

SOCIAL MEDIA DATA ANALYSIS USING BIG DATA AND HADOOP

Minor Project Report

**Submitted to Asansol Engineering College in partial fulfillment of the
requirements for the degree of**

**Bachelor of Technology
(Computer Science and Engineering)
of
Maulana Abul Kalam Azad University of Technology
Kolkata – 700064**

By

**Sourav Vishwakarma (Roll No.: 10800115114)
Shivam (Roll No.: 10800115105)
Ruchi Kumari (Roll No.: 10800115095)
Chandni Kumari (Roll No.:10800115033)**

**Under the guidance
Of**

**Dr. Sandip Roy Associate Professor,
Department of Computer Science and Engineering**



**Department of Computer Science & Engineering
Asansol Engineering College
Vivekananda Sarani**

Asansol-713305
2018

Department of Computer Science and Engineering
Asansol Engineering College
Asansol

Certificate

I hereby recommend that the thesis entitled “SOCIAL MEDIA DATA ANALYSIS USING BIG DATA AND HADOOP” submitted by Sourav Vishwakarma, Shivam, Ruchi Kumari, Chandni Kumari with roll nos.10800115114,10800115105,10800115095,10800115033 and Registration nos.151080110115, 151080110106, 151080110096, 151080110033 respectively, has been carried out under my guidance and supervision and may be accepted in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Maulana Abul Kalam Azad University of Technology, Kolkata-700064.

Mr. SANDIP ROY
(Associate Professor)
Department of Computer Science and Engineering
Asansol Engineering College
Asansol-713305, West Bengal.

Dr. MONISH CHATTERJEE
(Head and Associate Professor)
Department of Computer Science and Engineering
Asansol Engineering College
Asansol-713305, West Bengal.

ACKNOWLEDGEMENT

We express our sincere gratitude to **Dr. Sandip Roy**, our guide for his affectionate and valuable guidance without whose help the present work could not have been successful. We are also indebted to him as a teacher who introduced us to the topics related to the project.

We thank **Mr. Monish Chatterjee**, the Head of the Department of Computer Science and Engineering for his constant encouragement and permission to work in the Departmental laboratory and use the various resources of the Department of CSE whenever required without which our work would not have been possible.

We also thank Mr. Sandip Roy again of CSE Project Laboratory for his assistance in various implementation related issues whenever it was required.

We are also grateful to the other teachers of the Department of CSE who have taken the pain of teaching us various core subjects of Computer Science for the last few years. We are also thankful to all other staffs of the Department for clearing the various technical doubts during the Laboratory Sessions which boosted our confidence.

Sourav Vishwakarma

Roll No. - 1080015114

Registration No.-151080110115

Signature - _____

Chandni Kumari

Roll No. - 10800115033

Registration No. - 151080110033

Signature - _____

Shivam

Roll No. – 10800115105

Registration No. - 151080110106

Signature - _____

Ruchi Kumari

Roll No. - 10800115095

Registration No. - 151080110096

Signature - _____

CONTENTS

Chapters	Pages
1. Synopsis	5-6
2. Introduction to the work	7
3. Project Details	
3.1. Definitions and Theories	8-13
3.2. Data Flow Diagrams	14-15
3.3. Algorithms	16-23
3.4. System Requirements For implementation	24
4. Results	
4.1. Tables & Results	25-26
4.2. Screen Shots	27-29
5. Conclusion and Future Scope	30
6. Bibliography	31

SYNOPSIS

1. TITLE OF THE PROJECT

Title of the project is **“SOCIAL MEDIA DATA ANALYSIS USING BIG DATA AND HADOOP”**

2. PROJECT MOTTO

In this project we will mainly focus on Twitter because it is one of the biggest and popular social media platform now a day that's why it is intendedly comfortable to discuss about this website. We are using this twitter data for the business purpose and industrial or social purpose according to our data requirement and processing the data. It is very large amount of sized data increasing every second that is known as **big data**. Because of large amount of data increasing every day we cannot easily analysis this data. We are using here new technology is **HADOOP**. with the help of **HADOOP**, we can easily analysis the large amount of sized data. In this project, we are using **HADOOP** for the analyzing the twitter data.

3. FEATURES OF OUR PROJECT

- A. We can analyze the highly amount sized using Hadoop technology.**
- B. We can filter the highly unstructured, semi-structured and structured data using data mining.**
- C. We can process the sentimental analysis engine on Real-Time based on our choice of keyword.**
- D. We can target our audience using the data analyzed accordingly our marketing strategy.**
- E. We can also use the result of the project to know the opinion of the mass on the particular topic.**
- F. Using the Hadoop platform we can analyze the National Stock Exchange information.**

INTRODUCTION

Twitter is a widely used platform for posting comments and people can express their views and opinions. Sentiment analysis refers to use of natural language processing, text analysis to computational linguistics to identify and extract subjective information in source material. Number of tweets received every year is increased. It is hard to process this huge data.

To analyze this big data, we are using the Hadoop technology in the project. Hadoop is scalable open source framework where Hadoop Technology helps us to perform operations on distributed data in an efficient manner. Hadoop contains a programming model called Map Reduce where it provides an associated implementation for processing and generating big data sets with parallel, distributed algorithm on a cluster. In this Project, we are taking a opinions of the people on a well known person. People expressed their views about the person which helps us to analyze the positive, negative and neutral comments.

DEFINITIONS AND THEORIES

Our proposed architecture includes different methods/steps like:

- 1. Data Source:** There are around millions of twitter users in India as per statistics. Thus, posts tweeted about the service provider are the main source of data.
- 2. Hadoop:** Apache Hadoop is an efficient and scalable open source framework that processes big data in a distributed manner. It consists of HDFS file system and MapReduce engine.
- 3. Data Collection:** One year posted comments or data are taken to analyze the sentiments. A program is designed in Python/Java.
- 4. Naïve Bayes Classification:** Naïve Bayes gave an effective method to carry out the study of classification. It is used to know the word frequency.

5. Data Cleaning and Pre-Processing: Text pre-processing is an important phase for sentiment analysis. It contains Data Cleaning No Repetition Text Correction.

6. Data Analysis: The positive, negative or neutral tweets are analyzed based on key words. The classification is analyzed to find the results of sentiment analysis.

BIG DATA ECOSYSTEM

Big Data is an umbrella term for tools and technologies used to store, process and analyze huge volumes of structured, semi-structured or unstructured data ingested either through bulk transfer or generated in real time with high velocity, within an acceptable elapsed time.

Big Data Ecosystem can be defined as a complex network of interconnected systems.

Some of the major open source technologies in the big data ecosystem are,

Apache Hadoop (Distributed storage & processing framework)

- Distributed Storage: Hadoop Distributed File System (HDFS).
- Distributed Batch Processing: MapReduce 1.0 & 2.0
- Programmability: Apache Pig, Apache Hive
- Data Ingestion: Flume

THE HADOOP ECOSYSTEM

What is Apache Hadoop?

Hadoop is an open-source software platform and framework that helps in storing very large volume of data in a distributed way and it also helps in processing the data across the cluster in parallel for faster execution. It is written in Java.

All the modules in Apache Hadoop are designed with a generic assumption that hardware failures are common and should be automatically handled by the framework.

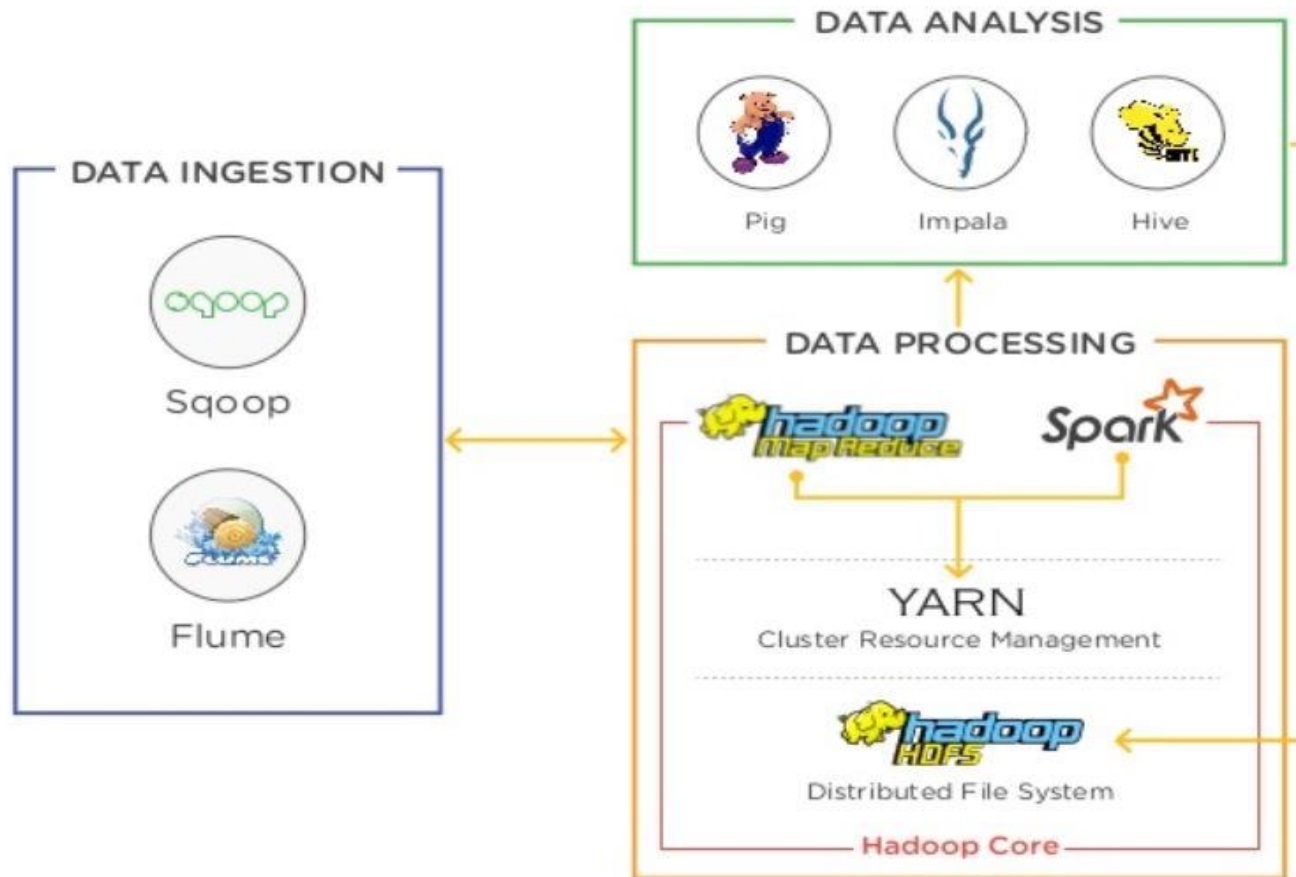
In lay man terms, Apache Hadoop is a open-source framework to run distributed applications on a cluster of machines built from commodity hardware.

We live in the data age. It's not easy to measure the total volume of data stored electronically, but an IDC estimate put the size of the "digital universe" at 0.18 zettabytes in 2006, and is forecasting a tenfold growth by 2011 to 1.8 zettabytes.* A zettabyte is 10^{21} bytes, or equivalently one thousand exabytes, one million petabytes, or one billion terabytes. That's roughly the same order of magnitude as one disk drive for every person in the world.

Hadoop provides us a platform where we can deal with this large amount of the data

Apache Hadoop has 2 major components :-

1. **Hadoop Distributed File System (HDFS)** - The Hadoop distributed file system (HDFS) is a distributed and scalable file system. It is highly flexible to scale up the file-system based on the need of the organisation. HDFS can be termed as distributed file system which is built to store and stream very large data sets in a reliable manner.
2. **MapReduce (MR)** - MapReduce can be termed as a software framework to process vast amount of data in parallel and fault-tolerant manner. It is also referred to as the heart of Hadoop.



Apache Flume

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log data, events (etc...) from various web serves to a centralized data store.

It is a highly reliable, distributed, and configurable tool that is principally designed to transfer streaming data from various sources to HDFS.

Flume Agent

An agent is an independent daemon process (JVM) in Flume. It receives the data (events) from clients or other agents and forwards it to its next destination (sink). Flume may have more than one agent. Following diagram represents a Flume Agent

Source

A source is the component of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.

Example – Avro source, Thrift source, twitter 1% source etc.

Channel

A channel is a transient store which receives the events from the source and buffers them till they are consumed by sinks. It acts as a bridge between the sources and the sinks.

These channels are fully transactional and they can work with any number of sources and sinks.

Example – JDBC channel, File system channel, Memory channel, etc.

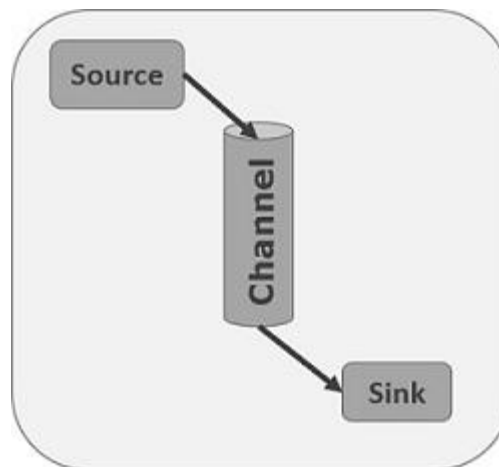
Sink

A sink stores the data into centralized stores like HBase and HDFS. It consumes the data (events) from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.

Example – HDFS sink

DATA FLOW DIAGRAMS

Create a Twitter App



Flume Agent

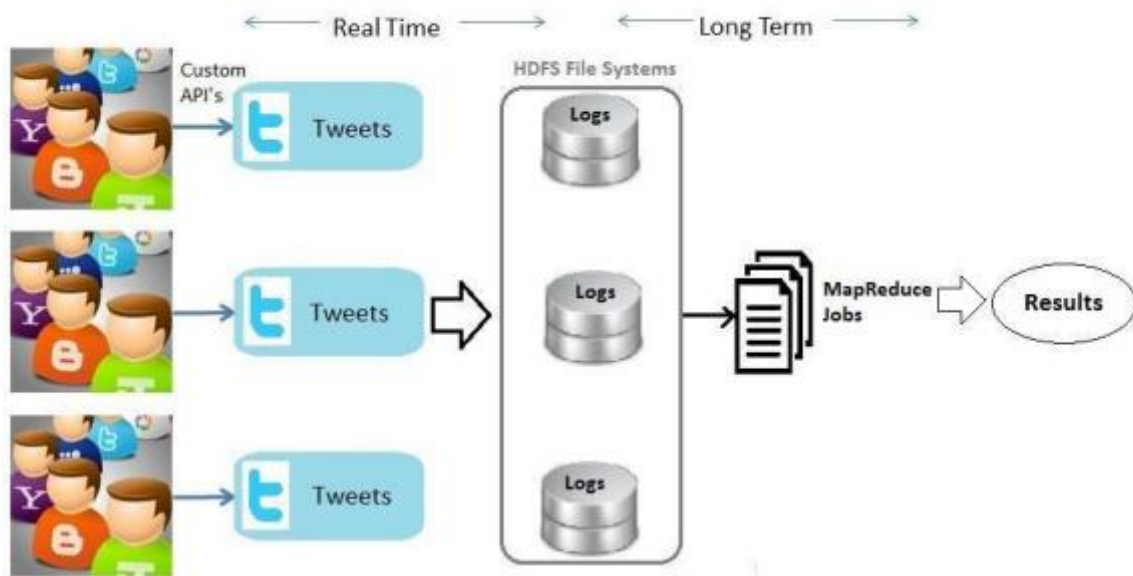
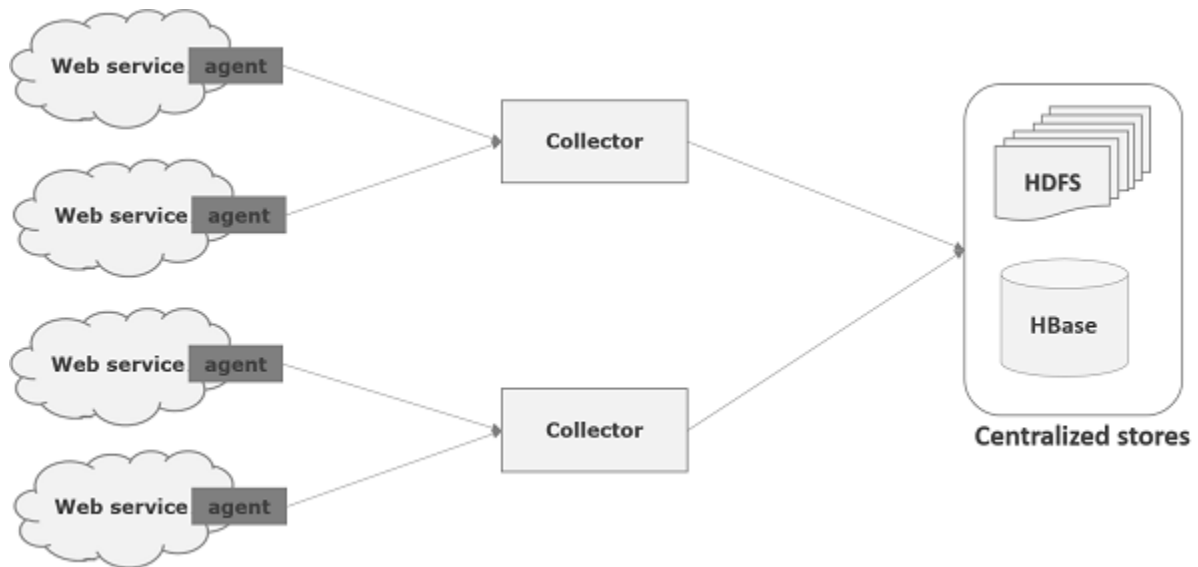


Fig 3: System Architecture

ALGORITHMS

FLUME INSTALLATION:-

For Installation:

<https://www.eduonix.com/blog/bigdata-and-hadoop/flume-installation-and-streaming-twitter-data-using-flume/>

Installation

Step 1: First download or copy the requisite tar file

cd /home/training/downloads

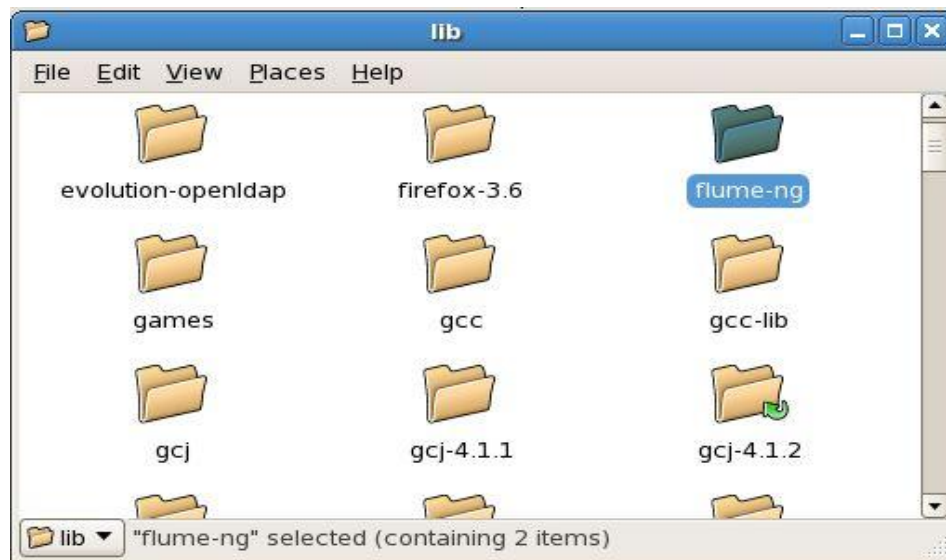


or,

If you have the file downloaded then just copy and paste it to
/home/training/downloads directory

Step 2:- Now create a directory in **/usr/lib/** and move the tar file into it

sudo mkdir /usr/lib/flume-ng



➤ **sudo mv /home/training/downloads/apache-flume-1.4.0-bin.tar.gz
/usr/lib/flume-ng/**



- `cd /usr/lib/flume-ng/`
- `sudo tar -xvf apache-flume-1.4.0-bin.tar.gz`
- `cd apache-flume-1.4.0-bin`

Step 3: Now copy the flume-sources-1.0-SNAPSHOT.jar file into downloads and copy it to lib directory of flume directory



We need to download flume-source-1.0-SNAPSHOT.jar file from internet or copy and paste it to downloads directory.

- `sudo mv /home/training/downloads/flume-sources-1.0-SNAPSHOT.jar /usr/lib/flume-ng/apache-flume-1.4.0-bin/lib/`
- `sudo chmod 777 /usr/lib/flume-ng/apache-flume-1.4.0-bin/lib/flume-sources-1.0-SNAPSHOT.jar`

Step 4: Now get into conf directory of flume and create the property file.

- `cd ..`
- `cd conf/`
- `ls -ltr`
- `sudo cp flume-env.sh.template flume-env.sh`
- `sudo gedit flume-env.sh`

Remove the # sign before java and that is

- `JAVA_HOME=/usr/lib/jvm/java-6-sun`
- `FLUME_CLASSPATH= '/usr/lib/flume-ng/apache-flume-1.4.0-bin/lib/flume-source-1.0-SNAPSHOT.jar'`

Step 5:

Now go to bin folder to execute the flume:

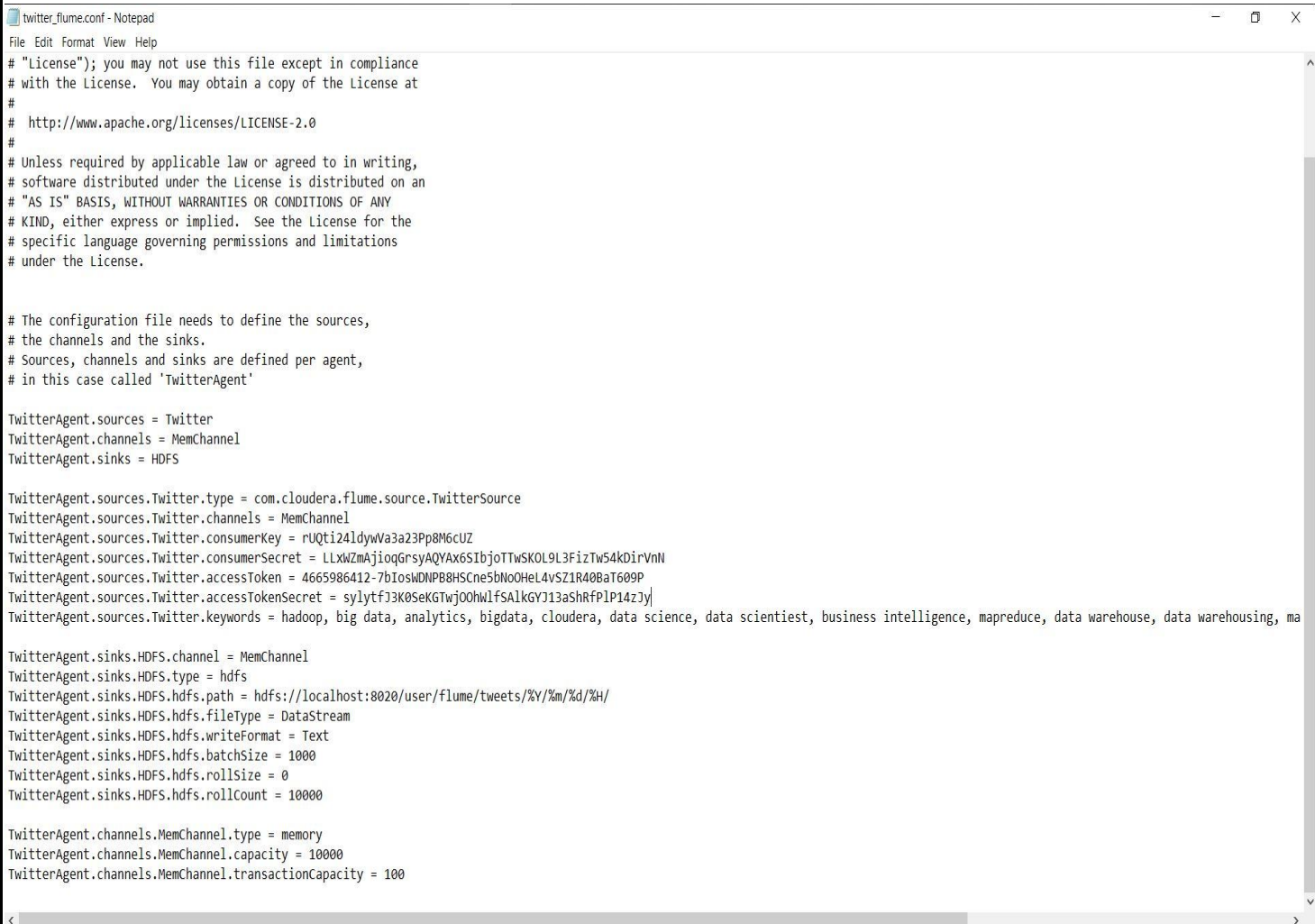
- `cd..`
- `cd bin/`

FLUME EXECUTION:-

- Always copy the execution file in the conf directory of flume in this case it will be:

`/usr/lib/flume-ng/apache-flume-1.4.0-bin/conf/`

- copy the file into conf directory----->twitter_flume.conf.txt



```
twitter_flume.conf - Notepad
File Edit Format View Help
# "License"); you may not use this file except in compliance
# with the License. You may obtain a copy of the License at
#
# http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing,
# software distributed under the license is distributed on an
# "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY
# KIND, either express or implied. See the license for the
# specific language governing permissions and limitations
# under the license.

# The configuration file needs to define the sources,
# the channels and the sinks.
# Sources, channels and sinks are defined per agent,
# in this case called 'TwitterAgent'

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

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = rUQtI24ldyWVa3a23Pp8M6cUZ
TwitterAgent.sources.Twitter.consumerSecret = LLxWZmAjloqGrSyAQYAX6SIbjoTTWSKOL9L3FizTw54kDirVnN
TwitterAgent.sources.Twitter.accessToken = 4665986412-7bIosWdNpPB8HSCne5bNo0Hel4vSZ1R40BaT609P
TwitterAgent.sources.Twitter.accessTokenSecret = sylytfj3K0SeKGTWj00HwlfSAlkGYJ13aShRfPlP14zJy
TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientiest, business intelligence, mapreduce, data warehouse, data warehousing, ma

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/flume/tweets/%Y/%m/%d/%H/
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.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Create a Twitter App



Apps / College_Project_Demo

App details **Keys and tokens** Permissions

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

rUQt24ldywVa3a23Pp8M6cUZ (API key)

NEW

LLxWZmAjioqGrsyAQYAx6SIbjoTTwSKOL9L3FizTw54kDirVnN (API secret key)

NEW

Regenerate

Access token & access token secret

4665986412-7blosWDNPB8HSCne5bNoOHeL4vSZ1R40BaT609P (Access token)

syhtfj3K0SeKGTwjOOHwIISAlkGY13aShRfIP14zhy (Access token secret)

Read and write (Access level)

Revoke

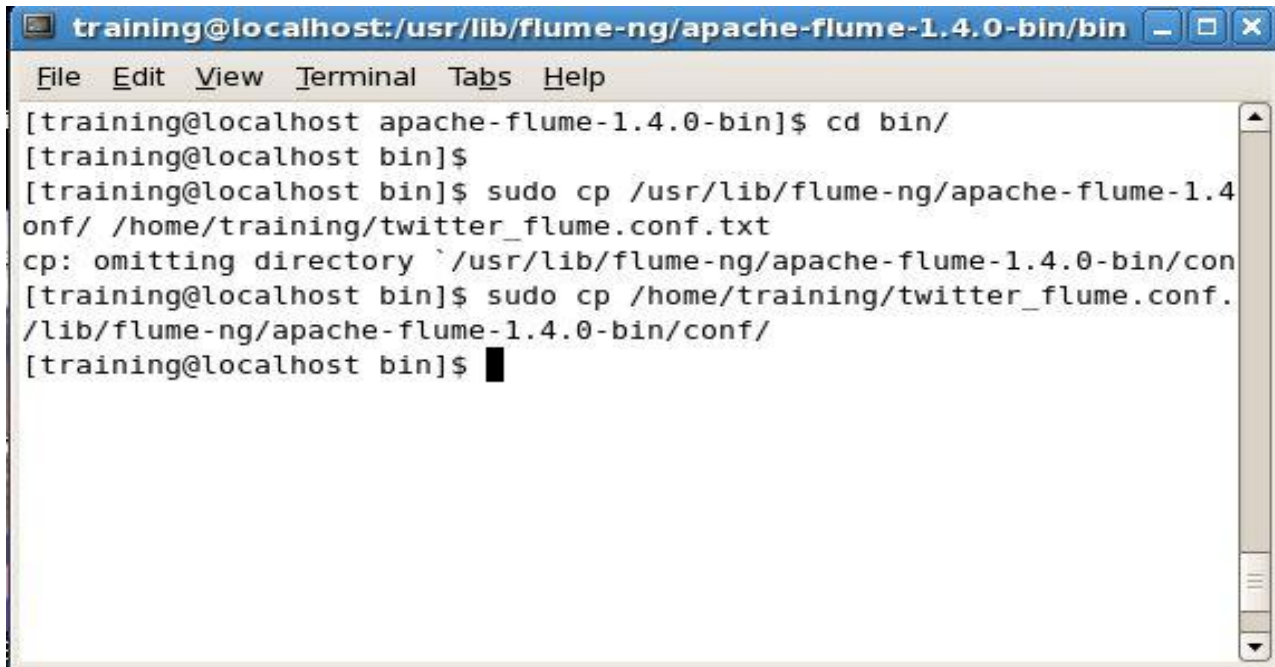
Regenerate

Developer policy and terms

Follow @twitterdev

Subscribe to developer news

- Then execute this command from /usr/lib/flume-ng/apache-flume-1.4.0-bin/bin/ directory

A terminal window titled 'training@localhost:/usr/lib/flume-ng/apache-flume-1.4.0-bin/bin' with standard window controls. The terminal shows the following commands and output:

```
[training@localhost apache-flume-1.4.0-bin]$ cd bin/
[training@localhost bin]$
[training@localhost bin]$ sudo cp /usr/lib/flume-ng/apache-flume-1.4
onf/ /home/training/twitter_flume.conf.txt
cp: omitting directory `/usr/lib/flume-ng/apache-flume-1.4.0-bin/con
[training@localhost bin]$ sudo cp /home/training/twitter_flume.conf.
/lib/flume-ng/apache-flume-1.4.0-bin/conf/
[training@localhost bin]$
```

- ./flume-ng agent -n TwitterAgent -c conf -f /usr/lib/flume-ng/apache-flume-1.4.0-bin/conf/twitter_flume.conf.txt

SYSTEM REQUIREMENTS FOR IMPLEMENTATION


➤ **HARDWARE REQUIREMENTS:**


1. Intel or AMD Quad core 1.5GHz Processor
2. 4GB RAM
3. 50GB Free Disk Space or more
4. Internet Access.

➤ **SOFTWARE REQUIREMENTS:**

1. Virtual Machine (VM Fusion).
2. Any Linux Operating System (Ex. CentOS).
3. Cloudera Platform.

RESULTS & TABLES



 HDFS:/user

Contents of directory /user

Goto:

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
flume	dir				2018-11-30 05:27	rwxf-xr-x	training	supergroup
hive	dir				2011-05-02 17:51	rwxf-xr-x	hue	supergroup
training	dir				2011-05-06 15:09	rwxf-xr-x	training	supergroup

[Go back to DFS home](#)

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1543584444869	file	37.89 KB	1	64 MB	2018-11-30 05:27	rw-r--r--	training	supergroup
FlumeData.1543584444870	file	35.68 KB	1	64 MB	2018-11-30 05:28	rw-r--r--	training	supergroup
FlumeData.1543584444871	file	44.51 KB	1	64 MB	2018-11-30 05:28	rw-r--r--	training	supergroup
FlumeData.1543584444872	file	61.54 KB	1	64 MB	2018-11-30 05:29	rw-r--r--	training	supergroup
FlumeData.1543584444873	file	87.05 KB	1	64 MB	2018-11-30 05:29	rw-r--r--	training	supergroup
FlumeData.1543584444874	file	46.27 KB	1	64 MB	2018-11-30 05:30	rw-r--r--	training	supergroup
FlumeData.1543584444875	file	61.41 KB	1	64 MB	2018-11-30 05:31	rw-r--r--	training	supergroup
FlumeData.1543584444876	file	32.79 KB	1	64 MB	2018-11-30 05:31	rw-r--r--	training	supergroup
FlumeData.1543584444877	file	74.81 KB	1	64 MB	2018-11-30 05:32	rw-r--r--	training	supergroup
FlumeData.1543584444878	file	31.96 KB	1	64 MB	2018-11-30 05:32	rw-r--r--	training	supergroup
FlumeData.1543584444879	file	60.29 KB	1	64 MB	2018-11-30 05:33	rw-r--r--	training	supergroup
FlumeData.1543584444880	file	34.67 KB	1	64 MB	2018-11-30 05:33	rw-r--r--	training	supergroup

File: /user/flume/tweets/2018/11/30/05/FlumeData.1543584444869

Goto:

[Go back to dir listing](#)

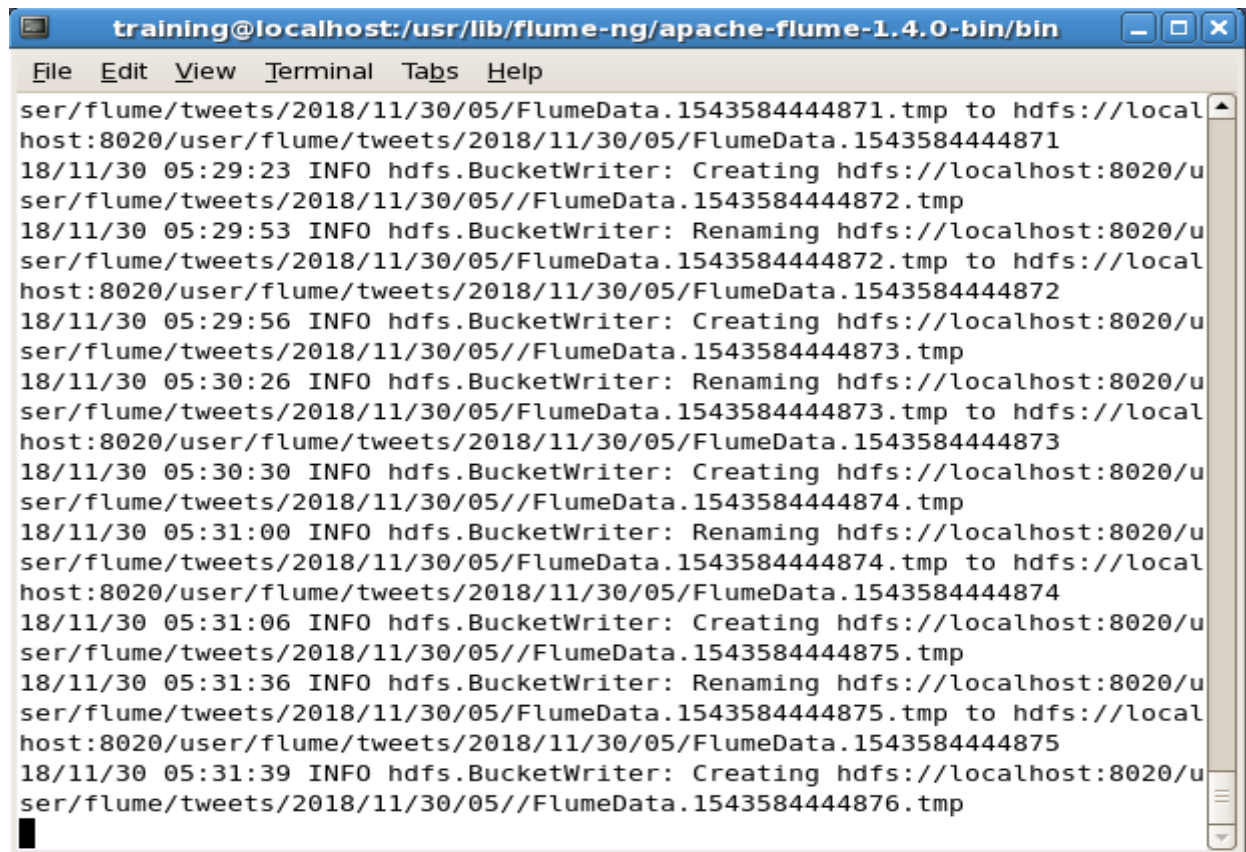
[Advanced view/download options](#)

[View Next chunk](#)

```
{*filter_level": "low", "quote_count": 0, "retweeted": false, "in_reply_to_screen_name": null, "truncated": false, "lang": "en", "in_reply_to_status_id_str": null, "id": 1068496699318915073, "in_reply_to_user_id_str": null, "timestamp_ms": "1543584439483", "in_reply_to_status_id": null, "created_at": "Fri Nov 30 13:27:19 +0000 2018", "favorite_count": 0, "place": null, "coordinates": null, "text": "RT @VicomComputer: Outstanding @VicomComputer event at Citi Field featuring Paul Zikopoulos (@BigData_paulz) of IBM discussing Big Data, as\u2026", "contributors": null, "retweeted_status": null, "filter_level": "low", "quote_count": 0, "retweeted": false, "in_reply_to_screen_name": null, "possibly_sensitive": false, "truncated": true, "lang": "en", "in_reply_to_status_id_str": null, "id": 1011644205981200384, "in_reply_to_user_id_str": null, "in_reply_to_status_id": null, "created_at": "Tue Jun 26 16:15:48 +0000 2018", "favorite_count": 6, "display_text_range": [0, 140], "place": null, "coordinates": null, "text": "Outstanding @VicomComputer event at Citi Field featuring Paul Zikopoulos (@BigData_paulz) of IBM discussing Big Data\u2026", "contributors": null, "geo": null, "reply_count": 0, "extended_tweet": {"media": [{"sizes": {"small": {"w": 680, "h": 355, "thumb": {"w": 150, "h": 150, "crop": "h": 150}, "medium": {"w": 1200, "h": 627, "thumb": {"w": 150, "h": 150, "crop": "h": 150}, "large": {"w": 1200, "h": 627, "thumb": {"w": 150, "h": 150, "crop": "h": 150}}, "id": 1011644203720499202, "media_url_https": "https://pbs.twimg.com/media/DgoVCxN4AI9JJ.jpg", "media_url": "http://pbs.twimg.com/media/DgoVCxN4AI9JJ.jpg", "expanded_url": "https://twitter.com/VicomComputer/status/1011644205981200384/photo/1", "indices": [281, 304], "id_str": "1011644203720499202", "type": "photo", "display_url": "pic.twitter.com/Z04unYH7GZ", "url": "https://t.co/Z04unYH7GZ"}, {"sizes": {"small": {"w": 680, "h": 355, "thumb": {"w": 150, "h": 150, "crop": "h": 150}, "large": {"w": 1200, "h": 627, "thumb": {"w": 150, "h": 150, "crop": "h": 150}}, "id": 1011644203766558720, "media_url_https": "https://pbs.twimg.com/media/DgoVCxYsAAUE_D.jpg", "media_url": "http://pbs.twimg.com/media/DgoVCxYsAAUE_D.jpg", "expanded_url": "https://twitter.com/VicomComputer/status/1011644205981200384/photo/1", "indices": [281, 304], "id_str": "1011644203766558720", "type": "photo", "display_url": "pic.twitter.com/Z04unYH7GZ", "url": "https://t.co/Z04unYH7GZ"}], "display_text_range": [0, 280], "entities": {"symbols": [], "urls": [], "hashtags": [], "media": [{"sizes": {"small": {"w": 680, "h": 355, "thumb": {"w": 150, "h": 150, "crop": "h": 150}, "medium": {"w": 1200, "h": 627, "thumb": {"w": 150, "h": 150, "crop": "h": 150}}, "id": 1011644203720499202, "media_url_https": "https://pbs.twimg.com/media/DgoVCxN4AI9JJ.jpg", "media_url": "http://pbs.twimg.com/media/DgoVCxN4AI9JJ.jpg", "expanded_url": "https://twitter.com/VicomComputer/status/1011644205981200384/photo/1", "indices": [281, 304], "id_str": "1011644203720499202", "type": "photo", "display_url": "pic.twitter.com/Z04unYH7GZ", "url": "https://t.co/Z04unYH7GZ"}, {"sizes": {"small": {"w": 680, "h": 355, "thumb": {"w": 150, "h": 150, "crop": "h": 150}, "large": {"w": 1200, "h": 627, "thumb": {"w": 150, "h": 150, "crop": "h": 150}}, "id": 1011644203766558720, "media_url_https": "https://pbs.twimg.com/media/DgoVCxYsAAUE_D.jpg", "media_url": "http://pbs.twimg.com/media/DgoVCxYsAAUE_D.jpg", "expanded_url": "https://twitter.com/VicomComputer/status/1011644205981200384/photo/1", "indices": [281, 304], "id_str": "1011644203766558720", "type": "photo", "display_url": "pic.twitter.com/Z04unYH7GZ", "url": "https://t.co/Z04unYH7GZ"}, "user_mentions": [{"id": 633212369, "name": "Vicom Computer Services, Inc.", "indices": [12, 26], "screen_name": "VicomComputer", "id_str": "633212369"}, {"id": 77449588, "name": "Paul Zikopoulos", "indices": [74, 88], "screen_name": "BigData_paulz", "id_str": "77449588"}], "full_text": "Outstanding @VicomComputer event at Citi Field featuring Paul Zikopoulos (@BigData_paulz) of IBM discussing Big Data, as well as IBM's compute & AI solutions to address today's exponential data growth. Excellent talk with very relevant content. Thanks to all that attended! https://t.co/Z04unYH7GZ", "entities": {"symbols": [], "urls": [{"expanded_url": "https://twitter.com/i/web/status/1011644205981200384", "indices": [117, 140], "display_url": "twitter.com/i/web/status/1011644205981200384", "url": "https://t.co/CL49wesGLQ"}], "hashtags": [], "user_mentions": [{"id": 633212369, "name": "Vicom Computer Services, Inc.", "indices": [12, 26], "screen_name": "VicomComputer", "id_str": "633212369"}, {"id": 77449588, "name": "Paul Zikopoulos", "indices": [74, 88], "screen_name": "BigData_paulz", "id_str": "77449588"}]}
```

```
100 "expanded_url": "https://twitter.com/VicomComputer/status/1011644205981200384/photo/1",
101 "indices": [281, 304],
102 "id_str": "1011644203766558720",
103 "type": "photo",
104 "display_url": "pic.twitter.com/Z04unYH7GZ",
105 "url": "https://t.co/Z04unYH7GZ"
106 }
107 },
108 "display_text_range": [0, 280],
109 "entities": {
110 "symbols": [],
111 "urls": [],
112 "hashtags": [],
113 "media": [{
```

SCREENSHOTS

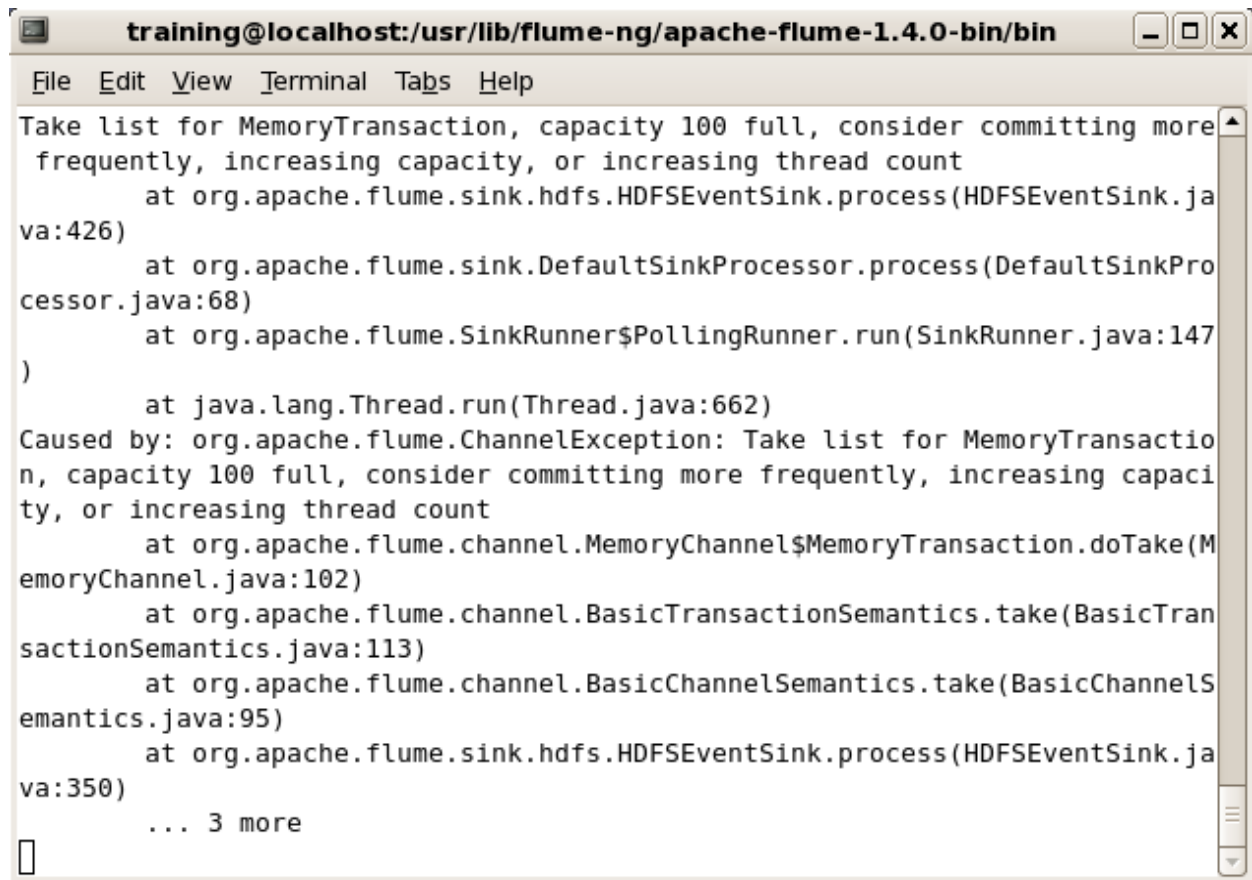


The screenshot shows a terminal window titled "training@localhost:usr/lib/flume-ng/apache-flume-1.4.0-bin/bin". The terminal displays a series of log messages from the HDFS BucketWriter, showing the creation and renaming of temporary files for tweet data. The messages are as follows:

```
ser/flume/tweets/2018/11/30/05/FlumeData.1543584444871.tmp to hdfs://local
host:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543584444871
18/11/30 05:29:23 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05//FlumeData.1543584444872.tmp
18/11/30 05:29:53 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05/FlumeData.1543584444872.tmp to hdfs://local
host:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543584444872
18/11/30 05:29:56 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05//FlumeData.1543584444873.tmp
18/11/30 05:30:26 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05/FlumeData.1543584444873.tmp to hdfs://local
host:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543584444873
18/11/30 05:30:30 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05//FlumeData.1543584444874.tmp
18/11/30 05:31:00 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05/FlumeData.1543584444874.tmp to hdfs://local
host:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543584444874
18/11/30 05:31:06 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05//FlumeData.1543584444875.tmp
18/11/30 05:31:36 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05/FlumeData.1543584444875.tmp to hdfs://local
host:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543584444875
18/11/30 05:31:39 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/u
ser/flume/tweets/2018/11/30/05//FlumeData.1543584444876.tmp
```

```
training@localhost:/usr/lib/flume-ng/apache-flume-1.4.0-bin/bin
File Edit View Terminal Tabs Help
nfiguration provider starting
18/11/30 05:27:13 INFO node.PollingPropertiesFileConfigurationProvider: Re
loading configuration file:/usr/lib/flume-ng/apache-flume-1.4.0-bin/conf/t
witter_flume.conf.txt
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: T
witterAgent
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Processing:HDFS
18/11/30 05:27:13 INFO conf.FlumeConfiguration: Post-validation flume conf
iguration contains configuration for agents: [TwitterAgent]
18/11/30 05:27:13 INFO node.AbstractConfigurationProvider: Creating channel
s
18/11/30 05:27:13 INFO channel.DefaultChannelFactory: Creating instance of
channel MemChannel type memory
18/11/30 05:27:13 INFO node.AbstractConfigurationProvider: Created channel
MemChannel
18/11/30 05:27:13 INFO source.DefaultSourceFactory: Creating instance of s
ource Twitter, type com.cloudera.flume.source.TwitterSource
18/11/30 05:27:13 INFO sink.DefaultSinkFactory: Creating instance of sink:
```

```
training@localhost:/usr/lib/flume-ng/apache-flume-1.4.0-bin/bin
File Edit View Terminal Tabs Help
r/flume/tweets/2018/11/30/05//FlumeData.1543583602039.tmp
18/11/30 05:29:24 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05/FlumeData.1543583602039.tmp to hdfs://localhost
:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543583602039
18/11/30 05:29:24 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05//FlumeData.1543583602040.tmp
18/11/30 05:29:54 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05/FlumeData.1543583602040.tmp to hdfs://localhost
:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543583602040
18/11/30 05:29:56 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05//FlumeData.1543583602041.tmp
18/11/30 05:30:26 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05/FlumeData.1543583602041.tmp to hdfs://localhost
:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543583602041
18/11/30 05:30:27 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05//FlumeData.1543583602042.tmp
18/11/30 05:30:57 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05/FlumeData.1543583602042.tmp to hdfs://localhost
:8020/user/flume/tweets/2018/11/30/05/FlumeData.1543583602042
18/11/30 05:30:58 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/use
r/flume/tweets/2018/11/30/05//FlumeData.1543583602043.tmp
```



```
training@localhost:/usr/lib/flume-ng/apache-flume-1.4.0-bin/bin
File Edit View Terminal Tabs Help
Take list for MemoryTransaction, capacity 100 full, consider committing more frequently, increasing capacity, or increasing thread count
    at org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:426)
    at org.apache.flume.sink.DefaultSinkProcessor.process(DefaultSinkProcessor.java:68)
    at org.apache.flume.SinkRunner$PollingRunner.run(SinkRunner.java:147)
)
    at java.lang.Thread.run(Thread.java:662)
Caused by: org.apache.flume.ChannelException: Take list for MemoryTransaction, capacity 100 full, consider committing more frequently, increasing capacity, or increasing thread count
    at org.apache.flume.channel.MemoryChannel$MemoryTransaction.doTake(MemoryChannel.java:102)
    at org.apache.flume.channel.BasicTransactionSemantics.take(BasicTransactionSemantics.java:113)
    at org.apache.flume.channel.BasicChannelSemantics.take(BasicChannelSemantics.java:95)
    at org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:350)
    ... 3 more
```

CONCLUSION & FUTURE SCOPE

This project concludes that various classes have been analyzed on tweets of well-known person. It is also useful in gauging the opinions of people when it comes to disserve topics related to any fields. In our Case study, we can further compare the service of various provides and judge which one is best.

As future Work, we can Further compare the reviews of various persons and judge who is best. Hadoop Map-Reduce and native algorithm, we can provide a simple automated method to evaluate what people information from social networks and analyzing it using Big Data techniques has left behind the traditional think.

BIBLIOGRAPHY

WEBSITES REFFERED:

- www.geeksforgeeks.org
- www.slideshare.net
- www.tutorialspoint.com
- www.cloudera.com/training
- www.youtube.com