Kalyan Big Data Projects – Project 3
How To Stream Twitter Data Into 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)

`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)

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

`kalyan-twitter-mongo-agent.conf` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project3-twitter-mongodb-json/kalyan-twitter-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-mongo-agent.conf**" file with below content

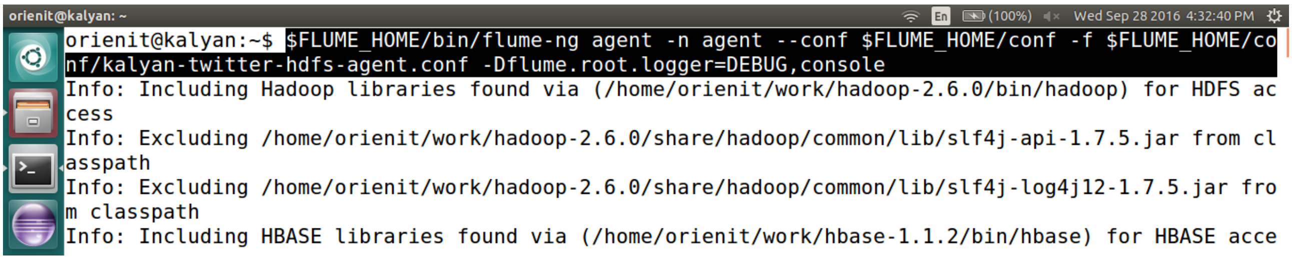
```
agent.sources = Twitter
agent.channels = MemChannel
agent.sinks =MongoDB

agent.sources.Twitter.type = com.orienit.kalyan.flume.source.KalyanTwitterSource
agent.sources.Twitter.channels = MemChannel
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.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 = MemChannel

agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

2. Copy "**kalyan-twitter-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

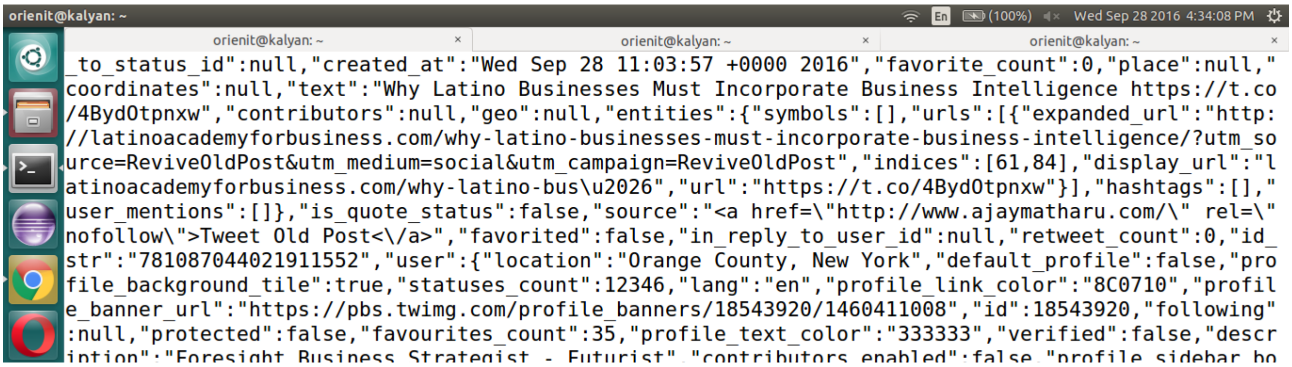
4. Execute the below command to `**Extract data fromTwitter into MongoDB using Flume**`

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



```
orienit@kalyan:~$ $FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/co
nf/kalyan-twitter-hdfs-agent.conf -Dflume.root.logger=DEBUG,console
Info: Including Hadoop libraries found via (/home/orienit/work/hadoop-2.6.0/bin/hadoop) for HDFS ac
cess
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from cl
asspath
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar fro
m classpath
Info: Including HBASE libraries found via (/home/orienit/work/hbase-1.1.2/bin/hbase) for HBASE acce
```

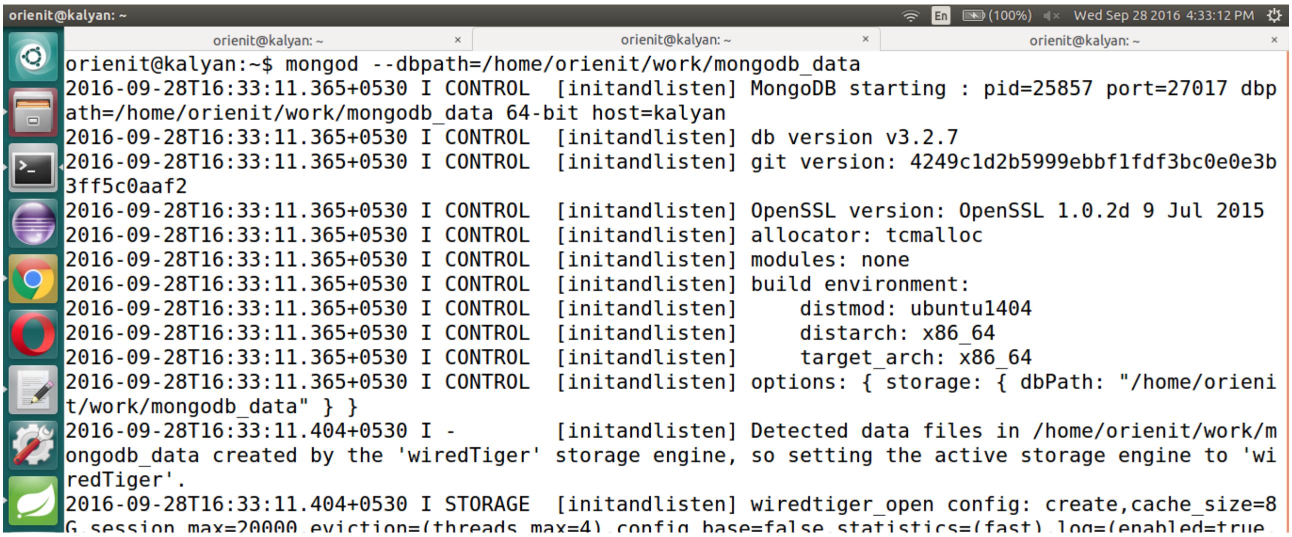
5. Verify the data in console



```
_to_status_id":null,"created_at":"Wed Sep 28 11:03:57 +0000 2016","favorite_count":0,"place":null,"
coordinates":null,"text":"Why Latino Businesses Must Incorporate Business Intelligence https://t.co
/4BydOtpnxw","contributors":null,"geo":null,"entities":{"symbols":[],"urls":[{"expanded_url":"http:
//latinoacademyforbusiness.com/why-latino-businesses-must-incorporate-business-intelligence/?utm_so
urce=ReviveOldPost&utm_medium=social&utm_campaign=ReviveOldPost","indices":[61,84],"display_url":"l
atinoacademyforbusiness.com/why-latino-bus\u2026","url":"https://t.co/4BydOtpnxw"}],"hashtags":[],"
user_mentions":[]},"is_quote_status":false,"source":"<a href=\"http://www.ajaymatharu.com/\" rel=\"
nofollow\">Tweet Old Post<\/a>","favorited":false,"in_reply_to_user_id":null,"retweet_count":0,"id_
str":"781087044021911552","user":{"location":"Orange County, New York","default_profile":false,"pro
file_background_tile":true,"statuses_count":12346,"lang":"en","profile_link_color":"8C0710","profil
e_banner_url":"https://pbs.twimg.com/profile_banners/18543920/1460411008","id":18543920,"following"
:null,"protected":false,"favourites_count":35,"profile_text_color":"333333","verified":false,"descr
iption":"Foresight Business Strategist - Futurist","contributors_enabled":false,"profile_sidebar_bo
```

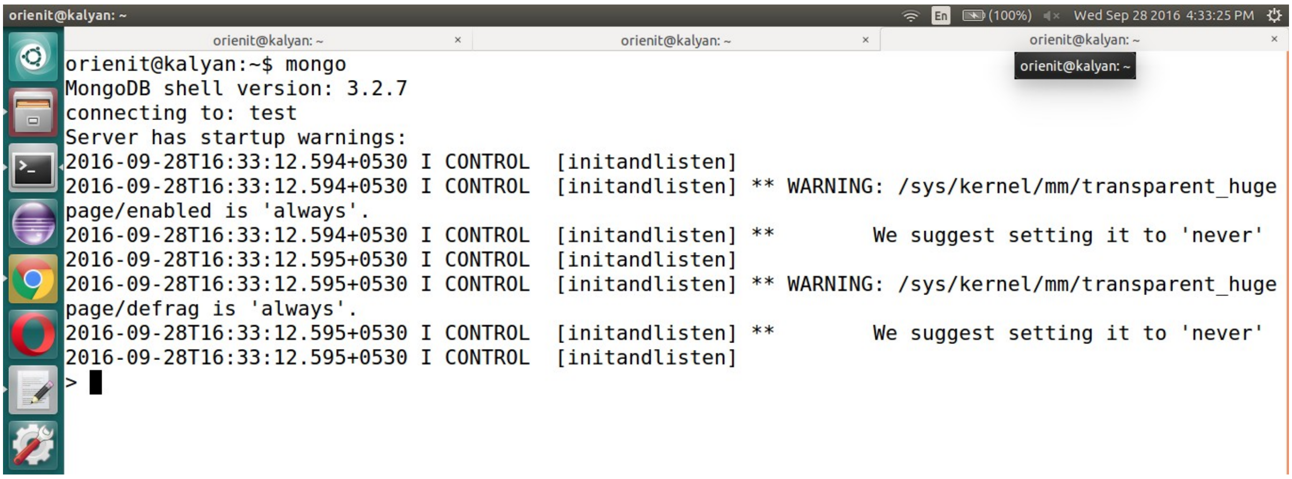
6. Verify the data in MongoDB

7. Start the MongoDB Server using below command



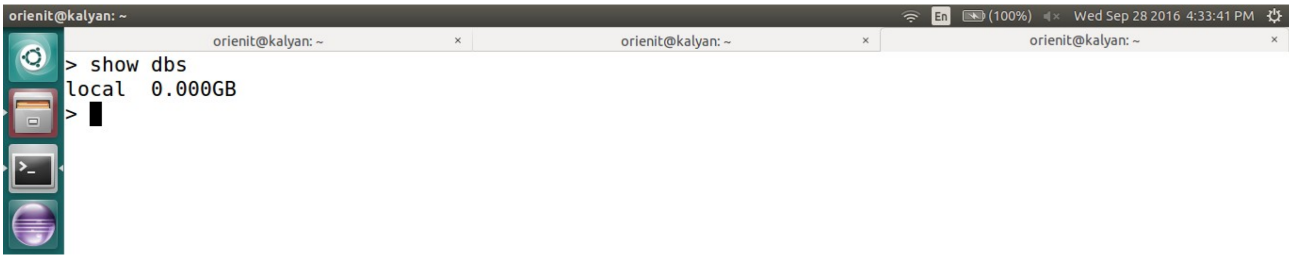
```
orienit@kalyan:~$ mongod --dbpath=/home/orienit/work/mongodb_data
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] MongoDB starting : pid=25857 port=27017 dbp
ath=/home/orienit/work/mongodb_data 64-bit host=kalyan
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] db version v3.2.7
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] git version: 4249c1d2b5999ebbf1fdf3bc0e0e3b
3ff5c0aaf2
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] OpenSSL version: OpenSSL 1.0.2d 9 Jul 2015
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] allocator: tcmalloc
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] modules: none
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] build environment:
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten]     distmod: ubuntu1404
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten]     distarch: x86_64
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten]     target_arch: x86_64
2016-09-28T16:33:11.365+0530 I CONTROL  [initandlisten] options: { storage: { dbPath: "/home/orieni
t/work/mongodb_data" } }
2016-09-28T16:33:11.404+0530 I -        [initandlisten] Detected data files in /home/orienit/work/m
ongodb_data created by the 'wiredTiger' storage engine, so setting the active storage engine to 'wi
redTiger'.
2016-09-28T16:33:11.404+0530 I STORAGE  [initandlisten] wiredtiger_open config: create,cache_size=8
G,session_max=20000,eviction=(threads_max=4),config_base=false,statistics=(fast),log=(enabled=true,
```

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



```
orienit@kalyan:~$ mongo
MongoDB shell version: 3.2.7
connecting to: test
Server has startup warnings:
2016-09-28T16:33:12.594+0530 I CONTROL  [initandlisten]
2016-09-28T16:33:12.594+0530 I CONTROL  [initandlisten] ** WARNING: /sys/kernel/mm/transparent_huge
page/enabled is 'always'.
2016-09-28T16:33:12.594+0530 I CONTROL  [initandlisten] **        We suggest setting it to 'never'
2016-09-28T16:33:12.595+0530 I CONTROL  [initandlisten]
2016-09-28T16:33:12.595+0530 I CONTROL  [initandlisten] ** WARNING: /sys/kernel/mm/transparent_huge
page/defrag is 'always'.
2016-09-28T16:33:12.595+0530 I CONTROL  [initandlisten] **        We suggest setting it to 'never'
2016-09-28T16:33:12.595+0530 I CONTROL  [initandlisten]
> 
```

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



```
> show dbs
local   0.000GB
> 
```

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

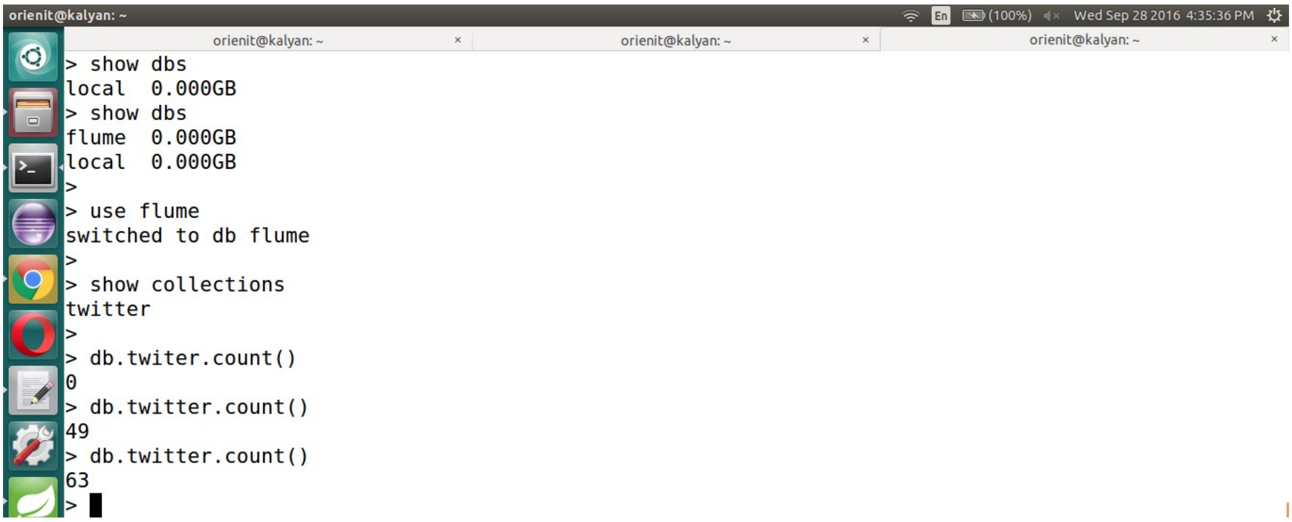
// list of databases
show dbs

// use flume database
use flume

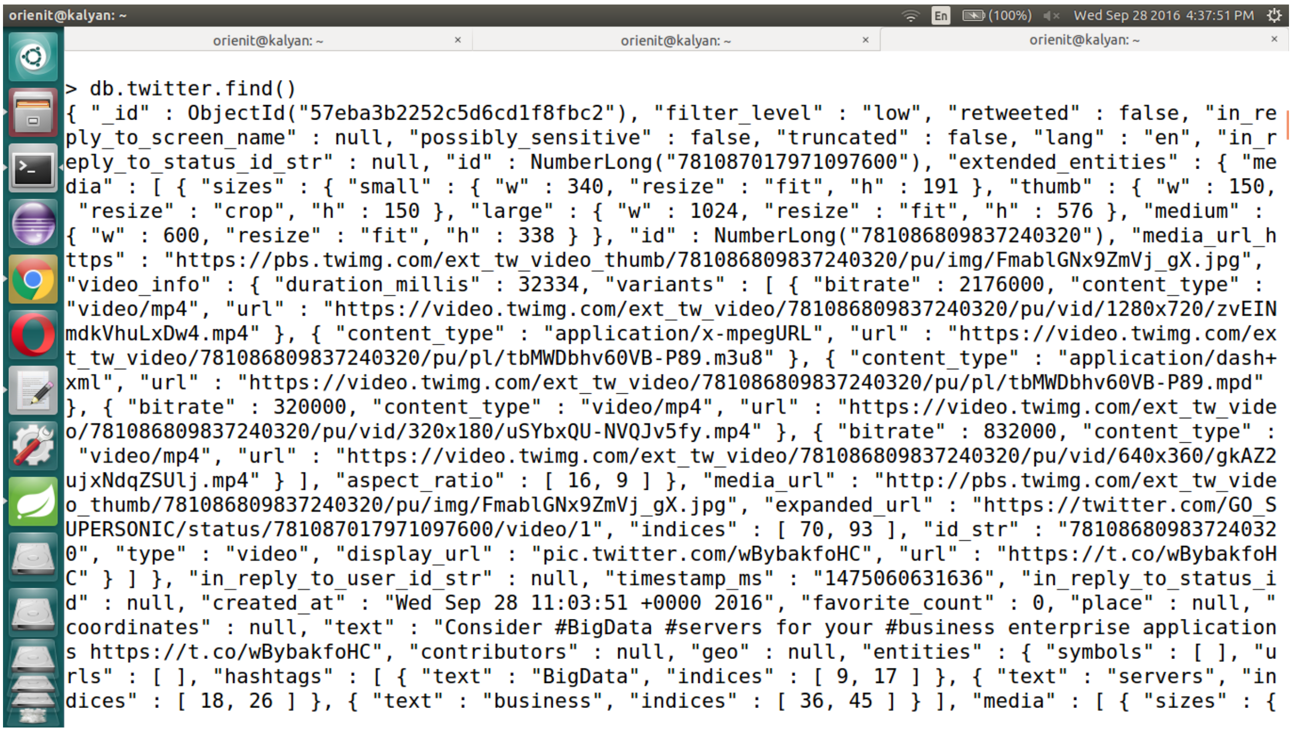
// list of collections
show collections

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

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



```
> 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
> 
```



```
> 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" : "78108680983724032
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" : {
```