Vivekanand Education Society's
Institute of Technology
Approved by AICTE & Affiliated to University of Mumbai

Artificial Intelligence and Data Science Department

BDA/Odd Sem 202323/Experiment 8

Name : Mayur Pimpude	**Class/Roll No. :D16AD/43**	**Grade :**

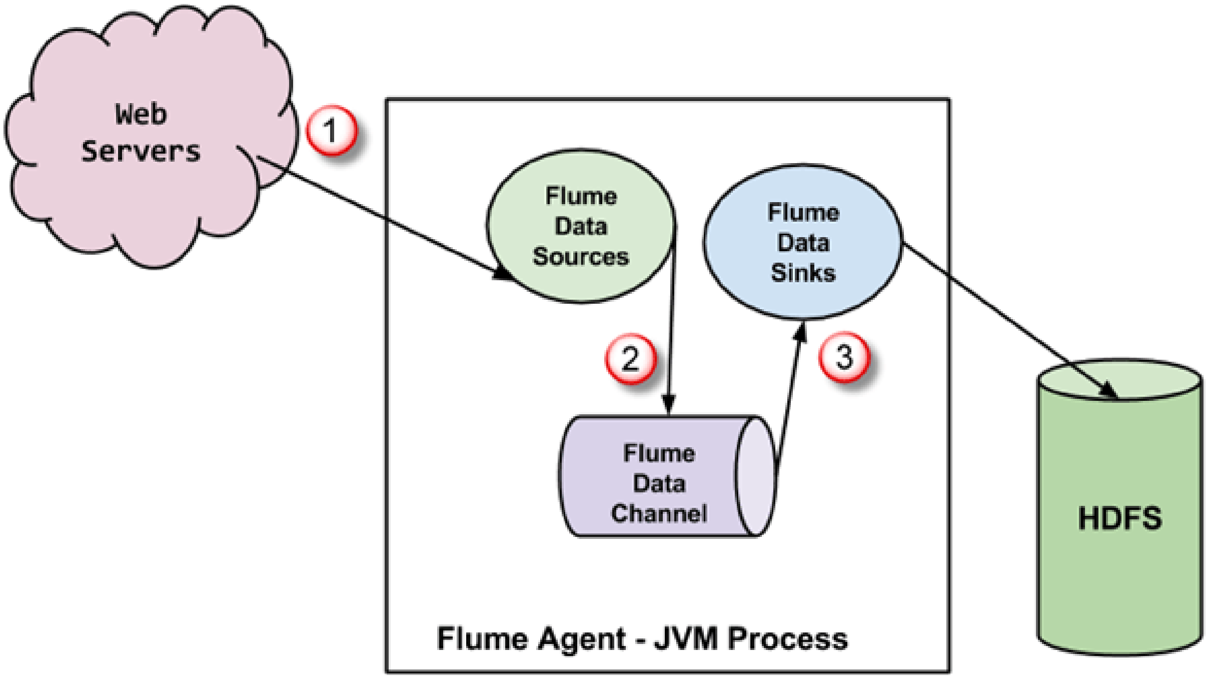
Aim: To study Twitter data analysis using flume.

What is Apache Flume in Hadoop?

Apache Flume is a reliable and distributed system for collecting, aggregating and moving massive quantities of log data. It has a simple yet flexible architecture based on streaming data flows. Apache Flume is used to collect log data present in log files from web servers and aggregate it into HDFS for analysis.

Flume Architecture

A **Flume agent** is a **JVM** process which has 3 components –**Flume Source, Flume Channel** and **Flume Sink**– through which events propagate after initiated at an external source.



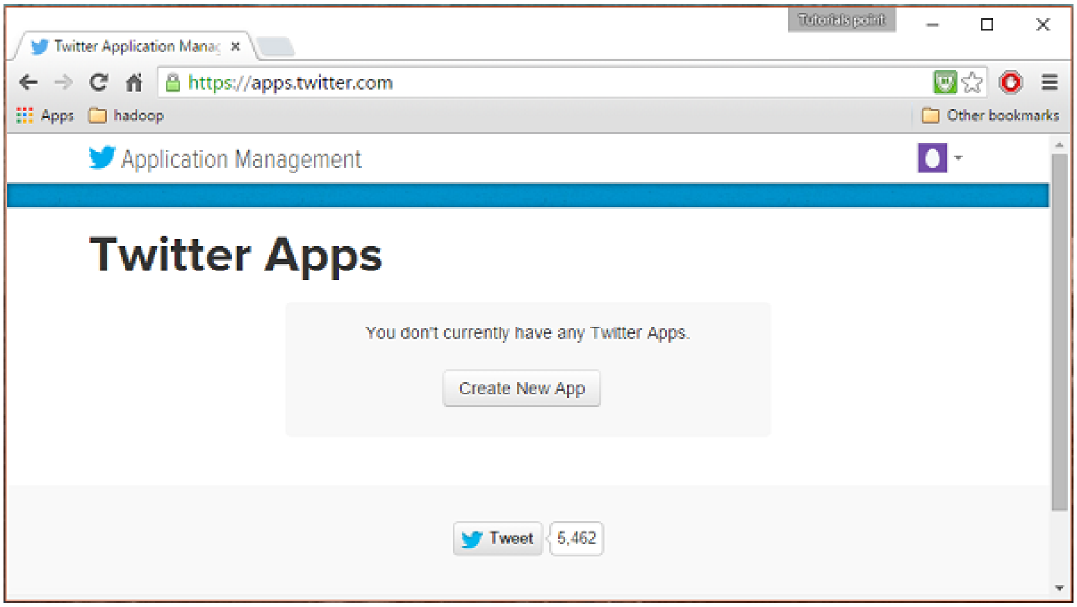
Flume Architecture

1. In the above diagram, the events generated by an external source (WebServer) are consumed by Flume Data Source. The external source sends events to the Flume source in a format that is recognized by the target source.
2. Flume Source receives an event and stores it into one or more channels. The channel acts as a store which keeps the event until it is consumed by the flume sink. This channel may use a local file system in order to store these events.
3. Flume sink removes the event from a channel and stores it into an external repository like e.g., HDFS. There could be multiple flume agents, in which case flume sinks forwards the event to the flume source of the next flume agent in the flow.

Output:

Step 1

To create a Twitter application, click on the following link https://apps.twitter.com/.

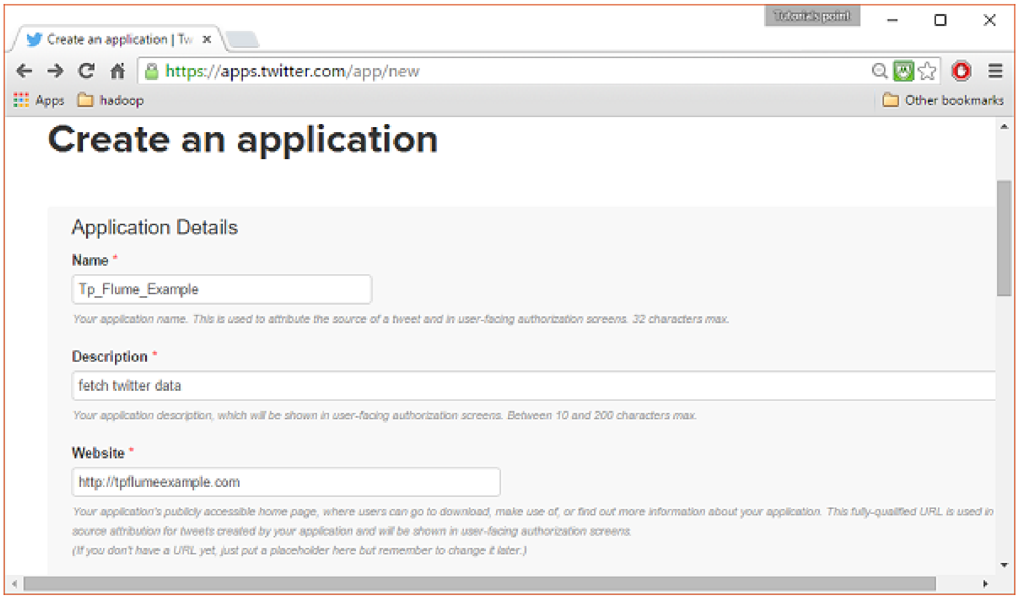


Twitter Apps

You don't currently have any Twitter Apps.

Step 2

Click on the Create New App button. You will be redirected to a window where you will get an application form in which you have to fill in your details in order to create the App.



Create an application

Application Details

Name *

Tp_Flume_Example

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

Description *

fetch twitter data

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

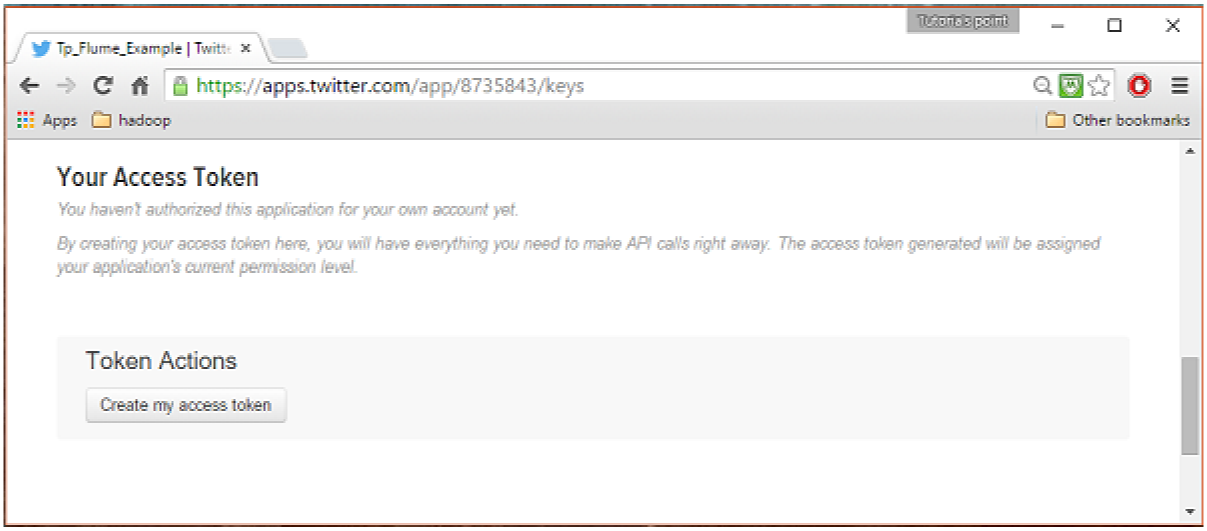
Website *

http://tpflumeexample.com

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Step 4

Under keys and Access Tokens tab at the bottom of the page,



Your Access Token

You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assigned your application's current permission level.

Token Actions

Step 1: Install / Verify Hadoop

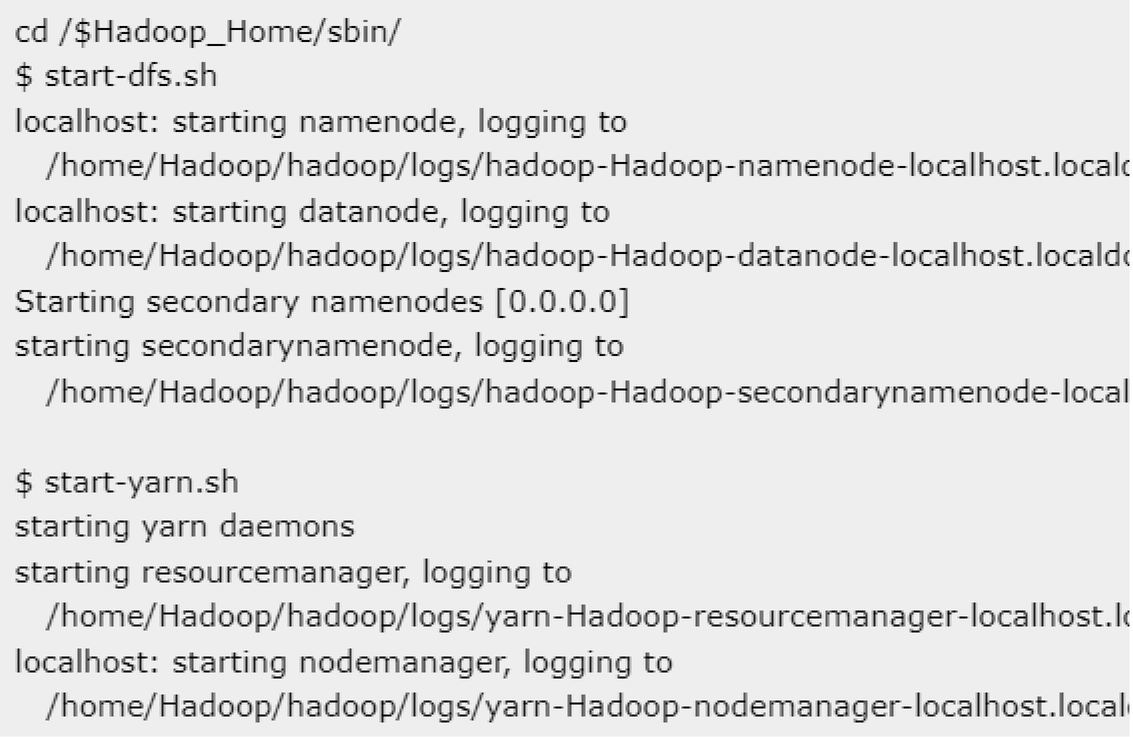
Install Hadoop. If Hadoop is already installed in your system,

```
$ hadoop version
```

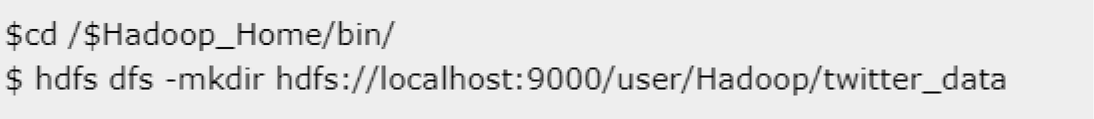
```
Hadoop 2.6.0
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r
e3496499ecb8d220fba99dc5ed4c99c8f9e33bb1
Compiled by jenkins on 2014-11-13T21:10Z
Compiled with protoc 2.5.0
From source with checksum 18e43357c8f927c0695f1e9522859d6a
This command was run using /home/Hadoop/hadoop/share/hadoop/comn
```

Step 2: Starting Hadoop

Browse through the sbin directory of Hadoop and start yarn and Hadoop dfs (distributed file system) as shown below.



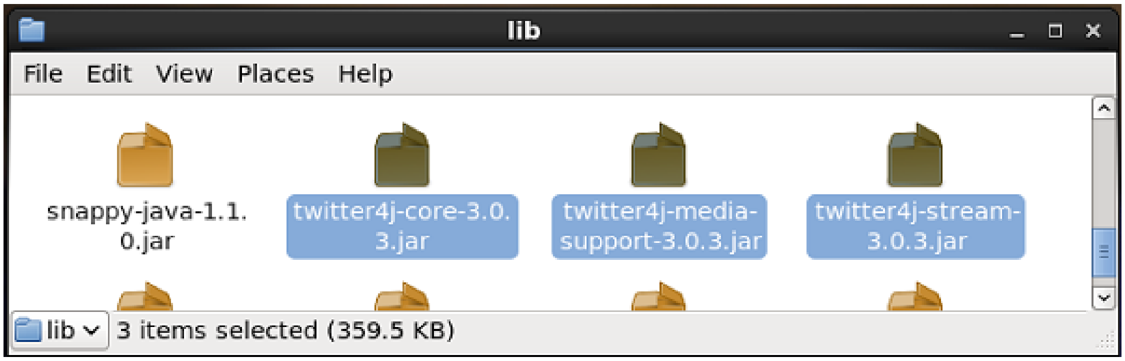
```
cd /$Hadoop_Home/sbin/
$ start-dfs.sh
localhost: starting namenode, logging to
   /home/Hadoop/hadoop/logs/hadoop-Hadoop-namenode-localhost.locald
localhost: starting datanode, logging to
   /home/Hadoop/hadoop/logs/hadoop-Hadoop-datanode-localhost.locald
Starting secondary namenodes [0.0.0.0]
starting secondarynamenode, logging to
   /home/Hadoop/hadoop/logs/hadoop-Hadoop-secondarynamenode-local

$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to
   /home/Hadoop/hadoop/logs/yarn-Hadoop-resourcemanager-localhost.lo
localhost: starting nodemanager, logging to
   /home/Hadoop/hadoop/logs/yarn-Hadoop-nodemanager-localhost.local
```



```
$cd /$Hadoop_Home/bin/
$ hdfs dfs -mkdir hdfs://localhost:9000/user/Hadoop/twitter_data
```

Configuring Flume

We have to configure the source, the channel, and the sink using the configuration file in the conf folder.



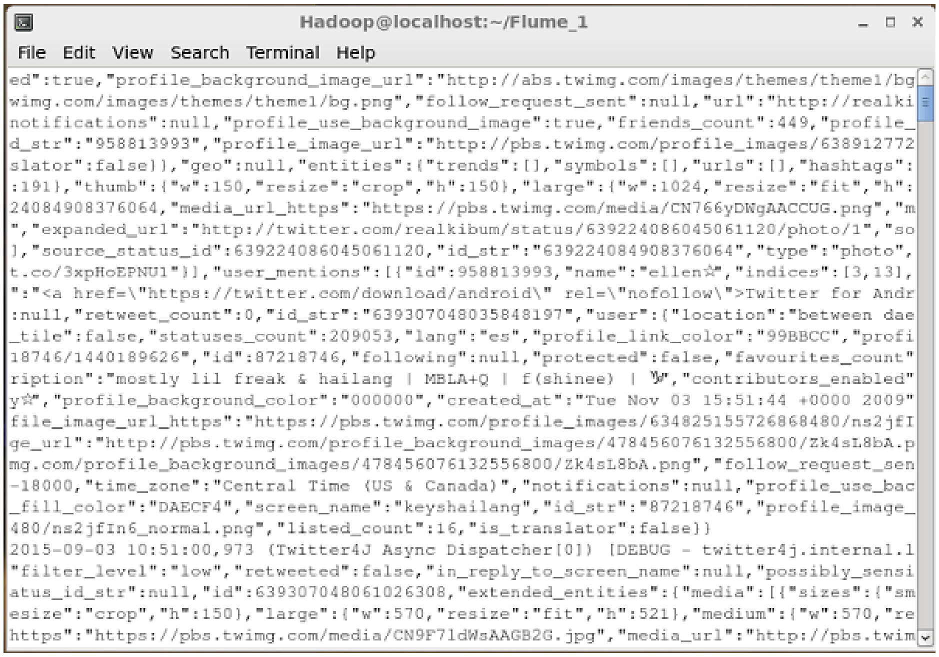
3 items selected (359.5 KB)

```
export CLASSPATH=$CLASSPATH:/FLUME_HOME/lib/*
```

Execution

Browse through the Flume home directory and execute the application as shown below.

```
$ cd $FLUME_HOME
$ bin/flume-ng agent --conf ./conf/ -f conf/twitter.conf
Dflume.root.logger=DEBUG,console -n TwitterAgent
```

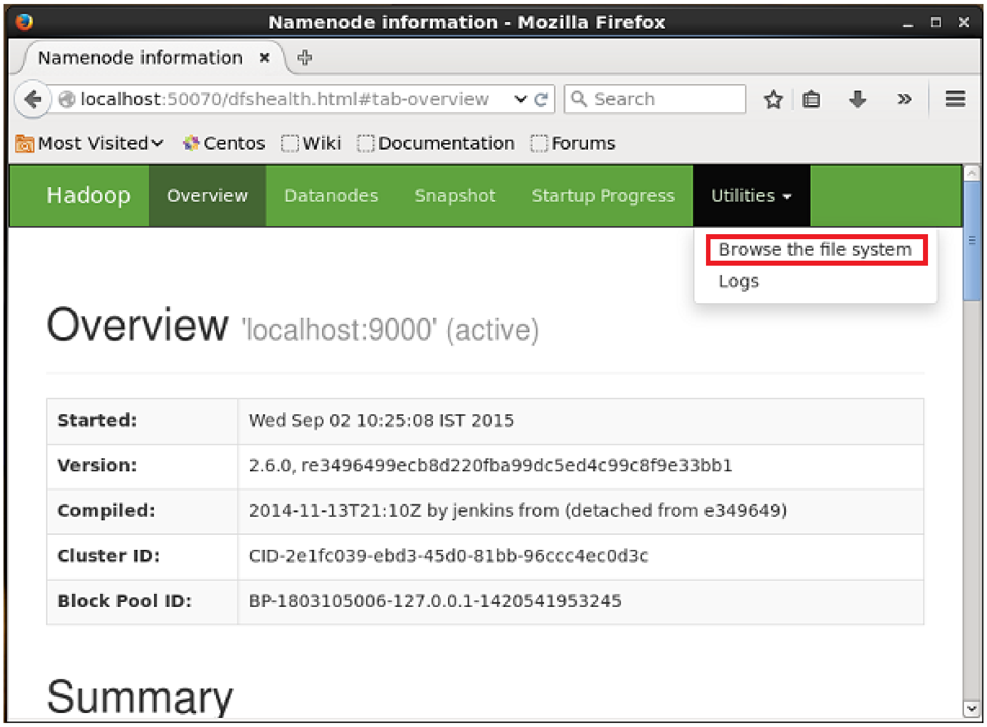


Hadoop@localhost:~/Flume_1
File Edit View Search Terminal Help

```
ed":true,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg
wimg.com/images/themes/theme1/bg.png","follow_request_sent":null,"url":"http://realki
notifications":null,"profile_use_background_image":true,"friends_count":449,"profile_
d_str":"958813993","profile_image_url":"http://pbs.twimg.com/profile_images/638912772
slator":false}},"geo":null,"entities":{"trends":[],"symbols":[],"urls":[],"hashtags":
:191},"thumb":{"w":150,"resize":"crop","h":150},"large":{"w":1024,"resize":"fit","h":
24084908376064,"media_url_https":"https://pbs.twimg.com/media/CN766yDWgAACCUG.png","m
","expanded_url":"http://twitter.com/realkibum/status/639224086045061120/photo/1","so
],"source_status_id":639224086045061120,"id_str":"639224084908376064","type":"photo",
t.co/3xpHoEPNU1"}],"user_mentions":[{"id":958813993,"name":"ellen☆","indices":[3,13],
":"<a href=\"https://twitter.com/download/android\" rel=\"nofollow\">Twitter for Andr
:null,"retweet_count":0,"id_str":"639307048035848197","user":{"location":"between dae
_tile":false,"statuses_count":209053,"lang":"es","profile_link_color":"99BBCC","profi
18746/1440189626","id":87218746,"following":null,"protected":false,"favourites_count"
ription":"mostly lil freak & hailang | MBLA+Q | f(shinee) | ❦","contributors_enabled"
y☆","profile_background_color":"000000","created_at":"Tue Nov 03 15:51:44 +0000 2009"
file_image_url_https":"https://pbs.twimg.com/profile_images/634825155726868480/ns2jfI
ge_url":"http://pbs.twimg.com/profile_background_images/478456076132556800/Zk4sL8bA.p
mg.com/profile_background_images/478456076132556800/Zk4sL8bA.png","follow_request_sen
-18000,"time_zone":"Central Time (US & Canada)","notifications":null,"profile_use_bac
_fill_color":"DAECF4","screen_name":"keyshailang","id_str":"87218746","profile_image_
480/ns2jfIn6_normal.png","listed_count":16,"is_translator":false}}
2015-09-03 10:51:00,973 (Twitter4J Async Dispatcher[0]) [DEBUG - twitter4j.internal.l
"filter_level":"low","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensi
atus_id_str":null,"id":639307048061026308,"extended_entities":{"media":[{"sizes":{"sm
esize":"crop","h":150},"large":{"w":570,"resize":"fit","h":521},"medium":{"w":570,"re
https":"https://pbs.twimg.com/media/CN9F71dWsAAGB2G.jpg","media_url":"http://pbs.twim
```

Verifying HDFS

```
http://localhost:50070/
```



Overview 'localhost:9000' (active)

Started:	Wed Sep 02 10:25:08 IST 2015
Version:	2.6.0, re3496499ecb8d220fba99dc5ed4c99c8f9e33bb1
Compiled:	2014-11-13T21:10Z by jenkins from (detached from e349649)
Cluster ID:	CID-2e1fc039-ebd3-45d0-81bb-96ccc4ec0d3c
Block Pool ID:	BP-1803105006-127.0.0.1-1420541953245

Summary

CONCLUSION:

The study concluded that Flume technology can be used to extract real-time data from Twitter and store it in HDFS for further analysis .