Session 11

Assignment 3

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 11 – Assignment 3 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 10 – Assignment 3** |
|  |  |
| **Start Date:** | 06/10/2017 |
|  |  |
|  |  |

# Contents

[Contents 2](#_Toc495044129)

[Change History 3](#_Toc495044130)

[1. Problem Statement 4](#_Toc495044131)

[2. Background 5](#_Toc495044132)

[2.1. Solution 5](#_Toc495044133)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 06/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

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

# Background

The component of the Hadoop ecosystem that lets us load data of type semi-structured and unstructured and live streaming data, from different sources, into the  HDFS cluster is **FLUME**.

Learning Flume will help users to collect from and store a large amount of data from different sources into the Hadoop cluster.

Apache Flume is a Hadoop ecosystem component used to collect, aggregate and moves a large amount of log data from different sources to a centralized data store.

It is an open source component which is designed to locate and store the data in a distributed environment and collects the data as per the specified input key(s).

The main components of the **Flume Architecture** are:

**Flume Event:**It is the main unit of the data that is transported inside the **Flume**(Typically a single log entry). It contains a payload of the byte array that is to be transported from the source path to the destination path which could be accompanied by optional headers.

**Flume Agent:** Is an independent Java virtual machine daemon process which receives the data (events) from clients and transports to the subsequent destination (sink or agent).

**Source:**Is the component of Flume agent which receives data from the data sources such as Twitter, Facebook, and weblogs from different sites and transfers this data to one or more channels in the form of Flume event.

**Channel:**Once, the Flume source receives an Event, it stores this data into one or more channel and buffers them till they are consumed by sinks. It acts as a bridge between the sources and sinks.

**Sink:**It stores the data into the centralized stores like HDFS and HBase.

## Solution

To stream data to a database from twitter the following pre-requisites are required.

* Twitter account
* Hadoop cluster

**Step 1:**

Login to Twitter Account and click the ‘create new app’ button at <https://apps.twitter.com/app>

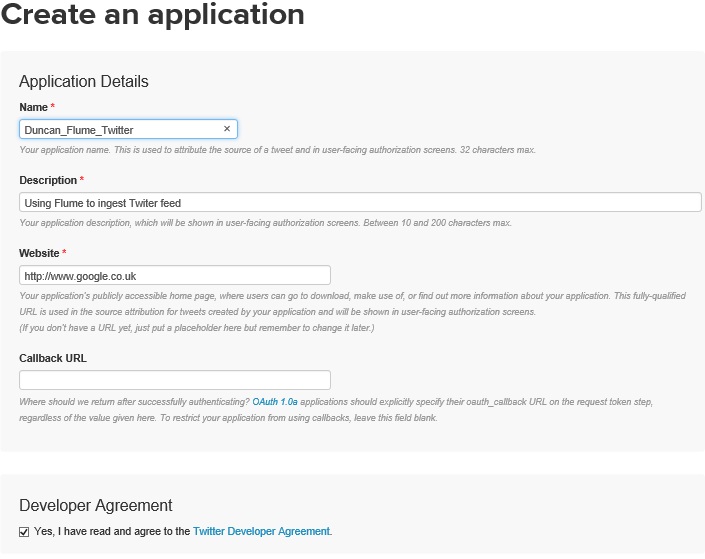
**Step 2:**

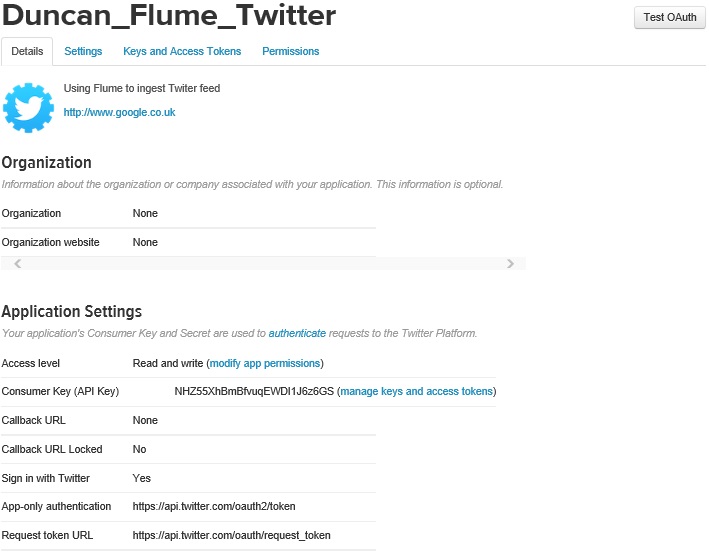
Create a Twitter application by entering the appropriate details.

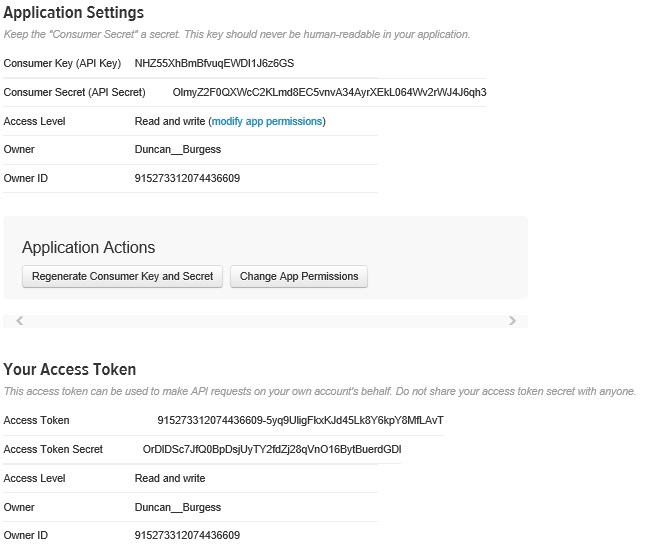
**Step 3:**

Configure **twitter app** and obtaining the generated consumer key and the consumer secret code and the access token details.

**Create application**







**Step 4:**

**Check flume environment**

export JAVA\_HOME=/usr/local/java

export PATH=$PATH:$JAVA\_HOME/bin

export HADOOP\_HOME=/usr/local/hadoop-2.6.0

export PATH=$PATH:$HADOOP\_HOME/bin

export FLUME\_HOME=/usr/local/flume

export PIG\_INSTALL=/usr/local/pig

export HIVE\_HOME=/usr/local/hive

export PATH=$HIVE\_HOME/bin:$PATH

export HADOOP\_USER\_CLASSPATH\_FIRST=true

export HBASE\_HOME=/usr/local/hbase

export SQOOP\_HOME=/usr/local/sqoop

export PATH=$PATH:$FLUME\_HOME/bin:$PIG\_INSTALL/bin:$HIVE\_HOME/bin:$HBASE\_HOME/bin:$SQOOP\_HOME/bin:$HADOOP\_HOME/sbin

**Path and homedir set correctly**

**Check for jars**

*[acadgild@localhost ~]$ cd /usr/local/flume/lib*

*[acadgild@localhost lib]$ ls twit\**

*twitter4j-core-3.0.3.jar*

*twitter4j-media-support-3.0.3.jar*

*twitter4j-stream-3.0.3.jar*

**Step 5:**

**Create flume config file flume.conf**

*TwitterAgent.sources= Twitter*

*TwitterAgent.channels= MemChannel*

*TwitterAgent.sinks=HDFS*

*TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource*

*TwitterAgent.sources.Twitter.channels=MemChannel*

*TwitterAgent.sources.Twitter.consumerKey=NHZ55XhBmBfvuqEWDI1J6z6GS*

*TwitterAgent.sources.Twitter.consumerSecret=OlmyZ2F0QXWcC2KLmd8EC5vnvA34AyrXEkL064Wv2rWJ4J6qh3*

*TwitterAgent.sources.Twitter.accessToken=915273312074436609-5yq9UligFkxKJd45Lk8Y6kpY8MfLAvT*

*TwitterAgent.sources.Twitter.accessTokenSecret=OrDlDSc7JfQ0BpDsjUyTY2fdZj28qVnO16BytBuerdGDl*

*TwitterAgent.sources.Twitter.keywords= FenlandRP,FRP,UHG,Optum*

*TwitterAgent.sinks.HDFS.channel=MemChannel*

*TwitterAgent.sinks.HDFS.type=hdfs*

*TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/user/acadgild/tweets*

*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.sinks.HDFS.hdfs.rollInterval=600*

*TwitterAgent.channels.MemChannel.type=memory*

*TwitterAgent.channels.MemChannel.capacity=10000*

*TwitterAgent.channels.MemChannel.transactionCapacity=5000*

The above file contains the configuration for the Flume Twitter App. Some important entries:

* Setting consumer keys and access token details obtained while creating the Twitter App.
* Setting the keywords that need to be searched/streamed from Twitter.
* Setting the path for storing the results from the Flume Twitter Stream
* Setting the properties for streaming channel and its capacity

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

**/user/acadgils/tweets**

**Run command**

* [acadgild@localhost home]$ **flume-ng agent -n TwitterAgent -f /usr/local/flume/conf/flume.conf**

*17/10/04 07:22:36 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting*

*17/10/04 07:22:36 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/usr/local/flume/conf/flume.conf*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: TwitterAgent*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Processing:HDFS*

*17/10/04 07:22:37 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAgent]*

*17/10/04 07:22:37 INFO node.AbstractConfigurationProvider: Creating channels*

*17/10/04 07:22:37 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory*

*17/10/04 07:22:37 INFO node.AbstractConfigurationProvider: Created channel MemChannel*

*17/10/04 07:22:37 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.TwitterSource*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Consumer Key: 'NHZ55XhBmBfvuqEWDI1J6z6GS'*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Consumer Secret: 'OlmyZ2F0QXWcC2KLmd8EC5vnvA34AyrXEkL064Wv2rWJ4J6qh3'*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Access Token: '915273312074436609-5yq9UligFkxKJd45Lk8Y6kpY8MfLAvT'*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Access Token Secret: 'OrDlDSc7JfQ0BpDsjUyTY2fdZj28qVnO16BytBuerdGDl'*

*17/10/04 07:22:37 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs*

*17/10/04 07:22:37 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]*

*17/10/04 07:22:37 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { source:org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@23dd8f79 counterGroup:{ name:null counters:{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name: MemChannel}} }*

*17/10/04 07:22:37 INFO node.Application: Starting Channel MemChannel*

*17/10/04 07:22:37 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.*

*17/10/04 07:22:37 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started*

*17/10/04 07:22:37 INFO node.Application: Starting Sink HDFS*

*17/10/04 07:22:37 INFO node.Application: Starting Source Twitter*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} ...*

*17/10/04 07:22:37 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered new MBean.*

*17/10/04 07:22:37 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started*

*17/10/04 07:22:37 INFO twitter.TwitterSource: Twitter source Twitter started.*

*17/10/04 07:22:37 INFO twitter4j.TwitterStreamImpl: Establishing connection.*

*17/10/04 07:22:38 INFO twitter4j.TwitterStreamImpl: Connection established.*

*17/10/04 07:22:38 INFO twitter4j.TwitterStreamImpl: Receiving status stream.*

*17/10/04 07:22:38 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false*

*17/10/04 07:22:39 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/acadgild/tweets/FlumeData.1507098158929.tmp*

*17/10/04 07:22:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable*

*17/10/04 07:22:41 INFO twitter.TwitterSource: Processed 100 docs*

*17/10/04 07:22:45 INFO twitter.TwitterSource: Processed 200 docs*

*17/10/04 07:22:48 INFO twitter.TwitterSource: Processed 300 docs*

*17/10/04 07:22:51 INFO twitter.TwitterSource: Processed 400 docs*

*17/10/04 07:22:55 INFO twitter.TwitterSource: Processed 500 docs*

*17/10/04 07:22:59 INFO twitter.TwitterSource: Processed 600 docs*

*17/10/04 07:23:03 INFO twitter.TwitterSource: Processed 700 docs*

*17/10/04 07:23:06 INFO twitter.TwitterSource: Processed 800 docs*

*17/10/04 07:23:10 INFO twitter.TwitterSource: Processed 900 docs*

*17/10/04 07:23:14 INFO twitter.TwitterSource: Processed 1,000 docs*

*17/10/04 07:23:14 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0*

*17/10/04 07:23:14 INFO twitter.TwitterSource: 27 docs/second*

*17/10/04 07:23:14 INFO twitter.TwitterSource: Run took 37 seconds and processed:*

*17/10/04 07:23:14 INFO twitter.TwitterSource: 0.007 MB/sec sent to index*

*17/10/04 07:23:14 INFO twitter.TwitterSource: 0.27 MB text sent to index*

*17/10/04 07:23:14 INFO twitter.TwitterSource: There were 0 exceptions ignored:*

*17/10/04 07:23:17 INFO twitter.TwitterSource: Processed 1,100 docs*

*17/10/04 07:23:21 INFO twitter.TwitterSource: Processed 1,200 docs*

*17/10/04 07:23:25 INFO twitter.TwitterSource: Processed 1,300 docs*

*17/10/04 07:23:28 INFO twitter.TwitterSource: Processed 1,400 docs*

*17/10/04 07:23:31 INFO twitter.TwitterSource: Processed 1,500 docs*

*17/10/04 07:23:35 INFO twitter.TwitterSource: Processed 1,600 docs*

*17/10/04 07:23:38 INFO twitter.TwitterSource: Processed 1,700 docs*

*17/10/04 07:23:42 INFO twitter.TwitterSource: Processed 1,800 docs*

*17/10/04 07:23:46 INFO twitter.TwitterSource: Processed 1,900 docs*

*17/10/04 07:23:50 INFO twitter.TwitterSource: Processed 2,000 docs*

*17/10/04 07:23:50 INFO twitter.TwitterSource: Total docs indexed: 2,000, total skipped docs: 0*

*17/10/04 07:23:50 INFO twitter.TwitterSource: 27 docs/second*

*17/10/04 07:23:50 INFO twitter.TwitterSource: Run took 73 seconds and processed:*

*17/10/04 07:23:50 INFO twitter.TwitterSource: 0.007 MB/sec sent to index*

*17/10/04 07:23:50 INFO twitter.TwitterSource: 0.537 MB text sent to index*

*17/10/04 07:23:50 INFO twitter.TwitterSource: There were 0 exceptions ignored:*

*17/10/04 07:23:54 INFO twitter.TwitterSource: Processed 2,100 docs*

*17/10/04 07:23:58 INFO twitter.TwitterSource: Processed 2,200 docs*

*17/10/04 07:24:01 INFO twitter.TwitterSource: Processed 2,300 docs*

*17/10/04 07:24:04 INFO twitter.TwitterSource: Processed 2,400 docs*

*17/10/04 07:24:07 INFO twitter.TwitterSource: Processed 2,500 docs*

*17/10/04 07:24:10 INFO twitter.TwitterSource: Processed 2,600 docs*

*17/10/04 07:24:13 INFO twitter.TwitterSource: Processed 2,700 docs*

*17/10/04 07:24:17 INFO twitter.TwitterSource: Processed 2,800 docs*

*17/10/04 07:24:20 INFO twitter.TwitterSource: Processed 2,900 docs*

*17/10/04 07:24:24 INFO twitter.TwitterSource: Processed 3,000 docs*

*17/10/04 07:24:24 INFO twitter.TwitterSource: Total docs indexed: 3,000, total skipped docs: 0*

*17/10/04 07:24:24 INFO twitter.TwitterSource: 28 docs/second*

*17/10/04 07:24:24 INFO twitter.TwitterSource: Run took 106 seconds and processed:*

*17/10/04 07:24:24 INFO twitter.TwitterSource: 0.008 MB/sec sent to index*

*17/10/04 07:24:24 INFO twitter.TwitterSource: 0.799 MB text sent to index*

*17/10/04 07:24:24 INFO twitter.TwitterSource: There were 0 exceptions ignored:*

*17/10/04 07:24:27 INFO twitter.TwitterSource: Processed 3,100 docs*

**Ctr+C will stop the feed**

**Results**

*[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/tweets*

*17/10/05 03:56:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable*

*Found 4 items*

*-rw-r--r-- 1 acadgild supergroup 8380250 2017-10-04 07:02 /user/acadgild/tweets/FlumeData.1507096342898*

*-rw-r--r-- 1 acadgild supergroup 3554221 2017-10-04 07:12 /user/acadgild/tweets/FlumeData.1507096945341*

*-rw-r--r-- 1 acadgild supergroup 8150960 2017-10-04 07:32 /user/acadgild/tweets/FlumeData.1507098158929*

*-rw-r--r-- 1 acadgild supergroup 2924270 2017-10-04 07:36 /user/acadgild/tweets/FlumeData.1507098762036*

This is the data that was streamed from twitter and stored in the HDFS directory **/user/acadgild/tweets**.

**To view example of the output**

*[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/tweets/FlumeData.1507096342898*

*Objavro.schema*

*$915454528870391808#Cape TownLead Developer @heyjudeapp, CEO TCT Digital, Founder and Writer @TechCapeTown.*

*ByronTheLifeOfByron(2017-10-04T06:52:22ZdImagine still manually retweeting someone in 2017?<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>nCdX\_6j*

*Objavro.schema"retweeted","type":["boolean","null"]},{"name":"in\_reply\_to\_user\_id","type":["long","null"]},{"name":"source","type":["string","null"]},{"name":"in\_reply\_to\_status\_id","type":["long","null"]},{"name":"media\_url\_https","type":["string","null"]},{"name":"expanded\_url","type":["string","null"]}]}hm#%A0$915454528882925568zcthutq bdfyjdrty3452(2017-10-04T06:52:22ZRT @KFM936: " -"<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>$915454528862003201THellPWhy do we torture ourselves with living.CatfishMilkCatfishMilk(2017-10-04T06:52:22ZR@idekarolina @culturemke @PostMalone Same<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>$915454528874364930N$Kurotora1023(2017-10-04T06:52:22ZskypeID*

*Kurotora1023*

*|<a href="http://twittbot.net/" rel="nofollow">twittbot.net</a>$915454528882987008="Mérida, Yucatán5 años de experiencia, encuentros Sw,Cuck-Old-Trios. Amante de la fotografía erótica.od \*FraternityEA\*tupac\_amaru9(2017-10-04T06:52:22ZRT @ALONDRAHOTSW: #NuevaFotoDePerfil ESPERO LES GUSTE ESTE NUEVO ÁNGULO FELIZ MARTES!!! https://t.co/e9yLQaxcWF<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>^https://pbs.twimg.com/media/DLP5cyXUEAA3FoU.jpghttps://twitter.com/ALONDRAHOTSW/status/915349441929936897/photo/1$9154545288703303683United StatesVideos of XStar Jessica Jaymes #jessicajaymes #porn #sex #xxxvideo watch all Jessica HD Videos on: xxxSTARSxxx ------- +18 only!$&Jessica Jaymes Vidsjessica\_jaymes1(2017-10-04T06:52:22Zd#jessicajaymes #porn\_stars https://t.co/b2FtNVMQNBh<a href="https://ifttt.com" rel="nofollow">IFTTT</a>$915454528866197504#txsu20)quaquaviacynae(2017-10-04T06:52:22Z<@slaysae I watched that movie.<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>$915454528878727168Bursa, Türkiye2EYHAN VAR M ULANNN...SANGEL FACE...kuskumelek1(2017-10-04T06:52:22ZBen zaten ölüm kardeimin nefesi , cennet gözleri yokken.*

*Kanserden ölürüm diye korkup îmansz ölürüm diye korkmayanlar yaayan ölüdürler...<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>$915454528887046146jR=VD . , . . , . ! . . . , !*

*! @youeun331R[ ]isyou0113(2017-10-04T06:52:22ZT@toy9054 @narae\_corpse (´`)<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>$915454528870338560*