

CAPSTONE PROJECT REPORT

(Project Term January-May 2021)

(Real Time Data Pipeline for Twitter Trends Analysis)

Submitted by

(MUKESH KUMAR SAH)

Registration Number : 11719859

(RISHABH SHARMA)

Registration Number : 11713811

(SHREYANS PRIYAM GUPTA)

Registration Number : 11705384

Project Group Number : KC036

Course Code : CSE445

Under the Guidance of

(AMRITPAL SINGH : 17673, ASSISTANT PROFESSOR)

School of Computer Science and Engineering



L OVELY
P ROFESSIONAL
U NIVERSITY

PAC FORM



TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering (SCSE)

Program : P132::B.Tech. (Computer Science & Engineering)

COURSE CODE : CSE445

REGULAR/BACKLOG : Regular

GROUP NUMBER : CSERGC0036

Supervisor Name : Amritpal Singh

UID : 17673

Designation : Assistant Professor

Qualification : _____

Research Experience : _____

SR.NO.	NAME OF STUDENT	Prov. Regd. No.	BATCH	SECTION	CONTACT NUMBER
1	Mukesh Kumar Sah	11719859	2017	K17RZ	9876074234
2	Rishabh Sharma	11713811	2017	K17QN	8290037461
3	Shreyans Priyam Gupta	11705384	2017	K17RZ	8233123101

SPECIALIZATION AREA : Database Systems

Supervisor Signature: _____

PROPOSED TOPIC : Real-Time Data Pipeline for Twitter Trends Analysis

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.72
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	8.44
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.72
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	8.72
5	Social Applicability: Project work intends to solve a practical problem.	7.72
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.56

PAC Committee Members		
PAC Member (HOD/Chairperson) Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member (Allied) Name: Gurpreet Singh	UID: 17671	Recommended (Y/N): Yes
PAC Member 3 Name: Savleen Kaur	UID: 18306	Recommended (Y/N): Yes

Final Topic Approved by PAC: Real-Time Data Pipeline for Twitter Trends Analysis

Overall Remarks: Approved

PAC CHAIRPERSON Name: 14770::Sawal Tandon

Approval Date: 12 Mar 2021

4/26/2021 8:24:16 PM

DECLARATION

We hereby declare that the project work entitled (“Real-Time Data Pipeline for Twitter Trends Analysis”) is an authentic record of our own work carried out as requirements of Capstone Project for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of Amritpal Singh, during January to May 2021. All the information furnished in this capstone project report is based on our own intensive work and is genuine.

Project Group Number : KC036

Name of Student 1: MUKESH KUMAR SAH

Registration Number: 11719859

Name of Student 2: RISHABH SHARMA

Registration Number: 11713811

Name of Student 3: SHREYANS PRIYAM GUPTA

Registration Number: 11705384

(Signature of Student 1)

Date: 27/04/2021

(Signature of Student 2)

Date: 27/04/2021

(Signature of Student 3)

Date: 27/04/2021

CERTIFICATE

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Capstone Project under my guidance and supervision. The present work is the result of their original investigation, effort, and study. No part of the work has ever been submitted for any other degree at any University. The Capstone Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, Punjab.

Signature and Name of the Mentor

AMRITPAL SINGH

Designation

ASSISTANT PROFESSOR

School of Computer Science and Engineering,

Lovely Professional University,

Phagwara, Punjab.

Date : 27/04/2021

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our mentor, Mr. Amritpal Singh for the guidance and support for our capstone project.

Name of Student 1: MUKESH KUMAR SAH

Registration Number: 11719859

Name of Student 2: RISHABH SHARMA

Registration Number: 11713811

Name of Student 3: SHREYANS PRIYAM GUPTA

Registration Number: 11705384

Place: Lovely Professional University

Date: 27/04/2021

TABLE OF CONTENTS

Cover Page.....	1
PAC Form.....	2
Declaration.....	3
Certificate.....	4
Acknowledgement.....	5
Table of Contents.....	6

1. INTRODUCTION
2. PROFILE OF THE PROBLEM. SCOPE OF THE STUDY (PROBLEM STATEMENT)
3. EXISTING SYSTEM
4. PROBLEM ANALYSIS
5. SOFTWARE REQUIREMENT ANALYSIS
6. DESIGN
7. TESTING
8. IMPLEMENTATION
9. PROJECT LEGACY
10. USER MANUAL
11. SOURCE CODE/ SYSTEM SNAPSHOTS
12. BIBLIOGRAPHY

1. 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 requires 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 make Big data non-identical from regular 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 large-scale parallel data processing, to build a real-time processing pipeline using open-source tools that can capture large amounts of data from numerous data sources, process, store, and analyze the large-scale data efficiently. 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 PHP dashboard through Apache Server and Finally visualization is done through PHP Web UI Dashboard.

2. SCOPE OF THE STUDY (PROBLEM STATEMENT)

The scope of the proposed project work intends to fill is to analyze the real-time unstructured data, which is growing exponentially with time, none of traditional data management tools can store it or process it efficiently. In today's world, Big data tools, frameworks and analytics tools helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

3. EXISTING SYSTEM

3.1. Introduction

In today's world, Data is growing exponentially and most of the data is in unstructured form, which means unstructured data cannot be stored in the traditional row or column form or in the tabular form. Unstructured data is everywhere. Examples of unstructured data are, text files, audios, videos, photos, webpages, and many other business documents. Most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated. Sources for unstructured data be, from Social Media Data, Satellite images, Web Content, IOT Data, Sensors Data, Mobile Data etc. Therefore, to process these types of data we have Big Data tools and Frameworks which can easily store and process the unstructured data.

3.2. Existing Software

In our project, we have chosen twitter social media as our source. As we know tweets are in unstructured form, so we have used Big Data tools which are as follows-

- Apache Flume
- Apache Kafka
- Apache Spark, RDD and DataFrames

3.3. DFD for present system

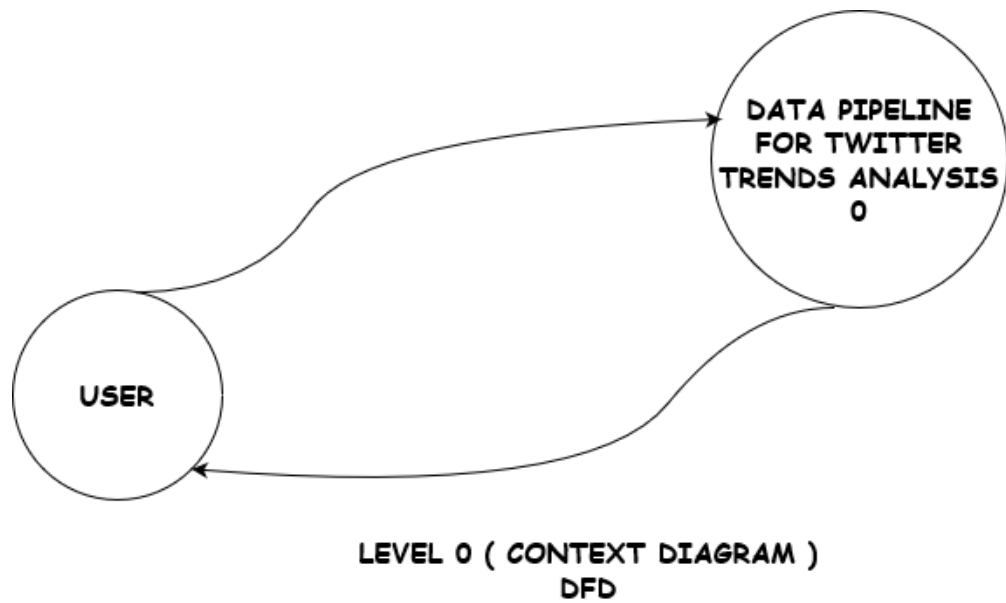


Fig 3.1: Level 0 (Context Diagram) DFD

- Level 1 DFD

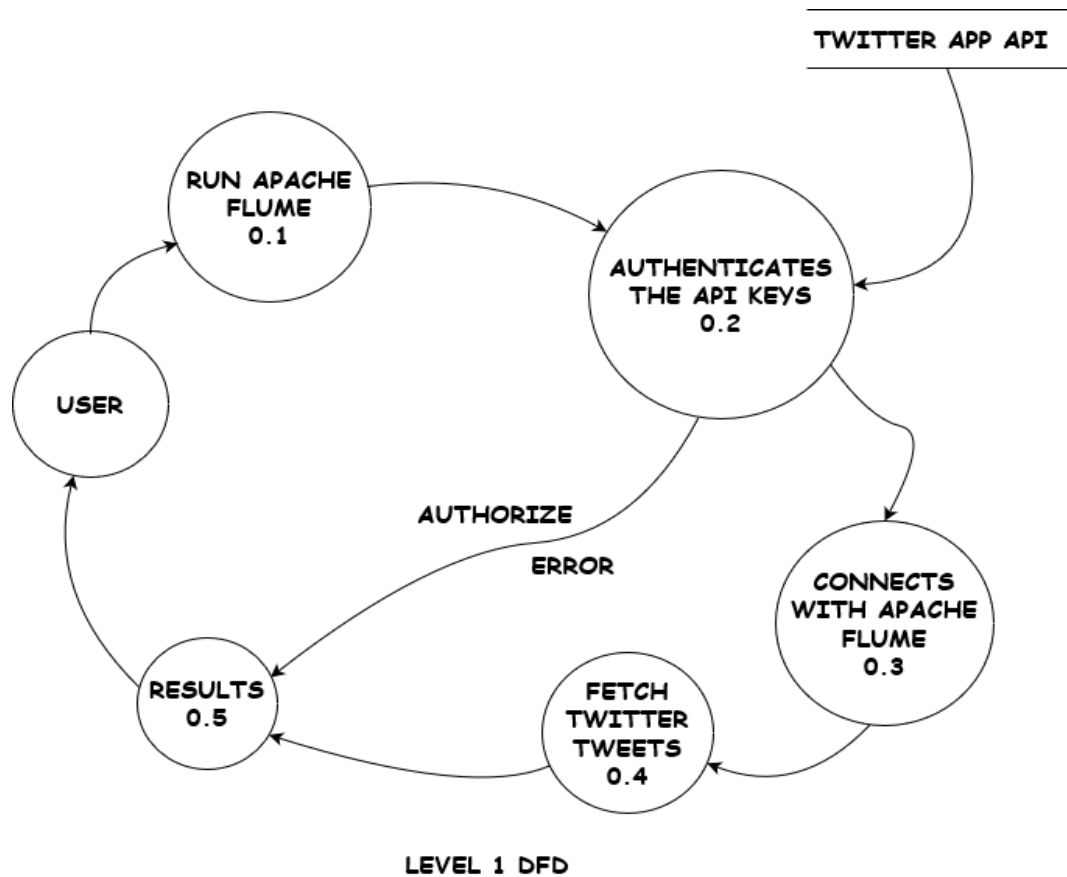
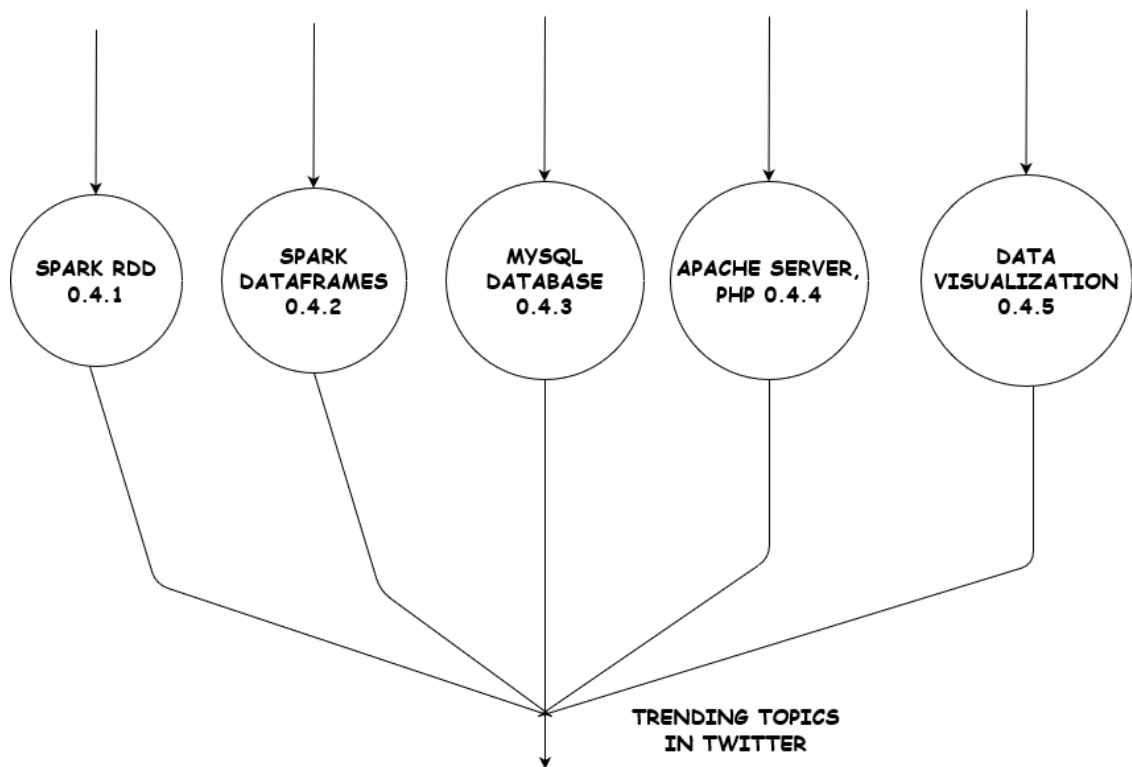


Fig 3.2: Level 1 DFD

- **Level 2 DFD**



LEVEL 2 DFD

Fig 3.3: Level 2 DFD

3.4. What's new in the System to be Developed.

The new in the system developed was to decrease the latency using big data tools and frameworks, processing the raw data more efficiently and storing our results in the database.

4. PROBLEM ANALYSIS

4.1. Product definition

Twitter streaming trends popularity 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 Kafka as a data ingestion system, Apache Spark as a real-time data processing system, MySQL for large datasets and retrieval, and MySQL with PHP through Apache Server for trending topics and real-time analytics.

The Twitter data is obtained by using Twitter Streaming API and is streamed to Apache Flume and Kafka which makes it available for Spark that performs data processing and sentiment classification and stores the results into MySQL. The popularity and sentiment of the trends are explored through an Apache Server and PHP live dashboard.

4.2. Tools and IDEs Used

- Twitter App API
- Apache Flume
- Apache Kafka
- Apache Spark
- Spark RDD and Spark DataFrames
- Java 8
- Scala Programming Language
- MySQL Database
- Apache Server
- PHP

4.3. Feasibility Analysis

A feasibility study is an analysis that takes all a project's relevant factors into account, including economic, technical, legal, and scheduling considerations to ascertain the likelihood of completing the project successfully. Project members use feasibility studies to know about the pros and cons of undertaking a project before they invest a lot of time and money into it.

Feasibility studies also can provide a company's management with crucial information that could prevent the company from entering blindly into risky businesses.

4.3.1. Technical Feasibility

Evaluating the 'Technical Feasibility' is the most tangled bit of plausibility. This occurs because, presently of time, not a lot of bare essential structures of the system, makes it confounded to get the issues like execution, costs us on, etc. Bundle of issues must be considered while doing a specific assessment. Grasp different developments related with the proposed system before starting the errand we should be clear about the advances that must be required for the improvement of the new structure. Check whether the relationship starting at now has the essential advances. Is the important development available with the affiliation?

4.3.2. Operational Feasibility

Our made errand is important just if it will in general be changed into information systems that will meet the affiliation's essentials. In a simpler decree, this preliminary of believability asks whether the structure will work when it is made and presented. Are there any immense limits to the execution? Here is the overview of specific request that helps in testing the operational reachability of an errand. Is there enough help for the undertaking from the executives from clients? On the off chance that the present framework is popular and used to the degree that people won't have the option to see purposes behind change.

- Are the present Business strategies worthy to client? In the event in which they are not, at that point clients may respect the change that realizes a progressively operational and helpful frameworks .

- Is there enough help for the project to the board from clients? On the off chances that the current framework is popular and uses the degree that people will not be see purposes behind change.
- Does the user involve in planning and development of the project?

Early inclusion diminishes the odds of the protection from system when everything is said in done and improving the probability of effective task.

Since the proposed system will help us to reduce the hardships that are encountered. In the current manual system, the naive system was operationally feasible.

4.3.3.Economic Feasibility

‘Economic Feasibility’ endeavors the gauge the expenditure of creation and actualizing another framework, against the advantages that would collect from having the new framework set up. This plausibility study gives the top administration financial legitimization for the new framework. A basic monetary examination which gives the genuine correlation of expenses and advantages are significantly more important for this situation. What's more, this demonstrates to be valuable perspective to think about real expenses as the undertaking advances. There could be different kinds of immaterial advantages because of mechanization. They may incorporate expanded consumer loyalty, improved precision of tasks, Advancement in the product quality, assisting exercises, increase view in basic leadership, practicality of the data, better documentation and record keeping, better representative assurance and quicker recovery if data.

Our Proposed project was technically and economically feasible. All our group members took and analyzed the factors and resources needed for this project.

4.4. Project Plan

The project took 4 months to complete; we have started doing this project on January 6th, 2021 and we finished the project on April 30th. Overall project was implemented module by module.

MODULES	START DATE	DAYS TO COMPLETE
Twitter App API	6-Jan	15
Apache Flume, Apache Kafka	21-Feb	21
Apache Spark, RDD, DataFrames	12-Mar	20
MySQL Database	2-Apr	8
Apache Server, PHP	10-Apr	20

Table: 4.1: Project Plan

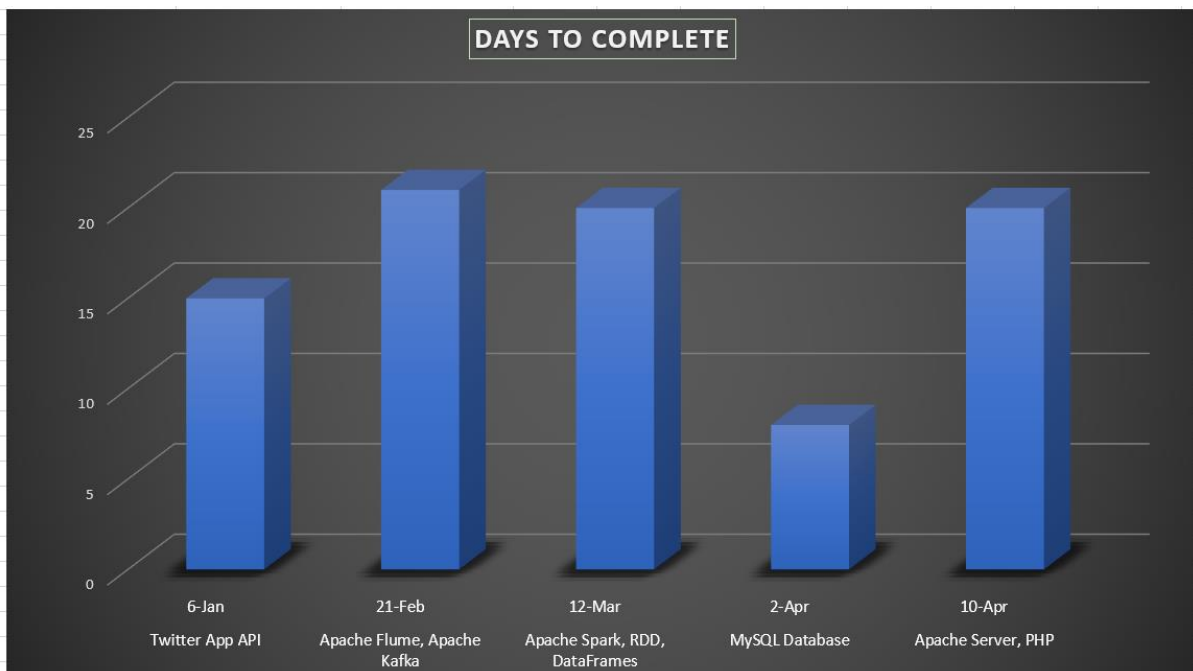


Fig 4.1 Gantt Chart

5. SOFTWARE REQUIREMENT ANALYSIS

5.1. Introduction

S.N.	Description	Comments
1	Users	Any user who has big data tools installed in their system can run this project easily.
2	Functional Requirements	<ul style="list-style-type: none">A. A module for user to get Twitter API keysB. A module to run Apache Flume and receive the tweets from the twitter.C. A module to send the incoming tweets of apache flume to apache Kafka broker.D. A module to store those incoming tweets in our local file system.E. A module to do sentiment analysis on the tweets.F. A module to store the results in the MySQL table.G. A module to integrate MySQL table and PHP using apache server.
3	Reporting Requirements	Users can be anyone who have basic knowledge of big data.
4	Security Requirements	The access to the system is very secure that no unauthorized person can access it. Overall project is secured and confidential.

5	Non-Functional Requirements	<p>A. Apache Spark uses Spark RDD and DataFrames which is more efficient and provides more accuracy to the results.</p> <p>B. Well-designed module which makes our system user-friendly.</p> <p>C. Enormous capacity to store the data and can be accessed easily.</p> <p>D. Well-designed performance with good functionalities</p>
---	-----------------------------	--

Table 5.1: SRS Requirements

5.2. General Description

1st we will get the Twitter APP API keys, then we fetch the data using Apache Flume in our system. After that we will integrate Apache Flume with Apache Kafka. After Getting the tweets data, we apply sentiment analysis on the tweets using spark RDD and DataFrames. Then, we will integrate MySQL with PHP using Apache Server to create a live Dashboard to get trending topics in twitter.

5.3. Specific Requirements

- Twitter App API Keys
- Apache Flume
- Apache Kafka
- Apache Spark
- Java 8
- Scala Programming Language
- MySQL Database
- Apache Server
- PHP

6. DESIGN

6.1. System Design

The various tools and frameworks of the big data, Apache Flume, Twitter App API Streaming, Apache Kafka, Apache Spark(RDD and DataFrames), MySQL, Apache Server and PHP were used and, all these were run locally for development.

6.1.1. 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.

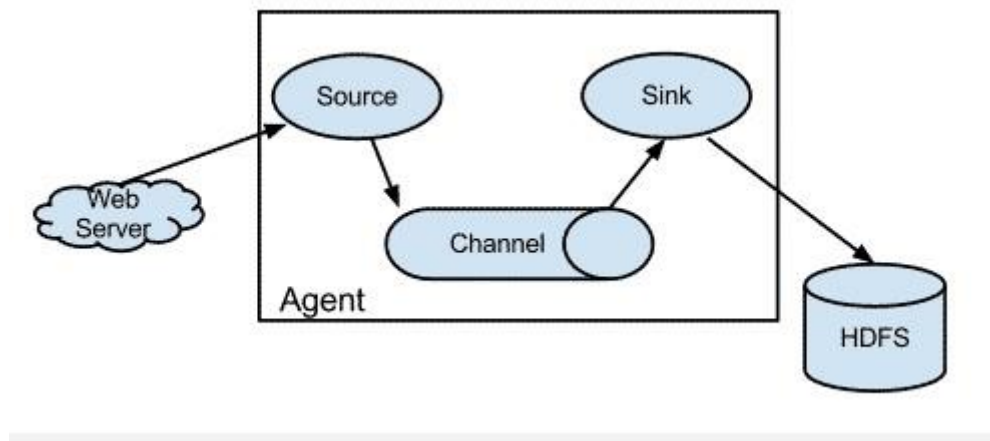


Fig 6.1: Data Flow Model in Apache Flume

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.

6.1.2. 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.

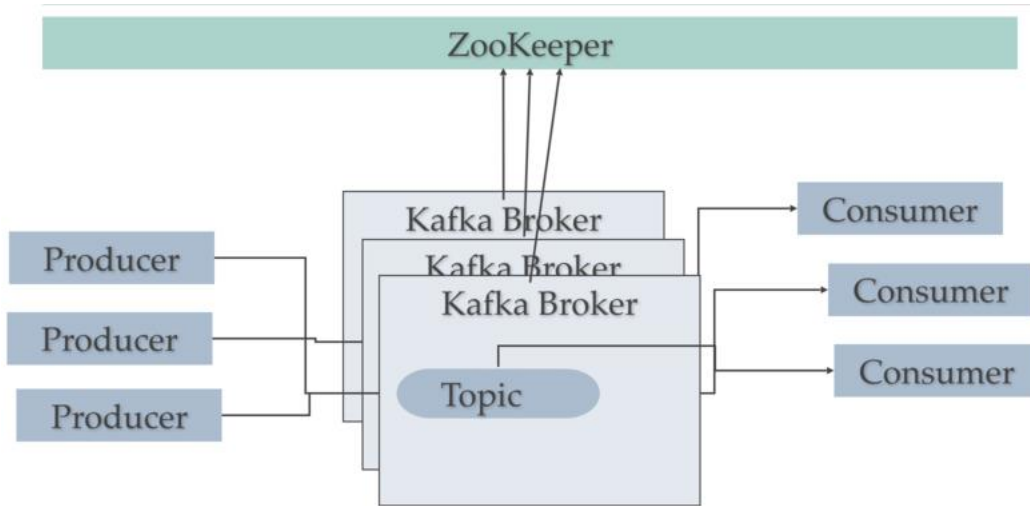


Fig 6.2: Kafka Architecture

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.

6.1.3. Apache Spark

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



Fig 6.3: Data Flow Model in Apache Spark

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 trending topics in the twitter.

6.1.4. 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.

6.1.5. Apache Server and PHP

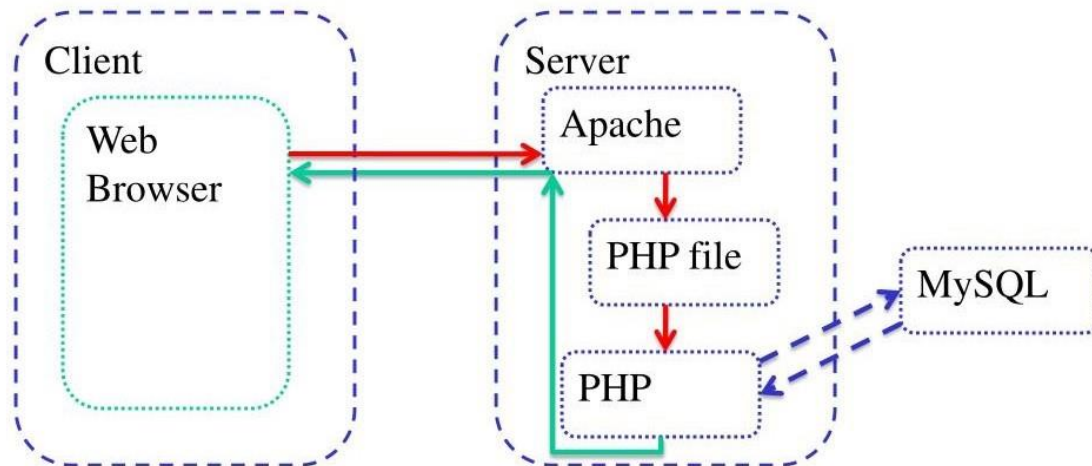


Fig 6.4: Data Flow Model in MySQL Database, Apache Server and PHP

After the results are stored in MySQL Database, the results of the sentiment analysis are fetched in PHP dashboard through apache server, and the live results of the trending topics and the trending HashTags of the twitter social media are shown in the WebUI.

6.2. Project Flowchart

Project was implemented by creating a data pipeline, which was followed by the following architecture-

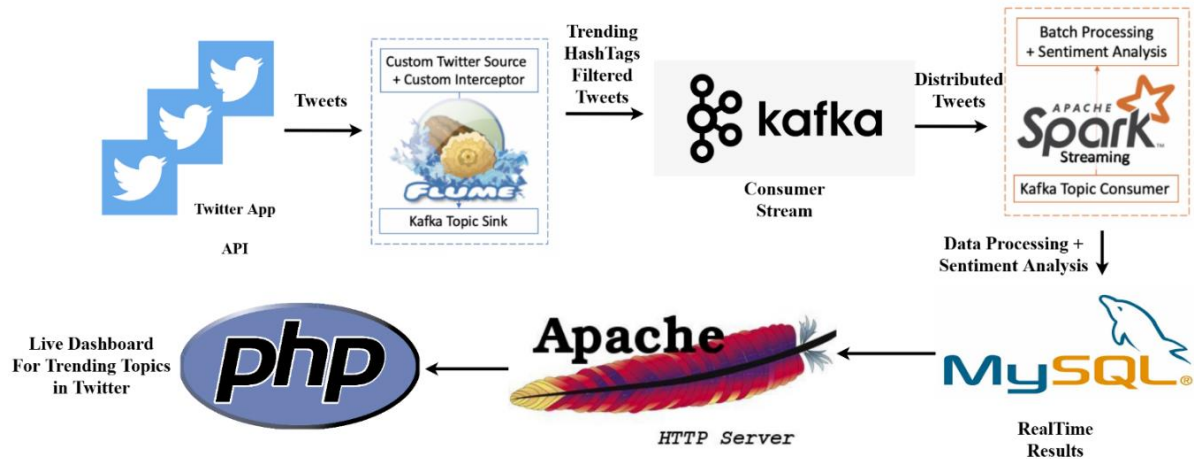


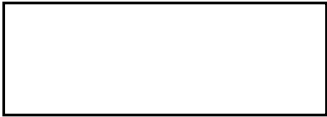

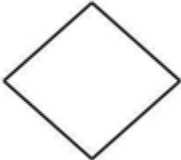
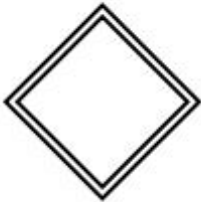
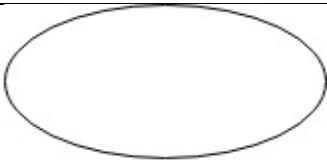
Fig 6.5: Overall Flow of Project

- Twitter App API is subscribed to Apache Flume and Twitter Streaming producer publishes streaming tweets to the 'twitterdata1' topic in an Apache Kafka broker.
- Then, Sentiment Analysis is done using Apache Spark RDD and DataFrames on the tweets which we receive from the 'twitterdata1' topic.
- The Spark engine performs batch processing on incoming tweets and performs sentiment classification before storing the processed results in the MySQL.
- After the results are stored in the MySQL database, PHP is connected through Apache Server, which creates a live dashboard to analyze popularity and sentiment of trending topics on Twitter.

6.3. Design Notations

Conceptual entity-relationship (ER) diagrams, conceptual data models form a broad perspective to be included in a model set. Rational ERD can be used as an installation for logical data models. Same way, they can be used to frame shared attribute connections between ER models for data-model integration.

ERD Entity Symbols- Units are articles or ideas that speak for essential information. The word entity is things in general. For Example, product, customer, region, or promotion. Three types of entities are commonly used as a part of entity-relationship diagrams.

Symbol	Name	Description
	Strong entity	These shapes are free from different entities and are regularly called parent entities since they will frequently have weak entities that rely upon them. They will likewise have a primary key, recognizing every event of the entity.
	Weak entity	Weak elements rely upon some other entity type. They don't have primary keys and have no importance in the chart without their parent element.
	Relationship	Association between different entities.
	Weak relationship	Have connections between a weak entity and it's an owner
	Attribute	Attributes are characteristics of an entity, a many-to-many relationship, or a one-to-one relationship.

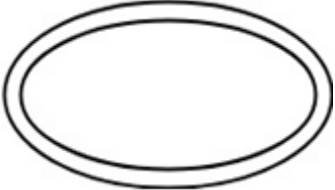

	<p>Multi-valued attribute</p>	<p>Values that can take more than one value as the name literally suggests.</p>
	<p>Derived attributes</p>	<p>Derived attributes are the attributes whose values can be calculated from the corresponding relying attributes.</p>

Table 6.1 : Description of ER entity symbols

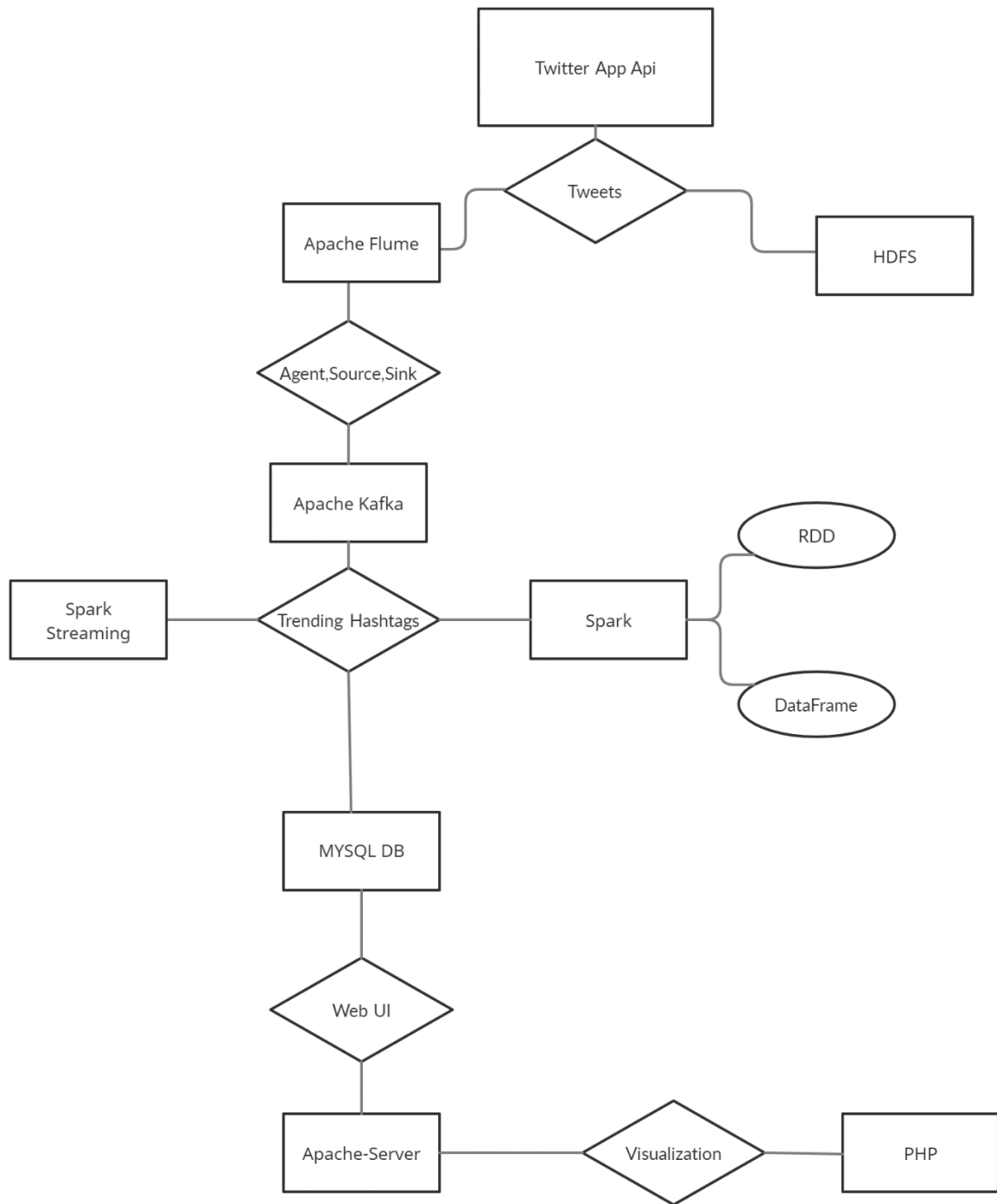


Fig 6.6 ER Diagram

6.4. Flowchart

The following are the notations used.


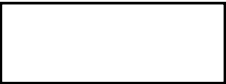
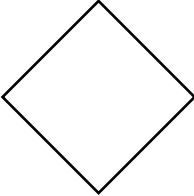

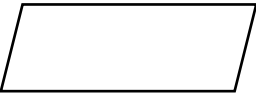
S.N.	Shapes	Name
1		Initial/ Terminal Symbol
2		Process
3		Decision Symbol
4		Data Flow
5		Data Object

Table 6.2: Flowchart shapes

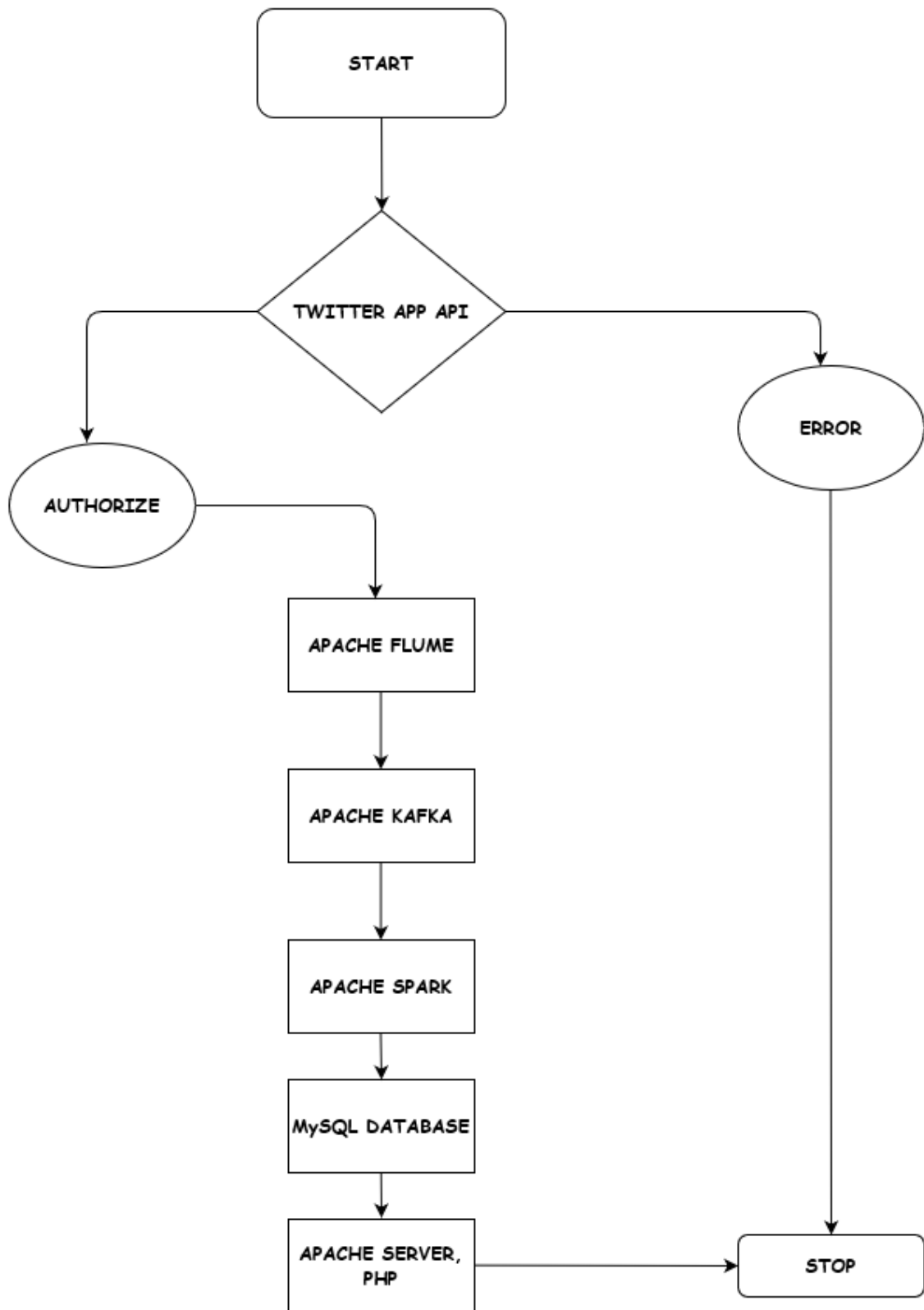


Fig 6.7: Flowchart

6.5. 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, Spark RDD and DataFrames-

- Initializing Spark Shell
- Loading the tweets to Spark RDD
- Applying Our Sentiment Analysis on Real-Time raw Tweets to get trending topics in twitter.

Initializing MySQL Database-

- Whatever trending results we get is finally stored in MySQL table.

Starting Apache Server-

- MySQL table is connected to PHP live dashboard using apache server.
- Apache server is initialized in the Shell, /etc/init.d/apache2 start.

Results in PHP Dashboard-

- Using PHP, Table data is fetched from MySQL table.
- And the trending topics in twitter is visualized in live dashboard.

7. TESTING

To evaluate the software application and its functionality we use software testing. We mainly do software testing to check whether the software has reached functional requirements which are mentioned in SRS and to check if any errors are there and rectify those errors in order to produce error-free products to customers. Software testing is a process to be executed for refinement of a program or application while searching for software bugs. It is the process that leads to verification and validation of the software to meet the business and technical requirements of the program or application.

7.1. Functional testing

As the name suggests itself, Functional testing is the one which we use for testing whether our system has met the functional requirements mentioned or not. It is done by giving input to system and verifying the output. It mainly concerns the output but not about the code. It involves Black Box Testing. It also explains the work done by system but not the functioning of the system. checking of User Interface, security, database etc. are involved in functional testing. Testing in this can be performed manually or automatically. It is the first type of testing that we perform. The stepwise procedure in functional testing is as,

- I. The identification of functions that the software is expected to perform.
- II. The creation of input data based on the function's specifications.
- III. The determination of output based on the function's specifications.
- IV. The execution of the test case.
- V. The comparison of actual and expected outputs.

7.2. Structural testing

Till now we have verified the inputs and outputs. Now we are concerned about the internal working of the system. The internal working of the system is verified in this testing. It can also be called as White-Box or Glass Box testing. It's more worried about how the system does than the functionality of the system.

The testers should have knowledge of internal implementations and the code of the internal system. They should be known to the complete working of the software and all of the ways in which code is been implemented. In 'White-Box Testing', an internal perspective of the system, as well as the programming skills, are being used in designing test cases. The Testing

Device will choose inputs to exercise paths through the code and determine the appropriate outputs. This is analogous to testing nodes in a circuit, e.g., in-circuit testing.

- I. Logic coverage statement
- II. Branch coverage
- III. Condition coverage
- IV. Dataflow coverage
- V. Path conditions and symbolic evaluation etc.

7.3. Levels of Testing

Various levels of testing help in identification of the areas which are left out, overlapped with any of other module or being repeated during the phase of the evolvment of life cycle. In development lifecycle models, there are different stages of the project which include requirement gathering and analysis, design, coding, or implementation, testing and deployment and hence, each phase goes through testing separately. Hence different levels of testing are as follows:

- A. **Unit Testing-** It is basic type of testing performed. Here the system is tested in units. It tests the system by dividing into parts and checking parts individually so that it can perform correctly without any errors and meeting the requirements for which it has been developed. It is a type of testing where each unit is tested separately as an independent. It needs to be done by the developers to make sure that their code is working fine and meet the user specifications. It tests each piece of code which includes classes, functions, interfaces, and procedures.
- B. **Component Testing-** In this testing, individual components are tested without combining them with other components. It can also be called module testing from the point of view of architectural. The one and the only differences among the unit and component testing id that in case of unit each piece of code is tested wherein component each module is tested which comprises many pieces of code together.
- C. **Integration testing-** Previously, we tested individual units and components, now we need to check the integrated one. In this testing, individual components are combined

and tested whether these components are working properly is not while working together. There are many ways in integration testing that can be used by testers for testing the system. Most of the time it is suggested to use bottom-up.

- Big-Bang Integration Testing
- Top-Down
- Bottoms-Up
- Functional incrementation

- D. **System Testing-** In System testing, the whole system is tested. The affinity of the application with the framework is tested.
- E. **Alpha-testing-** Alpha-testing is one that is to be done by the developers of the project. It is executed after all development processes.
- F. **Beta-testing-** Beta-testing is referred to as testing done by the customers of the project/product. It involves the acceptance of the customer of the developed project. It is done just before the launch of the product.

7.4. Testing the project

For this project, we have applied various levels of testing, in different system.

In all the system, our project worked very fine, we got the same results in all the testing phases.

8. IMPLEMENTATION

The implementation of the project was done by the group members timely. The topic of this project was chosen by our group members, but it was approved by our mentor Amritpal Sir and the project was able to implement by the guidance of our mentor and our teammates efforts. We have referred to the official documentation of all the big data frameworks and tools and, we have referred websites as mentioned in the references for the implementation of the project.

8.1. Implementation of the modules

The modules of this project were implemented step wise, i.e., one after another. 1st we got the Twitter APP API keys, then we implemented Flume in our system. After that we integrated Apache Flume with Apache Kafka. After Getting the tweets, we applied sentiment analysis on those tweets using spark RDD and DataFrames. Then, we integrated MySQL with PHP using Apache Server to create a live Dashboard to get trending topics in twitter.

9. PROJECT LEGACY

9.1. Current Status of the project

Currently, the project is working fine. We are able to get the tweets from twitter API. We are able to do sentiment analysis on those tweets. After processing the raw tweets, we are able to get the results and also, we are able to show the trending topics of twitter in our dashboard.

9.2. Remaining Areas of concern

The remaining areas of concern for this project is that complete pipeline of the data or we can say, automatic flow of data on streaming data was not created due to lack of time and the errors we got in this project. However, we were able to implement this project in some other ways.

9.3. Technical and Managerial lessons learnt.

From this project, we have learnt to apply big data tools and frameworks, we have also learnt the integration of various frameworks and tools, that are necessary in today's tech industry. From this project, we have got deep knowledge of all the tools and IDEs used for this project.

10. USER MANUAL FOR THE SOFTWARE DEVELOPED

- 1st we need to run Apache Flume and Twitter Streaming producer which publishes streaming tweets to the 'twitterdata1' topic in an Apache Kafka broker.
- Then we do Sentiment Analysis using Apache Spark RDD and DataFrames on the tweets which we receive from the 'twitterdata1' topic.
- The Spark engine performs batch processing on incoming tweets and performs sentiment classification before storing the processed results in the MySQL.
- After the results are stored in the MySQL database, PHP is connected through Apache Server, which creates a live dashboard to analyze popularity and sentiment of trending topics on Twitter.

11. SOURCE CODE (SYSTEM SNAPSHOTS)

- 1st, we created twitter developer account, after that we have configured our twitter app to get twitter API keys and those API keys, we used in Apache Flume.

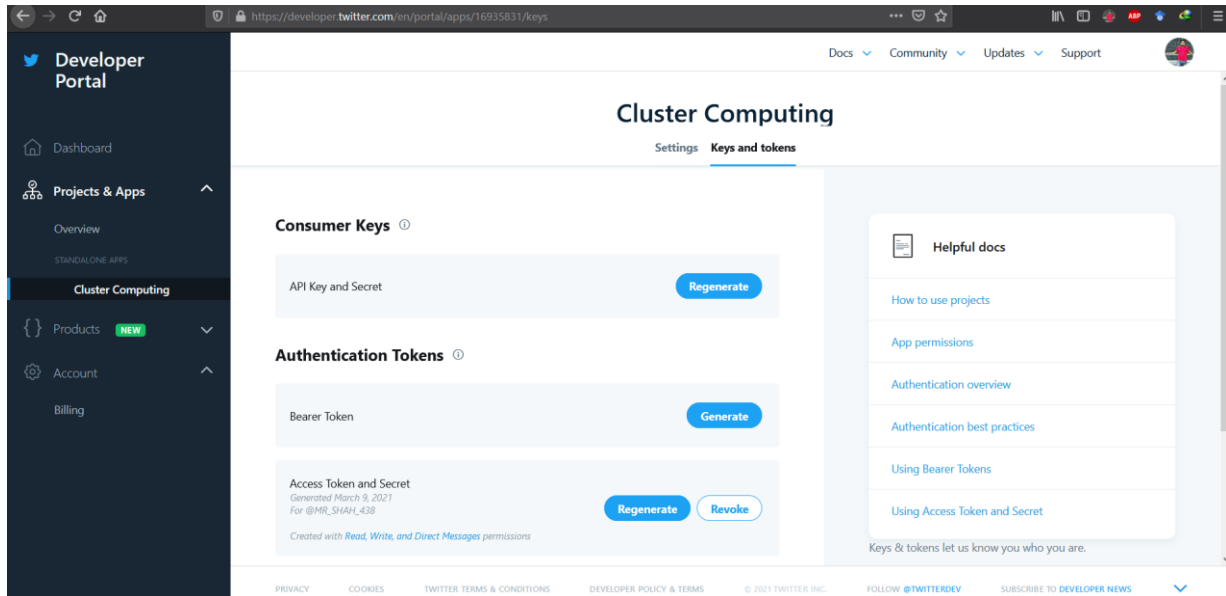


Fig 11.1 Twitter Developer Account for getting Twitter API keys

- Here, we have configured our Apache Flume conf file, we have taken source as Twitter Source, Channel as memory channel, sink as Apache Kafka Sink.

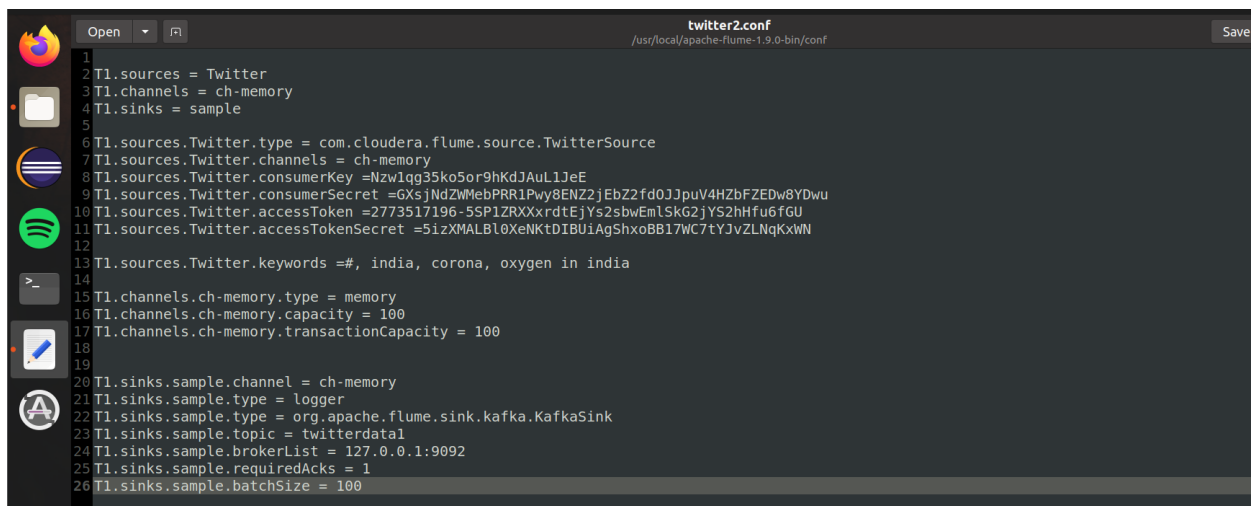


Fig 11.2: Conf file in Apache Flume

- Here, we have started Apache Zookeeper, so that we can run Apache Kafka because it works only with the coordination of Apache Zookeeper.

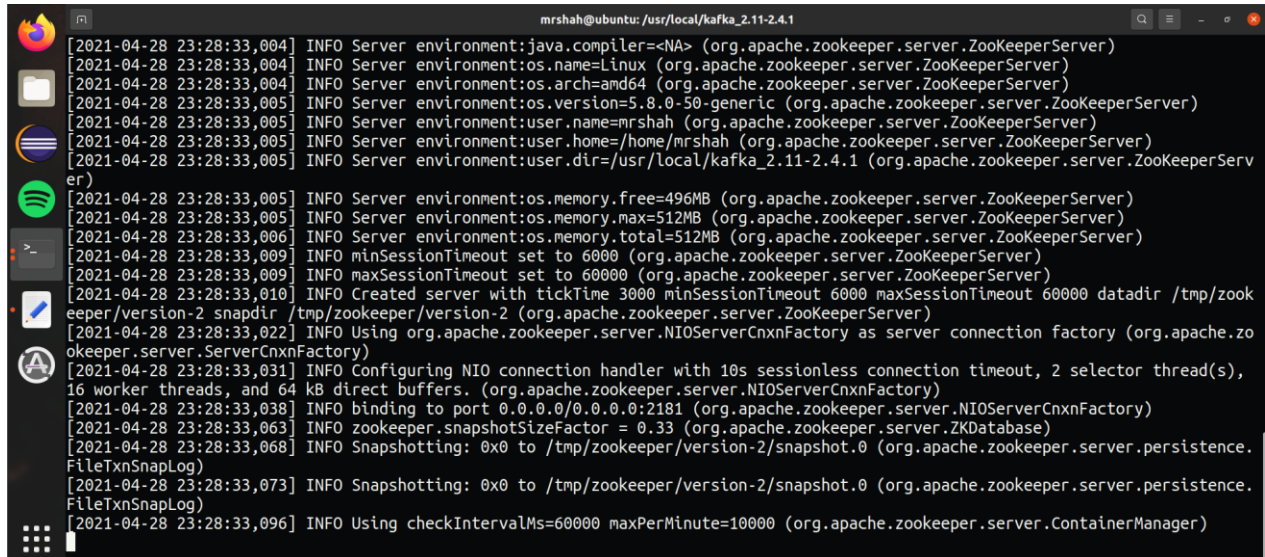
A terminal window titled 'mrshah@ubuntu: /usr/local/kafka_2.11-2.4.1' displays the logs for starting Apache Zookeeper. The logs show various environment variables being set, such as 'java.compiler', 'os.name', 'os.arch', 'os.version', 'user.name', 'user.home', and 'user.dir'. It also shows configuration details like 'minSessionTimeout', 'maxSessionTimeout', 'tickTime', 'snapshotSizeFactor', and 'snapshot'. The logs end with 'INFO Using checkIntervalMs=60000 maxPerMinute=10000 (org.apache.zookeeper.server.ContainerManager)'.

Fig 11.3: Starting Apache Zookeeper

- Here, we have started Apache Kafka server, all the incoming tweets to Apache Flume will ingested here only.

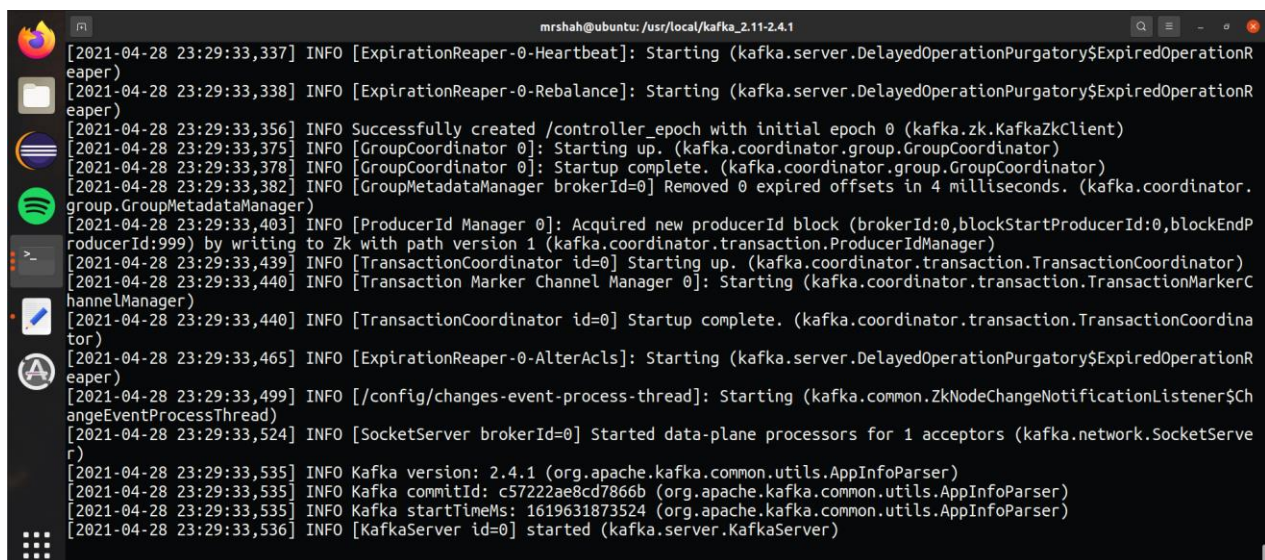
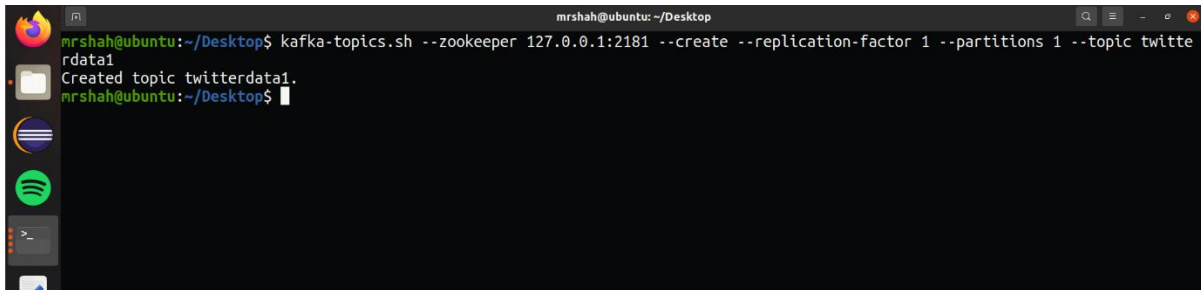
A terminal window titled 'mrshah@ubuntu: /usr/local/kafka_2.11-2.4.1' displays the logs for starting Apache Kafka. The logs show the startup of various components like 'ExpirationReaper-0-Heartbeat', 'ExpirationReaper-0-Rebalance', 'GroupCoordinator 0', 'GroupMetadataManager', 'ProducerId Manager', 'TransactionCoordinator', 'Transaction Marker Channel Manager', and 'SocketServer'. It also shows the Kafka version '2.4.1' and the commit ID 'c57222ae8cd7866b'. The logs end with 'INFO [KafkaServer id=0] started (kafka.server.KafkaServer)'.

Fig 11.4: Starting Apache Kafka

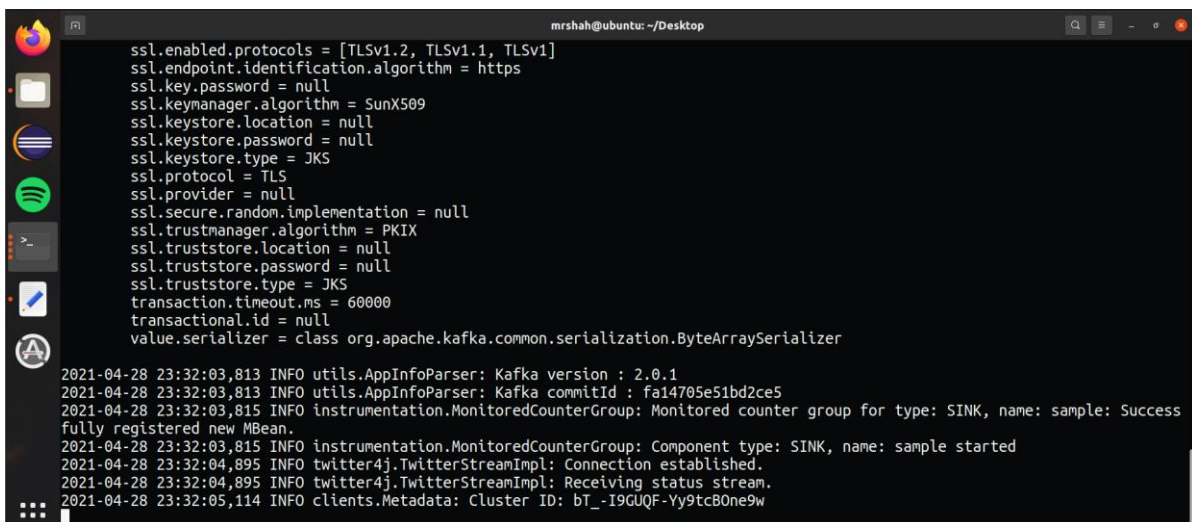
- Here, we have created Apache Kafka Topic, all the incoming tweets in the Kafka broker will be stored here.



```
mrshah@ubuntu: ~/Desktop
mrshah@ubuntu:~/Desktop$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --create --replication-factor 1 --partitions 1 --topic twitterdata1
Created topic twitterdata1.
mrshah@ubuntu:~/Desktop$
```

Fig 11.5 Creating Topic in Kafka broker as twitterdata1.

- Here, we are running our Apache Flume twitter.conf file, so that we can consume the tweets from Twitter App API.



```
mrshah@ubuntu: ~/Desktop
ssl.enabled.protocols = [TLSv1.2, TLSv1.1, TLSv1]
ssl.endpoint.identification.algorithm = https
ssl.key.password = null
ssl.keymanager.algorithm = SunX509
ssl.keystore.location = null
ssl.keystore.password = null
ssl.keystore.type = JKS
ssl.protocol = TLS
ssl.provider = null
ssl.secure.random.implementation = null
ssl.trustmanager.algorithm = PKIX
ssl.truststore.location = null
ssl.truststore.password = null
ssl.truststore.type = JKS
transaction.timeout.ms = 60000
transactional.id = null
value.serializer = class org.apache.kafka.common.serialization.ByteArraySerializer

2021-04-28 23:32:03,813 INFO utils.AppInfoParser: Kafka version : 2.0.1
2021-04-28 23:32:03,813 INFO utils.AppInfoParser: Kafka commitId : fa14705e51bd2ce5
2021-04-28 23:32:03,815 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: sample: Success
fully registered new MBean.
2021-04-28 23:32:03,815 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: sample started
2021-04-28 23:32:04,895 INFO twitter4j.TwitterStreamImpl: Connection established.
2021-04-28 23:32:04,895 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
2021-04-28 23:32:05,114 INFO clients.Metadata: Cluster ID: bT_-I9GUQF-Yy9tcB0ne9w
```

Fig 11.6: Running the twitter2.conf file in the shell.

- Here, we are consuming all the tweets from Apache Flume in our console using Apache Kafka Consumer.

```

translator_type":"none","id":2700083636,"geo_enabled":true,"profile_background_color":"000000","lang":null,"profile_sidebar_borde
r_color":"000000","profile_text_color":"000000","verified":false,"profile_image_url":"http://pbs.twimg.com/profile_images/1298271
709028466689/Iq_K825h_normal.jpg","time_zone":null,"url":"http://t2blive.com","contributors_enabled":false,"profile_background_til
e":false,"profile_banner_url":"https://pbs.twimg.com/profile_banners/2700083636/1617114364","statuses_count":29340,"follow_reque
st_sent":null,"followers_count":22405,"profile_use_background_image":false,"default_profile":false,"following":null,"name":"T2BLi
ve.COM","location":null,"profile_sidebar_fill_color":"000000","notifications":null},"retweet_count":0,"retweeted":false,"geo":n
ull,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":false,"id_str":"1387463676412657665","in_reply_to_user_
id":null,"favorite_count":0,"id":"1387463676412657665","text":"RT @T2BLive: India has reported 3,53,100+ New covid19 cases Today as
of now...\n\nStay Safe..Wear Mask 🧤","place":null,"lang":"en","quote_count":0,"favorited":false,"coordinates":null,"truncated
":false,"timestamp_ms":"1619632092942","reply_count":0,"entities":{"urls":[],"hashtags":[],"user_mentions":[{"indices":[3,11],"sc
reen_name":"T2BLive","id_str":"2700083636","name":"T2BLive.COM","id":"2700083636"}],"symbols":[]},"contributors":null,"user":{"utc
_offset":null,"friends_count":186,"profile_image_url_https":"https://abs.twimg.com/sticky/default_profile_images/default_profile_
normal.png","listed_count":0,"profile_background_image_url":"","default_profile_image":false,"favourites_count":1184,"description
":null,"created_at":"Sat Nov 21 08:43:35 +0000 2020","is_translator":false,"withheld_in_countries":[],"profile_background_image_u
rl_https":"","protected":false,"screen_name":"NaniDevotee","id_str":"1330069276905119744","profile_link_color":"1DA1F2","translat
or_type":"none","id":"1330069276905119744","geo_enabled":false,"profile_background_color":"F5F8FA","lang":null,"profile_sidebar_bor
der_color":"C0DEED","profile_text_color":"333333","verified":false,"profile_image_url":"http://abs.twimg.com/sticky/default_profi
le_images/default_profile_normal.png","time_zone":null,"url":null,"contributors_enabled":false,"profile_background_tile":false,"s
tatuses_count":2267,"follow_request_sent":null,"followers_count":14,"profile_use_background_image":true,"default_profile":true,"f
ollowing":null,"name":"NaniMBDevotee","location":null,"profile_sidebar_fill_color":"DDEEF6","notifications":null}}
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Wed Apr 28 17:48:12 +0000 2021","in_reply_to_user_i
d_str":null,"source":"<a href='\"http://twitter.com/download/iphone\"' rel='\"nofollow\"'>Twitter for iPhone</a>","retweet_count":0,"
retweeted":false,"geo":null,"filter_level":"low","in_reply_to_screen_name":null,"is_quote_status":false,"id_str":"138746367653432
9348","in_reply_to_user_id":null,"favorite_count":0,"id":"1387463676534329348","text":"Where are those couple who brought Corona vi
rus in SA from Italy?","place":null,"lang":"en","quote_count":0,"favorited":false,"coordinates":null,"truncated":false,"times
tamp_ms":"1619632092971","reply_count":0,"entities":{"urls":[],"hashtags":[],"user_mentions":[],"symbols":[]},"contributors":null
,"user":{"utc_offset":null,"friends_count":297,"profile_image_url_https":"https://pbs.twimg.com/profile_images/129864786511656550
5/Pa2N0Sno_normal.jpg","listed_count":0,"profile_background_image_url":"","default_profile_image":false,"favourites_count":3180,"

```

Fig 11.7: Getting the tweets

- Here, we are doing sentiment analysis on the raw tweets which is in unstructured form of data. We are applying Spark RDD and Spark DataFrames to filter out the data and getting the trending topics in twitter in real time.

```

57
58 val sRDD = sc.textFile("/home/mrshah/Desktop/s8.txt")
59 val tweetwords = sRDD.flatMap(tweetText => tweetText.split(" "))
60 val hashtags = tweetwords.filter(word => word.startsWith("#"))
61 val a = hashtags.map(hashtag => (hashtag, 1))
62 a.foreach(println)
63
64 val b = a.toDF()
65 b.show()
66 b.select("_1").show()
67 b.groupBy("_1").count().show()
68 val c = b.groupBy("_1").count()
69 c.show()
70
71
72 c.rdd.partitions.size
73 val c2 = c.coalesce(1)
74 c2.rdd.partitions.size
75 c2.show()
76 c2.write.csv("/home/mrshah/Desktop/PolularHashTags2")
77

```

Fig 11.8: Applying Spark RDD and DataFrames to raw tweets.

- Here, after doing the sentiment analysis on the tweets, we are getting our results, and the results are stored as CSV file in Local File System and then it is again loaded in MySQL Database.

```
scala> c2.show()
+-----+-----+
|_1|count|
+-----+-----+
|#TheSanatanDharma| 1|
|#nationalist| 4|
|#WhereIsPM| 1|
+-----+-----+

scala> c2.write.csv("/home/mrshah/Desktop/PolularHashTags.csv")
```

Fig 11.9: Saving the results and storing it to MySQL Database.

- After the results are stored in MySQL Database, PHP is connected to MySQL Database through Apache Server and Finally, the results are visualized in Web Console through PHP dashboard.

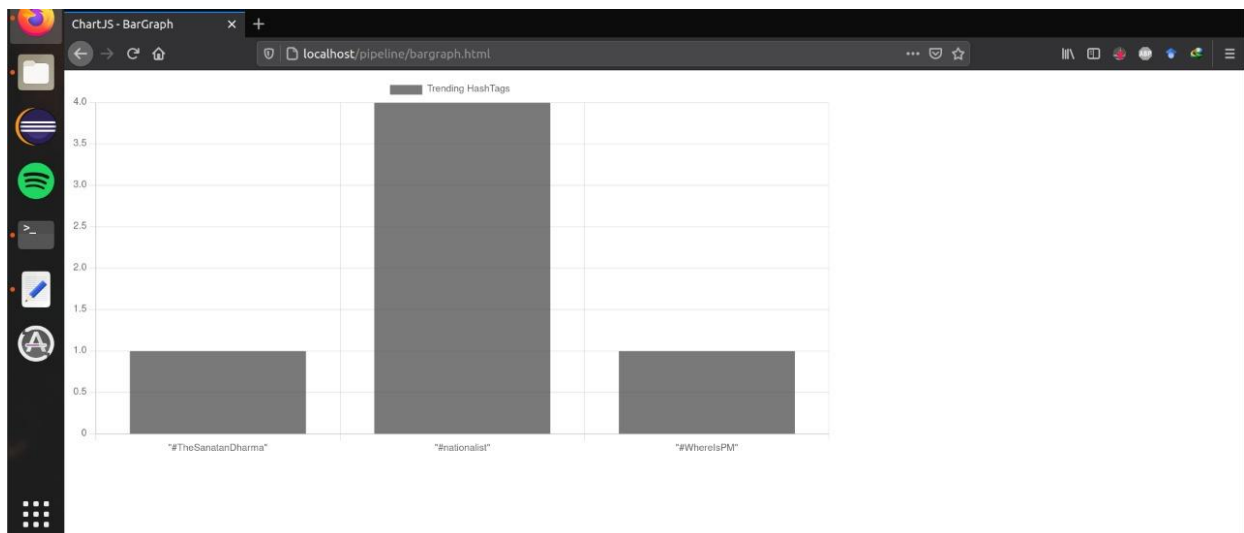


Fig 11.10: Visualizing the results in Web Console using PHP.

12. Bibliography

1. Getting Started for Twitter App API . *Twitter Developer Account*. [Online] [Cited: 06 17, 2020.] <https://developer.twitter.com/en/docs/twitter-ads-api/getting-started>.
2. Mr. SINGH, AMRITPAL. PHAGWARA : s.n., 2020. Big Data Fundamentals, Big Data Frameworks, Big Data Programming Tools.
3. SAH, MUKESH KUMAR and SHARMA, RISHABH. PHAGWARA, PUNJAB : s.n., 2020.
4. DAS, SUSHREE, et al. s.l. : Elsevier, 2018, *Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction*.
5. Tableau Overview. Tutorialspoint. [Online] [Cited: 03 31, 2021.] https://www.tutorialspoint.com/tableau/tableau_overview.htm.
6. Spark Streaming on DStreams. Apache Spark. [Online] [Cited: 03 28, 2021.] <https://spark.apache.org/docs/latest/streaming-programming-guide.html>.
7. Spark Streaming. DataBricks. [Online] [Cited: 03 28, 2021.] <https://databricks.com/glossary/what-is-spark-streaming>.
8. MySQL, MySQL Introduction, Installation. Tutorial Point. [Online] [Cited: 05 08, 2020.] <https://www.tutorialspoint.com/mysql/mysql-introduction.htm>.
9. Introduction, Key Concepts, APIs. Apache Kafka. [Online] [Cited: 03 28, 2021.] <https://kafka.apache.org/documentation/#introduction>.
10. Introduction to Apache Flume, Flume User Guide. Apache Flume. [Online] [Cited: 03 28, 2021.] <https://flume.apache.org/releases/content/1.9.0/FlumeUserGuide.html>.
11. Installing SBT on Linux. Scala SBT. [Online] [Cited: 03 16, 2020.] <https://www.scala-sbt.org/1.x/docs/Installing-sbt-on-Linux.html>.