# Big Data Hadoop Training

Session 11 Assignment 1 Solution

**Create a flume agent that streams data from Twitter and stores in the HDFS.**

1. **Flume Agent (configuration file) :** Here, agent name is **TwitterAgent.**

# Naming the components on the current agent.

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

# Describing/Configuring the source

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.consumerKey = u5i7vsgbaXu0tLfBGMYEXduxA

TwitterAgent.sources.Twitter.consumerSecret = 0AFOdoa1j68UKTiJKHGKvDH3Vroju6LYfMI7acYnbCFnbS0do6

TwitterAgent.sources.Twitter.accessToken = 847854748309049348-WYvJl2IPZG21JLoT6giZyjrVAEmoGyq

TwitterAgent.sources.Twitter.accessTokenSecret = YyqIamlXhpM1BTUp76enaw3fg9GvOzbckZAmKOnmxpcfU

TwitterAgent.sources.Twitter.keywords = hadoop,election,sports,cricket,Big data,Bahubali

# Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://pavan:8020/tmp/flume/twitter\_sentiment\_analysis

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.filePrefix = twitter-

TwitterAgent.sinks.HDFS.hdfs.rollInterval = 0

TwitterAgent.sinks.HDFS.hdfs.rollSize = 524288

TwitterAgent.sinks.HDFS.hdfs.rollCount = 0

TwitterAgent.sinks.HDFS.hdfs.idleTimeout = 0

TwitterAgent.sinks.HDFS.hdfs.batchSize = 100

TwitterAgent.sinks.HDFS.hdfs.threadsPoolSize = 2

TwitterAgent.sinks.HDFS.hdfs.round = true

TwitterAgent.sinks.HDFS.hdfs.roundUnit = hour

# Describing/Configuring the channel

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel

**Once we create a Flume Agent, we perform transferring Data Stream from Twitter into HDFS as follows:**

* We will be collecting real-time tweets from Twitter using Flume.

**Streaming Twitter Data Using Flume**

* Create the flume configuration code as shown below.
* Change the twitter api keys with the keys of our Twitter App
* Now, we have to decide which keywords tweet data to be collected from the twitter application, then change the keywords in the TwitterAgent.sources.Twitter.keywords command.

In our example, we are fetching tweet data related to Hadoop, election, sports, cricket, Big data and Bahubali.

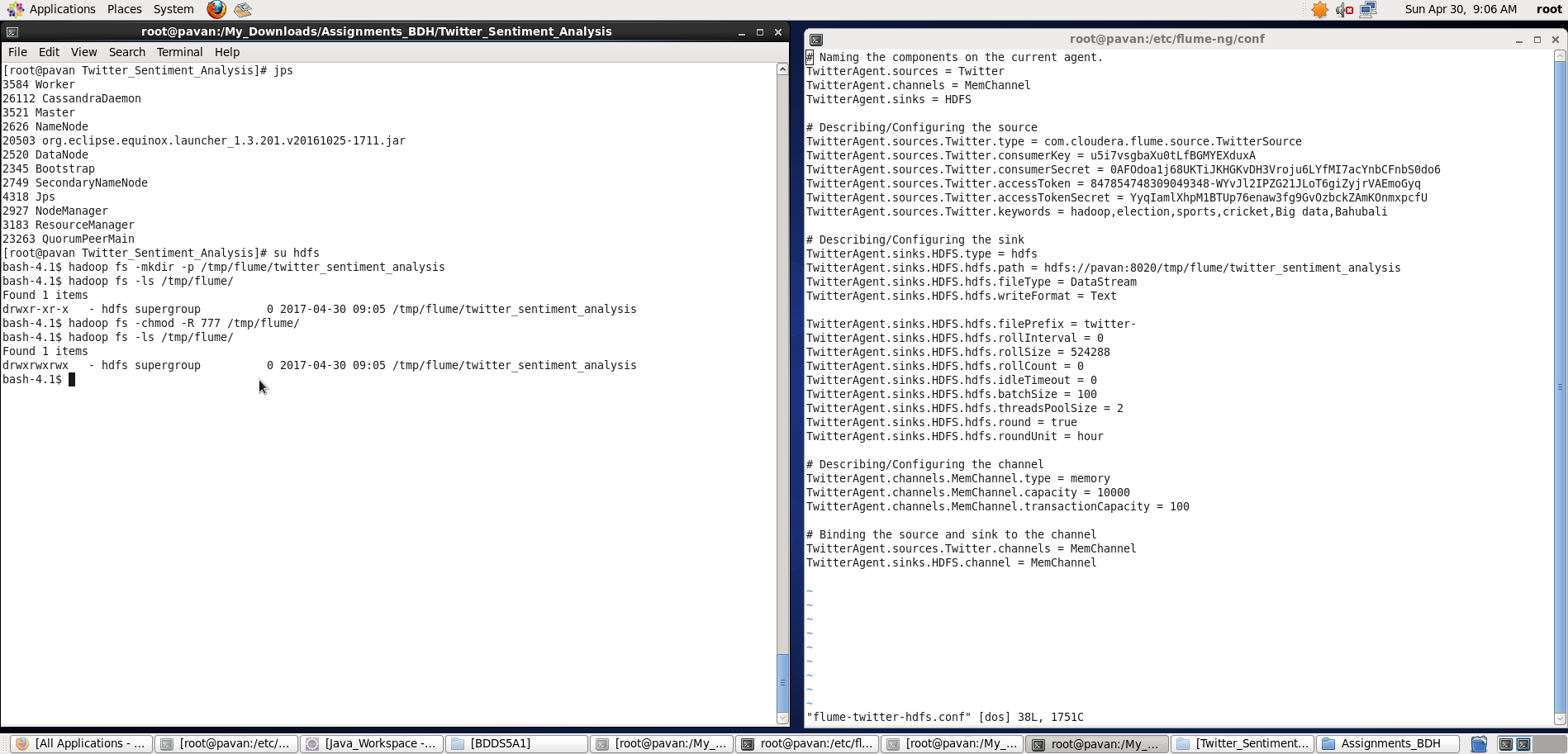
* Open a new terminal and start all the Hadoop daemons, before running the flume command to fetch the twitter data.

Use the ‘jps’ command to see the running Hadoop daemons.

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

**hadoop fs –mkdir –p /tmp/flume/twitter\_sentiment\_analysis**

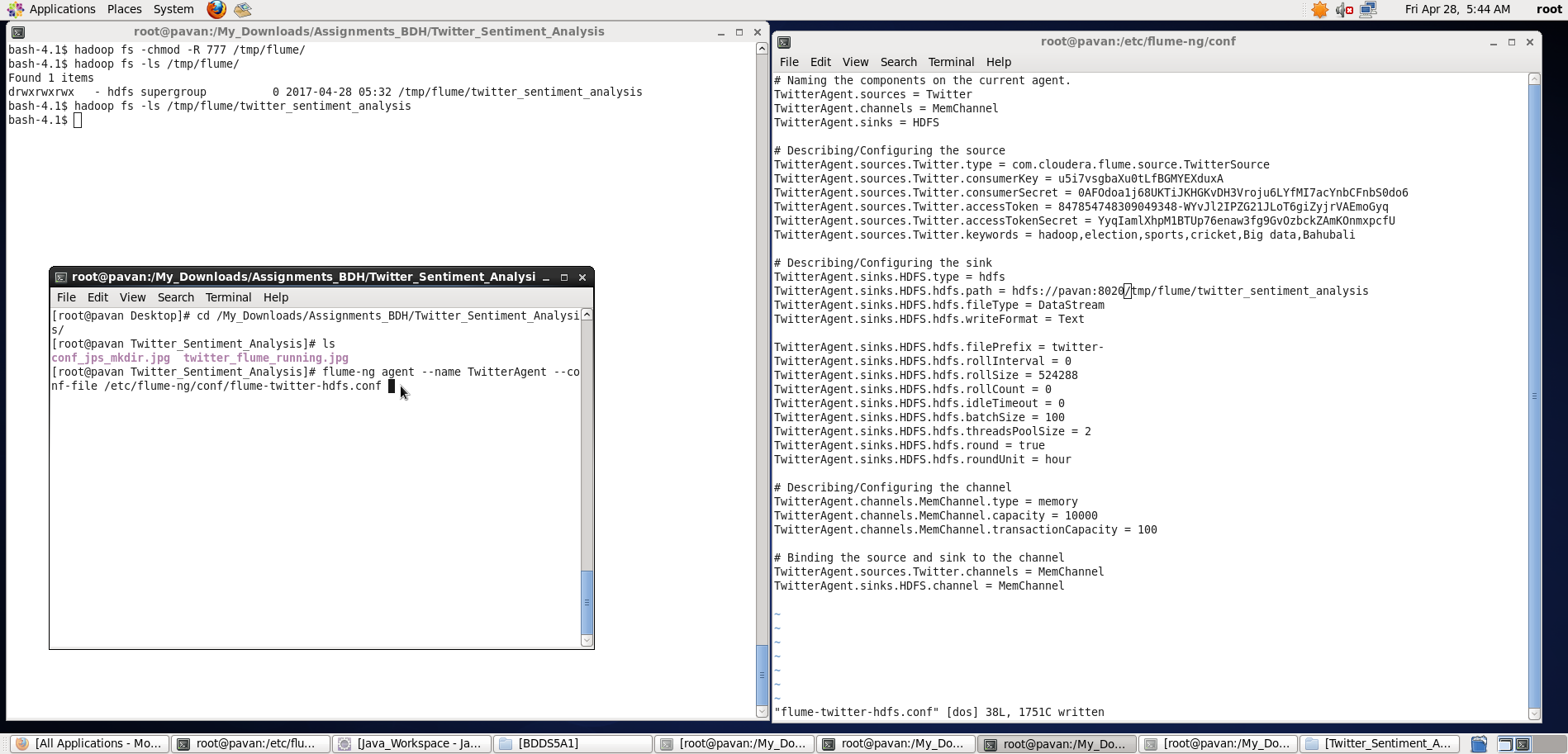
* **Note: Give appropriate read/write permissions for FLUME to write into the HDFS.**



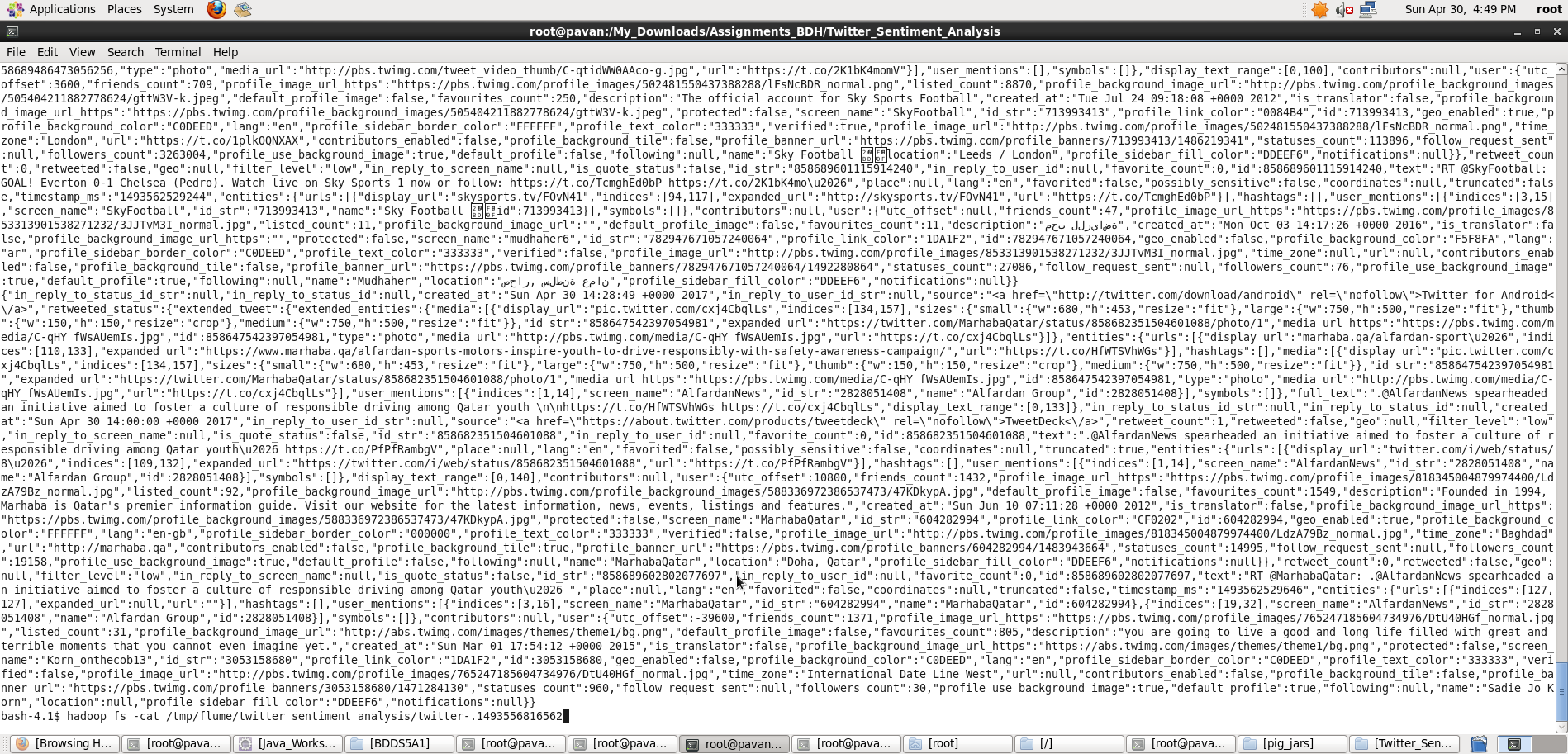
* For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

**flume-ng agent --name TwitterAgent –conf-file /etc/flume-ng/conf/flume-twitter-hdfs.conf**

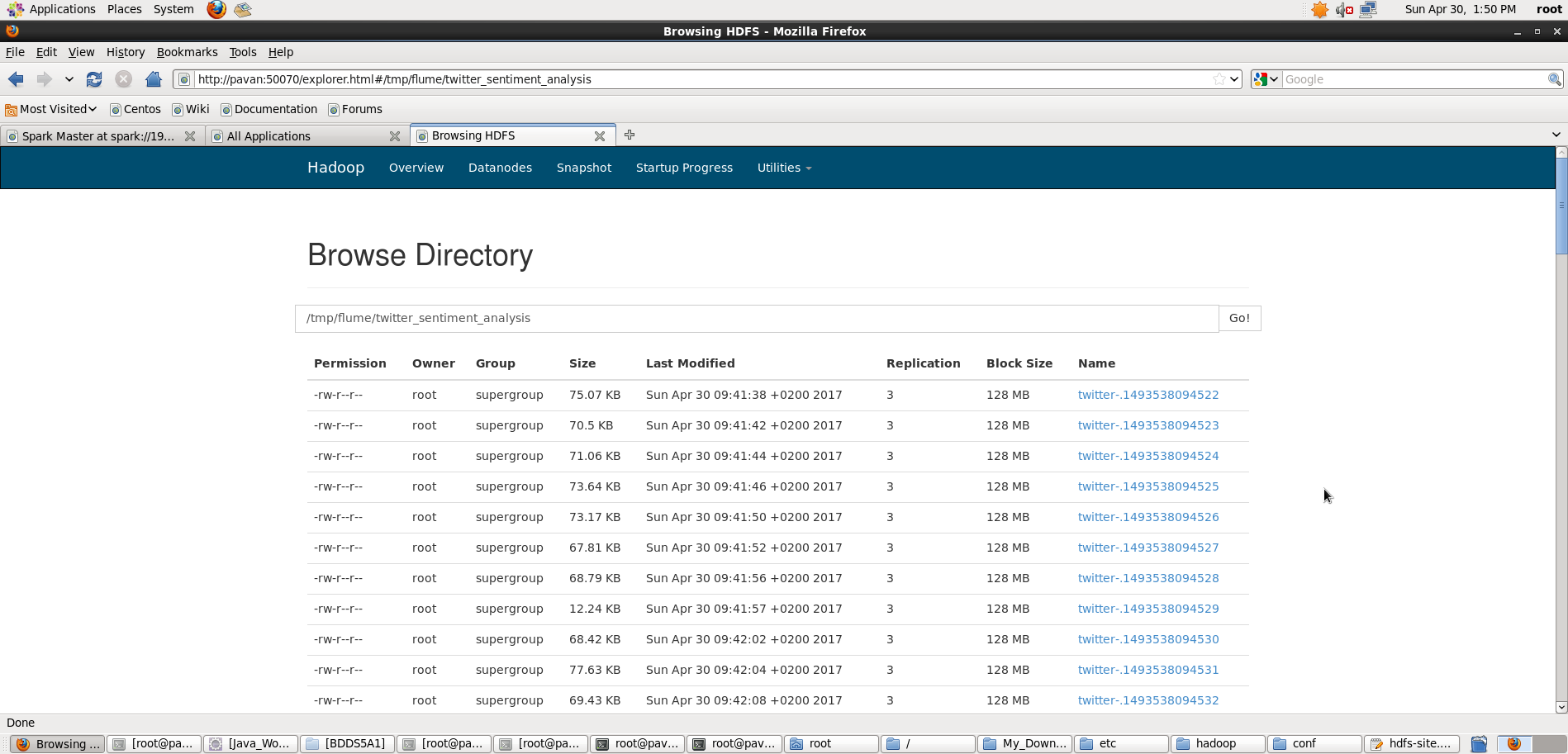
The above command will start fetching data from Twitter and steams it into the HDFS given path



* Once, the tweet data started streaming it into the given HDFS path we can use ‘Ctrl+c’ command to stop the streaming process.
* To check the contents of the tweet data we can use the following command:
* **hadoop fs –ls /tmp/flume/twitter\_sentiment\_analysis**
* We can use the *‘cat’* command to display the tweet  data inside the /tmp/flume/twitter\_sentiment\_analysis/twitter-149\* path.
* **hadoop fs –cat /tmp/flume/twitter\_sentiment\_analysis/<flumeData file name>**



* We can observe from the above image that we have successfully fetched twitter data into our HDFS cluster directory.  Once the tweets have been successfully stored in your database, we can manipulate the tweet data to fit the needs of our future projects.
* All the real-time tweets is kept it the location ‘**/tmp/flume/twitter\_sentiment\_analysis**‘ directory in HDFS. You can refer to the below screen shot for the same.



**Thus, we created a Flume Agent and transferred Data Stream from Twitter to HDFS.**