pig, hive, sqoop, flume, compared all these by using some parameters like data structures they operate on, required software, companies using, developed by, language supported .

Research paper on ”Social Media based Sentimental Analysis using Hive and Flume ” which was published by Rahul Deva, Garima Kulshreshtha on November 2019 , in this paper they have discussed about hive, flume, parameters of big data, hadoop, MapReduce, unstructured data, twitter, dataset they used , analysed the results they got from their experiment

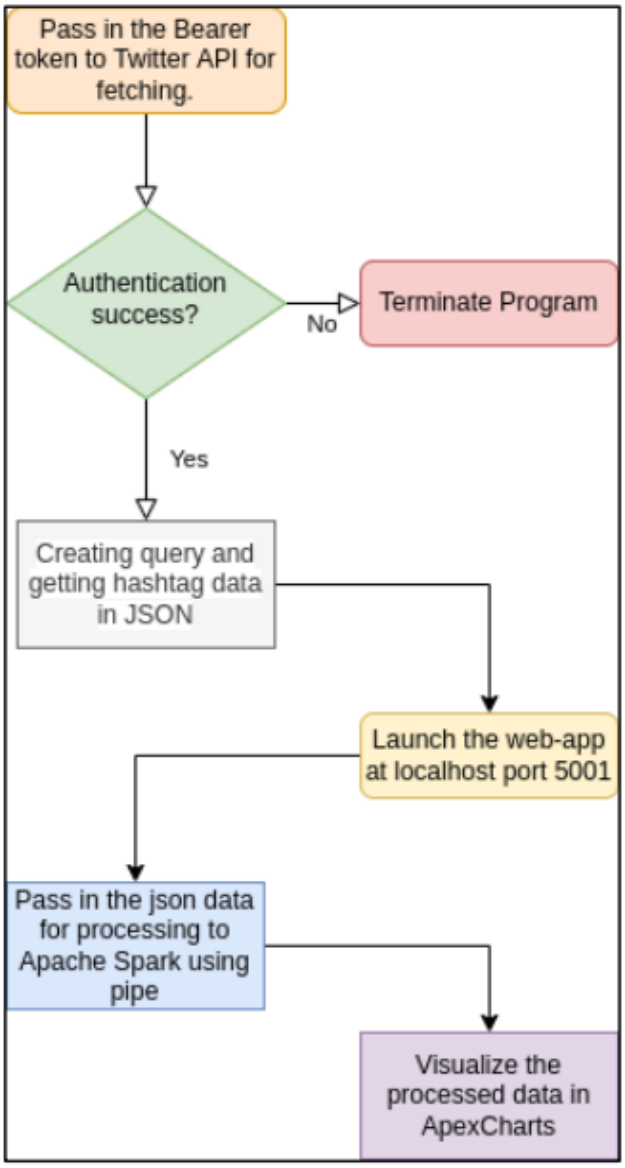
IV. METHODOLOGY

A. Using spark

We use Apache Spark streaming to process tweets retrieved from Twitter API and identify the trending hashtags from them based on a certain keywords and represent the data in a real-time dashboard using flask web framework In brief, This program uses fetch recent tweets based on “keywords” using Twitter API v2, filter hashtags from those tweets and give them to Apache Spark streaming for processing. After that it will launch a flask web server on localhost:5001 to view the data in a visual dashboard powered by ApexCharts.

The program will start to execute by passing parameters while running the python files with the help of a whole bash script which serves the purpose. twitter-app.py will establish a TCP connection with the twitter api and gets the data using requests HTTP library, and it is converted to json and sent to spark using send-tweets-to-spark() as a list. It will then read the keys.txt file which will have the bearer token for twitter developer access to the hashtags and creates a string object that contains the authentication information. And a url is created to do the query according to the parameters passed in.

The flask app gets executed afterwards and it will run the app on port 5001 and after sorting the data recieved in the request, it renders the ApexCharts As soon as the web app is up and running, the spark-app will start processing the data recieved and shows the realtime Analysis of the most popular hashtags published in tweets at that moment using the pyspark library. It also has the feature of reporting the errors if there is any problem in connecting to the webapp using localhost.



B. Using HIVE

FLUME steps

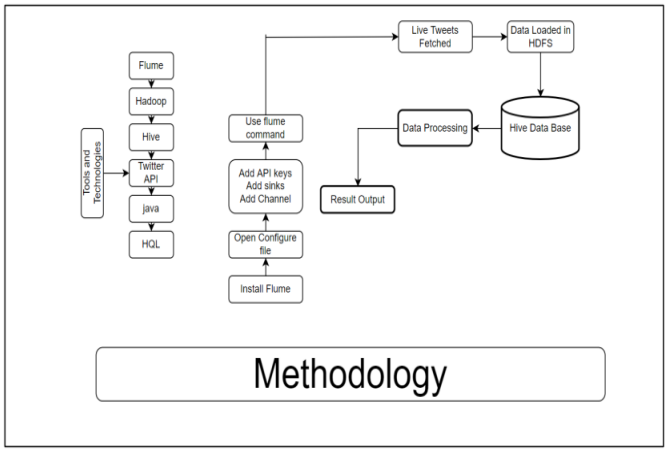
- Install Flume and set the Environment Variables.
- Edit the conf file.
- Add the API keys .
- Add source (from where we are fetching data. In our case it is twitter)
- Add sink (to which folder we are going to dump the data. In our case it is HDFS)
- Add Channel (Through which channel we are fetching data. In our case it is MemChannel)
- Run Flume command. Data is Fetched.

HIVE steps

- Install Hive
- Setup the environment
- Create warehouse folder in root/user/hive/warehouse
- The place where we store the tables is warehouse and we created that folder.
- Change the permissions to that folders such that hive can access that folder without an issue.
- Then start processing in hive.

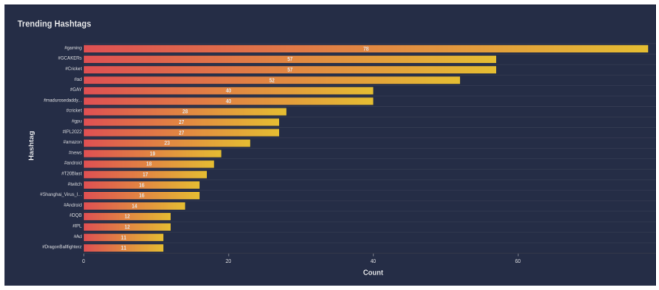
Hive Commands

- LOAD PATH INPATH ¡path¿ INTO TABLE ¡table-name¿ (loading hdfs data to hive table)
- Select * from tweets (displays the contents of table)



V. Results

A. Using Spark




```
Hashtag: #cryptonewshindi
Hashtag: #CryptoDFrog
Hashtag: #SoulMeta
Hashtag: #etherum
Hashtag: #stocks
Hashtag: #bitcoin
Hashtag: #HEX
No hashtag found in current tweet, moving on...
Hashtag: #Bitcoin
Hashtag: #Lightning
Hashtag: #Bitcoin
Hashtag: #THEFIRST
Hashtag: #BSC
Hashtag: #StarlinkInu
Hashtag: #BNB
Hashtag: #CRO
Hashtag: #CRORewards
Hashtag: #1000xgem
Hashtag: #Bitcoin
```

VI. Conclusion

There are several methods for analysing Data from twitter (unstructured data, semi-structured data) and analysing Twitter. Hive and Flume were utilised. Hive and Flume are Bigdata Hadoop solutions that are effective in extracting and loading data from both structured and unstructured data. Hive is a set of SQL-like tools for organising and queering unstructured data, whereas Flume is dependable and can gather, aggregate, and move massive amounts of streaming event data. There are several approaches for real-time data streams, such as utilising codes or Mapreduce. The Twitter data is stored in HDFS and it is analysed in this research. So, in comparison to the previous ways, the processing time is also quite short since Hadoop Map Reduce and Hive are the best methods for processing massive amounts of data in a short amount of time.

References

[1] Research paper on "Real-time Twitter data analysis using Hadoop ecosystem".

[2] Research paper on "Sentiment Analysis on Twitter Data Using Apache Flume and Hive" .

[3] Research paper on "Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive" .

[4] Research paper on "Social Media based Sentimental Analysis using Hive and Flume ".