

Session 12 - Oozie and Flume

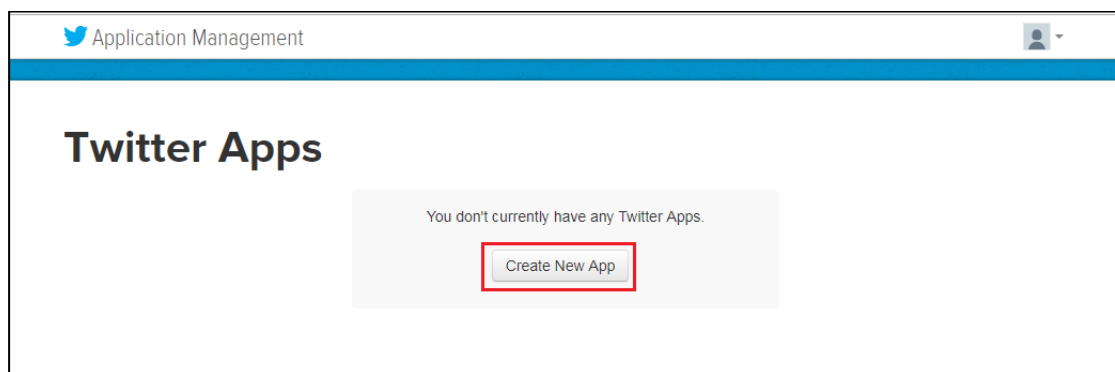
Assignment 1

Task 1:



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

PART 1 : Getting access keys from Twitter

Step 1 : Login to Twitter account and then go to '<https://apps.twitter.com/app>' to go to the Twitter developer site and click on 'Create New App'.



Step 2 : Then enter all the necessary details, click on the 'I Agree' button for Developer Agreement and then continue by clicking in the 'Create Twitter Application' button.

 Application Management 

Create an application

Application Details

Name *

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

Description *

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

Website *

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 the 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.)

Callback URLs

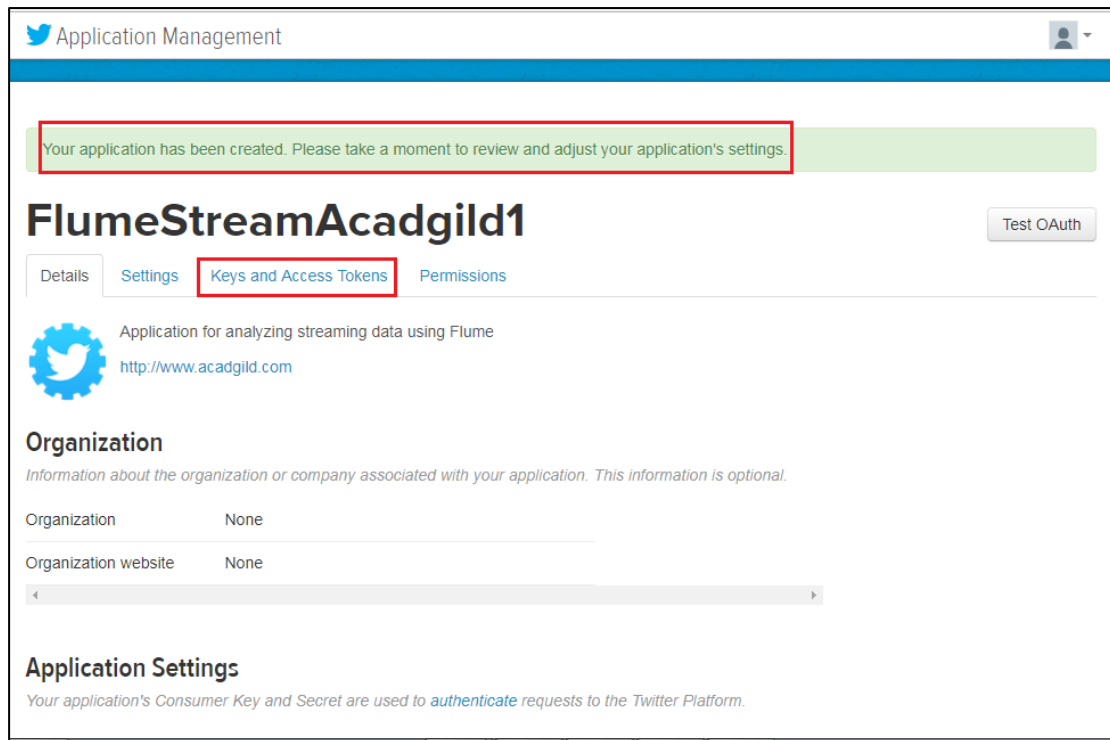
Where should we return after successfully authenticating? OAuth 1.0a applications must explicitly specify their oauth_callback URL(s) here, as well as include the one of the URLs below in the request token step. To restrict your application from using callbacks, leave this field blank.

[Add a Callback URL](#)

Developer Agreement

☒ Yes, I have read and agree to the [Twitter Developer Agreement](#).

Step 3 : Once the application is created, a confirmation message is sent. Then go to the 'Keys and Access Tokens' tab of the application.



Step 4 : From the 'Keys and Access Tokens' tab, copy the Consumer Key (API Key), Consumer Secret (API Secret) values and save them.

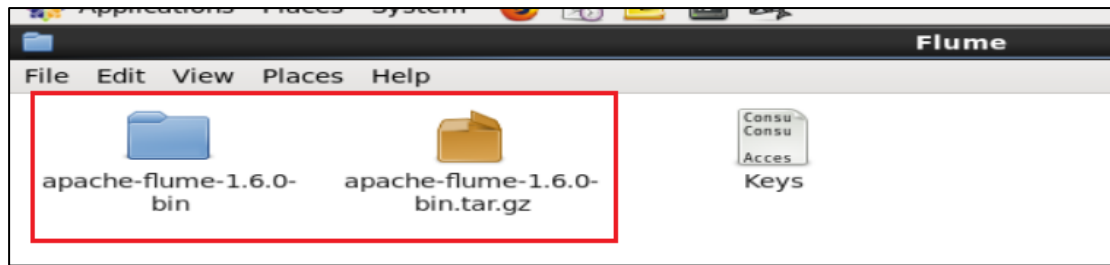
Then, click on the 'Generate Consumer Key and Secret' button below. Then, copy and save these values.

These values will be used to programmatically access Twitter from the Flume agent.

These values can be used to access the twitter account and do any activities from the account. So do not use these values publicly.

PART 2 : Setting Up a Flume Agent for streaming Twitter data

Step 1 : Install Flume by downloading the flume jar file and unzipping it in the installing directory.



Step 2 : Setting the path of the installed directory in .bashrc

Open .bashrc file : `sudo gedit .bashrc`

It will ask for the password of the current user.

Once the file is opened, go to the FLUME_HOME and change it's value to the directory that Flume was just installed.

```

.bashrc
export PATH=$PATH:$HIVE_HOME/bin

#Below 2 lines we have to add for SPARK Installation
export SPARK_HOME=/home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin

# Below 2 lines we have to add for SQOOP Installation
export SQOOP_HOME=/home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha
export PATH=$PATH:$SQOOP_HOME/bin

# BELOW 2 lines we have to add for HBASE Installation
export HBASE_HOME=/home/acadgild/install/hbase/hbase-1.2.6
export PATH=$PATH:$HBASE_HOME/bin

# Below 2 lines we have to add for kafka Installation
export KAFKA_HOME=/home/acadgild/install/kafka/kafka_2.12-0.10.1.1
export PATH=$PATH:$KAFKA_HOME/bin

# Below 2 lines we have to add for FLUME Installation
export FLUME_HOME=/home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin
export PATH=$PATH:$FLUME_HOME/bin

# Below 2 lines we have to add for zookeeper Installation
export ZOOKEEPER_HOME=/home/acadgild/install/zookeeper/zookeeper-3.4.10
export PATH=$PATH:$ZOOKEEPER_HOME/bin
  
```

Step 3 : Save the .bashrc file and run the following command to update the .bashrc file.

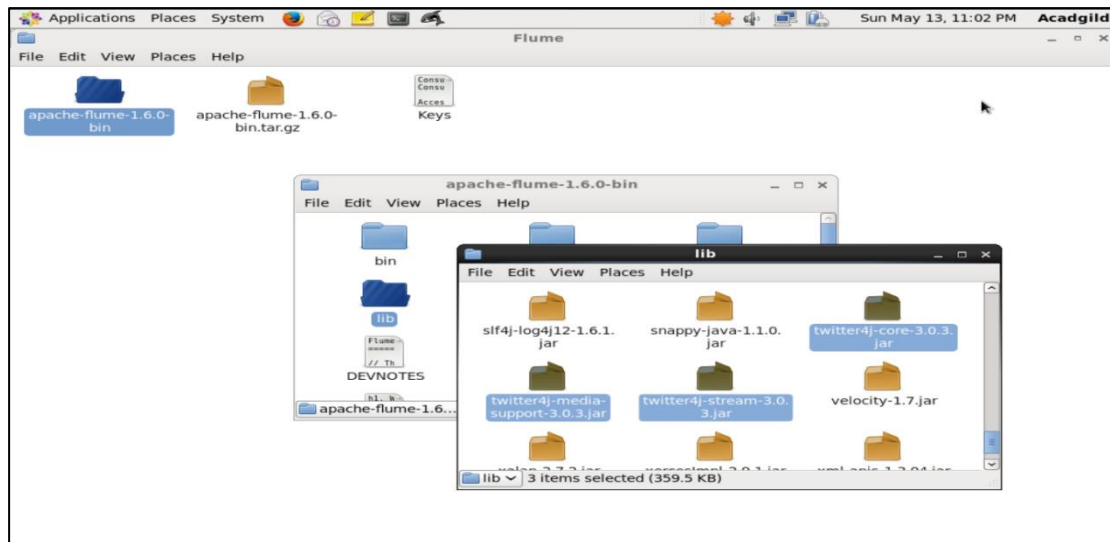
`source .bashrc`

```

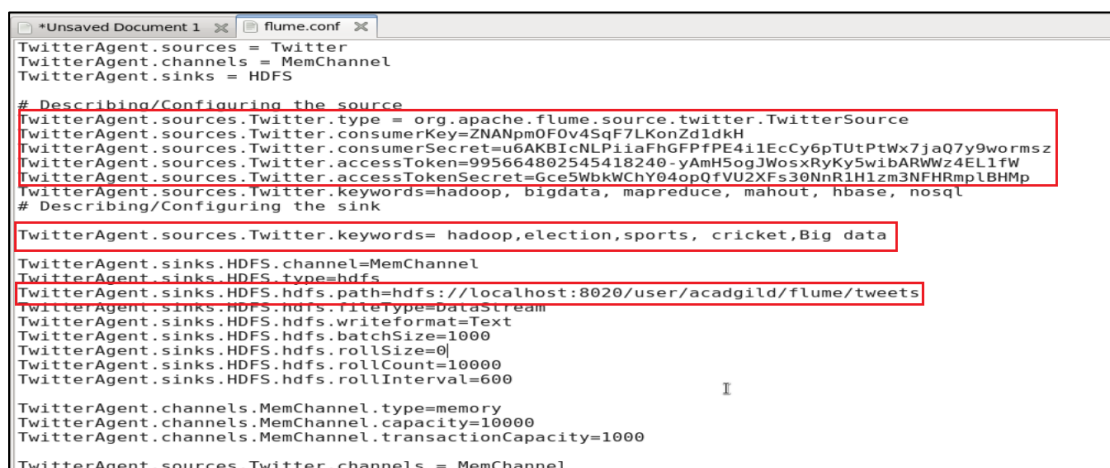
[acadgild@localhost ~]$ sudo gedit .bashrc
(gedit:21796): Gtk-WARNING **: Attempting to store changes into '/root/.local/share/recently-used.xbel', but failed: Failed to create file '/root/.local/share/recently-used.xbel.4ZVNI2': No such file or directory
(gedit:21796): Gtk-WARNING **: Attempting to set the permissions of '/root/.local/share/recently-used.xbel', but failed: No such file or directory
(gedit:21796): Gtk-WARNING **: Attempting to store changes into '/root/.local/share/recently-used.xbel', but failed: Failed to create file '/root/.local/share/recently-used.xbel.86Y2IZ': No such file or directory
(gedit:21796): Gtk-WARNING **: Attempting to set the permissions of '/root/.local/share/recently-used.xbel', but failed: No such file or directory
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ source .bashrc
[acadgild@localhost ~]$
  
```

Step 4 : Make sure you have below jars placed in your \$FLUME_HOME/lib directory:

1. twitter4j-core-X.XX.jar
2. twitter4j-stream-X.X.X.jar
3. twitter4j-media-support-X.X.X.jar



Step 5 : Create an empty file in the apache-flume-1.6.0-bin/conf folder named 'flume.conf' and copy paste the configuration parameters in the file.



Replace the existing values of consumer Key, Consumer Secret, Access Token, Access Token Secret with the values obtained from twitter w=in the earlier part.

'TwitterAgent.sources.Twitter.keywords' is the command to set the list of keywords that are to be searched for in the tweets.

'TwitterAgent.sinks.HDFS.hdfs.path' : set the HDFS path where the streaming feed will be stored on HDFS. Give the port (8020 or 9000) that is specified in the 'core-site.xml' of your hadoop installation.

Step 6 : Make sure the hadoop components are up by running a jps command. If not up, run start-all.sh to get them up and running.

Also, create the target HDFS folders specified in the flume.conf file.

```
[acadgild@localhost ~]$ jps
3104 NameNode
3380 SecondaryNameNode
3205 DataNode
26197 HQuorumPeer
3542 ResourceManager
22749 Jps
5258 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
3645 NodeManager
[acadgild@localhost ~]$ hadoop dfs -mkdir /user/acadgild/flume
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
18/05/13 23:22:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop dfs -mkdir /user/acadgild/flume/tweets
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
18/05/13 23:22:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$
```

Step 7 : Run the flume agent to get the twitter feed stream.

`flume-ng agent -n TwitterAgent -f <location of created/edited conf file>`

`flume-ng agent -n TwitterAgent -f`

`/home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin/conf/flume.conf`

```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin/conf/flume.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop) for HDFS access
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.2.6/bin/hbase) for HBASE access
Info: Excluding /home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-api-1.7.7.jar from classpath
Info: Excluding /home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/home/acadgild/install/hive/apache-hive-2.3.2-bin) for Hive access
+ exec /usr/java/jdk1.8.0_151/bin/java -Xmx20m -cp '/home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/etc/hadoop/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/activation-1.1.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/apacheds-lln-2.0.0-M15.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/install/hadoop/h
```

The connection with Twitter is established and the tweet stream is read and written to HDFS.

```

18/05/14 00:31:31 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered new MBean.
18/05/14 00:31:31 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/05/14 00:31:31 INFO twitter.TwitterSource: Twitter source Twitter started.
18/05/14 00:31:31 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/05/14 00:31:34 INFO twitter4j.TwitterStreamImpl: Connection established.
18/05/14 00:31:34 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/05/14 00:31:35 INFO hdfs.HDFSDataStream: Serializer = TEXT, UserRawLocalFileSystem = false
18/05/14 00:31:36 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/acadgild/flume/tweets/FlumeData.1526238095572.tmp
18/05/14 00:31:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/05/14 00:31:37 INFO twitter.TwitterSource: Processed 100 docs
18/05/14 00:31:40 INFO twitter.TwitterSource: Processed 200 docs
18/05/14 00:31:42 INFO twitter.TwitterSource: Processed 300 docs
18/05/14 00:31:45 INFO twitter.TwitterSource: Processed 400 docs
18/05/14 00:31:48 INFO twitter.TwitterSource: Processed 500 docs
18/05/14 00:31:50 INFO twitter.TwitterSource: Processed 600 docs
18/05/14 00:31:53 INFO twitter.TwitterSource: Processed 700 docs
18/05/14 00:31:53 INFO lifecycle.LifecycleSupervisor: Stopping lifecycle supervisor 10

```

The stream will be continuously read. To stop the process press ctrl+c.

Step 8 : Then, the stream will be read into the HDFS location specified.

```
hadoop dfs -ls /user/acadgild/flume/tweets
```

To see the contents of the file, do a cat on the file

```
hadoop dfs -cat /user/acadgild/flume/tweets/FlumeData.1526237236072
```

```

File Edit View Search Terminal Help
ملف تحرير عرض البحث الخطة مساعدة
ہیڈنگ گفامو او
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

The contents of the stream of tweets containing the keywords specified in the flume.conf file can be seen in the output.