ISSN (Online) : 2454 -7190 ISSN (Print) 0973-8975

J.Mech.Cont.& Math. Sci., Vol.-13, No.-5, November-December (2018) Pages 164-176

Real-time Data Streaming using Apache Spark on Fully Configured Hadoop Cluster

[1]Kashi Sai Prasad, [2]Dr. S Pasupathy

[1]Department of Computer Science and Engineering, MLR Institute of Technology,
Hyderabad, 500043.

[2]Associate Professor, Department of Computer Science and Engineering,
Annamalai University, Annamalainagar - 608002, Tamil Nadu.

https://doi.org/10.26782/jmcms.2018.12.00013

Abstract

Data plays a major role in today's Internet world.Analyzing historical data became easy due to advancement of analytical tools. Gathering data from social networking websites is a great challenge for today's data scientists. Many advancements and research has been conducted to gather streaming data(data generated every second) .Hadoop has provided acomponent called Apache Flume to ingest data into HDFS for processing using MapReduce. It has its own benefits,which made many analysis easy for social networking data,but Apache Flume requires a depthknowledge on configuration files and administration.

Our work proposes a framework for real-time data streaming of Twitter data. Apache spark which is an enhancement of Hadoop in terms of speed and faster processing provides much more insight than Apache flume.Spark is an in-memory distributed computing engine to increase processing speed over MapReduce, Spark is considered one of the most advanced ecosystem component for Batch and near-real time processing. We in our paper are explaining in detail about data ingestion using Apache Spark and Scala IDE. In our work the data will be directly ingested from Twitter website through tokens and access keys provided,which will be explained in chapter 3,4. Our GUI can also help a user to tweet into Twitter directly without moving on to Twitter website. We have also provided an option to categorize tweet of specific persons using '#' tags.The data thus obtained can be used for statistical analysis and generating reports.

Keywords*:* Apache Spark;BigData; Flume; Hadoop; MapReduce; Twitter data ingestion.

I. Introduction

Data ingestion and processing in parallel is very important because the amount of data is generally inexabytes and above. Currently there are different workflows offering real-time data analysis for Twitter, presenting general processing over streaming data. This study will attempt to develop an analytical framework with the ability of in-memory processing to extract and analyze structured and unstructured Twitter data [III].

Data processing can be classified into two types i.e., Batch Processing and Real-time processing. Batch processing dealing with huge amount of data which requires more time for processing. In Real-time processing is instant, results are generated hardly in milliseconds or seconds.

Data Processing Types

A. Batch Processing
B. Real-time Processing

A. Batch Processing

- Core Batch Hadoop, MapReduce
- Batch Interactive Hive,Impala [VIII]

B. Real-time Processing

Near Real-time	Spark Streaming
Core Real-time	Storm, flink

Apache Spark with many advanced features became the super fast mechanism for data processing. Twitter provides a tweet area with 140 character restricted short messages, where users can express their views and share their opinions, feelings, etc., Companies which capture click stream data of the users like Netflix, YouTube, Amazon etc., gives the scope to gather huge amount of streaming data which can be processed using Apache spark. In twitter hash tag is the convention of prefixing a word in the tweet with symbol '#' which indicates the keywords of a tweet. These words will be used for categorization of tweets based on topics and aids in searching. As per the Twitter Inc., statistics more than 500M tweets are generated every day. So handling such big streaming data using Apache spark is considered as next generation of Bigdata processing.

Problem in detail:

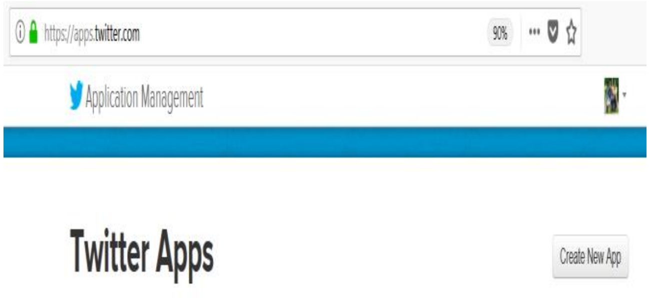
The streaming data generated through regular activities of users is voluminous to measure, which needs a very huge storage. Continuous triggering of data through an easy storage process can be achieved with a fully distributed cluster which in parallel can store and process the data. Twitter data ingestion through Apache Flume in Hadoop is a very big process. Apache streaming using Scala programming language can reduce the burden of writing the Flume queries through a simple process. The problem to be addressed is how Apache spark can be a solution for overcoming the process of Apache flume for data ingestion.

Streaming Twitter Data Using Apache Flume

Prerequisites: Twitter Account

I. Login to the twitter Account

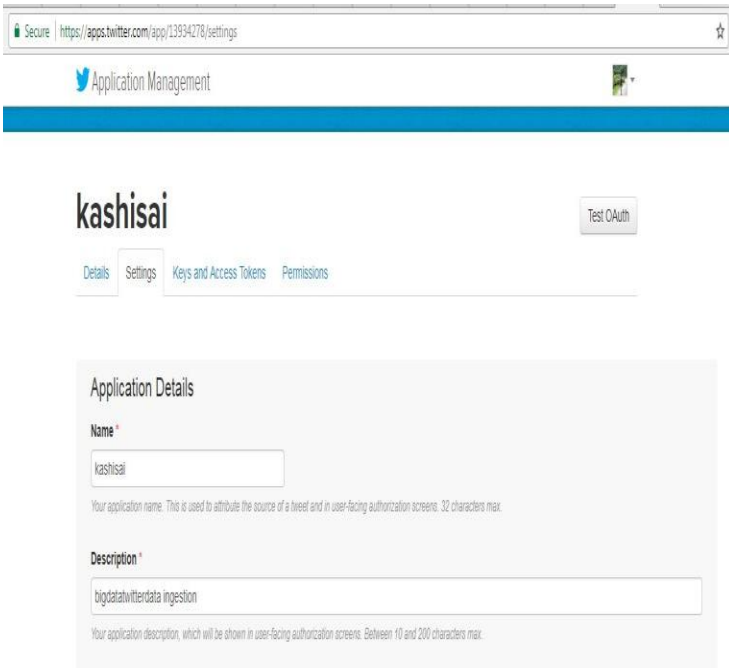
II. Go to the following link and create new App



https://apps.twitter.com

Figure 1

III. Enter the necessary details.



Secure https://apps.twitter.com/app/13934278/settings

Name *

kashisai

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

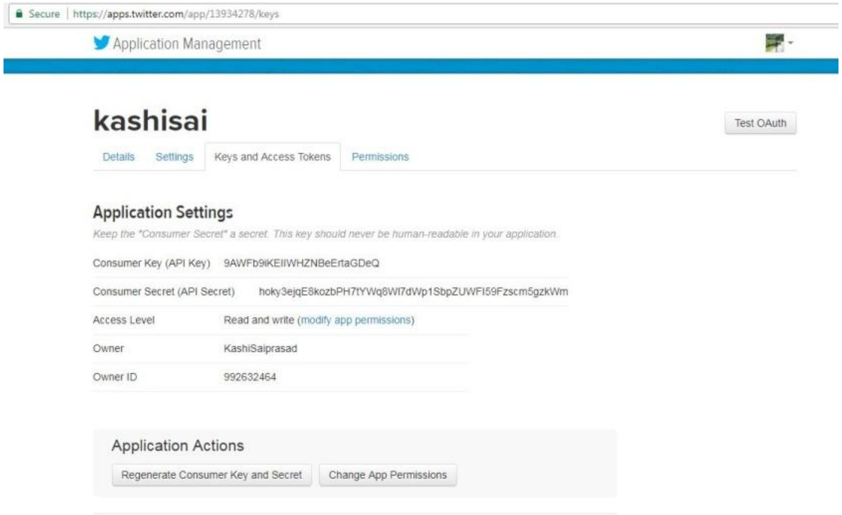
bigdatatwitterdata ingestion

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Figure 2

IV. Next we should accept the developer agreement and select a button: ‘create your Twitter application'.

V. Select the “Keys and Access Token” tab [VII].



kashisai

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) 9AWFb9iKEIIWHZNBeErtaGDeQ

Consumer Secret (API Secret) hoky3ejqE8kozbPH7tYWq8WI7dWp1SbpZUWFI59Fzscm5gzkWm

Access Level Read and write (modify app permissions)

Owner KashiSaiprasad

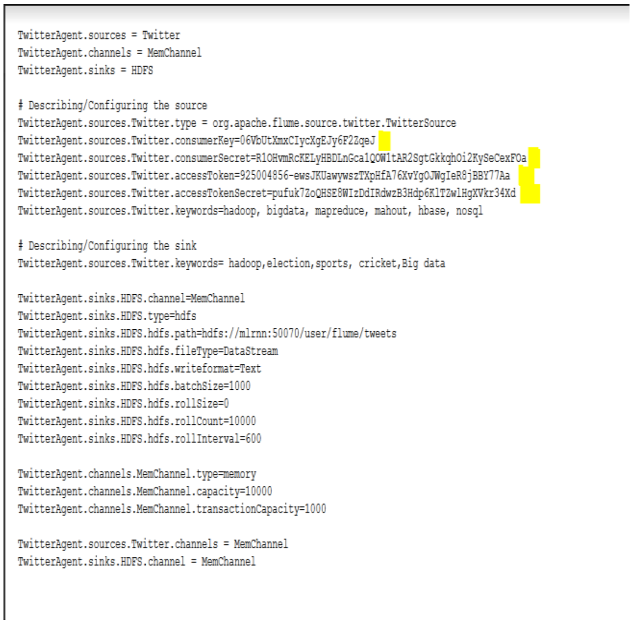
Owner ID 992632464

Application Actions

Figure 3

VI. Copy the consumer key and the consumer secret code and Copy the Access Token and Access token Secret code.

VII. We have to decide which Keywords tweet data to be collected from the twitter application. So, you can change the keywords in the TwitterAgent,sources.Twitter.keywordscommand.



```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=06VbUtXmxCIycXgEJy6F2ZqeJ
TwitterAgent.sources.Twitter.consumerSecret=R1OHvmRcKELyHBDLnGca1QOW1tAR2SgtGkkqhOi2KySeCexFOa
TwitterAgent.sources.Twitter.accessToken=925004856-ewsJKUawywszTXpHfA76XvYgOJWgIeR8jBBY77Aa
TwitterAgent.sources.Twitter.accessTokenSecret=pufuk7ZoQHSE8WIzDdIRdwzB3Hdp6KlTZwlHgXVkr34Xd
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql

# 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://mlrnn:50070/user/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
TwitterAgent.sinks.HDFS.hdfs.rollSize=0
TwitterAgent.sinks.HDFS.hdfs.rollCount=10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=1000

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Figure 4

VIII. Create a new directory inside HDFS path, where the Twitter tweet data should be stored.

```
hadoop fs –mkdir –p /usr/flume/tweet
```

IX. For fetching data from Twitter, Use the below command to fetch the twitter tweet data into HDFS cluster path.

```
Flume-ng agent –n TwitterAgent –f <location of created /edited conf file>
```

X. The above command will start fetching data from Twitter and streams in into the HDFS given path.

XI. Once, the tweet data started streaming it into the given HDS path we can use ‘Ctrl+c’ command to stop the streaming process.

As Flume has Complex structure, configuration and maintenance will be difficult. In Flume throughput depends on the backing store of the channel so scalability and reliability in not up to the mark.

Fully Configured Hadoop Cluster

We have access to a fully configured Hadoop cluster with 9 Nodes (3 Master nodes, 5 data nodes and an edge node). The data ingested from Twitter directly comes into HDFS which can be accessed by an user by providing the credentials.

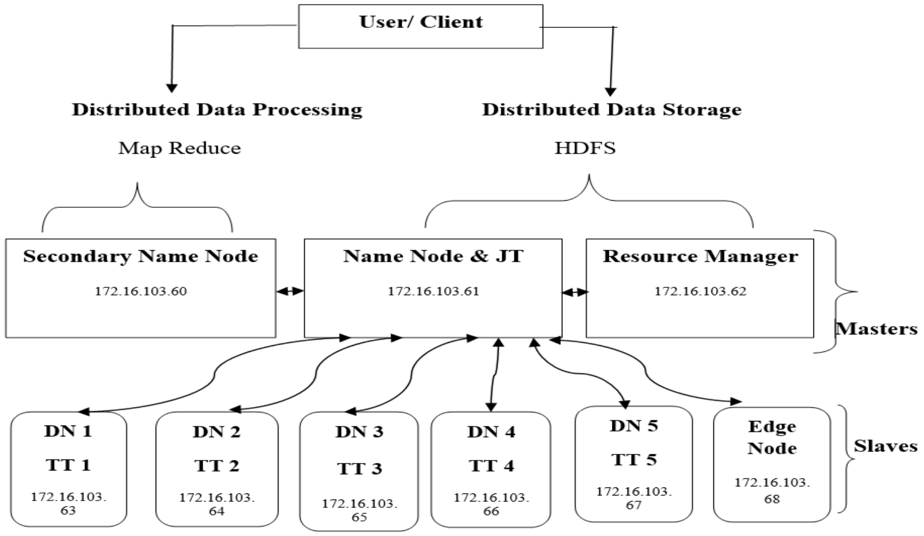


Fig 5: Multi Node Hadoop Cluster Architecture

II. Literature Survey

Hadoop though provides the best solution for the Bigdata requirements, there are much more requirements due to vast usage of networking sites. Apache Spark was built to prevail over the restrictions of Hadoop MapReduce cluster computing paradigm. Apache spark is built using the existing MapReduce paradigm but the computing power and processing strategy is completely different [IV].As the computing power of Spark is very fast it is said that spark is hundred times faster than Hadoop MapReduce because of the in memory processing of data the time spent in moving data in and out is not required [V].

As per Microsoft statistics by 2020 there will be 30B connected devices, each individual will generate 1.5GB of data per day, Smart homes will generate approximately 50 GB of data per day, Smart buildings with smart devices can generate 150GB data per day.

The limitations which are identified in terms of data processing by MapReduce are below:

1. MapReduce uses only Java for building an application.
2. Hadoop MapReduce uses disk-based processing and not suitable for performing streaming data.
3. Not suitable for Real-Time streaming data processing
4. High latency
5. Reading and writing data from disk (No in memory processing)
6. Reusability
7. Fault Tolerance
8. Complex to use
9. MapReduce needs an external job scheduler like Oozie to schedule complex jobs.
10. MapReduce can't cache the data in memory for future requirements.

III. System Architecture

Dataflow Diagram

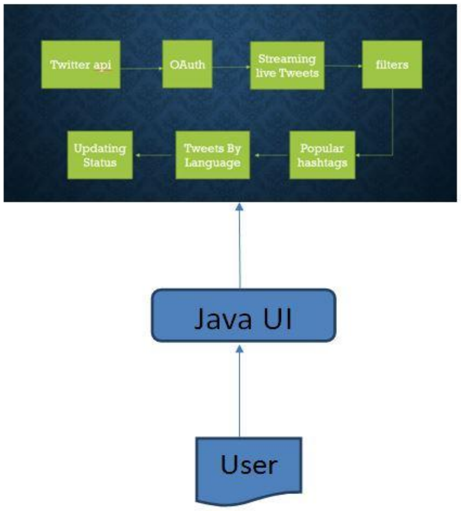


Figure 6

Apache spark is a general purpose engine to solve variety of data processing problems and also it can be used for traditional batch processing operations with additional support for Real-time

Processing. The primary purpose of Spark is to efficiently handle data of a large magnitude [II]. It's a fast In-Memory Data Processing engine. There are specific modules for several use cases like Spark Core, Spark SQL, Spark Streaming module for real-time processing[V].

Workflow:

An user uses the UI developed using Java and in the background we have used the Twitter API to connect to Twitter account and using OAuth standard our application provides a secure access to twitter data.

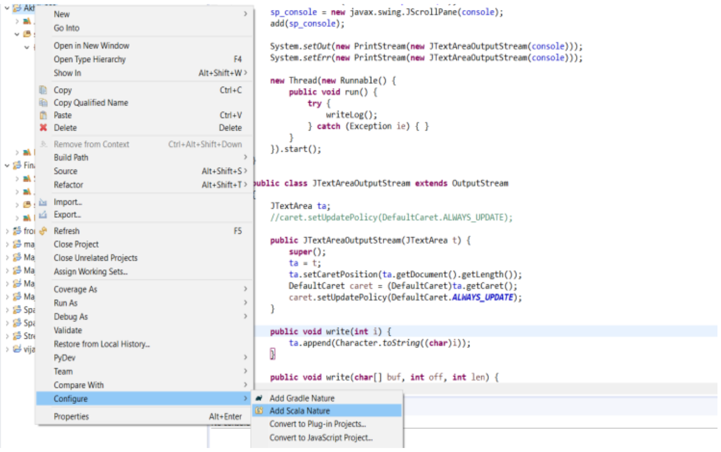
OAuth is a standard that application can use to provide client application with secured delegate access. OAuth can help the developer to create and debug applications which use Twitter's Application program Interface.If we want to use OAuth our application need to obtain the access tokens of an user. OAuth is a protocol that supports authorization workflows.

The streaming live tweets can be ingested into HDFS in a single click until we close the UI. Later filters are applied to get the Tweets using hash (#) tags, Tweets are also obtained from various languages using filters. The data thus obtained will be stored in the Hadoop cluster and using Scala we can process the data for insights.

IV. Implementation

A) Configuration

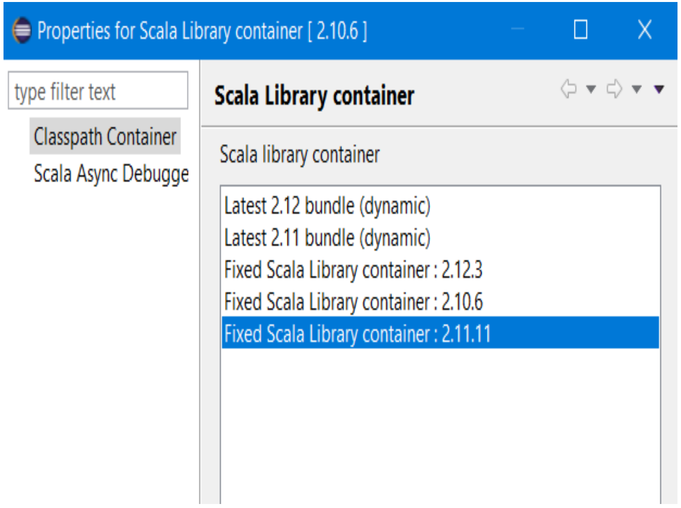
Adding Scala nature to already existing Java project:



```
    sp_console = new javax.swing.JScrollPane(console);
    add(sp_console);

    System.setOut(new PrintStream(new JTextAreaOutputStream(console)));
    System.setErr(new PrintStream(new JTextAreaOutputStream(console)));

    new Thread(new Runnable() {
        public void run() {
            try {
                writeLog();
            } catch (Exception ie) { }
        }
    }).start();
}

ublic class JTextAreaOutputStream extends OutputStream
{
    JTextArea ta;
    //caret.setUpdatePolicy(DefaultCaret.ALWAYS_UPDATE);

    public JTextAreaOutputStream(JTextArea t) {
        super();
        ta = t;
        ta.setCaretPosition(ta.getDocument().getLength());
        DefaultCaret caret = (DefaultCaret)ta.getCaret();
        caret.setUpdatePolicy(DefaultCaret.ALWAYS_UPDATE);
    }

    public void write(int i) {
        ta.append(Character.toString((char)i));
    }

    public void write(char[] buf, int off, int len) {
```

Figure 7

Setting Scala verion:



Properties for Scala Library container [2.10.6]

type filter text

Classpath Container

Scala Async Debugge

Scala Library container

Scala library container

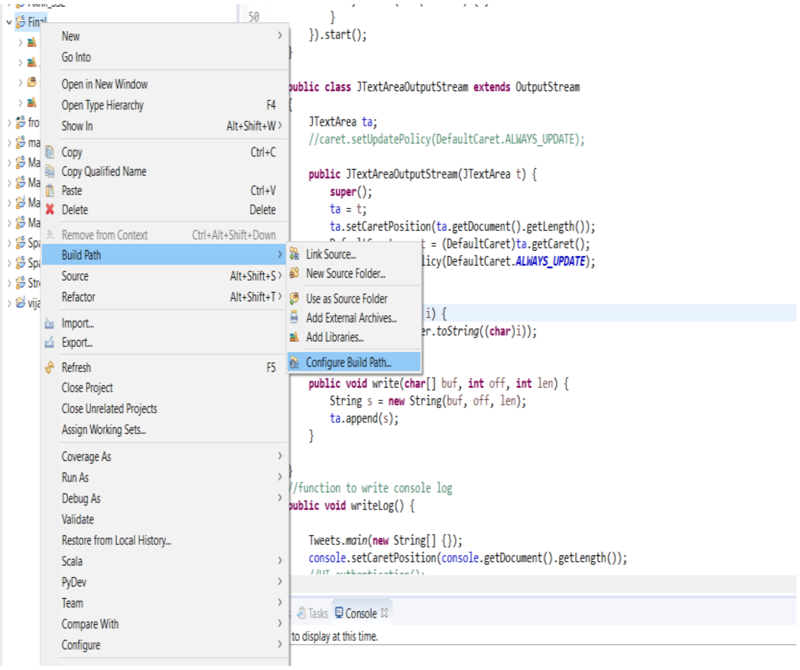
- Latest 2.12 bundle (dynamic)
- Latest 2.11 bundle (dynamic)
- Fixed Scala Library container : 2.12.3
- Fixed Scala Library container : 2.10.6
- Fixed Scala Library container : 2.11.11

Figure 8

Scala runs primarily on the Java Virtual Machine (JVM). As Scala and the JVM improve independently over time, Scala may drop compatibility with older JVM versions, in order to better take advantage of new JVM features.

B) Configuring build path for adding spark jars:

Download spark jars from
http://spark.apache.org/downloads.html
Add all jars from \spark-2.2.1-bin-hadoop2.7\jars to project build path.



```
        }
    }).start();
}

public class JTextAreaOutputStream extends OutputStream
{

    JTextArea ta;
    //caret.setUpdatePolicy(DefaultCaret.ALWAYS_UPDATE);

    public JTextAreaOutputStream(JTextArea t) {
        super();
        ta = t;
        ta.setCaretPosition(ta.getDocument().getLength());
                           t = (DefaultCaret)ta.getCaret();
                          licy(DefaultCaret.ALWAYS_UPDATE);

                           i) {
                          r.toString((char)i));

    public void write(char[] buf, int off, int len) {
        String s = new String(buf, off, len);
        ta.append(s);
    }

}
//function to write console log
public void writeLog() {

    Tweets.main(new String[] {});
    console.setCaretPosition(console.getDocument().getLength());
```

to display at this time.

Figure 9

The developed project consists of both Java and Scala IDE and we need to choose compatible JVM and Scala versions.

V. Results

A) Screenshots

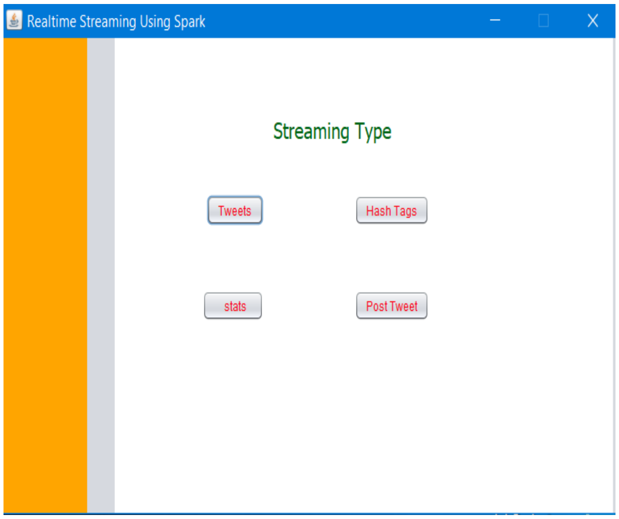
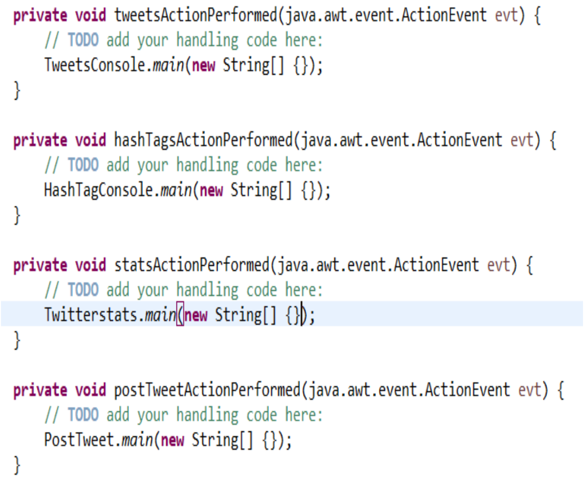


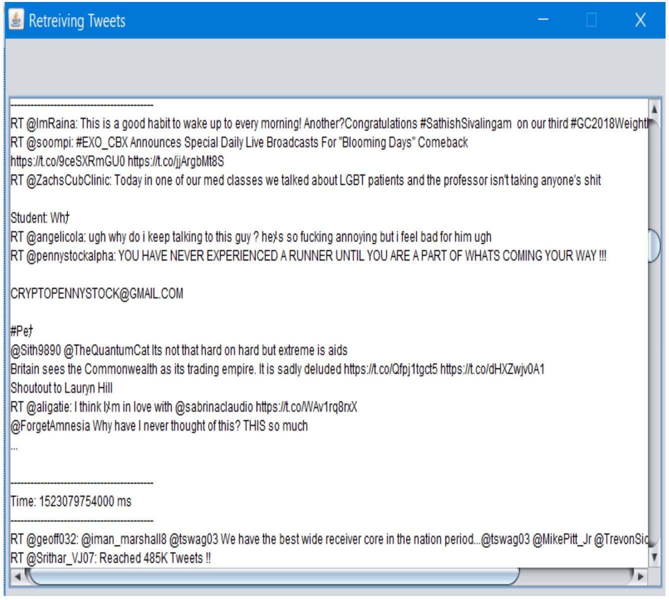
Figure 10



```
private void tweetsActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
    TweetsConsole.main(new String[] {});
}

private void hashTagsActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
    HashTagConsole.main(new String[] {});
}

private void statsActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
    Twitterstats.main(new String[] {});
}

private void postTweetActionPerformed(java.awt.event.ActionEvent evt) {
    // TODO add your handling code here:
    PostTweet.main(new String[] {});
}
```

Figure 11

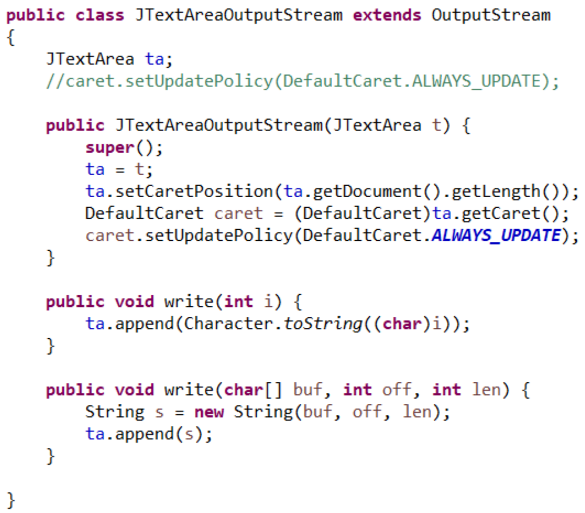
Streaming Tweets



```
----------------------------------------
RT @ImRaina: This is a good habit to wake up to every morning! Another?Congratulations #SathishSivalingam  on our third #GC2018Weight
RT @soompi: #EXO_CBX Announces Special Daily Live Broadcasts For "Blooming Days" Comeback
https://t.co/9ceSXRmGU0 https://t.co/jjArgbMt8S
RT @ZachsCubClinic: Today in one of our med classes we talked about LGBT patients and the professor isn't taking anyone's shit

Student: Wh†
RT @angelicola: ugh why do i keep talking to this guy ? he)s so fucking annoying but i feel bad for him ugh
RT @pennystockalpha: YOU HAVE NEVER EXPERIENCED A RUNNER UNTIL YOU ARE A PART OF WHATS COMING YOUR WAY !!!

CRYPTOPENNYSTOCK@GMAIL.COM

#Pe†
@Sith9890 @TheQuantumCat Its not that hard on hard but extreme is aids
Britain sees the Commonwealth as its trading empire. It is sadly deluded https://t.co/Qfpj1tgct5 https://t.co/dHXZwjv0A1
Shoutout to Lauryn Hill
RT @aligatie: I think I)m in love with @sabrinaclaudio https://t.co/WAv1rq8rxX
@ForgetAmnesia Why have I never thought of this? THIS so much
...

----------------------------------------
Time: 1523079754000 ms
----------------------------------------
RT @geoff032: @iman_marshall8 @tswag03 We have the best wide receiver core in the nation period...@tswag03 @MikePitt_Jr @TrevonSic
RT @Srithar_VJ07: Reached 485K Tweets !!
```

Figure 12

Here in this scenario we are redirecting the console output to a JTextArea as

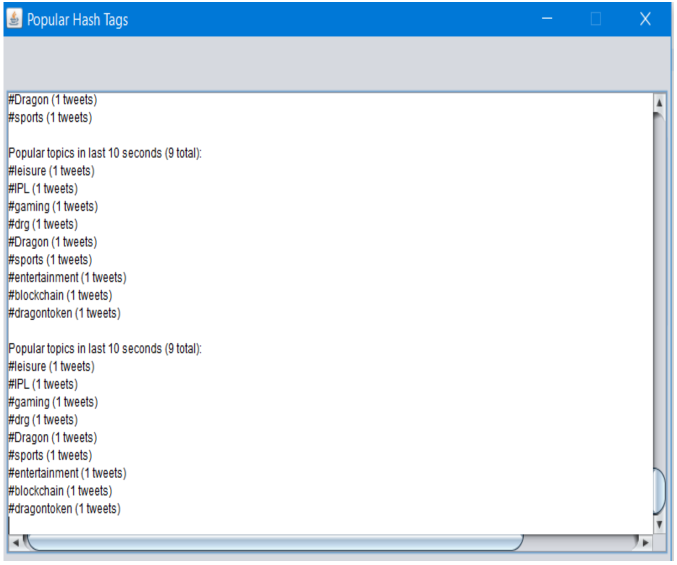


```
public class JTextAreaOutputStream extends OutputStream
{
    JTextArea ta;
    //caret.setUpdatePolicy(DefaultCaret.ALWAYS_UPDATE);

    public JTextAreaOutputStream(JTextArea t) {
        super();
        ta = t;
        ta.setCaretPosition(ta.getDocument().getLength());
        DefaultCaret caret = (DefaultCaret)ta.getCaret();
        caret.setUpdatePolicy(DefaultCaret.ALWAYS_UPDATE);
    }

    public void write(int i) {
        ta.append(Character.toString((char)i));
    }

    public void write(char[] buf, int off, int len) {
        String s = new String(buf, off, len);
        ta.append(s);
    }

}
```

Figure 13

System.setOut(new PrintStream(new JTextAreaOutputStream(console)));

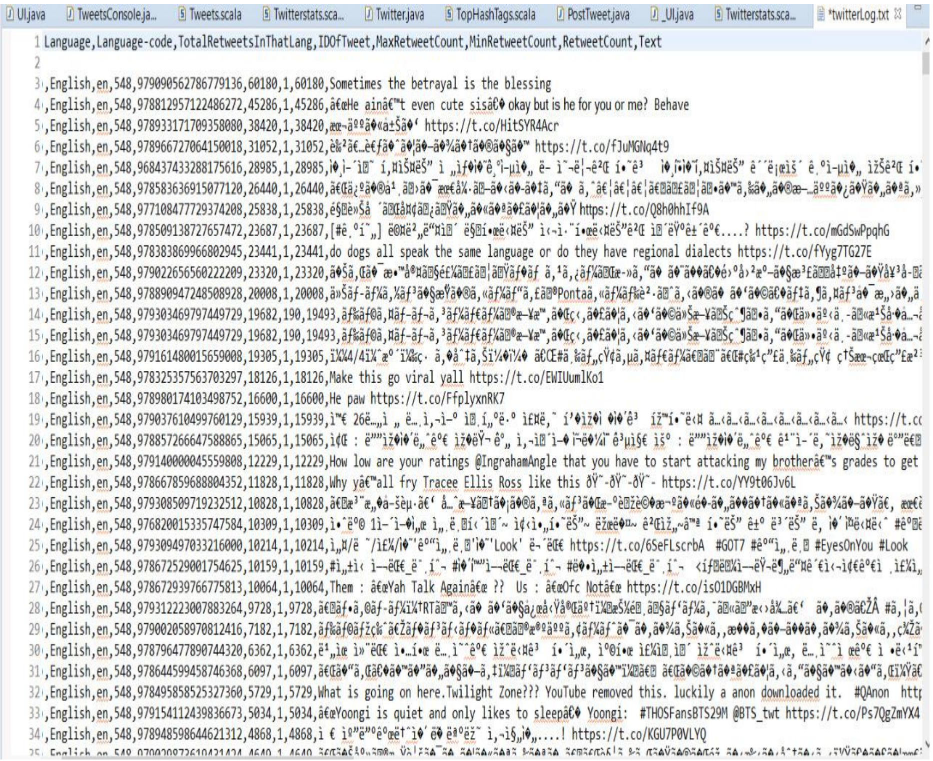
Popular HashTags



```
#Dragon (1 tweets)
#sports (1 tweets)

Popular topics in last 10 seconds (9 total):
#leisure (1 tweets)
#IPL (1 tweets)
#gaming (1 tweets)
#drg (1 tweets)
#Dragon (1 tweets)
#sports (1 tweets)
#entertainment (1 tweets)
#blockchain (1 tweets)
#dragontoken (1 tweets)

Popular topics in last 10 seconds (9 total):
#leisure (1 tweets)
#IPL (1 tweets)
#gaming (1 tweets)
#drg (1 tweets)
#Dragon (1 tweets)
#sports (1 tweets)
#entertainment (1 tweets)
#blockchain (1 tweets)
#dragontoken (1 tweets)
```

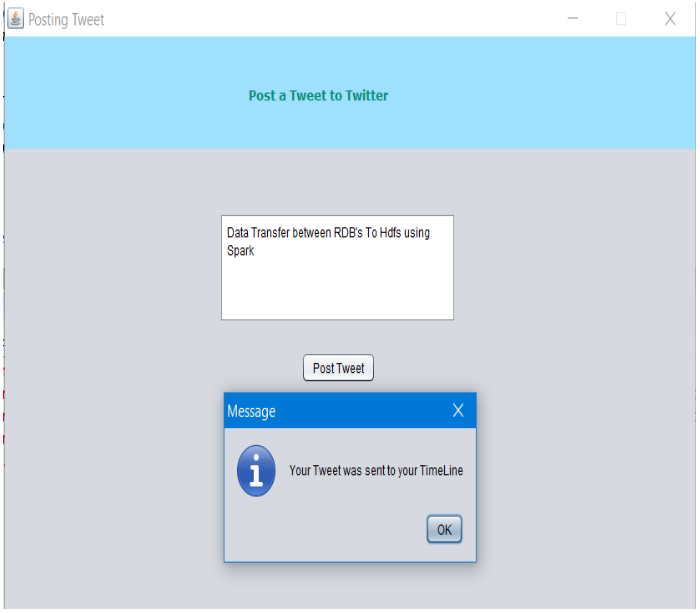
Figure 14

Twitter stats



```
Language,Language-code,TotalRetweetsInThatLang,IDOfTweet,MaxRetweetCount,MinRetweetCount,RetweetCount,Text

,English,en,548,979090562786779136,60180,1,60180,Sometimes the betrayal is the blessing
,English,en,548,978812957122486272,45286,1,45286,â€œHe ainâ€™t even cute sisâ€� okay but is he for you or me?  Behave
,English,en,548,978933171709358080,38420,1,38420,æœ¬ãººã�«å±Šã�‘ https://t.co/HitSYR4Acr
,English,en,548,978966727064150018,31052,1,31052,è‰²ã€…è€ƒã�ˆã�¦ã�–ã�¾ã�†ã�®ã�§ã�™ https://t.co/fJuMGNq4t9
,English,en,548,968437433288175616,28985,1,28985,ì�¸ì–´ì˜ í‚¤ìŠ¤ëŠ” ì „ìƒ�ì�˜ê¸°ì–µì�„ ë– ì˜¬ë¦¬ê²Œ í•˜ê³   ì�¸í•ì�˜í‚¤ìŠ¤ëŠ” ê´´ë¡œìš´ ê¸°ì–µì�„ ìžŠê²Œ í•
,English,en,548,978583636915077120,26440,1,26440,ã€Œä¿ºã�®å¹ ã�›ã�¯æœ€å¼·ã�–ã�‹ã�–ã�‡ã‚“ã� ã‚ˆâ€¦â€¦â€¦ã€�ã�£ã�¦ã�•ã�™ã‚‰ã�„ã�®æ–…äººã�¿ã�Ÿã�„ã�ªã‚»
,English,en,548,977108477729374208,25838,1,25838,é§�è»Šå ´ã�Œå¤¢ã�¿ã�Ÿã�„ã�«ã�ªã�£ã�¦ã�„ã�Ÿ https://t.co/Q8h0hhIf9A
,English,en,548,978509138727657472,23687,1,23687,[#ê¸°í˜„] ë©¤ë²„ë“¤ì�´ ë§�í•œë‹¤ëŠ” ì‹¬ì·¨í•œë‹¤ëŠ”ê²Œ ì�´ëŸ°ê±´ê°€....? https://t.co/mGdSwPpqhG
,English,en,548,978383869966802945,23441,1,23441,do dogs all speak the same language or do they have regional dialects https://t.co/fYyg7TG27E
,English,en,548,979022656560222209,23320,1,23320,ã�Šã‚Œã�¯æ•™å®¤ã�§é£¼ã�£ã�¦ã�Ÿãƒ�ãƒ ã‚¹ã‚¿ãƒ¼ã�Œæ»ã‚“ã� ã�¨ã��ã€�é›°å›²æ°—ã�§æ³£ã��å‡ºã�–ã�Ÿå¥³å�ã
,English,en,548,978890947248508928,20008,1,20008,ä»Šãƒãƒ¼ã‚½ãƒ³ã�§æŸã�®ã‚«ãƒ¼ãƒ‰ã‚£ã�®Pontaã‚«ãƒ¼ãƒ‰è²·ã�ˆã‚‹ã�®ã� ã�‘ã�©ã€�ãƒ‡ã‚¶ã‚¤ãƒ³å�¯æ„›ã�„ä
,English,en,548,979303469797449729,19682,190,19493,ãƒ‰ãƒ©ã‚¤ãƒ–ãƒ¬ã‚³ãƒ¼ãƒ€ãƒ¼ã�®æ˜ åƒ�ã�Œç‹‚ã�£ã�¦ã‚‹ã�‘ã�©ä»Šæ—¥ã�Šç¨¼ã�•ã‚“ã�Œä»•äº‹ä¸ã�«æ¹Šå�‘å…¬å
,English,en,548,979303469797449729,19682,190,19493,ãƒ‰ãƒ©ã‚¤ãƒ–ãƒ¬ã‚³ãƒ¼ãƒ€ãƒ¼ã�®æ˜ åƒ�ã�Œç‹‚ã�£ã�¦ã‚‹ã�‘ã�©ä»Šæ—¥ã�Šç¨¼ã�•ã‚“ã�Œä»•äº‹ä¸ã�«æ¹Šå�‘å…¬å
,English,en,548,979161480015659008,19305,1,19305,ï¼4/4ï¼ˆæ°´ï¼‰ç· ã‚�åˆ‡ã‚Šï¼�ï¼� ã€Œ#ä¸‰ãƒ„ç¢ºã‚µãƒ€ãƒ¼ã€�ã�¨ã€Œ#ç‰¹ç”£ä¸‰ãƒ„ç¢ ç†Šæœ¬çœŒç”£æ²
,English,en,548,978325357563703297,18126,1,18126,Make this go viral yall https://t.co/EWIUumlKo1
,English,en,548,978980174103498752,16600,1,16600,He paw https://t.co/FfplyxnRK7
,English,en,548,979037610499760129,15939,1,15939,ì™€ 26ë…„ì „ ë…¸ì‚¬ì—° ì�¸í„°ë·° ì£¤ë‚˜ í’�ìž�ì �ì�´ê³  íž™í•˜ë‹¤ ã…‹ã…‹ã…‹ã…‹ã…‹ã…‹ã…‹ã…‹ https://t.cc
,English,en,548,978857266647588865,15065,1,15065,ì¢Œ : ë””ìž�ì�´ë„ˆê°€ ìž�ëŸ¬ê°„ ì‚¬ì�´ì–� ì¬ë�¼ì˜ ê³µì§€ ìš° : ë””ìž�ì�´ë„ˆê°€ ê¹¨ì–´ë‚˜ìž�ë§ˆìž� ë°”ë€�
,English,en,548,979140000045559808,12229,1,12229,How low are your ratings @IngrahamAngle that you have to start attacking my brotherâ€™s grades to get
,English,en,548,978667859688804352,11828,1,11828,Why yâ€™all fry Tracee Ellis Ross like this ðŸ˜ðŸ˜ðŸ˜ https://t.co/YY9t06Jv6L
,English,en,548,979308509719232512,10828,1,10828,ã€�æ³¨æ„�å–šèµ·ã€‘ å…ˆæ—¥ã�†ã�¡ã�®ã‚ªã‚«ãƒ³ã�Œæ–°è�žè©�æ¬ºã�«é�ã�„ã��ã�†ã�«ã�ªã‚Šã�¾ã�–ã�Ÿã€‚ æœ€ã
,English,en,548,976820015335747584,10309,1,10309,ì•ˆë°© 1ì–´ì–�ì„œ ì„¸ë�í‹´ì�´ ì¢‹ì•„í•˜ëŠ” ëžœë�¤ ê²Œìž„ ~â™ª í•˜ëŠ” ê±° ë³´ëŠ” ë¸ ì�´ ì¡´ë‹¤ë‹ˆ #ê°�ë
,English,en,548,979309497033216000,10214,1,10214,ì„¸/ë¯/ì£¼/ì�˜'ê°“ì„¸ë�'ì�˜'Look' ë¬´ëŒ€ https://t.co/6SeFLscrbA  #GOT7 #ê°“ì„¸ë� #EyesOnYou #Look
,English,en,548,978672529001754625,10159,1,10159,#ì„±ì‹ ì—¬ëŒ€_ë¨¸í�¬ #ì�´í™”ì—¬ëŒ€_ë¨¸í�¬ #ë�•ì„±ì—¬ëŒ€_ë¨¸í�¬  <íƒ�ë�¼ì—¬ëŸ¬ë¶„ë“¤ê´€ì‹¬ì¢€ê°€ì ¸ì£¼ì„¸ì
,English,en,548,978672939766775813,10064,1,10064,Them : â€œYah Talk Againâ€œ ??  Us : â€œOfc Notâ€œ https://t.co/isO1DGBMxH
,English,en,548,979312223007883264,9728,1,9728,ã€�ãƒ•ã‚©ãƒãƒ¼ï¼†RTã�™ã‚‹ã� ã�‘ã�§å¿œå‹Ÿå®Œäº†ï¼�æŠ½é�¸ã�§ãƒ‘ãƒ¼ã‚¯ã�«ã�”æ‹›å¾…ã€‘ ã�‚ã�®ã€ŽÂ #ã‚¦ã‚
,English,en,548,979002058970812416,7182,1,7182,ãƒ‰ãƒ©ãƒžç‰ˆã€Žãƒ�ãƒ³ãƒ‹ãƒ�ãƒ«ã€�ã�®æ®ºäººã‚¢ãƒ¼ãƒˆã�¯ã�‚ã�¾ã‚Šã�«ã‚‚æ��ã‚�ã�–ã��ã�‚ã�¾ã‚Šã�«ã‚‚ç¾Žã
,English,en,548,978796477890744320,6362,1,6362,ë¹„ìœ ì»¨ëŒ€ ì•…í•œ ë…¸ì˜ˆê°€ ìž˜ë‹¤ê³  í•´ì„œ, ì°©í•œ ìž¥ì�¸ì�´ ìž˜ë‹¤ê³  í•´ì„œ, ë…¸ì˜ˆì œê°€ ì •ë‹¹í
,English,en,548,978644599458746368,6097,1,6097,ã€Œã�“ã‚Œã€�ã�™ã�”ã�„ã�§ã�–ã‚‡ï¼�ãƒ‘ãƒ³ãƒ‘ãƒ³ã�§ã�™ï¼�ã€� ã€Œã�©ã�†ã�ªã�£ã�¦ã‚‹ã‚“ã�§ã�™ã�‹ã�“ã‚Œï¼Ÿã€
,English,en,548,978495858525327360,5729,1,5729,What is going on here.Twilight Zone??? YouTube removed this. luckily a anon downloaded it.  #QAnon  http
,English,en,548,979154112439836673,5034,1,5034,â€œYoongi is quiet and only likes to sleepâ€� Yoongi:  #THOSFansBTS29M @BTS_twt https://t.co/Ps7QgZmYX4
,English,en,548,978948598644621312,4868,1,4868,ì € ìº”ë”°ê°œë†ˆì�´ ë� ëª°ëž˜ ì‚¬ì§„ì�„....! https://t.co/KGU7P0VLYQ
,English,en,548,979020872610421424,4640,1,4640,
```

Figure 15



Data Transfer between RDB's To Hdfs using Spark

Your Tweet was sent to your TimeLine

Figure 16

Checkthe Twitter Timeline so that the post will be seen on the account who have logged in.

VI. Conclusion

Apache Hadoop became a synonym for Bigdata due to its scalability, cost effectiveness, replication etc., to process voluminous data. Even though Hadoop is very fast in processing and generating efficient results, there is still incompleteness. Data ingestion into Hadoop cluster is a crucial task in today's fast moving world.The components of Hadoop Ecosystem like Apache Sqoop, Apache Flume, KAFKA, Cassandra can complete the task but the streaming data generated needs a much more reliable and efficient tool like Apache Spark which is 10 times faster than Hadoop. Our work concludes explaining the procedure to connect to a Twitter using OAuth, Access token and ingest data into cluster, and process to find out insights from generated data.

Apache Spark is preferred on top of Apache flume for its varied applications.A Hadoop user can easily understand the working of Apache Spark and Scala in the explained work.

Future Work

Technology is in continuous advancement due to large amount of data generated from online networking sites. Gathering data and storing it in Hadoop cluster is explained in our work. This work can be more precisely implemented using Apache Oozie which can gather data at regular time intervals allotted. As we have feasibility to increase the nodes in the developed cluster, as per data requirements we can parallely process the data.

Below are some other interesting dimensions which we can add to this project in the future.

- Creating dashboards to monitor all incoming tweets, user interaction and present results in a visually appealing manner.
- Geo-spatial maps could be plotted, signifying the location (and sentiment) of the tweets
- Creating a pipeline to dump all processed tweets into HDFS, Databases like Cassandra, Mongo DB, PostgreSQL or MySQL.

References

I. Altti Ilari Maarala, Mika Rautiainen, Miikka Salmi, Susanna Pirttikangas and Jukka Riekki", Low latency analytics for streaming traffic data with Apache Spark" IEEE International Conference on Big Data (2015).

II. Anand Gupta, Hardeo Kumar Thakur " A Big Data Analysis Framework Using Apache Spark and Deep Learning", IEEE International Conference on Data Mining Workshops (2017).

III. Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabrizi "Developing a Real-time Data Analytics Framework For Twitter Streaming Data"**,** IEEE 6th International Congress on Big Data (2017).

IV. Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari, Shuja Ur Rehman Baig " Real-time Text Analytics Pipeline Using Open-source Big Data Tools", arXiv:1712.04344, Dec (2017).

V. Marouane Birjalia, Abderrahim Beni-Hssane, Mohammed Erritali "Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume ", The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks Elsevier (2017).

VI. Ramkrushna C. Maheshwar, D. Haritha "Survey on High Performance Analytics of Bigdata with Apache Spark", International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) (2016).

VII. Sangeeta "Twitter Data Analysis Using FLUME & HIVE on Hadoop Framework", Special Issue on International Journal of Recent Advances in Engineering & Technology (IJRAET) V-4 I-2February (2016).

VIII. S. Cha and M. Wachowicz. "Developing a real-time data analytics framework using Hadoop", IEEE International Congress on Big Data June (2015).