Kalyan Big Data Projects – Project 4
How To Stream Twitter Data Into Hadoop and MongoDB in JSON format Using Apache Flume

Pre-Requisites of Flume Project:

hadoop-2.6.0
flume-1.6.0
mongodb-3.2.7
java-1.7

NOTE: Make sure that install all the above components

Flume Project Download Links:

`hadoop-2.6.0.tar.gz` ==> link (https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)

`apache-flume-1.6.0-bin.tar.gz` ==> link (https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)

`kalyan-bigdata-examples.jar` ==> link (https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-bigdata-examples.jar)

`mongodb-linux-x86_64-ubuntu1404-3.2.7.tgz` ==> link (http://downloads.mongodb.org/linux/mongodb-linux-x86_64-ubuntu1404-3.2.7.tgz?_ga=1.51737257.1298711466.1475055109)

`kalyan-flume-project-0.1.jar` ==> link (https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-flume-project-0.1.jar)

`mongodb-driver-core-3.3.0.jar` ==> link (http://central.maven.org/maven2/org/mongodb/mongodb-driver-core/3.3.0/mongodb-driver-core-3.3.0.jar)

`mongo-java-driver-3.3.0.jar` ==> link (http://central.maven.org/maven2/org/mongodb/mongo-java-driver/3.3.0/mongo-java-driver-3.3.0.jar)

`kalyan-twitter-hdfs-mongo-agent.conf` ==> link (https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project4-twitter-hadoop-mongodb-json/kalyan-twitter-hdfs-mongo-agent.conf)

Learnings of this Project:

- We will learn Flume Configurations and Commands
- Flume Agent
  1. Source (Twitter Source)
  2. Channel (Memory Channel)
  3. Sink (MongoDB Sink)
- Major project in Real Time `Social Media (Twitter) Sentiment Analysis`
  1. We are extracting the data from twitter using twitter api credentials
  2. This data will be useful to do setiment analysis on twitter tweets
  3. JSON is the output format
- We can use mongodb / hive / pig / mapreduce to analyze this data
  1. explore mongodb to analysis
  2. explore hive query to analysis
  3. explore pig scripts to analysis
  4. explore mapreduce to analysis

---

1. create "**kalyan-twitter-hdfs-mongo-agent.conf**" file with below content

```
agent.sources = Twitter
agent.channels = MemChannel1 MemChannel2
agent.sinks = HDFS MongoDB

agent.sources.Twitter.type = com.orienit.kalyan.flume.source.KalyanTwitterSource
agent.sources.Twitter.channels = MemChannel1 MemChannel2
agent.sources.Twitter.consumerKey = ********
agent.sources.Twitter.consumerSecret = ********
agent.sources.Twitter.accessToken = ********
agent.sources.Twitter.accessTokenSecret = ********
agent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data
scientiest, business intelligence, mapreduce, data warehouse, data warehousing, mahout, hbase,
nosql, newsql, businessintelligence, cloudcomputing

agent.sinks.HDFS.type = hdfs
agent.sinks.HDFS.channel = MemChannel1
agent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/flume/tweets
agent.sinks.HDFS.hdfs.fileType = DataStream
agent.sinks.HDFS.hdfs.writeFormat = Text
agent.sinks.HDFS.hdfs.batchSize = 100
agent.sinks.HDFS.hdfs.rollSize = 0
agent.sinks.HDFS.hdfs.rollCount = 100
agent.sinks.HDFS.hdfs.useLocalTimeStamp = true

agent.sinks.MongoDB.type = com.orienit.kalyan.flume.sink.KalyanMongoSink
agent.sinks.MongoDB.hostNames = localhost
agent.sinks.MongoDB.database = flume
```

agent.sinks.MongoDB.collection = twitter
agent.sinks.MongoDB.batchSize = 10
agent.sinks.MongoDB.channel = MemChannel2

agent.channels.MemChannel1.type = memory
agent.channels.MemChannel1.capacity = 1000
agent.channels.MemChannel1.transactionCapacity = 100

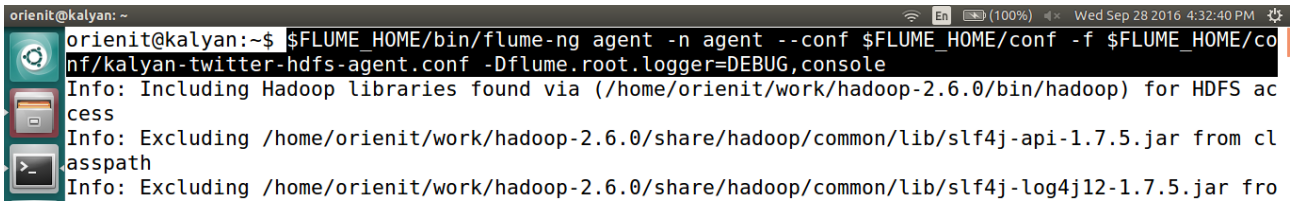
agent.channels.MemChannel2.type = memory
agent.channels.MemChannel2.capacity = 1000
agent.channels.MemChannel2.transactionCapacity = 100

2. Copy "**kalyan-twitter-hdfs-mongo-agent.conf**" file into "**$FUME_HOME/conf**" folder

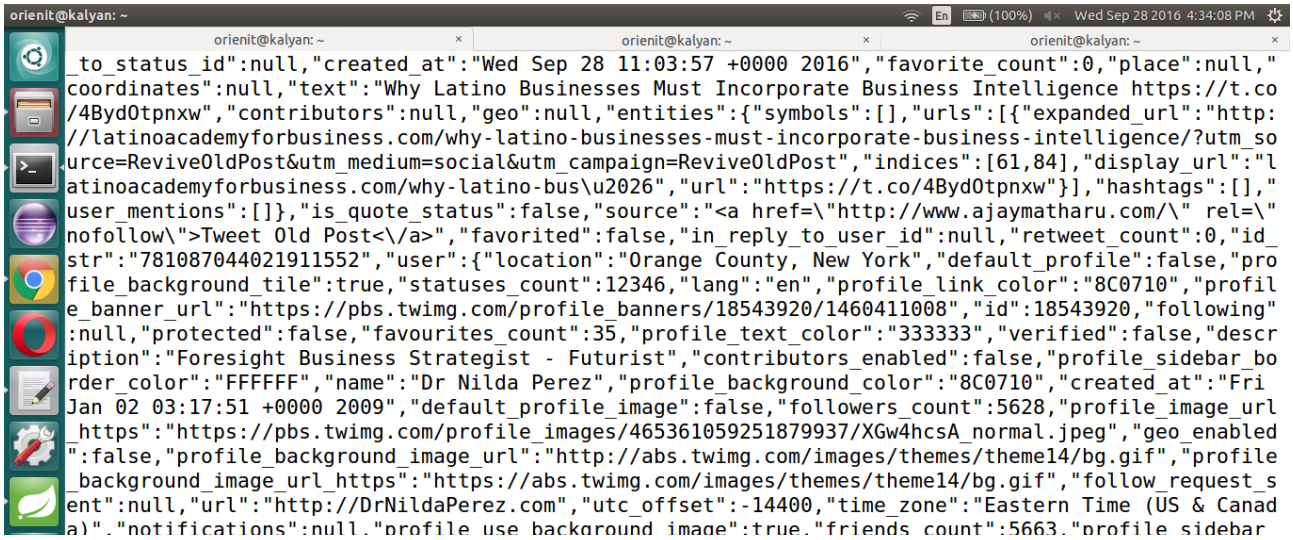
3. Copy "**kalyan-flume-project-0.1.jar, mongodb-driver-core-3.3.0.jar and mongo-java-driver-3.3.0.jar**" files into "**$FLUME_HOME/lib**" folder

4. Execute the below command to `**Extract data from Twitter into HDFS & MongoDB using Flume**`

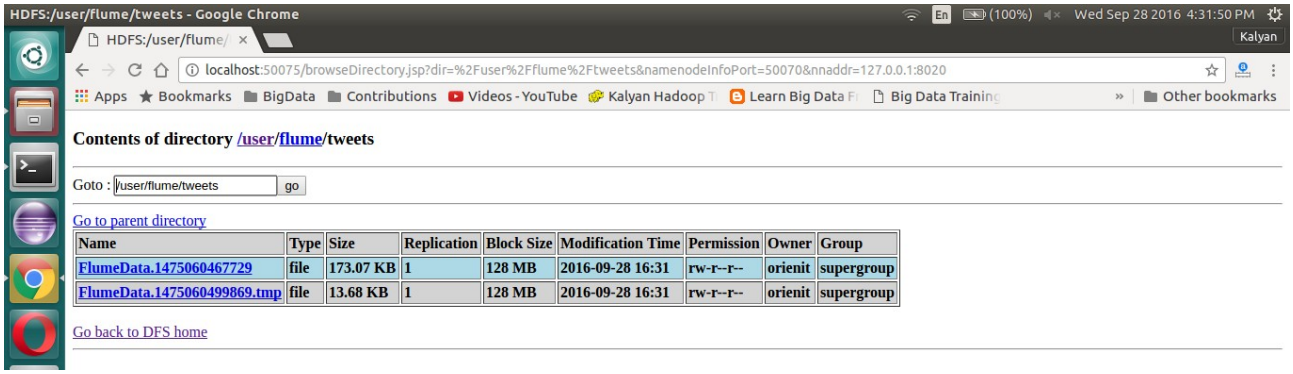
$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/conf/kalyan-twitter-hdfs-mongo-agent.conf -Dflume.root.logger=DEBUG,console

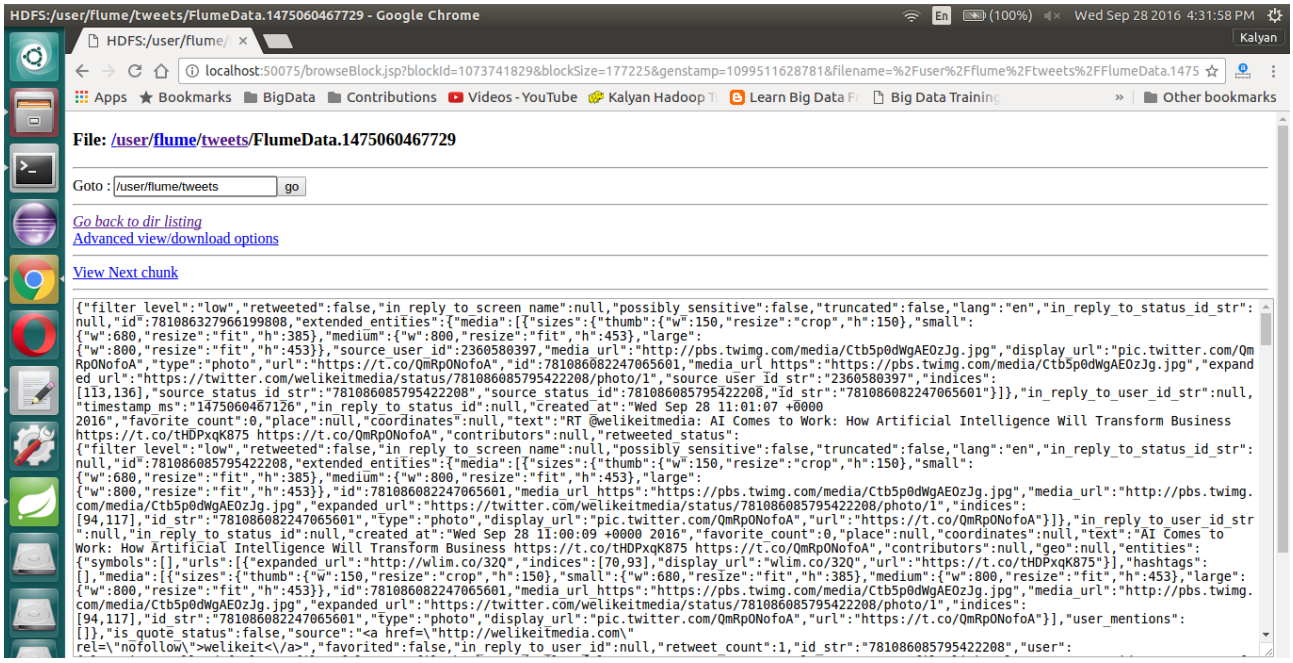


5. Verify the data in console

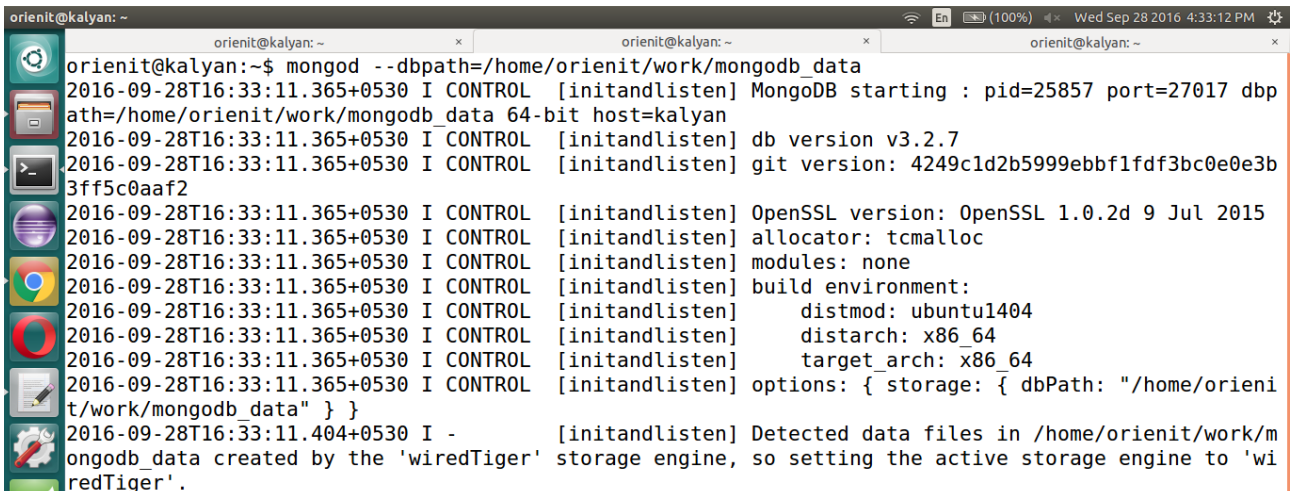


6. Verify the data in HDFS and MongoDB

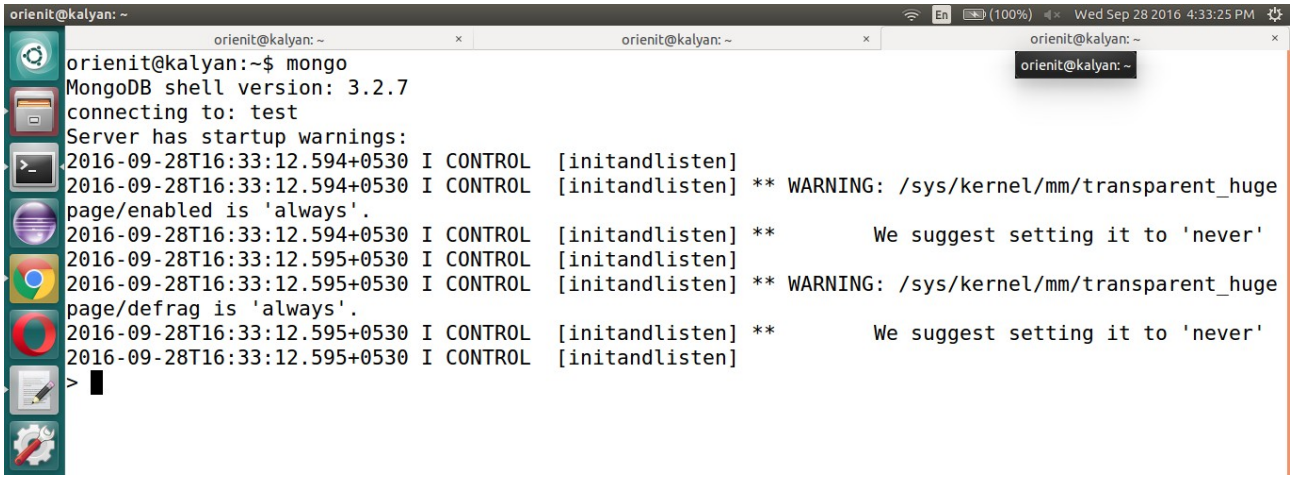




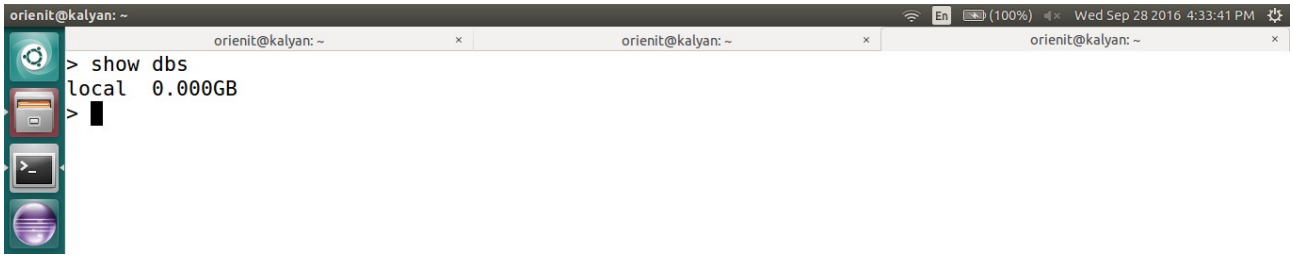
7. Start the MongoDB Server using below command



8. Start the MongoDB client using below command (mongo)



9.Verify the List of DataBases in MongoDB using below command (show dbs)



10.Verify the List of Operations in MongoDB using below commands

// list of databases
show dbs

// use flume database
use flume

// list of collections
show collections

// find the count of documents in 'twitter' collection
db.twitter.count()

// display list of documents in 'twitter' collection
db.twitter.find()

```
orienit@kalyan: ~
> show dbs
local  0.000GB
> show dbs
flume  0.000GB
local  0.000GB
>
> use flume
switched to db flume
>
> show collections
twitter
>
> db.twiter.count()
0
> db.twitter.count()
49
> db.twitter.count()
63
>
```

```
orienit@kalyan: ~
> db.twitter.find()
{ "_id" : ObjectId("57eba3b2252c5d6cd1f8fbc2"), "filter_level" : "low", "retweeted" : false, "in_re
ply_to_screen_name" : null, "possibly_sensitive" : false, "truncated" : false, "lang" : "en", "in_r
eply_to_status_id_str" : null, "id" : NumberLong("781087017971097600"), "extended_entities" : { "me
dia" : [ { "sizes" : { "small" : { "w" : 340, "resize" : "fit", "h" : 191 }, "thumb" : { "w" : 150,
 "resize" : "crop", "h" : 150 }, "large" : { "w" : 1024, "resize" : "fit", "h" : 576 }, "medium" :
{ "w" : 600, "resize" : "fit", "h" : 338 } }, "id" : NumberLong("781086809837240320"), "media_url_h
ttps" : "https://pbs.twimg.com/ext_tw_video_thumb/781086809837240320/pu/img/FmablGNx9ZmVj_gX.jpg",
"video_info" : { "duration_millis" : 32334, "variants" : [ { "bitrate" : 2176000, "content_type" :
"video/mp4", "url" : "https://video.twimg.com/ext_tw_video/781086809837240320/pu/vid/1280x720/zvEIN
mdkVhuLxDw4.mp4" }, { "content_type" : "application/x-mpegURL", "url" : "https://video.twimg.com/ex
t_tw_video/781086809837240320/pu/pl/tbMWDbhv60VB-P89.m3u8" }, { "content_type" : "application/dash+
xml", "url" : "https://video.twimg.com/ext_tw_video/781086809837240320/pu/pl/tbMWDbhv60VB-P89.mpd"
}, { "bitrate" : 320000, "content_type" : "video/mp4", "url" : "https://video.twimg.com/ext_tw_vide
o/781086809837240320/pu/vid/320x180/uSYbxQU-NVQJv5fy.mp4" }, { "bitrate" : 832000, "content_type" :
 "video/mp4", "url" : "https://video.twimg.com/ext_tw_video/781086809837240320/pu/vid/640x360/gkAZ2
ujxNdqZSUlj.mp4" } ], "aspect_ratio" : [ 16, 9 ] }, "media_url" : "http://pbs.twimg.com/ext_tw_vide
o_thumb/781086809837240320/pu/img/FmablGNx9ZmVj_gX.jpg", "expanded_url" : "https://twitter.com/GO_S
UPERSONIC/status/781087017971097600/video/1", "indices" : [ 70, 93 ], "id_str" : "7810868098372403 2
0", "type" : "video", "display_url" : "pic.twitter.com/wBybakfoHC", "url" : "https://t.co/wBybakfoH
C" } ] }, "in_reply_to_user_id_str" : null, "timestamp_ms" : "1475060631636", "in_reply_to_status_i
d" : null, "created_at" : "Wed Sep 28 11:03:51 +0000 2016", "favorite_count" : 0, "place" : null, "
coordinates" : null, "text" : "Consider #BigData #servers for your #business enterprise application
s https://t.co/wBybakfoHC", "contributors" : null, "geo" : null, "entities" : { "symbols" : [ ], "u
rls" : [ ], "hashtags" : [ { "text" : "BigData", "indices" : [ 9, 17 ] }, { "text" : "servers", "in
dices" : [ 18, 26 ] }, { "text" : "business", "indices" : [ 36, 45 ] } ], "media" : [ { "sizes" : {
```