Step-by-step streaming Twitter Data Into Hadoop

Jim Sun, BI Consultant

ATCG Solutions

Jim.sun@atcgsolutions.com

ATCG Solutions
core changes, proven results™

About the Author

Jim Sun is a Business Intelligence consultant focused on SAP and Hadoop Big Data Implementations at ATCG Solutions. He is certified in multiple BI and big data tools. One of Jim's biggest big data project was working with Chinese government to help them design and optimize the biggest logistics park in Hubei Province. Jim graduated from top 5 ranked universities in China and holds a master degree in US.

To Share Click Here

ATCG Solutions
core changes, proven results™

What is Flume

- Flume is a reliable service integrated with Hadoop that collects, aggregates and streams large amount of log data. Also it can be used to stream live data from various data source
- Flume is basis of a lot of data streaming tools

ATCG Solutions
core changes, proven results™

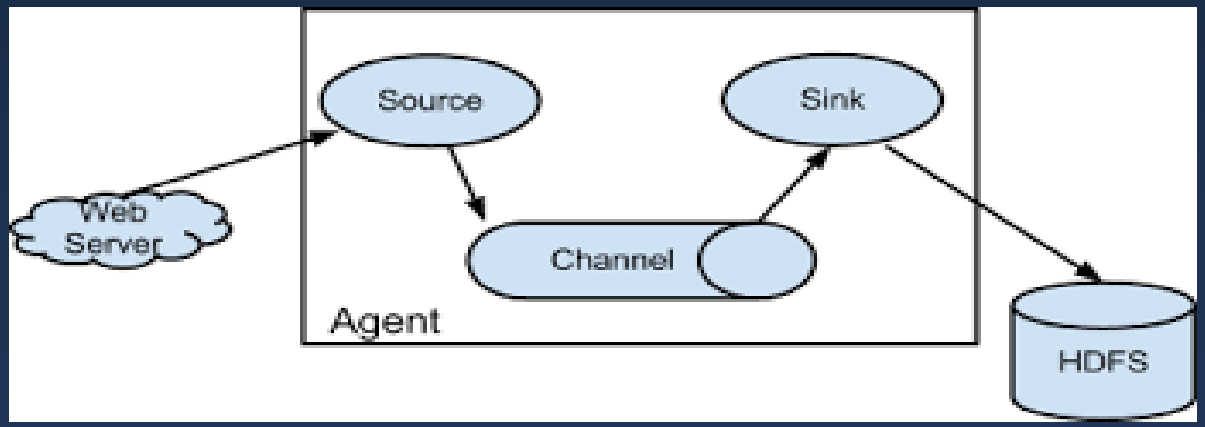
Flume Key Component

- Event---the data to be transferred
- Source---data input
- Sink---data output to destination (HDFS)
- Channel---connection between Source and Sink
- Agent---the physical JAVA VM
- Client---produce and transmit Event

ATCG Solutions
core changes, proven results™

Flume Topology

- For end user, we only need to understand 4 key components: Agent, Source, Channel and Sink. We will set up a Flume agent later and configure Source, Channel and Sink in it.



To Share Click Here

ATCG Solutions
core changes, proven results™

Install Flume

```
[root@sandbox ~]# yum install flume_
```

```
[root@sandbox ~]# yum install flume
Loaded plugins: fastestmirror, priorities
Determining fastest mirrors
epel/metalink                                            |  14 kB     00:00
 * base: centos-mirror.jchost.net
 * epel: kdeforge2.unl.edu
 * extras: centos.
 * updates: centos
HDP-2.1
HDP-UTILS-1.1.0.16
HDP-UTILS-1.1.0.17
Updates-ambari-1.5
ambari-1.x
base
base/primary_db
epel
epel/primary_db
Transaction Summary
================================================================================
Install       1 Package(s)

Total download size: 64 M
Installed size: 71 M
Is this ok [y/N]: y
Downloading Packages:
flume-1.4.0.2.1.1.0-385.el6.noarch.rpm                   |  64 MB     03:58
Running rpm_check_debug
Running Transaction Test
Transaction Test Succeeded
Running Transaction
  Installing : flume-1.4.0.2.1.1.0-385.el6.noarch                           1/1
  Verifying  : flume-1.4.0.2.1.1.0-385.el6.noarch                           1/1

Installed:
  flume.noarch 0:1.4.0.2.1.1.0-385.el6

Complete!
```

ATCG Solutions
core changes, proven results™

Create a Twitter APP as Agent

All programs that try to connect to Twitter and use Twitter data will all be defined as a “Twitter APP”, so is our Flume Agent. So first things first, we need to set up our Twitter APP.

- Go to http://apps.twitter.com and click “Create New APP”
- Fill in the form and create APP and access token
- Take notes on “Consumer Key”, “Consumer Secret”, “Access token” and “Acess token secret”. We will use them later

ATCG Solutions
core changes, proven results™

Configure Flume

- Download a WinSCP to help you securely and easily transfer files between your host machine and virtual machine
- Connect to sandbox using WinSCP and navigate to file /root/etc/flume/conf/flume.conf
- Download a sample flume.conf file for Twitter data streaming from here: https://github.com/cloudera/cdh-twitter-example/blob/master/flume-sources/flume.conf

ATCG Solutions
core changes, proven results™

Configure flume.conf

- Copy the “Consumer Key”, “Consumer Secret”, “Access token” and “Acess token secret” and paste to flume.conf accordingly.

```
TwitterAgent.sources.Twitter.consumerKey = 8gGIYfg6L0Sx
TwitterAgent.sources.Twitter.consumerSecret = GMxYNr2Ej
TwitterAgent.sources.Twitter.accessToken = 2953817984-m
TwitterAgent.sources.Twitter.accessTokenSecret = oKUNim
```

- And pick up the key words and local host:

```
TwitterAgent.sources.Twitter.keywords = hadoop, ATCG Solutions, ATCG
```

```
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://sandbox.hortonworks.com:8020/root/flume/tweets/%Y/%m/%d/%H/
```

Note: Here I use Hortonworks Sandbox as demo, your localhost should be different.

ATCG Solutions
core changes, proven results™

Agent

- You can develop your own agent jar file or download it from here: http://files.cloudera.com/samples/flume-sources-1.0-SNAPSHOT.jar
- This file is developed by Cloudera but also applicable to other Hadoop distributions
- Copy this file to direction: /usr/lib/flume/lib

ATCG Solutions
core changes, proven results™

Start Flume

Up to now everything is set, next we just need to start our agent and track the data inflow:

```
[root@sandbox ~]# nohup flume-ng agent --conf-file /etc/flume/conf/flume.conf --
name TwitterAgent >flume_twitteragent.log &_
```

Note: add “nohup” at the beginning to keep our agent running even when we close our VM session.

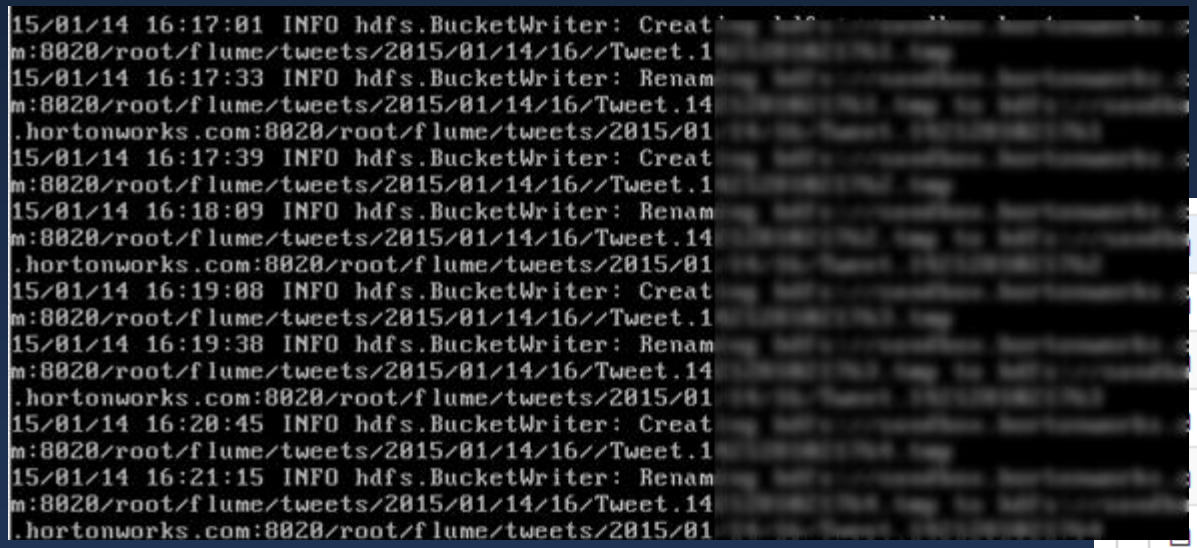
Track the Progress

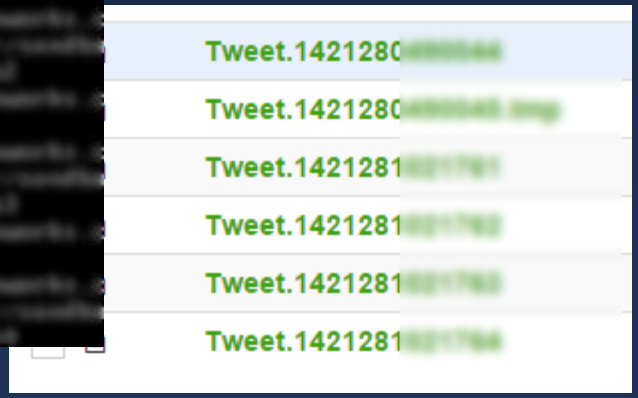
```
[root@sandbox ~]# tail -f flume_twitteragent.log
15/01/14 16:01:40 INFO node.Application: Starting Channel MemChannel
15/01/14 16:01:40 INFO instrumentation.MonitoredCounterGroup: Monitoried counter
 group for type: CHANNEL, name: MemChannel, registered successfully.
15/01/14 16:01:40 INFO instrumentation.MonitoredCounterGroup: Component type: CH
ANNEL, name: MemChannel started
15/01/14 16:01:40 INFO node.Application: Starting Sink HDFS
15/01/14 16:01:40 INFO node.Application: Starting Source Twitter
15/01/14 16:01:40 INFO instrumentation.MonitoredCounterGroup: Monitoried counter
 group for type: SINK, name: HDFS, registered successfully.
15/01/14 16:01:40 INFO instrumentation.MonitoredCounterGroup: Component type: SI
NK, name: HDFS started
15/01/14 16:01:40 INFO twitter4j.TwitterStreamImpl: Establishing connection.
15/01/14 16:02:14 INFO twitter4j.TwitterStreamImpl: Connection established.
15/01/14 16:02:14 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
```


ATCG Solutions
core changes, proven results™

Track our data

You can either track the data streaming process in VM or Hue:





To Share Click Here

ATCG Solutions
core changes, proven results™

Future Steps

- Develop a JAVA SerDe file to de-serialize those data
- Develop Hive Script and generate tables in Hive
- Connect to Hive through ODBC/JDBC connection and visualize the Twitter data

ATCG Solutions
core changes, proven results™

About ATCG Solutions: We are a Technology Agnostic BI Consulting Company focused on providing business intelligence solutions.

www.atcgsolutions.com
sales@atcgsolutions.com
Tel: 866.337.2252

Questions? More Information?

We offer an Assessment Service to help you determine your need for a Real-Time Database Environment. **Click Here**

Get the Assessment