Study on Twitter Big Data Solution

Iman Najmi   
Department of Faculty Computing   
Universiti Malaysia Pahang Al-Sultan Abdullah  
CB20138@student.umpsa.edu.my

Nur Izzlin  
Department of Faculty Computing  
Universiti Malaysia Pahang Al-Sultan Abdullah  
CB20012@student.umpsa.edu.my

*Abstract* — Twitter, which is now known as ‘X’, one of the most prominent and extensive social media platforms, accumulates millions of tweets daily covering a diverse range of crucial topics. This substantial volume of unprocessed data holds significant potential for various applications, including industrial, social, economic, government policies, or business purposes, contingent upon tailored organization and processing. Hadoop emerges as an optimal tool for the analysis of Twitter data, effectively handling distributed big data, streaming data, time-stamped data, and text data. This article explores the utilization of Flume to extract Twitter data, subsequently storing it in HDFS for analysis. Apache Hive is then employed to analyze the data, focusing on determining the geographical distribution of tweets and identifying the keywords associated with both the maximum and minimum tweet frequencies.

# Introduction

As we all know, X (formerly known as Twitter) has changed how news articles are shared on the social media platform, with the headlines no longer appearing on posts. Billionaire owner Elon Musk confirmed that he was behind the change in August, stating that it would "greatly improve" the look of news items being shared [1] . Elon Musk, a self-proclaimed defender of free speech, has made major changes to the social media platform since taking over last year. He renamed the X platform over the summer, replacing the famous blue bird associated with Twitter

Twitter provides the possibility to extract 1% of the tweets that are tweeted at that particular time. Based on the keywords given to the Twitter API the tweets are downloaded and the tweets are sources of various information [2]. Being one of the most important social media of our time, X contains a wealth of information. Millions of tweets are sent and received daily on X hence managing the analyzing tweets is a bit complicated due to the limited character length, and they often involve abbreviations, emoticons, slang, and many others.

There is also a case study of Twitter’s integration of machine learning tools into its existing Hadoop-based [3] The authors have identified stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data [3] This explains that X has enhanced its data processing capabilities by incorporating machine learning tools, specifically leveraging stochastic gradient descent for online ensemble and learning methods for processing. Efficiently bulk data into Hadoop-based systems

Next, to build classifiers for tweets processing we need to collect training data so that we can apply proper learning algorithms [4]. Twitter data is used as source, for example, in sentiment analysis where the task is to classify messages into two categories depending on whether they convey positive or negative feelings, since labeling tweets manually as positive or negative is a laborious and expensive task [4]. As stated this shows that the case study aiming to implement a suitable learning algorithm in order to manage or ranked the datas follows by the subject categories

# Brief Problem Statement

During the South by Southwest Interactive (SXSWi) conference in 2007, Twitter observed an increase in popularity, with the number of daily tweets gaining from 20,000 to a shocking 60,000 [4]. This significant increase in usage, however, brought with it several issues. The character restriction on Twitter, along with the common usage of acronyms, emoticons, idiom phrases, presented a severe challenge to efficiently organizing and analyzing the massive amount of tweets being sent. The variety of data included not only plain text but also emoticons and numerous multimedia features, resulting in a complex and varied dataset that faced considerable storage, managing, and inspection challenges. The complexity of the problem created a significant data engineering challenge, the likes of which had never been seen before in the field of social media

Innovative solutions were necessary to solve these difficult challenges and optimize the development potential of the social media platform Twitter. Introducing Hadoop, an open-source distributed computing framework that has proven to be a game changer. Hadoop's strengths in distributing, storing, and processing huge data sets made it a perfect candidate for managing the massive volume of tweets flowing in from all corners of Twitter [6]. Integrating Hadoop into Twitter's infrastructure is justified due to its scalability, distributed storage (HDFS), and parallel processing capabilities. Hadoop's strengths address the challenges posed by Twitter's massive data volume, enabling efficient management, storage, and processing of tweets from a growing user base. This strategic move optimizes Twitter's development potential by providing a robust and scalable solution for handling the vast amounts of data generated on the platform.

The method of using Hadoop for this purpose is complex and diverse. The massive repository of tweet data had to be stored and then effectively accessed and retrieved from the Hadoop cluster. Given the constantly growing amount of Twitter's data stream, this was a challenge in and of itself. However, by adopting the Hadoop technique, Twitter may begin on the task of data mining, sorting through the platform's massive and, frequently, unstructured data. This data mining method could yield significant insights, identifying patterns, sentiments, and information that could be used to propel Twitter's growth and improve the user experience [5].

Implementing Hadoop for Twitter's data management involves a complex and diverse process. The challenge lies in storing and efficiently retrieving the vast repository of tweet data within the dynamic Hadoop cluster. Despite the constant growth of Twitter's data stream, Hadoop facilitates data mining, enabling the platform to analyze large, often unstructured datasets. This data mining approach has the potential to uncover valuable insights, including patterns, sentiments, and information that can drive Twitter's growth and enhance the overall user experience. The adoption of Hadoop thus positions Twitter to leverage its data effectively for strategic development and optimization

Hadoop comprises two essential elements: the Hadoop Distributed File System (HDFS) and the workflow of the MapReduce algorithm[8]. Twitter's extensive data, including Tweets, will be stored within HDFS. Once HDFS accumulates a substantial amount of Twitter data, MapReduce will collect and transmit it to the necessary tools in the form of key-value pairs, which represent the specific task at hand [7]. Twitter employs Hadoop, consisting of HDFS and the MapReduce algorithm, to manage its extensive data. Tweets are stored in HDFS, and when a substantial amount accumulates, MapReduce efficiently processes and transmits the data as key-value pairs. This integration optimizes data storage, retrieval, and processing, showcasing the collaborative strength of Hadoop's components in handling Twitter's vast and dynamic dataset.

Before analysis we have to collect twitter data on which we can perform analysis, for collecting twitter data we have to create a twitter app through which we can generate a consumer and token access keys through which we can fetch real-time twitter data using flume [9].Apache flume is integrated over Hadoop which can fetch data from source which is a twitter data and through memory channel workflow stored it into the HDFS which is a sink in our flume configuration [9].

After  generating a  consumer and  access token  keys  we can configure flume on which we can configure web sources from where the data is coming in our paper it’s a twitter and intermediate channels which  provide  workflow  and than configure  a  sink from  which  the  data  is  stored and  that  is HDFS in our paper [9]. And the consumer and access token keys are also written in configuration file [9]. To collect Twitter data for analysis, a Twitter app is created to generate consumer and access token keys. Apache Flume is then integrated with Hadoop for data processing. Using these keys, Flume is configured with web sources, like Twitter, as the data source. The workflow involves intermediate channels for data flow, and the configuration includes a sink, HDFS in this case, for storing the collected data. Hence, this comprehensive setup enables the real-time collection and storage of Twitter data, facilitating subsequent analysis within the Hadoop ecosystem

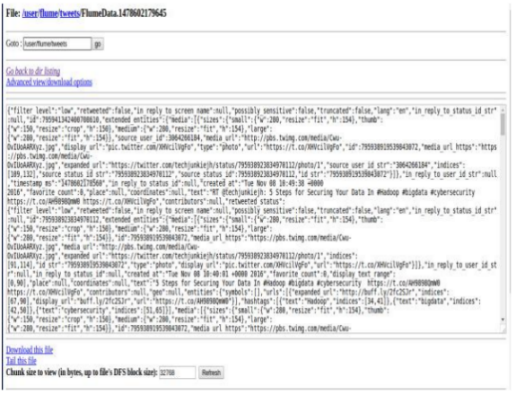


Figure 1

After storing all twitter data into the HDFS we are performing the analysis part for these we use hive through which we can convert the un-structure complex data into readable or understandable structure form [9]. The JSON data comes from the twitter shown in figure 1 are in the unstructured form, so we need a powerful analytical tool to analyze this type of data, so we can use apache hive [9]. Tweets are preprocesses for removing noise and meaningless symbols. And then the data is available in the form of schema oriented , and using hive we analyze the data by writing different queries for decision making [9]. After storing Twitter data in HDFS, the analysis phase involves using Apache Hive to convert the unstructured JSON data into a readable and structured format. Given that Twitter data is inherently complex and unstructured, Hive serves as a powerful analytical tool. The tweets undergo preprocessing to remove noise and meaningless symbols, resulting in schema-oriented data. Utilizing Hive, various queries are written to analyze the data, facilitating decision-making processes. This approach streamlines the analysis of Twitter data, making it more accessible and insightful through the structured format achieved with the help of Hive

# ETL Process

HDFS, MapReduce, HBase, YARN, Pig, Hive, and other components involved in the Hadoop Ecosystem. For distributed storage and computation, it employs a master/slave architecture. The NameNode serves as the storage master, while the DataNodes are the subordinate nodes. In the world of distributed computation, the Job Tracker serves as the master, while Task Trackers serve as the servants. These subordinates are in charge of data storage and complex calculations. Each subordinate node is outfitted with a Task Tracker daemon that communicates with the Job Tracker and NameNode. Hadoop's master and subordinate systems may be implemented on-premises or in the cloud, allowing for distributed storage and processing [11]. The ETL process was also implemented in the Hadoop Ecosystem. There are some big data tooling that are involved in the ETL process of the Hadoop Ecosystem. Figure 2 shows the list of Hadoop Ecosystem Components and Architecture.

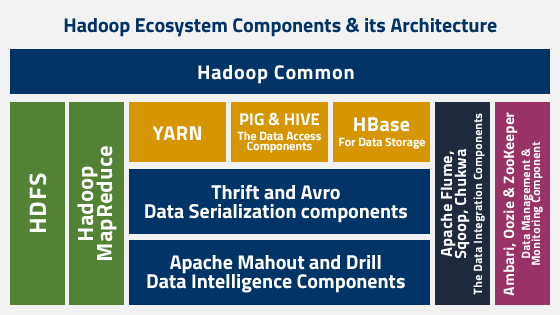


Figure 2

## Extract

### The Twitter Data

The raw data of Twitter which is tweets will extract the data into the HBASE or Hadoop Distributed File System (HDFS) that were data storage techniques for storing data into the file system[10]. This extract process also involves retrieving data which it utilizes by MapReduce where it will  distribute into tuples that represent as key-value pairs for the given task to the big data tools that are involved in the solutions[5].

### NoSQL database

For doing twitter data analysis first data is uploaded using FLUME in local HDFS [9]. The twitter API used in Flume , through which all the tweets are directly fetch from the twitter site and stored it into the HDFS [9]. While the article does not explicitly mention the use of a NoSQL database, the mention of Hadoop suggests the use of distributed storage systems, and Hadoop often works with NoSQL databases for certain types of data processing.

## Transform

### Programming Tool

The JSON data coming from the twitter shown in figure 5 are in the unstructured form, so we need a powerful analytical tool to analyse these type of data, so we can use apache hive which is a open source data warehouse analytical too [9]. So we can integrate hive with hadoop to analyse the data and along with configure a hive web interface shown in figure 6 in which we connect hadoop with hive web interface [9]. Apache Hive is used for analyzing Twitter data, which involves transforming the raw data into a structured and analyzable format. Hive uses a SQL-like language called HiveQL for querying and analyzing data stored in Hadoop.

### Business intelligence

The article discusses performing analysis on Twitter data to find the number of tweets posted location-wise and identifying keywords with maximum and minimum tweets. This falls under the realm of business intelligence, understanding and deriving insights from the data. We perform analysis on twitter data to find the number of tweets are posted location wise and also finds the keywords on which maximum and minimum tweets are Posted [9].

## Load

### Data technology

Apache Flume is mentioned as a tool for extracting Twitter data. Flume is often used for efficiently collecting and moving large amounts of streaming data. Hadoop and its Ecosystems, for getting raw data from the Social Network, we may use Hadoop online streaming toolusing Apache Flume [9]. By utilizing this tool only, we are going to configure everything, which we wanted to get (data) from the Social Network [9]. For doing twitter data analysis first data is uploaded using FLUME in local HDFS [9]. The twitter API used in Flume , through which all the tweets are directly fetch from the twitter site and stored it into the HDFS  [9].

### Cloud provider

While a specific cloud provider is not mentioned in the provided text, the use of Hadoop and the mention of Hadoop Distributed File System (HDFS) suggest that a distributed storage system is utilized. Hadoop is often deployed on cloud platforms like AWS, Azure, or Google Cloud for distributed processing. Hadoop was derived from Google File System (GFS) and Google's Map Reduce [9].

### Analytics & visualization

Although not explicitly mentioned, the use of Apache Hive for analysis implies some level of analytics. Visualization tools may also be employed, but specific tools like Tableau or Power BI are not mentioned. It is taking sentiment analysis, for this it is using Hive and its queries to give the sentiment data based up on the groups that have defined in the HQL (Hive Query Language) [9]. Flume is used to fetching real time twitter data and stored in HDFS and after the data storage we are performing analysis of these complex data using hive [9]. After storing all twitter data into the HDFS we are performing the analysis part for these we use hive through which we can convert the un-structure complex data in to readable or understandable structure form [9].

# Demostrate ETL Process

## Extract

In first step We are creating a twitter app using a twitter streaming API for fetching real time twitter data [9].  for collecting twitter data we have to create a twitter app through which we can generate a consumer and token access keys through which we can fetch real time twiiter data using flume [9].  Figure 3 in the article illustrates the generation of consumer and access token keys through the Twitter application. These keys serve as authentication credentials, allowing Apache Flume to access and fetch real-time Twitter data. This initial step is crucial for setting up the extraction process and ensuring a secure and authorized connection between the data source (Twitter) and the data processing tool (Flume) in the subsequent steps of the ETL process.

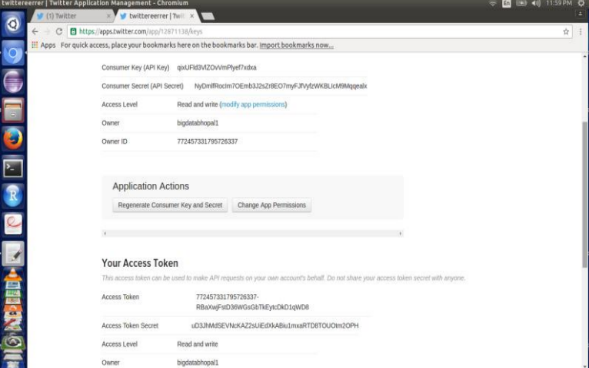


Figure 3

After generating a consumer and access token keys we can configure flume on which we can configure web sources from where the data is coming in our paper it‟s a twitter and intermediate channels which provide workflow and than configure a sink from which the data is stored and that is HDFS in our paper [9]. After obtaining the necessary Twitter authentication keys, the configuration of Apache Flume involves setting up web sources to fetch data from Twitter, configuring intermediate channels for a structured workflow, and specifying a sink to store the data in HDFS. This configuration establishes a seamless data pipeline for the extraction and storage of real-time Twitter data.



Figure 4

After configuring the flume-twitter.conf file we can start a twiiter agent through terminal which can a twitter API which provide a conncetion establishment between HDFS and twitter and the tweets are start downloading from the twitter and stored it into the the HDFS. Figure 4 shows the twitter JSON data coming from the twitter and stored into HDFS.

## Transform

Once the Extract procedure is complete, we will begin the transform step, which is where the analysis of Twitter JSON Data began. Figure 4 displays the JSON data received from Twitter. We will require an intelligent analytical tool to analyze Twitter data using Apache Hive. We will connect Hive with Hadoop in order to analyze data and configure the Hive web interface. Figure 5 shows that we have already connected Hadoop to the Hive web interface [9].

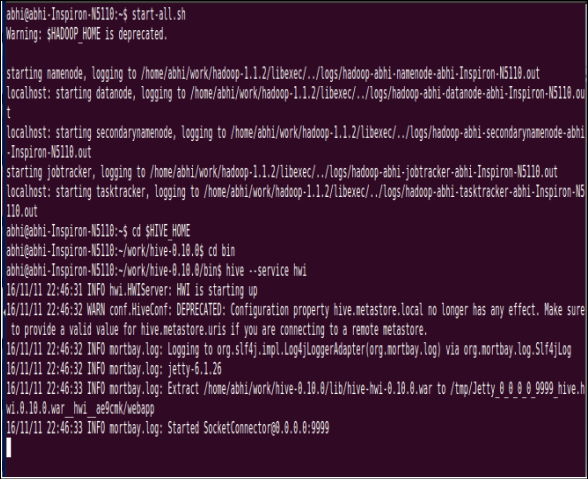


Figure 5

After connecting the Hadoop and Hive web interfaces (HWI), we will open the HWI using the localhost:9999/hwi url displayed in Figure 6. Following that, we will setup the session in order to do the analysis. We cannot do the analysis without first creating a session in the HWI. Once the session has been properly created, we can build the hive table to contain the twitter JSON data [9]. For a better understanding, we may use the json serde to convert unstructured data into structured data. Figure 7 illustrates that we have already established the tweets\_raw hive table [9].

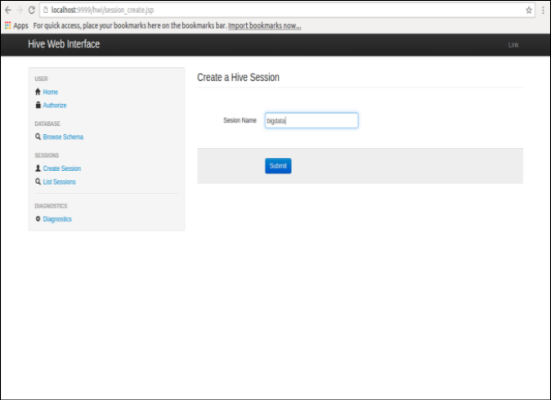


Figure 6

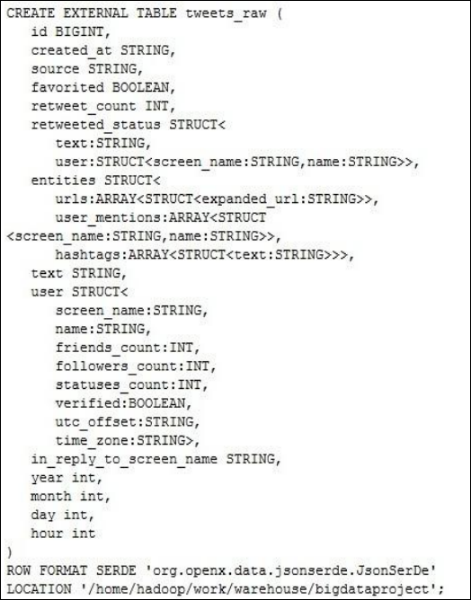


Figure 7

We will load the data inpath '/user/flume/tweets/' overwrite into table tweets\_raw; once we have created the table. Figure 8 shows how data is stored in an organized manner. We can now execute the analysis after the JSON data has been loaded into the hive table. We must write SQL queries for analyzing the data using the session that we have built in figure 6. We must fill in the session details, which contain the result file of the place where the analysis result will be stored as a text file, as shown in Figure 9. In order to analyze the query field, we must construct the SQL query code

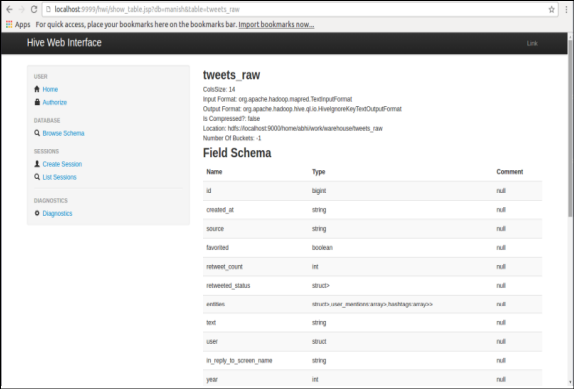


Figure 8

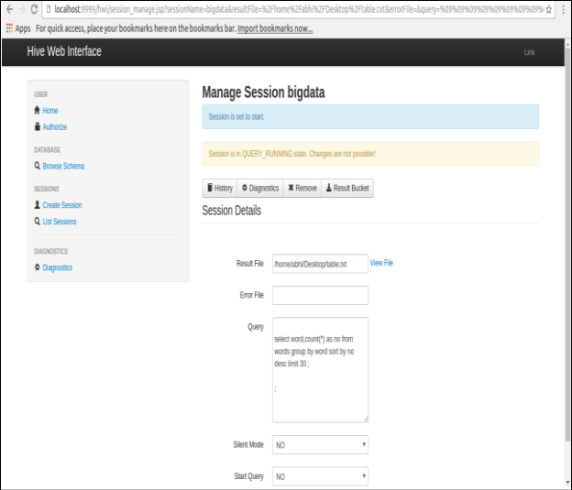


Figure 9

Once the SQL query on the HWI runs successfully, the MapReduce job will be launched for the hive query on the backend of the interface which is shown in the figure 10. It showed that analyze in transform process is successful.

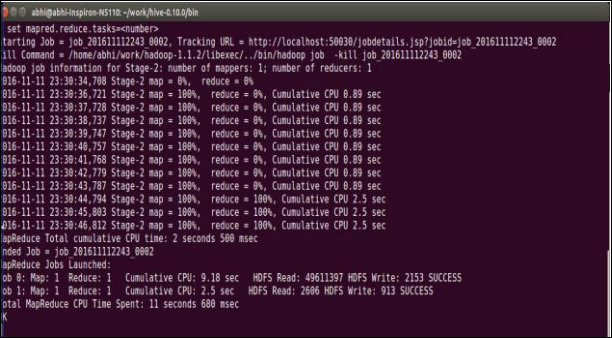


Figure 10

## Load

Once the Transform process is completed successfully, the query will be stored as a text file, as seen in the session detail in Figure 9. Figure 11 shows that the query result is successfully store in the hive web interface.

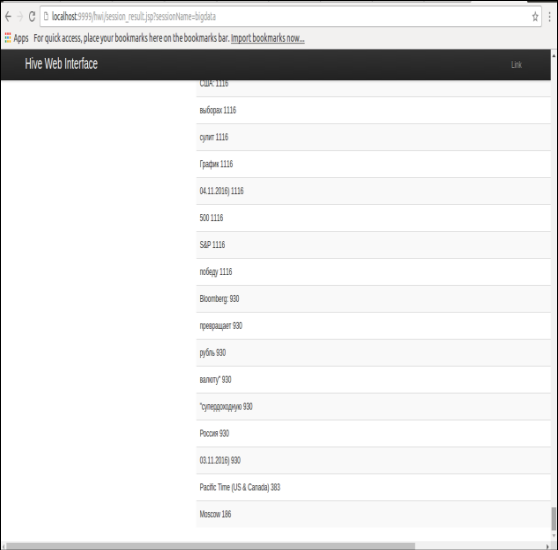


Figure 11

# Conclusion

Upon thoroughly examining the process of analyzing social data, it becomes evident that traditional analytical tools are insufficient for handling extensive and intricate datasets. To address this challenge, a robust solution has been adopted in the form of Hadoop, a powerful tool designed for in-depth analysis. Integration with complementary tools like Flume and Hive, both operating atop Hadoop, enhances the analytical capabilities.

In this integrated ecosystem, Flume plays a crucial role in fetching data, particularly from social media sources like Twitter, and efficiently storing it in HDFS. Subsequently, Hive is employed for the comprehensive analysis of these vast and intricate datasets. The focus if the analysis, as outlined in the article, includes determining the geographical distribution of tweets and identifying keywords associated with both maximum and minimum tweet frequencies.

Looking ahead, there is a prospect for further enhancements in the analysis for this data, future iterations of the study could explore additional parameters to provide richer insights, including to more informed decision-making process.

# References

[1] Euronews, “The social media platform’s billionaire owner Elon Musk had said the change would give news articles a better look.,” Oct. 05, 2023. Accessed: Nov. 06, 2023. [Online]. Available: https://www.euronews.com/next/2023/10/05/musks-x-formerly-twitter-has-changed-how-it-displays-news-articles-by-scrapping-headlines#:~:text=X%20

[2] A. Barskar and A. Phulre, “Opinion Mining of Twitter Data using Hadoop and Apache Pig,” *Int J Comput Appl*, vol. 158, no. 9, 2017, doi: 10.5120/ijca2017912854.

[3] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The Hadoop distributed file system,” in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010*, 2010. doi: 10.1109/MSST.2010.5496972.